

David C. Wyld
Jan Zizka
Dhinaharan Nagamalai (Eds.)

Advances in Computer Science, Engineering and Applications

Proceedings of the Second International
Conference on Computer Science,
Engineering and Applications (ICCSEA 2012),
May 25–27, 2012, New Delhi, India, Volume 2

Editor-in-Chief

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

David C. Wyld, Jan Zizka,
and Dhinaharan Nagamalai (Eds.)

Advances in Computer Science, Engineering and Applications

Proceedings of the Second International
Conference on Computer Science,
Engineering and Applications (ICCSEA 2012),
May 25–27, 2012, New Delhi, India, Volume 2



Editors

David C. Wyld
Southeastern Louisiana University
Hammond
USA

Dhinaharan Nagamalai
Wireilla Net Solutions PTY Ltd
Melbourne
Australia

Jan Zizka
Mendel University
Brno
Czech Republic

ISSN 1867-5662
ISBN 978-3-642-30110-0
DOI 10.1007/978-3-642-30111-7
Springer Heidelberg New York Dordrecht London

e-ISSN 1867-5670
e-ISBN 978-3-642-30111-7

Library of Congress Control Number: 2012937233

© Springer-Verlag Berlin Heidelberg 2012

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

Preface

The Second International Conference on Computer Science, Engineering and Applications (ICCSEA-2012) was held in Delhi, India, during May 25–27, 2012. ICCSEA-2012 attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West. The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The ICCSEA-2012 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the conference. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer-review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer Science and Engineering research.

In closing, ICCSEA-2012 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research.

It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come. We hope that you will benefit from the fine papers from the ICCSEA-2012 conference that are in this volume and will join us at the next ICCSEA conference.

David C. Wyld
Jan Zizka
Dhinaharan Nagamalai

Organization

General Chairs

David C. Wyld	Southeastern Louisiana University, USA
Michal Wozniak	Wroclaw University of Technology, Poland
Henrique Joao Lopes Domingos	University of Lisbon, Portugal

Steering Committee

Jose Enrique Armendariz-Inigo	Universidad Publica de Navarra, Spain
John Karamitsos	University of the Aegean, Samos, Greece
Dhinaharan Nagamalai	Wireilla Net Solutions PTY LTD, Australia
Chih-Lin Hu	National Central University, Taiwan
Salah M. Saleh Al-Majeed	University of Essex, United Kingdom
Jan Zizka	SoNet/DI, FBE, Mendel University in Brno, Czech Republic

Program Committee Members

A. Kannan	K.L.N. College of Engineering, India
A.P. Sathish Kumar	PSG Institute of Advanced Studies, India
Abdul Aziz	University of Central Punjab, Pakistan
Abdul Kadhir Ozcan	Karatay University, Turkey
Ahmed Nada	Al-Quds University, Palestinian
Alejandro Regalado Mendez	Universidad del Mar - México, USA
Ali M.	University of Bradford, United Kingdom
Ali Maqousi	Petra University, Jordan
Anand Sharma	IMITS -Rajasthan, India
Andy Seddon	Asia Pacific Institute of Information Technology, Malaysia
Anjan K.	RVCE-Bangalore, India
Ankit Thakkar	Nirma University, India

VIII Organization

Anthony Atayero	Covenant University, Nigeria
Ashok Kumar Das	IIT Hyderabad, India
B. Srinivasan	Monash University, Australia
Balasubramanian K.	KLefke European University, Cyprus
Balasubramanian Karuppiah	MGR University, India
Beatrice Cynthia Dhinakaran	Hannam University, South Korea
Bela Genge	European Commission Joint Research Centre, Belgium
Bobby Barua	Ahsanullah University of Science and Technology, Bangladesh
Bong-Han Kim	Chonju University, South Korea
Boo-Hyung Lee	KongJu National University, South Korea
Brajesh Kumar Kaushik	Indian Institute of Technology, India
Carlos E. Otero	University of South Florida Polytechnic, USA
Ch.V. Rama Rao	Gudlavalleru Engineering College, India
Charalampos Z. Patrikakis	National Technical University of Athens, Greece
Chih-Lin Hu	National Central University, Taiwan
Chin-Chih Chang	Chung Hua University, Taiwan
Cho Han Jin	Far East University, South Korea
Danda B. Rawat	Old Dominion University, USA
Khamish Malhotra	University of Glamorgan, UK
M. Rajarajan	City University, UK
Mohammad Momani	University of Technology Sydney, Australia
Raja Kumar M.	National Advanced IPv6 Center (NAv6), Universiti Sains Malaysia
Salman Abdul Moiz	Centre for Development of Advanced Computing, India
Carlos E. Otero	The University of Virginia's College at Wise, USA
Wojciech Mazurczyk	Warsaw University of Technology, Poland
David C. Wyld	Southeastern Louisiana University, India
David W. Deeds	Shingu College, South Korea
Debasis Giri	Haldia Institute of Technology, India
Dhinaharan Nagamalai	Wireilla Net Solutions PVT Ltd, Australia
Dimitris Kotzinos	Technical Educational Institution of Serres, Greece
E. Martin	University of California, Berkeley, USA
Emmanuel Bouix	iKlux Media, France
Ermatita Zuhairi	Sriwijaya University, Indonesia
Farhat Anwar	International Islamic University, Malaysia
Firkhan Ali Bin Hamid Ali	Universiti Tun Hussein Onn Malaysia, Malaysia
Ford Lumban Gaol	University of Indonesia
Ghalem Belalem	University of Oran, Algeria
Giovanni Schembra	University of Catania, Italy
Girija Chetty	University of Canberra, Australia
Gomathi Kandasamy	Avinashilingam Deemed University for Women, India

H.V. Ramakrishnan	Dr. MGR University, India
Hao-En Chueh	Yuanpei University, Taiwan
Henrique Joao Lopes Domingos	University of Lisbon, Portugal
Ho Dac Tu	Waseda University, Japan
Hoang Huu Hanh	Hue University, Vietnam
Hussein Al-Bahadili	Petra University, Jordan
Hwangjun Song	Pohang University of Science and Technology, South Korea
Intisar Al-Mejibli	University of Essex, United Kingdom
J.K. Mandal	University of Kalyani, India
Jacques Demerjian	Communication & Systems, Homeland Security, France
Jae Kwang Lee	Hannam University, South Korea
Jan Zizka	SoNet/DI, FBE, Mendel University in Brno, Czech Republic
Jeong-Hyun Park	Electronics Telecommunication ResearchInstitute, South Korea
Jeyanth N.	VIT University, India
Jifeng Wang	University of Illinois at Urbana Champaign, USA
Jivesh Govil	Cisco Systems Inc., USA
Johann Groschdl	University of Bristol, UK
John Karamitsos	University of the Aegean, Greece
Johnson Kuruvila	Dalhousie University, Canada
Jose Enrique Armendariz-Inigo	Universidad Publica de Navarra, Spain
Jungwook Song	Konkuk University, South Korea
K.P. Thooyamani	Bharath University, India
Kai Xu	University of Bradford, United Kingdom
Kamalrulnizam Abu Bakar	Universiti Teknologi Malaysia, Malaysia
Khamish Malhotra	University of Glamorgan, UK
Krishnamurthy E.V.	ANU College Of Engg & Computer Science, Australia
Krzysztof Walkowiak	Wroclaw University of Technology, Poland
Krzysztof Walkowiak	Wroclaw University of Technology, Poland
Lu Yan	University of Hertfordshire, UK
Lus Veiga	Technical University of Lisbon, Portugal
M. Aqeel Iqbal	FUIEMS, Pakistan
Mahesh Goyani	G.H. Patel College of Engineering and Technology, India
Maragathavalli P.	Pondicherry Engineering College, India
Marco Folli	University of Pavia, Italy
Marco Roccetti	Universty of Bologna , Italy
Martin A.	Pondicherry University, India
Massimo Esposito	ICAR-CNR, Italy
Michal Wozniak	Wroclaw University of Technology, Poland
Mohammad Ali Jabreil Jamali	Islamic Azad University, Iran
Mohammad Zaidul Karim	Daffodil International University, Bangladesh

Mohsen Sharifi	Iran University of Science and Technology, Iran
Moses Ekpennyong	University of Uyo, Nigeria
Muhammad Sajjadur Rahim	University of Rajshahi, Bangladesh
Murugan D.	Manonmaniam Sundaranar University, India
N. Kaliammal	NPR College of Engg & Tech, India
N. Krishnan	Manonmaniam Sundaranar University, India
Nabendu Chaki	University of Calcutta, India
Naohiro Ishii	Aichi Institute of Technology, Japan
Nasrollah M. Charkari	Tarbiat Modares University, Iran
Natarajan Meghanathan	Jackson State University, USA
Nicolas Sklavos	Technological Educational Institute of Patras, Greece
Nidaa Abdual Muhsin Abbas	University of Babylon, Iraq
Olakanmi Oladayo	University of Ibadan, Nigeria
P. Ashok Babu	Narsimhareddy Engineering college, India
Patrick Seeling	University of Wisconsin - Stevens Point, USA
PESN Krishna Prasad	Aditya Engineering College, India
Phan Cong Vinh	London South Bank University, UK
Ponpit Wongthongtham	Curtin University of Technology, Australia
Premanand K. Kadbe	Vidya Pratishthan's College of Engineering, India
Rafael Timoteo	University of Brasilia - UnB, Brazil
Raja Kumar M.	Universiti Sains Malaysia
Rajagopal Palsonkennedy	Dr. MGR University, India
Rajarshi Roy	IIT- Kharagpur, India
Rajendra Akerkar	Technomathematics Research Foundation, India
Rajesh Kumar P.	The Best International, Australia
Rajeshwari Hegde	BMS College of Engineering, India
Rajeswari Balasubramaniam	Dr. MGR University, India
Rajkumar Kannan	Bishop Heber College, India
Rakhesh Singh Kshetrimayum	Indian Institute of Technology, Guwahati, India
Ramayah Thurasamy	Universiti Sains Malaysia, Malaysia
Ramayah Thurasamy	Universiti Sains Malaysia, Malaysia
Razvan Deaconescu	University Politehnica of Bucharest, Romania
Reza Ebrahimi Atani	University of Guilan Iran
Rohitha Goonatilake	Texas A&M International University, USA
S. Geetha	Anna University - Tiruchirappalli, India
S. Hariharan	B.S. Abdur Rahman University, India
Sagarmay	Deb Central Queensland University, Australia
Sajid Hussain	Acadia University, Canada
Salah M. Saleh Al-Majeed	University of Essex, United Kingdom
Salim Lahmiri	University of Québec at Montreal, Canada
Samarendra Nath Sur	Sikkim Manipal University, India
Sarmistha Neogy	Jadavpur University, India
Sattar B. Sadkhan	University of Babylon, Iraq
Sergio Ilarri	University of Zaragoza, Spain

Serguei A. Mokhov	Concordia University, Canada
Sharvani G.S.	RV College of Engineering, India
Shivan Haran	Arizona State University, USA
Shobha Shankar	Vidya Vardhaka College of Engineering, India
Shubhamoy Dey	Indian Institute of Management Indore, India
Sriman Narayana Iyengar	VIT University, India
Sundarapandian Vaidyanathan	VelTech Dr. RR & Dr. SR Technical University, India
SunYoung Han	Konkuk University, South Korea
Susana Sargento	University of Aveiro, Portugal
Virgil Dobrota	Technical University of Cluj-Napoca, Romani
Vishal Sharma	Metanoia Inc., USA
Wei Jie	University of Manchester, UK
Wichian Sittiprapaporn	Mahasarakham University, Thailand
William R. Simpson	Institute for Defense Analyses, USA
Xiaohong Yuan	North Carolina A & T State University, USA
Xin Bai	The City University of New York, USA
Yannick Le Moullec	Aalborg University, Denmark
Yaser M. Khamayseh	Jordan University of Science and Technology, Jordan
Yeong Deok Kim	Woosong University, South Korea
Yuh-Shyan Chen	National Taipei University, Taiwan
Yung-Fa Huang	Chaoyang University of Technology, Taiwan
Yung-Fa Huang	Chaoyang University of Technology, Taiwan
Zakaria Moudam	Université sidi mohammed ben Abdellah, Morocco
Nicolas Sklavos	Technological Educational Institute of Patras, Greece
Roberts Masillamani	Hindustan University, India

External Reviewers

Amit Choudhary	Maharaja Surajmal Institute, India
Abhishek Samanta	Jadavpur University, Kolkata, India
Anjan K.	MSRIT, India
Nana Patil	NIT Surat, Gujarat
Mydhili Nair	M.S. Ramaiah Institute of Technology, India
Padmalochan Bera	Indian Institute of Technology, Kharagpur, India
Osman B. Ghazali	Universiti Utara Malaysia, Malaysia
Suparna DasGupta	suparnadasguptait@gmail.com
Cauvery Giri	RVCE, India
Pradeepini Gera	Jawaharlal Nehru Technological University, India
Reshma Maulik	University of Calcutta, India
Soumyabrata Saha	Guru Tegh Bahadur Institute of Technology, India
Srinivasulu Pamidi	V.R. Siddhartha Engineering College Vijayawada, India

Suhaidi B. Hassan	Office of the Assistant Vice Chancellor, Economics Building
Mahalinga V. Mandi	Dr. Ambedkar Institute of Technology, Bangalore, Karnataka, India
Omar Almomani	College of Arts and Sciences Universiti Utara Malaysia
Sara Najafzadeh	University Technology Malaysia
Ramin Karimi	University Technology Malaysia
Samodar Reddy	India School of Mines, India
Ashutosh Gupta	MJP Rohilkhand University, Bareilly
Jayeeta Chanda	jayeeta.chanda@gmail.com
Rituparna Chaki	rituchaki@gmail.com
Durga Toshniwal	Indian Institute of Technology, India
Mohammad Mehdi Farhangia	Universiti Teknologi Malaysia (UTM), Malaysian
S. Bhaskaran	SAASTRA University, India
Bhupendra Suman	IIT Roorkee (India)
Yedehalli Kumara Swamy	Dayanand Sagar College of Engineering, India
Swarup Mitra	Jadavpur University, Kolkata, India
R.M. Suresh	Mysore University
Nagaraj Aitha	I.T., Kamala Institute of Tech & Science, India
Ashutosh Dubey	NRI Institute of Science & Technology, Bhopal
Balakannan S.P.	Chonbuk Nat. Univ., Jeonju
Sami Ouali	ENSI, Compus of Manouba, Manouba, Tunisia
Seetha Maddala	CBIT, Hyderabad
Reena Dadhich	Govt. Engineering College Ajmer
Kota Sunitha	G. Narayananamma Institute of Technology and Science, Hyderabad
Parth Lakhya	parth.lakhya@einfochips.com
Murty, Ch. A.S.	JNTU, Hyderabad
Shriram Vasudevan	VIT University, India
Govardhan A.	JNTUH College of Engineering, India
Rabindranath Bera	Sikkim Manipal Inst. of Technol., India
Sanjay Singh	Manipal Institute of Technology, India
Subir Sarkar	Jadavpur University, India
Nagamanjula Prasad	Padmasri Institute of Technology, India
Rajesh Kumar Krishnan	Bannari Amman Inst. of Technol., India
Sarada Prasad Dakua	IIT-Bombay, India
Tsung Teng Chen	National Taipei Univ., Taiwan
Balaji Sriramulu	drsbalaji@gmail.com
Chandra Mohan	Bapatha Engineering College, India
Saleena Ameen	B.S. Abdur Rahman University, India
Babak Khosravifar	Concordia University, Canada
Hari Chavan	National Institute of Technology, Jamshedpur, India
Lavanya	Blekinge Institute of Technology, Sweden
Pappa Rajan	Anna University, India
Rituparna Chaki	West Bengal University of Technology, India

Salini P.	Pondichery Engineering College, India
Ramin Karimi	University Technology Malaysia
P. Sheik Abdul Khader	B.S. Abdur Rahman University, India
Rajashree Biradar	Ballari Institute of Technology and Management, India
Scsharma	IIT - Roorkee, India
Kaushik Chakraborty	Jadavpur University, India
Sunil Singh	Bharati Vidyapeeth's College of Engineering, India
Doreswamyh Hosahalli	Mangalore University, India
Debdatta Kandar	Sikkim Manipal University, India
Selvakumar Ramachandran	Blekinge Institute of Technology, Sweden
Naga Prasad Bandaru	PVP Siddartha Institute of Technology, India
HaMeEm sHaNaVaS	Vivekananda Institute of Technolgy, India
Gopalakrishnan Kaliaperumal	Anna University, chennai
Ankit	BITS, PILANI India
Aravind P.A.	Amrita School of Engineering India
Subhabrata Mukherjee	Jadavpur University, India
Valli Kumari Vatsavayi	AU College of Engineering, India

Contents

Networks and Communications

Partitioning and Internetworking Wireless Mesh Network with Wired Network for Delivery Maximization and QoS Provisioning	1
<i>Soma Pandey, Vijay Pande, Govind Kadambi, Stephen Bate</i>	
A New Secret Key Cipher: C128	15
<i>Indrajit Das, R. Saravanan</i>	
Optimal Bandwidth Allocation Technique in IEEE 802.11e Mobile Ad Hoc Networks (MANET)	25
<i>R. Mynuddin Sulthani, D. Sreenivasa Rao</i>	
Trusted AODV for Trustworthy Routing in MANET	37
<i>Sridhar Subramanian, Baskaran Ramachandran</i>	
Dynamic Fuzzy Based Reputation Model for the Assurance of Node Security in AODV for Mobile Ad-Hoc Network	47
<i>Arifa Azeez, K.G. Preetha</i>	
Policy Based Traffic in Video on Demand System	55
<i>Soumen Kanrar</i>	
Distance Aware Zone Routing Protocol for Less Delay Transmission and Efficient Bandwidth Utilization	63
<i>Dhanya Sudarsan, P.R. Mahalingam, G. Jisha</i>	
Mobile Data Offloading: Benefits, Issues, and Technological Solutions	73
<i>Vishal Gupta, Mukesh Kumar Rohil</i>	
Performance Analysis of Gigabit Ethernet Standard for Various Physical Media Using Triple Speed Ethernet IP Core on FPGA	81
<i>V.R. Gad, R.S. Gad, G.M. Naik</i>	

Assortment of Information from Mobile Phone Subscribers Using Chronological Model [IGCM]: Application and Management Perspective	91
<i>Neeraj Kumar, Raees A. Khan</i>	
Modeling Soft Handoffs' Performance in a Realistic CDMA Network	103
<i>Moses E. Ekpenyong, Enobong Umana</i>	
A Security Approach for Mobile Agent Based Crawler	119
<i>Vimal Upadhyay, Jai Balwan, Gori Shankar, Amritpal</i>	
Prospects and Limitations of Organic Thin Film Transistors (OTFTs)	125
<i>B.K. Kaushik, Brijesh Kumar, Y.S. Negi, Poornima Mittal</i>	
Active Learning with Bagging for NLP Tasks	141
<i>Ruy Luiz Milidiú, Daniel Schwabe, Eduardo Motta</i>	
Mining Queries for Constructing Materialized Views in a Data Warehouse	149
<i>T.V. Vijay Kumar, Archana Singh, Gaurav Dubey</i>	
Similarity Based Cluster Analysis on Engineering Materials Data Sets	161
<i>Doreswamy, K.S. Hemanth</i>	
A Chaotic Encryption Algorithm: Robustness against Brute-Force Attack	169
<i>Mina Mishra, V.H. Mankar</i>	
Introducing Session Relevancy Inspection in Web Page	181
<i>Sutirtha Kumar Guha, Anirban Kundu, Rana DattaGupta</i>	
Way Directing Node Routing Protocol for Mobile Ad Hoc Networks	193
<i>M. Neelakantappa, A. Damodaram, B. Satyanarayana</i>	
Web-Page Prediction for Domain Specific Web-Search Using Boolean Bit Mask	211
<i>Sukanta Sinha, Rana DattaGupta, Debajyoti Mukhopadhyay</i>	
Security Enhanced Digital Image Steganography Based on Successive Arnold Transformation	221
<i>Minati Mishra, Sunit Kumar, Subhadra Mishra</i>	
Impact of Bandwidth on Multiple Connections in AODV Routing Protocol for Mobile Ad-Hoc Network	231
<i>K.G. Preetha, A. Unnikrishnan, K. Paulose Jacob</i>	
Conceptualizing an Adaptive Framework for Pervasive Computing Environment	241
<i>Akhil Mohan, Nitin Upadhyay</i>	

Dynamic DCF Backoff Algorithm(DDBA) for Enhancing TCP Performance in Wireless Ad Hoc Networks	257
<i>B. Nithya, C. Mala, B. Vijay Kumar, N.P. Gopalan</i>	
Hybrid Cluster Validation Techniques	267
<i>Satish Gajawada, Durga Toshniwal</i>	
Energy Efficient and Minimal Path Selection of Nodes to Cluster Head in Homogeneous Wireless Sensor Networks	275
<i>S. Taruna, Sheena Kohli, G.N. Purohit</i>	
Texel Identification Using K-Means Clustering Method	285
<i>S. Padmavathi, C. Rajalaxmi, K.P. Soman</i>	
Single and Multi Trusted Third Party: Comparison, Identification and Reduction of Malicious Conduct by Trusted Third Party in Secure Multiparty Computing Protocol	295
<i>Zulfa Shaikh, Poonam Garg</i>	
Ubiquitous Medical Learning Using Augmented Reality Based on Cognitive Information Theory	305
<i>Zahra Mohana Gebril, Imam Musa Abiodunde Tele, Mohammed A. Tahir, Behrang Parhizkar, Anand Ramachandran, Arash Habibi Lashkari</i>	
A Secured Transport System by Authenticating Vehicles and Drivers Using RFID	313
<i>C.K. Marigowda, J. Thriveni, Javid K. Karangi</i>	
Virtualization of Large-Scale Data Storage System to Achieve Dynamicity and Scalability in Grid Computing	323
<i>Ajay Kumar, Seema Bawa</i>	
Behavioral Profile Generation for 9/11 Terrorist Network Using Efficient Selection Strategies	333
<i>S. Karthika, A. Kiruthiga, S. Bose</i>	
A Quantitative Model of Operating System Security Evaluation	345
<i>Hammad Afzali, Hassan Mokhtari</i>	
Energy Management in Zone Routing Protocol (ZRP)	355
<i>Dilli Ravilla, Chandra Shekar Reddy Putta</i>	
A New Approach for Vertical Handoff in Wireless 4G Network	367
<i>Vijay Malviya, Praneet Saurabh, Bhupendra Verma</i>	
An Analysis on Critical Information Security Systems	377
<i>Sona Kaushik, Shalini Puri</i>	

Emergency Based Remote Collateral Tracking System Using Google's Android Mobile Platform	391
<i>Prabhu Dorairaj, Saranya Ramamoorthy, Ashok Kumar Ramalingam</i>	
Performance Analysis of (AIMM-I46) Addressing, Inter-mobility and Interoperability Management Architecture between IPv4 and IPv6 Networks	405
<i>Gnana Jayanthi Joseph, S. Albert Rabara</i>	
Wireless Mobile Networks	
Strong Neighborhood Based Stable Connected Dominating Sets for Mobile Ad Hoc Networks	415
<i>Natarajan Meghanathan, Michael Terrell</i>	
OERD - On Demand and Efficient Replication Dereplication	425
<i>Vardhan Manu, Gupta Paras, Kushwaha Dharmender Singh</i>	
A Mathematical Model for Performance Evaluation and Comparison of MAP Selection Schemes in n Layer HMIPv6 Networks	435
<i>Abhishek Majumder</i>	
Utilizing Genetic Algorithm in a Sink Driven, Energy Aware Routing Protocol for Wireless Sensor Networks	447
<i>Hosny M. Ibrahim, Nagwa M. Omar, Ali H. Ahmed</i>	
Load Balancing with Reduced Unnecessary Handoff in Hierarchical Macro/Femto-cell WiMAX Networks	457
<i>Prasun Chowdhury, Anindita Kundu, Iti Saha Misra, Salil K. Sanyal</i>	
A Study on Transmission-Control Middleware on an Android Terminal in a WLAN Environment	469
<i>Hiromi Hirai, Kaori Miki, Saneyasu Yamaguchi, Masato Oguchi</i>	
A Study of Location-Based Data Replication Techniques and Location Services for MANETs	481
<i>C.B. Chandrakala, K.V. Prema, K.S. Hareesa</i>	
A Comparative Analysis of Modern Day Network Simulators	489
<i>Debjyoti Pal</i>	
Design and Implementation of a RFID Based Prototype SmArt LibRARY (SALARY) System Using Wireless Sensor Networks	499
<i>K.S. Kushal, H.K. Muttanna Kadal, S. Chetan, Shivaputra</i>	
Optimal Route Life Time Prediction of Dynamic Mobile Nodes in Manets	507
<i>Ajay Kumar, Shany Jophin, M.S. Sheethal, Priya Philip</i>	

Reachability Analysis of Mobility Models under Idealistic and Realistic Environments	519
<i>Chirag Kumar, C.K. Nagpal, Bharat Bhushan, Shailender Gupta</i>	
Chaotic Cipher Using Arnolds and Duffings Map	529
<i>Mina Mishra, V.H. Mankar</i>	
Effect of Sound Speed on Localization Algorithm for Underwater Sensor Networks	541
<i>Samedha S. Naik, Manisha J. Nene</i>	
An Analytical Model for Power Control B-MAC Protocol in WSN.....	551
<i>V. Ramchand, D.K. Lobiyal</i>	
Enterprise Mobility – A Future Transformation Strategy for Organizations	559
<i>Jitendra Maan</i>	
Allocation of Guard Channels for QoS in Hierarchical Cellular Network ...	569
<i>Kashish Parwani, G.N. Purohit</i>	
Application Development and Cost Analysis for Content Based Publish Subscribe Model in Mobile Environment	579
<i>Medha A. Shah, P.J. Kulkarni</i>	
Energy Aware AODV (EA-AODV) Using Variable Range Transmission	589
<i>Pinki Nayak, Rekha Agarwal, Seema Verma</i>	
Automatic Speech Recognizer Using Digital Signal Processor	599
<i>Raghavendra M. Shet, Raghunath S. Holambe</i>	
Pre Decision Based Handoff in Multi Network Environment	609
<i>Manoj Sharma, R.K. Khola</i>	
A Swarm Inspired Probabilistic Path Selection with Congestion Control in MANETs	617
<i>Subhankar Joardar, Vandana Bhattacherjee, Debasis Giri</i>	
A Hierarchical CPN Model for Mobility Analysis in Zone Based MANET	627
<i>Moitreyee Dasgupta, Sankhayan Chaudhury, Nabendu Chaki</i>	
Sensor Deployment for Mobile Object Tracking in Wireless Sensor Networks	637
<i>Yingchi Mao, Ting Yin</i>	
ELRM: A Generic Framework for Location Privacy in LBS	647
<i>Muhamed Ilyas, R. Vijayakumar</i>	

Energy Efficient Administrator Based Secure Routing in MANET	659
<i>Himadri Nath Saha, Debika Bhattacharyya, P.K. Banerjee</i>	
Effect of Mobility on Trust in Mobile Ad-Hoc Network	673
<i>Amit Kumar Raikwar</i>	
Communications Security and Information Assurance	
Securing Systems after Deployment	685
<i>David (DJ) Neal, Syed (Shawon) Rahman</i>	
On the Security of Two Certificateless Signature Schemes	695
<i>Young-Ran Lee</i>	
Information Security Using Chains Matrix Multiplication	703
<i>Ch. Rupa, P.S. Avadhani</i>	
Formal Security Verification of Secured ECC Based Signcryption Scheme	713
<i>Atanu Basu, Indranil Sengupta, Jamuna Kanta Sing</i>	
Universal Steganalysis Using Contourlet Transform	727
<i>V. Natarajan, R. Anitha</i>	
Algorithm for Clustering with Intrusion Detection Using Modified and Hashed K – Means Algorithms	737
<i>M. Varaprasad Rao, A. Damodaram, N.Ch. Bhatra Charyulu</i>	
Z Transform Based Digital Image Authentication Using Quantization Index Modulation (Z-DIAQIM)	745
<i>Nabin Ghoshal, Soumit Chowdhury, Jyotsna Kumar Mandal</i>	
Secret Data Hiding within Tolerance Level of Embedding in Quality Songs (DHTL)	753
<i>Uttam Kr. Mondal, J.K. Mandal</i>	
A Novel DFT Based Information Embedding for Color Image Authentication (DFTIECIA)	763
<i>J.K. Mandal, S.K. Ghosal</i>	
A Real Time Detection of Distributed Denial-of-Service Attacks Using Cumulative Sum Algorithm and Adaptive Neuro-Fuzzy Inference System	773
<i>R. Anitha, R. Karthik, V. Pravin, K. Thirugnanam</i>	
Encryption of Images Based on Genetic Algorithm – A New Approach	783
<i>Jalesh Kumar, S. Nirmala</i>	

Content Based Image Retrieval Using Normalization of Vector Approach to SVM	793
<i>Sumit Dhariwal, Sandeep Raghuvanshi, Shailendra Shrivastava</i>	
Modified Grøstl: An Efficient Hash Function	803
<i>Gurpreet Kaur, Vidyavati S. Nayak, Dhananjoy Dey, S.K. Pal</i>	
Iris Recognition Systems with Reduced Storage and High Accuracy Using Majority Voting and Haar Transform	813
<i>V. Anitha, R. Leela Velusamy</i>	
Recovery of Live Evidence from Internet Applications	823
<i>Ipsita Mohanty, R. Leela Velusamy</i>	
Face Detection Using HMM-SVM Method	835
<i>Nupur Rajput, Pranita Jain, Shailendra Shrivastava</i>	
High Capacity Lossless Semi-fragile Audio Watermarking in the Time Domain	843
<i>Sunita V. Dhavale, R.S. Deodhar, L.M. Patnaik</i>	
Key Management Protocol in WIMAX Revisited	853
<i>Noudjoud Kahya, Nacira Ghoualmi, Pascal Lafourcade</i>	
Image Authentication Technique Based on DCT (IATDCT)	863
<i>Nabin Ghosal, Anirban Goswami, Jyotsna Kumar Mondal, Dipankar Pal</i>	
Survey on a Co-operative Multi-agent Based Wireless Intrusion Detection Systems Using MIBs	873
<i>Ashvini Vyavhare, Varsharani Bhosale, Mrunal Sawant, Fazila Girkar</i>	
A Binary Vote Based Comparison of Simple Majority and Hierarchical Decision for Survivable Networks	883
<i>Charles A. Kamhoua, Kevin A. Kwiat, Joon S. Park</i>	
A Novel Way of Protecting the Shared Key by Using Secret Sharing and Embedding Using Pseudo Random Numbers	897
<i>P. Devaki, G. Raghavendra Rao</i>	
Security Assurance by Efficient Non-repudiation Requirements	905
<i>S.K. Pandey, K. Mustafa</i>	
Poor Quality Watermark Barcodes Image Enhancement	913
<i>Mohammed A. Atiea, Yousef B. Mahdy, Abdel-Rahman Hedar</i>	
Hiding Data in FLV Video File	919
<i>Mohammed A. Atiea, Yousef B. Mahdy, Abdel-Rahman Hedar</i>	
Taxonomy of Network Layer Attacks in Wireless Mesh Network	927
<i>K. Ganesh Reddy, P. Santhi Thilagam</i>	

Implementing Availability State Transition Model to Quantify Risk Factor	937
<i>Shalini Chandra, Raees Ahmad Khan</i>	
Performance Analysis of Fast DOA Estimation Using Wavelet Denoising over Rayleigh Fading Channel on MIMO System	953
<i>A.V. Meenakshi, R. Kayalvizhi, S. Asha</i>	
Grid Computing	
DAGITIZER – A Tool to Generate Directed Acyclic Graph through Randomizer to Model Scheduling in Grid Computing	969
<i>D.I. George Amalarethinam, P. Muthulakshmi</i>	
A New Fault Tolerant Routing Algorithm for Advance Irregular Alpha Multistage Interconnection Network	979
<i>Ved Prakash Bhardwaj, Nitin</i>	
Comparing and Analyzing the Energy Efficiency of Cloud Database and Parallel Database	989
<i>Jie Song, Tiantian Li, Xuebing Liu, Zhiliang Zhu</i>	
A Grid Fabrication of Traffic Maintenance System	999
<i>Avula Anitha, Rajeev Wankar, C. Raghavendra Rao</i>	
Intrusion Detection and QoS Security Architecture for Service Grid Computing Environment	1009
<i>Raghavendra Prabhu, Basappa B. Kodada, K.M. Shivakumar</i>	
Service Composition Design Pattern for Autonomic Computing Systems Using Association Rule Based Learning	1017
<i>Mohammed A.R. Quadri, Vishnuvardhan Mannava, T. Ramesh</i>	
An Enhancement to AODV Protocol for Efficient Routing in VANET – A Cluster Based Approach	1027
<i>M.C. Aswathy, C. Tripti</i>	
Human Emotion Recognition and Classification from Digital Colour Images Using Fuzzy and PCA Approach	1033
<i>Shikha Tayal, Sandip Vijay</i>	
A New Process Placement Algorithm in Multi-core Clusters Aimed to Reducing Network Interface Contention	1041
<i>Ghobad Zarrinchian, Mohsen Soryani, Morteza Analoui</i>	
Resource Based Optimized Decentralized Grid Scheduling Algorithm	1051
<i>Piyush Chauhan, Nitin</i>	

Web-Based GIS and Desktop Open Source GIS Software: An Emerging Innovative Approach for Water Resources Management	1061
<i>Sangeeta Verma, Ravindra Kumar Verma, Anju Singh, Neelima S. Naik</i>	
A Design Pattern for Service Injection and Composition of Web Services for Unstructured Peer-to-Peer Computing Systems with SOA	1075
<i>Vishnuvardhan Mannava, T. Ramesh, Mohammed A.R. Quadri</i>	
A Policy Driven Business Logic Change Management for Enterprise Web Services	1085
<i>M. Thirumaran, P. Dhavachelvan, G. Naga Venkata Kiran</i>	
Author Index	1095

Partitioning and Internetworking Wireless Mesh Network with Wired Network for Delivery Maximization and QoS Provisioning

Soma Pandey¹, Vijay Pande², Govind Kadambi³, and Stephen Bate⁴

¹ CMR Institute of Technology, Visveshwarya Technology University, Bangalore, India
soma.p@cmrit.ac.in

² Essel Adi Smart Grid Limited, Mumbai, India
vijay@esseladi.com

³ MS Ramaiah School of Advanced Studies, Bangalore, India
govind@msrsas.org

⁴ Coventry University, Priory Street, Coventry, U.K
esx064@coventry.ac.uk

Abstract. Wireless mesh architecture is a first step towards providing high-bandwidth network coverage. This architecture has major drawback of losing the bandwidth over multiple hops thereby resulting in poor quality of service (QoS) at nodes separated by more than two hops. This paper proposes a three step approach to guarantee bandwidth demand at each node of the network thereby providing high quality of service even to nodes separated by large distances from each other. The authors have presented a novel method for clustering the nodes and load sharing amongst the clusters based on graph partitioning approach. This work also presents a system and method of integrating Wireless Mesh Networks (WMN) with wired network for further increase in the QoS.

Keywords: Partitioning, Internetworking, Wireless mesh network, IEEE 802.11s.

1 Introduction

The wireless mesh network (WMN) is an emerging technology to extend the use of wireless communication. Mesh architecture sustains signal strength by breaking long distances into a series of shorter hops. Intermediate nodes not only boost the signal, but cooperatively make the forwarding decisions based on their knowledge of the network. Such architecture provides high network coverage, spectral efficiency, and economic advantage. Throughout the paper we use the IEEE 802.11s standard for infrastructure mode WMN. The authors have chosen IEEE802.11s as the WMN because major part of this work focuses on providing a wired backup to mesh nodes, and this standard already has a protocol defined for internetworking between the 802.11 and non 802.11 networks. But this work can be generalized to optimize any infrastructure based wireless mesh network. For this reason this work does not differentiate between a Mesh Point (MP) and Mesh Access Point (MAP) as separate entities as both are sources of bandwidth demand. Therefore hereafter both these entities will be called nodes whereas the Mesh Portal Point (MPP) will be called the gateway node.

At Layer 2 of WMN the crucial QoS parameter that can be delivered is the bandwidth demand of a node. It is a well known fact that wireless networks yield low throughput and poor QoS because they are bandwidth starved due to radio spectrum limitations. The authors suggest that if bandwidth demand at a node can be met; QoS constraints can be satisfied.

In subsequent sections the authors use a graph model of WMN to present a novel method of QoS provisioning in WMN. This work provides a three step approach to satisfy the bandwidth demand of all the nodes in a WMN.

- **Step I:** In order to share the load amongst gateways there has to be well defined clusters around the gateways. In [5] the authors have already provided a partitioning mechanism to create well defined clusters around gateways. This paper moves further and maps these partitions on to the Adjacency matrix of the graph model of WMN
- **Step II:** Using the concept of ‘*Supergraphs*’ authors proceed to share the load amongst the partitions dynamically. The load sharing algorithm ensures that the under loaded partitions share the load of their neighboring overloaded partitions under certain mathematically validated constraints. This algorithm also ensures that there is no need to re-compute the partitions every time a node transits to a neighboring partition for load sharing.
- **Step III:** In case the constraints defined for load sharing in step II are not satisfied and nodes are still bandwidth starved, then the authors provide the partitions with a wired network backup. This step defines the set of constraints to be observed while transiting a node to the wired network. Although the authors’ choice of network is Broadband over Powerline (BPL) but this work is not limited to BPL and is applicable to any wired Internet Protocol (IP) network. For more on BPL-WMN inter-networking refer [12]

Note: Due to space limitations the authors have kept the contents of this paper limited to defining the mathematical constraints and presenting an algorithm for partitioning, load sharing and wired network interworking. In their coming paper authors have presented a detailed protocol between the gateways, core router and Dynamic Host Configuration Protocol (DHCP) server which is to be observed while implementing these algorithms. The authors have presented a centralized protocol to implement these algorithms

2 Motivation and Related Work

WMN suffer from the limitation of throughput drop and bandwidth loss over multiple hops, Li et al. [4]. Reducing the distance between nodes in the WMN loses the very purpose of mesh networking which is to provide wider coverage area with minimal infrastructure. Robinson et al.[2] and Aoun et al. [3] propose to increase throughput by introducing multiple gateways . Placement of multiple gateways throughout the mesh does not always result in more throughput as proved by [5]. In previous literature Xie et al.,[7] and Bejerano et al.,[8] have suggested to improve this shortcoming by creating clusters around each gateway and then make provisions for load sharing amongst these clusters. Nandiraju et al.,[6] and Bejerano et al., [8] have pointed out

that by clustering the nodes in to non overlapping clique increases the throughput of networks. But partitioning the graphs in itself is an NP hard problem. With every change in load or transition of nodes amongst partitions, there is need to re-compute the partitions. This situation gets worse when nodes move to another network rather than another partition as the nodes are no longer a part of the same WMN. This is called ‘loss of wireless neighborhood’ problem. Current literature addresses this by emphasizing on the spanning tree computation periodically, thereby identifying all the nodes belonging to the same network. Partitioning the network every time with a changing load and node scenario is a serious problem as the whole network remains non-operational throughout this computation thereby reducing the network throughput. In this paper, authors have made a novel attempt to represent the network graph model in its adjacency matrix form. The adjacency matrix representation of the graph model of WMN preserves the neighborhood of each and every node, irrespective of whether it moves to a neighboring partition or to another network. The adjacency matrix representation eliminates the need to recreate partitions, even when the WMN is interworked with another network. This method increases the network throughput because it creates the partitions only once during the network design phase. Once the partitions are created, they are simply mapped onto the adjacency matrix of the graph model of WMN, which is done by our graph partitioning algorithm. Thereafter there is no need to continuously partition the network with changing load as in [7] and [8].

Contributions of this paper are

- One time partitioning of WMN.
- System and method to map partitions of a WMN on to the adjacency matrix of its graph model.
- System and method to transit nodes amongst these partitions directly using the adjacency matrix.
- System and method to interwork the WMN with another wired network and mapping the same onto the global adjacency matrix.
- Defining a set of mathematical bounds and constraints for load sharing and node transitions amongst partitions.
- A locally recursive algorithm for node selection and transition to the BPL or any other wired network.
- Nodes continue to remain part of WMN with their neighborhood preserved in the adjacency matrix, irrespective of whether they transit to another partition or to another wired network.

3 Notations and Assumptions

Let undirected planar Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represent a WMN ‘W’ with n number of nodes and gateways. The graph nodes and gateways represent the vertices and wireless links are represented by edges between the nodes. Self loops are not permitted. Let \mathcal{V} be the set of vertices v_1, v_2, \dots, v_n such that $|\mathcal{V}| = n$ and \mathcal{E} be the set of edges e_1, e_2, \dots, e_m such that $|\mathcal{E}| = m$. Total number of gateways/MPP is assumed to be k . Then, $\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_k$ will be the k distinct partitions of graph \mathcal{G} , each with one gateway.

Let P_i denote a node corresponding to the contracted subgraph \mathcal{G}_i in the supergraph of the partitioned \mathcal{G} . Let n_i be the number of vertices in \mathcal{G}_i . Let V_i be vertex set for \mathcal{G}_i $\forall i = 1 \dots k$. Let $\mathcal{A}(\mathcal{G})$ be the Adjacency matrix corresponding to graph \mathcal{G} and $\mathcal{A}(\mathcal{G}_i)$ be the adjacency matrix for partition \mathcal{G}_i .

$\mathcal{B}(\mathcal{G})$: Incidence matrix corresponding to graph \mathcal{G}

$\mathcal{B}(\mathcal{G}_i)$: Incidence matrix corresponding to the partition i of graph \mathcal{G}

$\mathcal{C}(\mathcal{G})$: Cycle matrix corresponding to graph \mathcal{G}

$\mathcal{C}(\mathcal{G}_i)$: Cycle matrix corresponding to the partition i of graph \mathcal{G}

Let C_i be the capacity of i^{th} gateway in partition i . $Q\{\mathcal{A}(\mathcal{G}_i)\}$ denotes QoS available at partition \mathcal{G}_i

R_i : current bandwidth demand (load) of partition \mathcal{G}_i

U_i : Upper working demand limit of QoS for partition \mathcal{G}_i

L_i : Lower working demand limit of QoS for partition \mathcal{G}_i

Under normal load conditions if demand of node n_i is d_i then total load of partition \mathcal{G}_i with n_i number of nodes is

$$R_i = \sum_{j=1}^{n_i} d_j \quad \forall i = 1, \dots, k$$

Overload of a partition is given by

$$\text{overload}(\mathcal{G}_i) = \begin{cases} 0, & \text{if } R_i < C_i \\ R_i - C_i, & \text{Otherwise} \end{cases}$$

4 Step I: Selective Partitioning

Selective Partitioning is called so because a graph is partitioned with certain constraints. The constraint in our case is that each partition must have exactly one gateway. This algorithm assumes that initial partitioning of WMN is already done. A WMN can be partitioned using any of the graph partitioning procedures available in literature [10]. Alternatively researchers can also use the node marking and partitioning algorithms presented by authors in [5] and [9]. First we present an observation on \mathcal{A}

$\mathcal{A}(\mathcal{G})$ can be written in block diagonal form as

$$\mathcal{A}(\mathcal{G}) = \begin{bmatrix} [\mathcal{A}(\mathcal{G}_1)] & & \cdots & & 0 \\ \vdots & [\mathcal{A}(\mathcal{G}_2)] & & & \vdots \\ 0 & \cdots & & \ddots & [\mathcal{A}(\mathcal{G}_k)] \end{bmatrix}_{n \times n}$$

Based on this observation we present the Algorithm I for selective graph partition

1. From WMN create $\mathcal{G}(v, e)$ with one gateway
2. From $\mathcal{G}(v, e)$ construct $\mathcal{A}(\mathcal{G})$.
3. Now take $[\mathcal{A}(\mathcal{G})]_{n \times n}$ and identify the 1st gateway of the WMN represented by $\mathcal{G}(v, e)$

4. Around the first gateway create a partition $[\mathcal{A}_{11}]_{n_1 \times n_1}$ by relabeling / visiting nodes and demand augmentation of nodes in $[\mathcal{A}(\mathcal{G})]$ against gateway capacity C_1 . Refer authors' paper [9] for complete procedure on node marking and relabeling.
5. Now identify the second gateway in $\mathcal{A}_2 = [\mathcal{A}] - \begin{bmatrix} \mathcal{A}(\mathcal{G}_1) & 0 \\ 0 & \ddots \end{bmatrix}$ now create $[\mathcal{A}_{22}]_{n_2 \times n_2}$. Relabel $\mathcal{A}(\mathcal{G})$ such that all the nodes adjacent to the second gateway have their corresponding rows and columns next to the second gateway row and column thereby creating the second matrix $[\mathcal{A}_{22}]$.
6. Now get $\mathcal{A}_3 = [\mathcal{A}] - \begin{bmatrix} \mathcal{A}(\mathcal{G}_1) & \cdots & 0 \\ \vdots & \mathcal{A}(\mathcal{G}_2) & \vdots \\ 0 & \cdots & \ddots \end{bmatrix}$ and create $[\mathcal{A}_{33}]_{n_3 \times n_3}$ by relabeling / node visiting fundamental on $[\mathcal{A}_3]$
7. Repeat step 4 to 6 till the last partition $[\mathcal{A}_k]$ is formed such that $\sum_{i=1}^k n_i = n$, this leads to k disjoint initial partitions
8. Hence WMN in the initial partition looks like

$$\mathcal{A}(\mathcal{G}) = \begin{bmatrix} [\mathcal{A}(\mathcal{G}_1)] & \cdots & & \\ \vdots & [\mathcal{A}(\mathcal{G}_2)] & & \vdots \\ & & \ddots & \\ & \cdots & & [\mathcal{A}(\mathcal{G}_k)] \end{bmatrix}$$
 obtained by relabeling and node visiting. Thus the adjacency matrix has all the partitions defined by their own adjacency matrix right across the diagonal. The other elements of the adjacency matrix are as before corresponding to the various rows and columns as in the initial adjacency matrix. Thus initial set of disjoint partitions in \mathcal{A} matrix is created
9. End

Note: The partitioned adjacency matrix of step 8 has to be created only once

From Algorithm I we get the WMN partitioned in k partitions each having one gateway. We denote the i^{th} partition by subgraph \mathcal{G}_i and its adjacency matrix by $\mathcal{A}(\mathcal{G}_i)$. It can be seen that $\mathcal{V}_i \cap \mathcal{V}_j = \emptyset \forall i, j \in [1 \dots k] \exists i \neq j$. Hence any vertex in \mathcal{G}_i can be made as gateway. Each \mathcal{G}_i will have n_i number of nodes. All these partitioned matrices must satisfy the condition $\mathcal{B}(\mathcal{G}_i) \times [\mathcal{C}(\mathcal{G}_i)]^T = \mathcal{C}_i \times [\mathcal{B}(\mathcal{G}_i)]^T = 0 \pmod{2}$ where superscript T denotes the transposed matrix [11].

5 Step II QoS Provisioning by Load Sharing

Before moving to the load sharing algorithm we define a few terms and formulate some theorems.

Definition I: A 'Cut set' \mathcal{E}_{ij} is set of all the edges between two partitioned subgraphs \mathcal{G}_i and \mathcal{G}_j of \mathcal{G} such that for each edge both its incident vertices belong to two different partitions \mathcal{G}_i and \mathcal{G}_j

Definition II: A 'Supergraph' \mathcal{G}^2 of \mathcal{G} is the graph obtained such that each vertex P_i of \mathcal{G}^2 represents the partition subgraph \mathcal{G}_i and edge e_i is the edge belonging to the cut set \mathcal{E}_{ij} . Fig1

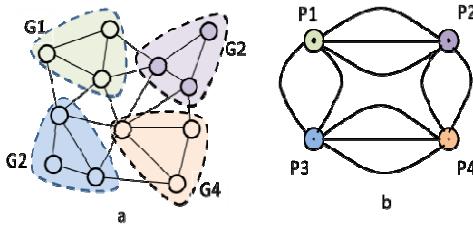


Fig. 1. a: A Graph G with its 4 partitions. **b:** Supergraph G^2 of G .

We present following set of properties of the Supergraphs

Property I: Since a WMN with k gateways will have k partitions therefore number of vertices in G^2 will be k .

Property II: G^2 will be a multigraph because there can be many edges for each pair of partition cut set.

Property III: G^2 can also be a complete graph of k vertices i.e. K_k graph. Consequently the following proof is needed.

Lemma 1: If G^2 is a super graph of G and if G is planar G^2 will also be planar

Proof: By contradiction let us assume that G^2 is non planar. Then G^2 will have intersecting edges. Now since G is contracted to G^2 . This implies that G also has intersecting edges. As a result G has to be non planar. Thus by contradiction since $G(v, e)$ is a planar therefore $G^2 \cong G(v, e)$ is also planar. **(Q.E.D)**

The implication of the above lemma is for the reinforcement of fact that any node transition from one partition to another does not contradict the planar structure of the graph. The transition of nodes can happen from one partition to another if the two partitions are neighbors. Partition/node at one hop distance are called neighbors.

In \mathcal{A} we retain only the partitions on the diagonal and replace remaining elements by 0. This matrix we call as \mathcal{A}' . Then,

$$\Rightarrow \mathcal{A}'(\mathcal{G}) = \begin{bmatrix} & \leftarrow n_1 \rightarrow & n_2 & \rightarrow \dots \rightarrow & n_k & \leftarrow \\ & [\mathcal{A}(\mathcal{G}_1)] & & & & \frac{n_1}{n_1} \\ & & [\mathcal{A}(\mathcal{G}_2)] & & & \frac{n_2}{n_2} \\ & & & \ddots & & \vdots \\ & & & & [\mathcal{A}(\mathcal{G}_k)] & \frac{n_k}{n_k} \end{bmatrix}$$

Here n_i is number of vertices in partition subgraph \mathcal{G}_i . Let, $\mathcal{A}''(\mathcal{G}) = \mathcal{A}(\mathcal{G}) - \mathcal{A}'(\mathcal{G})$. Then,

$$\mathcal{A}''(\mathcal{G}) = \mathcal{A}(\mathcal{G}) - \mathcal{A}'(\mathcal{G}) = \begin{bmatrix} [0] & [\mathcal{A}]_{n_1 \times n_2} & [\mathcal{A}]_{n_1 \times n_3} & \dots & [\mathcal{A}]_{n_1 \times n_k} \\ [\mathcal{A}]^T_{n_1 \times n_2} & [0] & \ddots & & \vdots \\ \vdots & & & \ddots & \\ [\mathcal{A}]^T_{n_k \times n_k} & \dots & & [\mathcal{A}]_{n_{k-1} \times n_k} & [0] \end{bmatrix}$$

Proposition 1: \mathcal{P}_i has ‘ q ’ paths to \mathcal{P}_j iff there are q number of non zero entries in $[\mathcal{A}]_{n_i \times n_j}$ of matrix $\mathcal{A}''(\mathcal{G})$

Proof: The number of non zero entries in $[\mathcal{A}]_{n_i \times n_j}$ is the edge cut set of \mathcal{P}_i , and \mathcal{P}_j . Now without loss in generality, \mathcal{P}_i can be termed as node, $\forall i = 1 \dots k$ and if \mathcal{P}_i is neighbor of \mathcal{P}_j then it can be joined by edges from their edge cut set.

If $\mathcal{A}(\mathcal{G}_i)$ is operating at its limit U_i then node need to be transited to $\mathcal{A}(\mathcal{G}_j)$ operating at limit L_j . In the next proposition we define the operation in order to balance the load on the \mathcal{G}^2 graph.

Proposition 2: For any partition \mathcal{P}_i if the cumulative QOS requirements are not met then following operations can be performed

Transit the node to the neighboring partition.

Create a partition in the sub partition

Proof: Consider the k partitions as created earlier. These partitions operating under normal load must satisfy following condition

$$[Q(\mathcal{A}(\mathcal{G}_i))]_{n_i} \geq Q(\mathcal{A}(\mathcal{G}_i))|_{n_i} \geq [Q(\mathcal{A}(\mathcal{G}_i))] \quad \forall i \in [1 \dots k]$$

This is same as $U_i \geq R_i \geq L_i$ (normal load constraint)

To satisfy the QoS requirement if a node is shifted to powerline network, then it is same as further partitioning of partition $\mathcal{A}(\mathcal{G}_i)$. Let us assume that $\mathcal{A}(\mathcal{G}_i)^j$ is the j^{th} partition of \mathcal{G}_i . Likewise, if there are ω sub partitions of $\mathcal{A}(\mathcal{G}_i)$ then the following condition holds good

$$Q(\mathcal{A}(\mathcal{G}_i)) = \sum_{j=1}^{\omega} Q(\mathcal{A}(\mathcal{G}_i^j))$$

where

$$\sum_{i=1}^k n_i = n = \text{total number of nodes}$$

Now if $Q(\mathcal{A}(\mathcal{G}_i))|_{n_i} \geq [Q(\mathcal{A}(\mathcal{G}_i))]_{n_i}$ then following can be carried out

1. Transiting q nodes of partition i to neighboring partition such that following condition holds good

$$Q(\mathcal{A}(\mathcal{G}_i))|_{n_i-q} \leq [Q(\mathcal{A}(\mathcal{G}_i))]_{n_i}$$

2. Create a partition in \mathcal{G}_i on n_i nodes such that $[Q(\mathcal{A}(\mathcal{G}_i))]^{n_i} \geq Q(\mathcal{A}(\mathcal{G}_i))|_{n_i-p}$ and $[Q(\mathcal{A}(\mathcal{G}_i^p))] > Q(\mathcal{A}(\mathcal{G}_i))|_p$ (For example: This means p nodes in n_i partition are put on to power line network.)

Each node now can be represented in the \mathcal{G}^2 graph with k nodes as $\mathcal{P}_1, \dots, \mathcal{P}_k$ and corresponding \mathcal{A} matrix as $\mathcal{A}(\mathcal{G}_i)$ and dynamic load as $Q(\mathcal{A}(\mathcal{G}_i))$ where $i = 1 \dots k$. This means that $Q(\mathcal{A}(\mathcal{G}_i)) = Q(\mathcal{P}_i)$. Now each \mathcal{P}_i can either be underloaded or over-loaded as mentioned before. So \mathcal{P}_i belongs to either U_i or L_i hence in graph $\mathcal{G}^2(Q(\mathcal{P}_i) \in \{U_i, L_i\}, e) \forall i = \{1 \dots k\}$

Theorem 1: Consider Graph \mathcal{G}^2 ($\{U,L\}$, e), where U is set of vertices operating at overloaded condition and L is set of nodes operating at under load condition with P_1, \dots, P_k nodes and also consider that for P_1, \dots, P_k , theorem 2 is satisfied; then with k partitions the load can be balanced by approach of node transition iff their bipartite graph with U and L exist for \mathcal{G}^2 .

Proof: If overloaded nodes (partitions) can transit nodes inside the partition such that the under loaded partition will have more to accommodate as compared to the loaded partition hence such condition becomes the necessary condition. The proof of sufficiency follows from the contradiction. Consider that \mathcal{G}^2 ($\{U,L\}$, e) is not bipartite then it means that one of the overloaded partition nodes P_i is in neighborhood of another overloaded partition P_j . Thus the transition of nodes from one overloaded partition (node) to the other overloaded partition can be expected. Hence **Bipartite graph formation between U_i and L_i nodes is necessary and sufficient condition for node transitions in \mathcal{G}^2 graph.**

5.1 Constraints to Be Followed for Node Transitions from One Partition to Another

There are three major constraints which must be followed to enable movement of a node from one partition \mathcal{G}_i to another neighboring partition \mathcal{G}_j

1. $R_i \geq U_i$
2. $R_j \leq L_j$
3. \mathcal{G}^2 ($\{U,L\}$, e) must be bipartite between U and L

Observation I: A node p transiting from one partition i to another partition j needs only relabeling within the global adjacency matrix \mathcal{A} such that row and column corresponding to node p in \mathcal{A} moves from $\mathcal{A}(\mathcal{G}_i)$ to $\mathcal{A}(\mathcal{G}_j)$. Since this transition is only affecting the active/passive table entries of partitions i and j gateways the whole network need not be defunct, only partitions i and j can stop their operations. There is also no need to re-compute the partitions as in earlier cases.

6 Step III: Mathematical Constraints for Node Transition to Powerline

In this step we prove that introducing a powerline network to a node within the WMN is analogous to the partitioning procedure performed recursively. The introduction of the powerline network to any node within the WMN is defined as two part process. First we define constraints on identifying the node which can be moved to powerline. Secondly we explain how the node remains connected to the WMN and preserves its integrity even when it is on another wired network. This will ensure that the node can be recalled to exactly the same location within same neighborhood in spite of its association to a wired network which in this case is BPL.

6.1 Part I: Which Node to Be Moved to Wired Network?

Constraint I: The node must be chosen from a partition \mathcal{G}_i such that $R_i \geq U_i$ (\mathcal{G}_i is overloaded)

Constraint II: Moving a vertex to powerline/wired network means deletion of all edges incident on it. After deletion of all such edges the following condition must hold true

$$\forall i, j = 1 \text{ to } k \& i \neq j, \mathcal{E}_{ij} \neq \emptyset$$

Moving a node p to power line is analogous to rendering its wireless links inactive. When this action is mapped onto the graph model it results in removing the edges incident on node p . Thereafter if this removal of edges, results in a null cut set then the operation cannot be performed because this results in creation of disconnected components within the supergraph \mathcal{G}^2 . A disconnected supergraph implies a disintegrated WMN. With this we present our next proposition.

Lemma 2: *A node can be shifted to the powerline network iff its removal does not create disintegrated WMN having disconnected partitions.*

Proof: Proof lies in the above mentioned explanation of constraint II. Let a node n_p have t edges incident on to t number of nodes in partition i . If there are q paths from partition i to partition j then removal of a node n_p to powerline network may result in $q-t$ paths. If $q-t \neq 0$ then this operation is valid else the node n_p has to remain on wireless medium i.e. on the mesh network.

6.2 Part II: How Will This Node Move to the Wired Network While Preserving the WMN Integrity?

We need to create a framework to accommodate the Powerline network (any other wired network). It can be seen in the above formulation that the same exercise can be carried out for each partition. Hence the implication for the internetwork is presented below.

Proposition 3: *A partitioned matrix of form*

$$\begin{array}{c} \leftarrow n_1 \rightarrow n_2 \rightarrow \dots \rightarrow n_k \leftarrow \\ \left[\begin{array}{c} \mathcal{A}(\mathcal{G}_1) \\ \vdots \\ \mathcal{A}(\mathcal{G}_2) \\ \vdots \\ \mathcal{A}(\mathcal{G}_k) \end{array} \right] \end{array} \quad \begin{array}{c} \frac{n_1}{n_2} \\ \vdots \\ \frac{n_k}{n_k} \end{array}$$

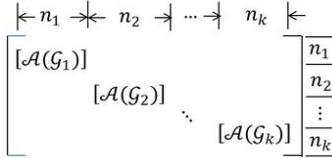
Can be partitioned recursively to the following form

$$\mathcal{A}'(\mathcal{G}) = \left[\begin{array}{ccccc} \leftarrow & n_1 & \rightarrow & n_2 & \rightarrow \dots \leftarrow & n_k & \rightarrow \\ \leftarrow & \xrightarrow{p} & \xrightarrow{q} & \xrightarrow{r} & \dots & \xrightarrow{s} & 0 \\ \left[\begin{array}{c} \mathcal{A}(\mathcal{G}_1^1) \\ \vdots \\ \mathcal{A}(\mathcal{G}_1^2) \end{array} \right] & \left[\begin{array}{c} \mathcal{A}(\mathcal{G}_1^2) \\ \vdots \\ \mathcal{A}(\mathcal{G}_2) \end{array} \right] & \dots & \left[\begin{array}{c} \mathcal{A}(\mathcal{G}_2) \\ \vdots \\ \mathcal{A}(\mathcal{G}_k) \end{array} \right] & \dots & \left[\begin{array}{c} \mathcal{A}(\mathcal{G}_k) \\ \vdots \\ 0 \end{array} \right] \end{array} \right]$$

Without loss of generality and integrity of the original adjacency matrix of graph \mathcal{G}

Proof:

We get partitioned $\mathcal{A}'(\mathcal{G})$ in following form



Now consider that even after partitioning, as per Algorithm I and proposition 2, QoS in $\mathcal{A}(\mathcal{G}_1)$ is not achieved, then without any structural changes matrix of figure 2

$$\mathcal{A}'(\mathcal{G}) = \begin{bmatrix} \leftarrow n_1 \rightarrow \leftarrow n_2 \rightarrow \cdots \leftarrow n_k \rightarrow \\ \leftarrow p \rightarrow \begin{matrix} [\mathcal{A}(\mathcal{G}_1^1)] & [\mathcal{A}(\mathcal{G}_1^2)] & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & [\mathcal{A}(\mathcal{G}_k)] & \end{matrix} \end{bmatrix}$$

Fig. 2. Matrix after moving q nodes of partition 1 to powerline

is derived.

Which means within $\mathcal{A}(\mathcal{G}_1)$ further partition is created such that q nodes of \mathcal{G}_1 are shifted to the powerline network such that $p + q = n_1$. Here we emphasize the fact that moving selected nodes from WMN to a wired network is as simple as creating a sub-partition within a partition which is overloaded. This can be performed recursively taking into consideration only the overloaded partition in question. More precisely moving ω nodes of partition \mathcal{G}_i to powerline is similar to creating ω partitions within \mathcal{G}_i such that

$$Q(\mathcal{A}(\mathcal{G}_i)) = \sum_{j=1}^{\omega} Q(\mathcal{A}(\mathcal{G}_i^j))$$

Where n_i is the total no of nodes within the overloaded partition \mathcal{G}_i .

(QED)

This method does not have the drawbacks listed by the authors for methods given by [6],[7] and [8]. As can be seen here an overloaded partition can very simply move selected nodes to the powerline network without having to compute the global spanning tree. This approach provides *locally recursive partitioning* method where in, computations are limited to only n_i vertices (number of vertices in i^{th} partition) as against total number of n vertices in normal methods.

7 Final Algorithm II

Before stating the algorithm we define a node to be ‘heavy’ if its demand is the highest within a partition. On the contrary a node is ‘light’ if its bandwidth demand is least

within the partition. Likewise a partition is heavy or light if its bandwidth demand is greater than U_i or less than L_i respectively. This definition can be generalized throughout \mathcal{G} .

7.1 Algorithm II

1. Partition the WMN using the adjacency matrix and Algorithm I
2. Transit heavy nodes to lighter partitions as per constraints defined in section 5.1 and maintain the mesh balance.
3. Repeat step 2 until the *node transition constraints* (section 5.1) are not satisfied.
4. From the overloaded partition select a node which is heaviest and satisfies the Broadband Over Powerline (BPL) *BPL constraints* (section 6.1)
5. Move this node to the BPL and check the *overload constraint* defined in proposition 2 (section 5).
6. Repeat steps 4 to 6 till the partition under consideration does not remain overloaded.
7. Go to Step 2.

Time Complexity

Since partitioning and mapping the partitions on to the \mathcal{A} matrix is a onetime operation performed only initially, we only consider the active time complexity of the whole procedure. Once the WMN has its partitioned \mathcal{A} matrix in place (algorithm I), the only operation taking place is relabeling which occurs every time there is a node transition from one partition to another. **Time complexity for relabeling operation is only O(n)!** Therefore this research has reduced the partitioning and load balancing procedure time complexity from NP hard to a simple polynomial time complexity. The main reason behind this reduction of time complexity is the requirement of the NP hard partitioning procedure to take place only once. Thereafter the whole procedure is based only on relabeling.

8 Results and Comparison

The simulations were performed on Matlab and Simulink. Nearest Gateway (NGW) solution is the current method used in the multihop multi gateway WMN models where nodes attach to their nearest gateways calculated by shortest path. Minimum Load Index (MLI) builds upon the NGW solution, improving it [15]. Our method with multiple gateways differs from the above two in terms of well defined clustering/ partitioning. Hence each gateway has table which is well defined in size due to the partitions created around them. We assume that the DHCP (Dynamic Host Configuration Protocol) router has significantly large bandwidth in comparison with the gateway bandwidth. Instead of nodes polling for the gateways and calculating the shortest paths to each gateway, here the gateway scheduling is centralized from the DHCP core router. This minimizes the backend traffic because of the elimination of computation of shortest path to all gateways at each packet delivery and reduction of route request packets to all gateways.

Table 1. Simulation Parameters

Parameter	Value
Number of gateways	Varied from 1 to 5
Number of nodes	Varying from 10 to 50
Maximum number of mesh clients per node	250 to 300
Mean Packet Arrival Rate	0.01s
Mean hop delay	0.01s
Retry Threshold	0.01s
Flow Rate (CBR/UDP) Flow half rate 67.2 kbps	Lognormal distribution, $\mu = \log(67)$, $\sigma = 0.4$
Simulation consideration	one flow rate with all nodes
Packet Size	512 bytes

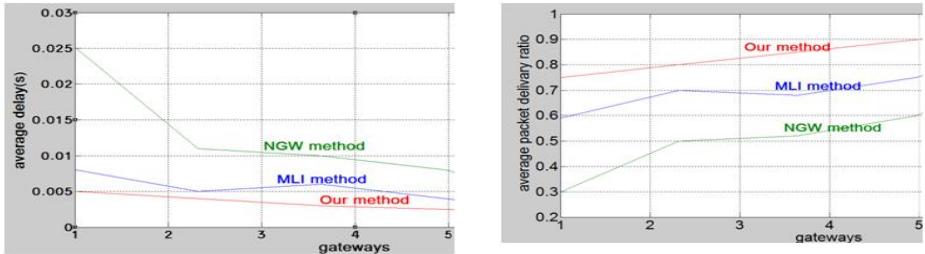
**Fig. 3.** Average delay and packet delivery ratio in WMN with increasing number of gateways, and fixed number of 50 nodes.

Figure 3 presents the average delay and packet delivery ratio in WMN with increasing number of gateways keeping number of nodes fixed to 50. Our method is presented by the red color whereas the green color represents the NGW technique and the blue color is the MLI technique. Figure 4 presents the average delay and packet delivery ration with increasing number of nodes keeping the number of gateways fixed to 5. As a consequence of the increase in the number of nodes there will be more number of nodes in each partition.

Multiple gateways without clustering result in managing bigger tables at gateways hence backend traffic becomes significantly higher across the WMN. In contrast in our method backend traffic is restricted within the gateway span/partition/cluster. The active routing table entry at the gateways is significantly small because it manages only the details of nodes within the gateways partition. All other global entries of the mesh can be moved over to the passive table of each gateway.

It can be also be seen that in our method that neighborhood is managed using only adjacency matrix hence search direction is restricted and scoped by the partition/cluster range. Similarly a partition's neighborhood is also fixed and well defined in the supergraph. This ensures integrity and connectivity of the WMN. Because of these salient issues the results outperforms other methods.

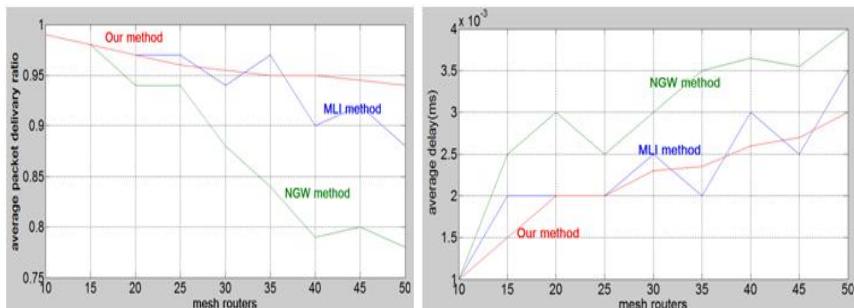


Fig. 4. Average delay and packet delivery ratio in WMN with increasing number of nodes and fixed number of 5 gateways

9 Conclusions

The authors have provided a very simple load sharing and node transition method to achieve QoS in WMN. This method fares better than all the previous methods due to the following factors

1. It reduces computation from global graph topology to very small local partitions.
2. Nodes moving to powerline are managed within local partitions thereby eliminating any chances of lost nodes.
3. At any point of time, nodes can be recalled back to WMN and the original topology of WMN can be restored, thereby ensuring the integrity of WMN.
4. It establishes a mechanism which changes the perspective of an NP hard partitioning and load sharing problem to that of a Polynomial time complexity one.

References

1. Mase, K.: Layer 3 Wireless Mesh Networks: Mobility Management Issues. *IEEE Communications Magazine*, 156–163 (July 2011)
2. Robinson, J., Uysal, M., Swaminathan, R., Knightly, E.: Adding capacity points to a wireless mesh network using local search. In: *Proceedings of IEEE INFOCOM*, Phoenix, AZ (April 2008)
3. Aoun, B., Boutaba, R., Iraqi, Y., Kenward, G.: Gateway Placement Optimization in Wireless Mesh Networks With QoS Constraints. *IEEE Journal on Selected Areas in Communications* 24(11), 2127–2136 (2006). doi:10.1109/JSAC.2006.881606
4. Li, J., Blake, C., De Couto, D., Lee, H.I., Morris, R.: Capacity of Ad Hoc Wireless Networks. In: *International Conference on Mobile Computing and Networking* (2001)
5. Pande, S., Kadambi, G., Bates, S., Pande, V.: A Load Sharing and Partitioning System for Multihop Wireless Mesh Network with Multiple Gateways. In: *IEEE International Conference on Open Systems* (September 2011)
6. Nandiraju, D., Santhanam, L., Nandiraju, N., Agrawal, D.P.: Achieving Load Balancing in Wireless Mesh Networks Through Multiple Gateways. In: *IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS)*, pp. 807–812 (2006)

7. Xie, B., Yu, Y., Kumar, A., Agrawal, D.P.: Load-balancing and Interdomain Mobility for Wireless Mesh Networks. In: Global Telecommunications Conference, GLOBECOM 2006, pp. 409–414. IEEE (2006)
8. Bejerano, Y., Han, S.-J., Kumar, A.: Efficient load-balancing routing for wireless mesh networks. *Comput. Netw.* 51(10), 2450–2466 (2007)
9. Pande, S., Pande, V.: A Node Marking Algorithm for Partitioning Wireless Mesh Networks. In: IEEE Conference on Open Systems (2011)
10. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for partitioning graphs. *Bell Systems Technical Journal* 49(2), 291–308 (1970)
11. Pandey, S., Pande, V., Kadambi, G.: Optimizing Delivery Mechanisms in Wireless Mesh Network with QOS Constraints. *International Journal of Computer Applications* (April 2010), <http://www.ijcaonline.org/archives/number14/310-477>
12. Pande, S., Pande, V., Kadambi, G.: Integrating Power line and Wireless for Intelligent and Opportunistic Networking. *International Journal of Computer Applications* (April 2010), <http://www.ijcaonline.org/archives/number25/448-749>
13. Salem, N.B., Hubaux, J.P.: A Fair scheduling for wireless mesh networks. In: Proceedings of the 1st IEEE Workshop on Wireless Mesh Networks (2005)
14. Ernst, J.B., Denko, M.K.: Fair scheduling with multiple gateways in WMN. In: International Conference on Advanced Information Networking and Applications, AINA2009 (2009)
15. Huang, C.F., Lee, H.W., Tseng, Y.C.: A two tier heterogeneous mobile adhoc network architecture and its load-balance routing problem. *Mobile Networks and Applications* 9(4), 379–391 (2004)

A New Secret Key Cipher: C128

Indrajit Das¹ and R. Saravanan²

¹ SCSE, VIT University, Vellore-632014, TN, India

² SITE, VIT University, Vellore-632014, TN, India

indrajit.das23@gmail.com, rsaravanan@vit.ac.in

Abstract. This paper describes a new secret key cryptosystem named CIPHER128 (C128). The algorithm is based on a feistel structure which encrypts a 128 bit block. It demonstrates few effective features like multiple S-Boxes, variable plain text size depending on data and key, padding by random variables, data and key dependent cyclic shift operations and variable length key.

Keywords: Symmetric Key Cryptography, S-Box, Substitution, Permutation, Feistel Networks, Confusion, Diffusion, Avalanche Effect.

1 Introduction and Motivation

Our aim of a new secret key cryptosystem was primarily to design an algorithm which satisfies the following conditions which would ensure quality encryption.

- The algorithm should be simple and easy to implement.
- It should use a variable size key. The key space should be large compared to existing algorithms.
- It should be able to encrypt a block of variable plaintext size. The size of the plaintext to be encrypted should be determined on a dynamic basis.
- It should guarantee substitution and permutation for all bits in each round.
- It should guarantee some randomness which will cause an unpredictable change on every encryption.
- It should be able to use data dependent operations.
- It should be a word-oriented algorithm.
- It should use operations which are comparatively efficient on processors.
- Bring in more and more parallelism in the algorithm for increasing the efficiency.

2 Proposed Design: C128

The C128 algorithm is a 128 bit block cipher. These 128 bits are a combination of plain text and random values selected by the sender. To explain this the block size of 128 bits, is divided into two parts

- *Part 1:* It comprises of the x higher order bits of the 128 bit block. These are plain text bits.
- *Part 2:* This part includes the remaining $128-x$ bits which are a combination of plain text bits and random bits selected by the sender.

The selection of the lower order $128-x$ i.e. *part 2* bits depends on the x higher order bits i.e. *part 1* and the key. We calculate N where

$$N = (\text{plain text bits} + \text{Key}) \text{ MOD } (128-x).$$

In the next $128-x$ bits i.e. *part 2* of the block, N bits are subsequent plaintext bits and rest $128-x-N$ bits are random bits selected by the encoder. The same calculation is done by the decoder and the last $128-x-N$ bits are discarded after decryption.

The value of x is pre determined by the sender and receiver. Theoretically the value of x lies in the range $0 \leq x \leq 128$ but if we select $x = 0$ then no plaintext bits are selected for *part 1*. The size of *part 2* will be 128 bits. $N = (\text{Key}) \text{ MOD } 128$ will result in N bit plaintext and $128-N$ random bits. Since the key is same for subsequent encryptions, the size of plain text and random data will be the same. Further, the value of x will determine the speed of the algorithm. For e.g. if $x = k$, it would mean that the algorithm will always encrypt at least k plain text bits and the larger value of $128-k$ will result in greater security.

This feature increases the security of the algorithm since the attacker, despite of the knowledge of the algorithm, is not able to predict the actual size of the plain text being encrypted. It can be observed that during encryption, every time the algorithm encrypts a block of different size and the random bits used for padding results in a random change in the value of data in every round. When the algorithm is used in *Cipher Block Chaining* mode (CBC) the *Initialization Vector* (IV) can be used to influence the value of N .

There are total 10 rounds and each round uses 8 sub keys hence 80 sub keys are required each for both encryption and decryption. The minimum length of key is 256 bits and the upper limit is 2560 bits which is very large making it unrealistic for a brute force attack.

The keys are calculated before encryption or decryption.

- Each Sub Keys are of 32 bit and are named as:
 $\text{SK}_0, \text{SK}_1, \text{SK}_2, \text{SK}_3, \dots, \text{SK}_{79}$
- There are four 32-bit S-boxes with 256 entries each:
 $\text{S}_{1,0}, \text{S}_{1,1}, \text{S}_{1,2}, \dots, \text{S}_{1,255}$
 $\text{S}_{2,0}, \text{S}_{2,1}, \text{S}_{2,2}, \dots, \text{S}_{2,255}$
 $\text{S}_{3,0}, \text{S}_{3,1}, \text{S}_{3,2}, \dots, \text{S}_{3,255}$
 $\text{S}_{4,0}, \text{S}_{4,1}, \text{S}_{4,2}, \dots, \text{S}_{4,255}$

2.1 Encryption Process

For encryption there are 10 identical rounds. The 128 bit block is divided into four 32 bit words. These 4 words are subjected to a feistel network where a sequence of Substitution and Permutation operations are carried out. The algorithm uses four

S-Boxes for substitution. A non-reversible functions F is used to select 32 bit values from the four S-Boxes. The permutation process is carried out using two shift operations 2WordMix and 4WordMix.

2WordMix: It takes as input the 128 bit (4 words) block and uses two sub keys to perform permutation on 1st, 2nd and 3rd, 4th words separately. The 128 bit block is viewed as a concatenation of 16 bytes. These bytes are named as B1, B2, B3... B16. These 16 bytes are used to form 4 words by concatenation as follows

- Word P = B1 || B3 || B5 || B7
- Word Q = B2 || B4 || B6 || B8
- Word R = B9 || B11 || B13 || B15
- Word S = B10 || B12 || B14 || B16

These 4 words P, Q, R and S then undergo a series of cyclic shift operations.

- P = (P XOR SK8i+2) << (SK8i+3 XOR Q XOR R XOR S)
- Q = (Q XOR SK8i+3) << (SK8i+2 XOR P XOR R XOR S)
- R = (R XOR SK8i+2) << (SK8i+3 XOR P XOR Q XOR S)
- S = (S XOR SK8i+3) << (SK8i+2 XOR P XOR Q XOR R)

The cyclic shift operation enables a mixing of bits between corresponding bytes of the words A, B and C, D used in the feistel networks.

4WordMix: It takes as input the 128 bit (4 words) block and performs byte permutation for the complete block. The 128 bit block is viewed as a concatenation of 16 bytes. These bytes are named as B1, B2, B3... B16. These 16 bytes are used to form 4 words by concatenation as follows

- Word P = B1 || B5 || B9 || B13
- Word Q = B2 || B6 || B10 || B14
- Word R = B3 || B7 || B11 || B15
- Word S = B4 || B8 || B12 || B16

These 4 words P, Q, R and S then undergo a series of cyclic shift operations.

- P = (P XOR SK8i+6) << (SK8i+7 XOR Q XOR R XOR S)
- Q = (Q XOR SK8i+7) << (SK8i+6 XOR P XOR R XOR S)
- R = (R XOR SK8i+6) << (SK8i+7 XOR P XOR Q XOR S)
- S = (S XOR SK8i+7) << (SK8i+6 XOR P XOR Q XOR R)

The cyclic shift operation enables a mixing of bits between corresponding bytes of the words A, B, C and D used in the feistel networks.

The algorithm for encryption round is as follows.

Encryption_C128 (128 bit data block X) /*Round i*/.

{

Step1. 128 bit data block X is divided into four 32 bit blocks: A, B, C and D.

Step2. A = A XOR SK8i+0 and B = B XOR SK8i+1

Step3. C = C XOR F (A) and D = D XOR F (B)

Step4. 2WordMix (A, B, C, D, SK8i+2, SK8i+3)

Step5. $C = A \text{ XOR } SK8i+4$ and $D = D \text{ XOR } SK8i+5$

Step6. $B = B \text{ XOR } F(A)$ and $C = C \text{ XOR } F(D)$

Step7. 4WordMix ($A, B, C, D, SK8i+6, SK8i+7$)

Step8. Return X

}

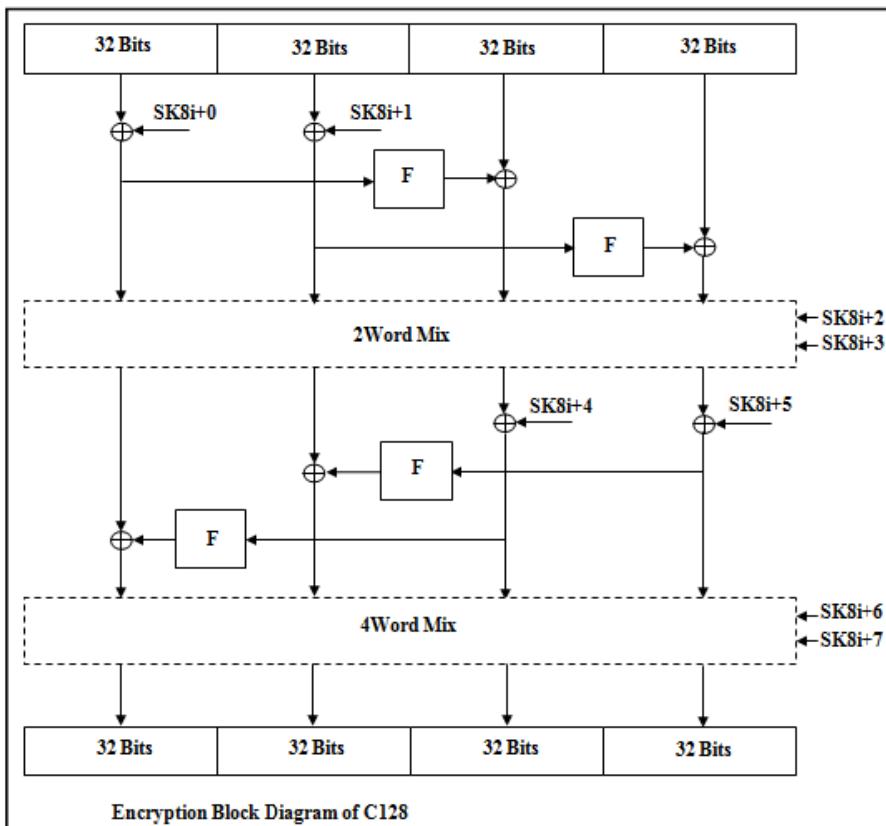


Fig. 1.

2.2 Decryption Process

For decryption there are 10 identical rounds. The permutation process that was carried out using two shift operations 2WordMix and 4WordMix are now reversed using Inv2WordMix and Inv4WordMix.

Inv2WordMix: It takes as input the 128 bit block. The 128 bit block is viewed as four words P, Q, R and S. These 4 words then undergo a series of cyclic shift operations.

- $S = (S \gg (SK8i+2 \text{ XOR } P \text{ XOR } Q \text{ XOR } R)) \text{ XOR } SK8i+3$
- $R = (R \gg (SK8i+3 \text{ XOR } P \text{ XOR } Q \text{ XOR } S)) \text{ XOR } SK8i+2$
- $Q = (Q \gg (SK8i+2 \text{ XOR } P \text{ XOR } R \text{ XOR } S)) \text{ XOR } SK8i+3$
- $P = (P \gg (SK8i+3 \text{ XOR } Q \text{ XOR } R \text{ XOR } S)) \text{ XOR } SK8i+2$

The 128 bit block is viewed as a concatenation of 16 bytes named as B1, B2, B3...B16. The bytes are used to form four words by concatenation as follows:

- Word P = B1 || B5 || B2 || B6
- Word Q = B3 || B7 || B4 || B8
- Word R = B9 || B13 || B10 || B14
- Word S = B11 || B15 || B12 || B16

The output is the concatenation of the words P, Q, R and S.

Inv4WordMix: It takes as input the 128 bit block. The 128 bit block is viewed as four words P, Q, R and S. These 4 words undergo a series of cyclic shift operations.

- $S = (S \gg (SK8i+6 \text{ XOR } P \text{ XOR } Q \text{ XOR } R)) \text{ XOR } SK8i+7$
- $R = (R \gg (SK8i+7 \text{ XOR } P \text{ XOR } Q \text{ XOR } S)) \text{ XOR } SK8i+6$
- $Q = (Q \gg (SK8i+6 \text{ XOR } P \text{ XOR } R \text{ XOR } S)) \text{ XOR } SK8i+7$
- $P = (P \gg (SK8i+7 \text{ XOR } Q \text{ XOR } R \text{ XOR } S)) \text{ XOR } SK8i+6$

The 128 bit block is viewed as a concatenation of 16 bytes. These bytes are named as B1, B2, B3... B16. These 16 bytes are used to form 4 words by concatenation as follows

- Word P = B1 || B5 || B9 || B13
- Word Q = B2 || B6 || B10 || B14
- Word R = B3 || B7 || B11 || B15
- Word S = B4 || B8 || B12 || B16

The output is the concatenation of the words P, Q, R and S.

The algorithm for decryption round is as follows

```
Decryption_C128 (128 bit data block X) /*Round i*/
```

```
{
```

Step1. 128 bit data block B is divided into four 32 bit blocks: A, B, C and D

Step2. Inv4WordMix (A, B, C, D, SK8i+6, SK8i+7)

Step3. A = A XOR F (C) and B = B XOR F (D)

Step4. D = D XOR SK8i+5 and C = A XOR SK8i+4

Step5. Inv2WordMix (A, B, C, D, SK8i+2, SK8i+3)

Step6. D = D XOR F (B) and C = A XOR F (A)

Step7. B = B XOR SK8i+1 and A = A XOR SK8i+0

Step8. Return X

```
}
```

The block diagram for encryption is given in figure 3

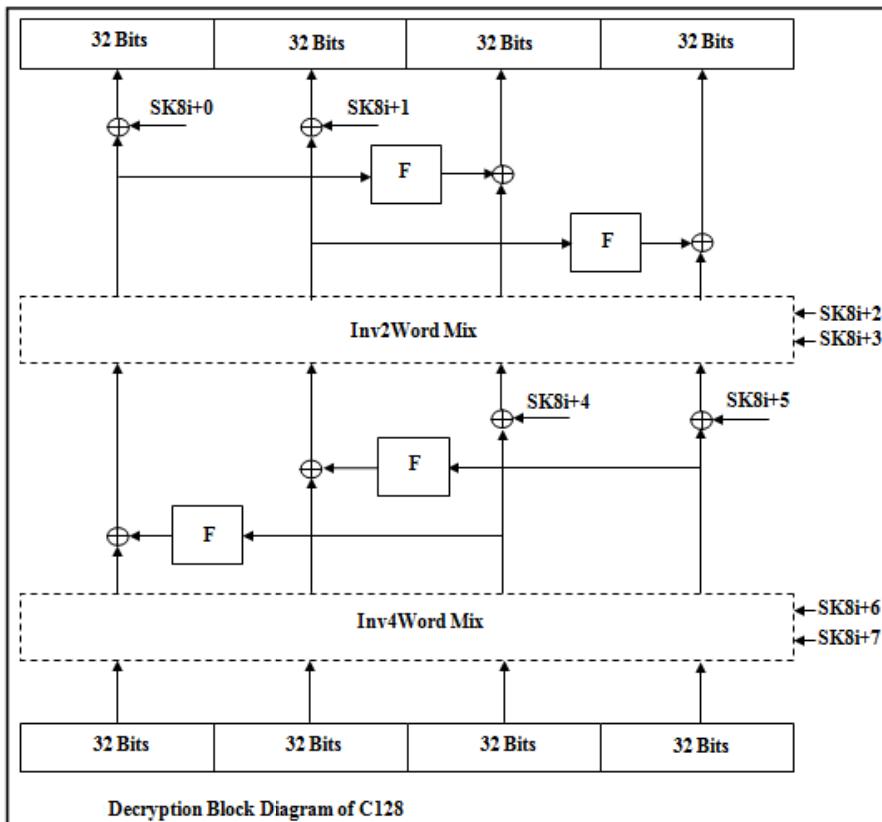


Fig. 2.

2.4 Function F

The most important part of a Feistel Cipher is the non-reversible function. This function uses simple operations like XOR and Modular Addition. The algorithm for function F is given below

```
Function_F_C128 (32 bit block)
{
    32 bit block is divided into four 8 bit blocks: A, B, C and D
    Val_A= S_BOX_1 [A]
    Val_B= S_BOX_2 [B]
    Val_C= S_BOX_3[C]
    Val_D= S_BOX_4 [D]
    Val_AB= Val_A XOR Val_B
    Val_CD= Val_C XOR Val_D
    Val_F = (Val_AB+Val_CD) mod 32
    Return Val_F
}
```

2.5 Function G

It is a non reversible function which takes as input a 32 bit block and gives two 32 bit values as output. This function is used in generating the sub keys and S-Boxes.

```
Function_G_C128 (32 bit block)
{
    32 bit block is divided into four 8 bit blocks: A, B, C and D
    Val_A= S_BOX_1 [A]
    Val_B= S_BOX_2 [B]
    Val_C= S_BOX_3[C]
    Val_D= S_BOX_4 [D]
    Val_G_1 = (Val_A + Val_B) mod 32
    Val_G_2 = (Val_C + Val_D) mod 32
    Return Val_G_1 and Val_G_2
}
```

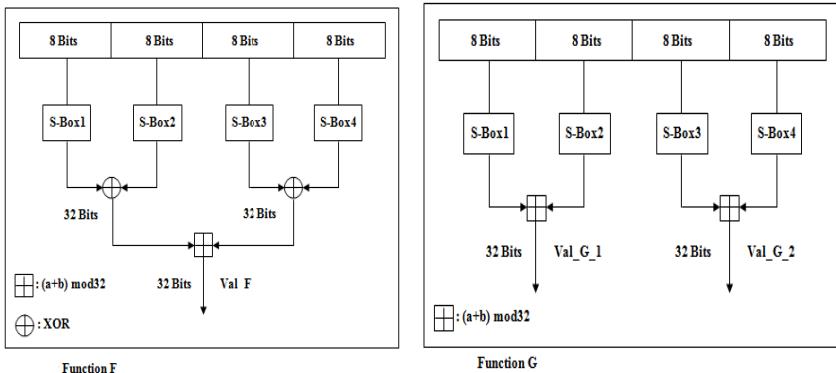


Fig. 3.

2.5 Sub Key and S-Box Generation

Sub Key and S-Box generation play a very important role providing high security to a symmetric key cryptosystem. The key schedule should satisfy completely or partially Strict Avalanche Criterion (SAC) [3, 6] and Bit Independence Criterion (BIC) [3, 6] in order to avoid certain key clustering attacks. Grossman and Tuckerman [7] showed that cryptosystems like the Data Encryption Standard (DES) where the key does not vary with successive rounds can be broken. It is therefore required that the primary key bits used to create sub key for a particular round i are different from those used in round $i+1$. C128 uses 80 Sub Keys and four S-Boxes of dimension $[1 \times 256]$ of 32 bit each for encryption and decryption. The Sub key and S-Box generation algorithm is uses two irreversible functions F and G. The algorithm for Sub Key and S-box is as follows

SK_SBox_C128 (variable size key)

{

Step1. Initialize each element of the four [1x256] S_Box and 80 sub keys with decimal values of e.

Step2. Create an array of 1104 words called ARR.

Step3. Select the first 32 bits from key and give it as input to non reversible function G. The first word from function G (Val_G_1) is XORed with the first element of S-Box. The first word of ARR is the second word from function G(Val_G_2).

Step4. Select next 32 bits from key and give it as input to function G. Val_G_1 is XORed with the next element of S-boxes. Val_G_2 is the value of next ARR element.

Step5. Repeat step4 till all the elements of S-Boxes and sub keys are XORed. Repeat the key when necessary.

Step5. Select element from ARR (starting from first to last) and give it as input to function G. XOR the first output of G (Val_G_1) with subsequent S box/Sub key element. Replace the ARR word used with the second output of function G(Val_G_2).

Step6. Repeat *step5* six times.

Step7. Select each 32 bit word from ARR (starting from first to last) and give it as input to function F. XOR the output of F with subsequent S box/Sub key element.
}

3 Analysis

- The algorithm encrypts variable plain text size on every encryption even if the key is same because the size of plain text to be encrypted depends upon both key and plain text.
- Data and Key dependent cyclic shift operations makes the algorithm more secured.
- Function F is non-reversible in nature which gives C128 quality avalanche effect.
- Use of four different S-boxes avoids symmetries when input bytes are equal.
- The key-dependent S-boxes protect against differential and linear cryptanalysis. The structure of the S-boxes is unknown to the attacker. Key-dependent S-boxes are easier to implement and can be created on demand, reducing the need for large data structures stored with the algorithm.
- The sub key generation process is designed in such a way that all the sub keys are affected by the key. One major advantage of the sub key generation algorithm is its parallel execution.
- Padding is done by random values known only to the sender which causes random change in the data which adds to the strength of the algorithm.

4 Future Work

- The algorithm could be modified for variable number of rounds.
- The function F could be modified for better speed and security.

5 Conclusion

In this paper a new cryptosystem based on Feistel network has been proposed. The algorithm exploits Feistel structure to provide better security than regular designs. Its variable size key makes the key space very large and hence brute force attack is not possible. Variable block size makes it difficult for the attacker to guess the size of the plaintext. Padding by random variables causes an undeterministic change in data in all rounds of encryption. Its testing against Differential and Linear attacks is yet to be done.

6 Notes

- i. *Confusion*: It refers to making the relation between cipher text, plain text and key as complex as possible. It is achieved by substitution [iii].
- ii. *Diffusion*: It refers to dissipating the plain text and key over cipher text such that each plain text block or key effects each cipher text bit. It is achieved by permutation [iv].
- iii. *Substitution*: Process of mixing linear and nonlinear operation in order to get an output by replacing the input by some value.
- iv. *Permutation*: Process of shuffling the bits so that each bit of cipher text should depend on each bit of plain text.
- v. *Substitution-Permutation Network (SPN)*: Series of linked mathematical operations used to design block ciphers [2, 3]. It is based on Shannon's concept of confusion and diffusion [1].
- vi. *S-Box*: It is used for substitution because the most fundamental property of an S-Box is that the output bits cannot be represented as linear operations on the input bits.
- vii. *Strict Avalanche Criterion (SAC)*: It is a generalization of avalanche effect. It states that the output bits of S-Box changes with probability $\frac{1}{2}$ when a single input bit is complemented.[4]
- viii. *Bit Independence Criterion (BIC)*: It states that the output bits j, k should change independently when any single input bit i is inverted for all i, j and k .[4]

References

- [1] Shannon, C.E.: Communication theory of secret systems. Bell Systems Technical Journal 28, 656–715 (1949)
- [2] Feistel, H.: Cryptography and Computer Privacy. Scientific American 228(5), 15–23 (1973)

- [3] Feistel, H., Notz, H.W., Lynn Smith, J.: Some cryptographic techniques for machine-to-machine data communications. IEEE Proceedings 63(11), 1545–1554 (1975)
- [4] Webster, A.F., Tavares, S.E.: On the Design of S-Boxes. In: Williams, H.C. (ed.) CRYPTO 1985. LNCS, vol. 218, pp. 523–534. Springer, Heidelberg (1986)
- [5] Evertse, J.-H.: Linear Structures in Block Ciphers. In: Price, W.L., Chaum, D. (eds.) EUROCRYPT 1987. LNCS, vol. 304, pp. 249–266. Springer, Heidelberg (1988)
- [6] Adams, C.M.: Constructing Symmetric Ciphers using Cast Design Procedure
- [7] Grossman, E., Tuckerman, B.: Analysis of a Feistel-like cipher weakened by having no rotating key. Technical Report RC 6375, IBM (1977)
- [8] Biham, E., Shamir, A.: Differential Cryptanalysis of DES-like Cryptosystems. In: Menezes, A., Vanstone, S.A. (eds.) CRYPTO 1990. LNCS, vol. 537, pp. 2–21. Springer, Heidelberg (1991)
- [9] Matsui, M.: Linear cryptanalysis method for DES cipher. In: Helleseth, T. (ed.) EUROCRYPT 1993. LNCS, vol. 765, pp. 386–397. Springer, Heidelberg (1994)
- [10] Rivest, R.L.: The RC5 Encryption Algorithm
- [11] Schneier, B.: Description of a New Variable-Length Key, 64-Bit Block Cipher (Blowfish)
- [12] Schneier, B.: Applied Cryptography. John Wiley & Sons, New York (1994)

Optimal Bandwidth Allocation Technique in IEEE 802.11e Mobile Ad Hoc Networks (MANET)

R. Mynuddin Sulthani¹ and D. Sreenivasa Rao²

¹ Department of ECE,

Sreenivasa Institute of Technology & Management Studies (SITAMS),
Chittoor, A.P (India)

rmsulthan@yahoo.co.in

² Department of ECE, JNTU College of Engineering,
JNTUH, Kukatpally, Hyderabad

Abstract. In mobile ad hoc networks (MANET), achieving QoS guarantees by over-provisioning of the resources is typically infeasible because in these networks, the overall available bandwidth is quite limited. In this paper, we propose an optimal bandwidth allocation technique in IEEE 802.11e MANET. In this approach, initially available bandwidth is estimated using the bandwidth probing technique. The estimated available bandwidth is shared in reserved and shared region by the bandwidth sharing scheme for the real-time and non-real time flows respectively. The bandwidth allocation policy allocates additional bandwidth from the reserved region of available bandwidth, when the mobile host needs additional bandwidth for real-time flows. In case of excess bandwidth utilization of mobile hosts, the excess bandwidth is restored to the shared region of available bandwidth thus minimizing the reserved region.. By simulation results, we show that the proposed approach offers optimal bandwidth for MANET

Keywords: Mobile Ad Hoc Networks (MANET), IEEE 802.11e, QoS (Quality of service), Resource Allocation.

1 Introduction

1.1 Mobile Ad Hoc Networks (MANET)

The network that depends on the principle of cooperation where every node pretends to be terminal as well as router is termed as mobile ad hoc networks. The performance of the network depends on how well the participants of the network collaborate with each other. The threat of misbehavior arises when nodes (or rather: the users controlling them) decide to maximize their own benefit rather than work together as a group. These gains can be measured for example in terms of throughput or battery life. The detection and mitigation of such behavior is important for the functioning of the ad-hoc network [1].

1.2 IEEE 802.11

The IEEE 802.11 standard came into existence to offer wireless local area networks (WLANs) within various environments, for example, public networks, enterprise networks, etc. In recent years, there has been gigantic growth in the popularity of wireless services and applications. In order to withstand such growth, standardization organizations such as the IEEE have decided to standardize the features by providing increased QoS and higher throughputs for IEEE 802.11 [3]. Different technologies of IEEE 802.11 namely, IEEE 802.11a, IEEE 802.11b, IEEE 802.11g offer error free performance, and as such they have been made the choice for WLANs and MANETs [2]. Currently, the IEEE 802.11 family of standards is most often being used to deploy MANETs. However, the MAC layer provided by these standards was designed for cooperation. Nodes contend for the medium using a distributed mechanism, which assumes that all participants behave properly [1].

1.3 IEEE 802.11 e

IEEE 802.11e was proposed to complement IEEE 802.11 MAC with the purpose of offering service differentiation in WLAN. The 802.11e draft brings out the Hybrid Coordination Function (HCF) that defines two new MAC methods. They are HCF controlled channel access (HCCA), and enhanced distributed channel access (EDCA), in order to substitute PCF and DCF modes in 802.11 [5].

EDCA is a distributed channel access method which can be used in ad hoc networks and provides QoS by delivering traffic based on differentiating user priorities [6]. EDCA brings in four new priority queues, one for every access category and thus achieving service differentiation. And by employing different parameter sets each priority queue has its own backoff entity [5].

1.4 Resource Allocation

The major role of resource management scheme is to map the service requirements of various applications to network resources so that the QoS requirements of the various users are met [14]. Reservation of resources is an important means to achieve quality of service guarantees in computer and communication networks. In continuous media communications with real-time requirements, such as audio/video communications, it is typical to reserve the resources during connection setup of a stream [15] .

Some networks try to satisfy QoS requirements simply by over-provisioning of resources; in mobile networks resource over-provisioning is typically infeasible because in these networks the available overall bandwidth (BW) is quite limited. As alternatives to resource over-provisioning one could try to achieve QoS guarantees by means of a priori reservation of resources or by means of prioritizing data units, e.g. the packets transmitted, in combination with access control for high-priority traffic.

Static reservation of resources may be quite inefficient, however, if significant variations of load exist during the life-time of a connection for which resources have been reserved. Dynamic reservation allocation is an efficient technique. [11]

1.5 Problem Identification

In paper [12], a scalable and reliable QoS architecture for mobile ad hoc networks is proposed which comprises multi-path routing protocol, a call admission control scheme and congestion control mechanism.

In paper [13], an EDCA scheduling algorithm is presented that allocates transmission opportunities (TXOP) for fluctuating VBR traffic depending on their queue length estimations for mobile ad hoc networks.

Both the papers [12] and [13], focuses only on QoS architecture and its estimation in IEEE 802.11e mobile ad hoc networks. By the method of over-provisioning the resources, some networks try to satisfy the QoS requirements. The over-provisioning of the resources in mobile networks is typically infeasible because in these networks the available overall bandwidth is quite limited [11].

In this paper, we propose an optimal bandwidth allocation technique in IEEE 802.16 Mobile ad hoc networks.

2 Related Works

Ali Hamidian et al [6] proposed a scheme called enhanced distributed channel access with resource reservation (EDCA/RR) with the aim of providing QoS guarantees.

Yimeng Yang et al [7] proposed a centralized feedback control model for resource management which is adapted to a specific application for WLANs with a centralized medium access method. This approach provides channel resource management in an efficient and flexible way. The drawback of this approach is that it is not modeled in the fully distributed manner for modeling multi-hop ad hoc networks.

A.Floros [8] proposed an Effective Bandwidth Control Policies for QoS enabled Wireless Networks. An admission control policy is introduced that is suitable for efficiently controlling traffic admissions under the EDCA method.

C. T. Calafate et al [9] proposed a novel QoS architecture for MANETs that seeks to alleviate the effects of both congestion and mobility on real-time applications. The architecture is highly modular and combines distributed admission control for MANET environments (DACME) with the IEEE 802.11e technology to offer soft QoS support to MANETs heavily loaded by both best effort and QoS traffic. The proposed architecture relies on modified dynamic source routing (MDSR) to reduce the impact of mobility on real-time sessions, also offering good performance when the routing protocols used are able to quickly respond to topology changes.

Yang Xiao et al [10] proposed a bandwidth sharing scheme for multimedia traffic in the IEEE 802.11e contention-based WLANs. A guard period is proposed to prevent bandwidth allocation from over-provisioning and is for best effort data traffic.

3 Optimum Bandwidth Allocation Scheme

3.1 Overview

We propose an efficient bandwidth allocation technique for IEEE 802.11e mobile ad hoc networks (MANET). In this approach, the source and destination node

communicates to assess the available bandwidth (AB) using the bandwidth probing technique. Based on the assessed bandwidth, the source node decides whether to admit the connection or not. The bandwidth sharing scheme is implemented which involves sharing the AB into reserved and shared region. This scheme utilizes bandwidth initially in the reserved region and further in the shared region. The bandwidth allocation policy allocates additional bandwidth from the reserved region of available bandwidth, when the mobile host needs additional bandwidth for high priority flows. In case of excess bandwidth utilization of mobile hosts, the excess bandwidth is restored to the shared region of available bandwidth thus minimizing the reserved region.

3.2 Estimation of Available Bandwidth

In the admission control scheme, the source and destination node communicates in order to assess the available bandwidth (AB). The source node (S) communicates by sending the probe packets to the destination node (D). These packets are generated back-to-back and should be followed by the probe reply. S keeps a timer to detect probe reply losses. D upon receiving all the probe packets sends a single reply packet with the measured value for the available end-to-end bandwidth. This value is defined as:

$$AB = \frac{8 \times Z}{\Delta t_r} \cdot (P_r - 1) \quad (\text{bits/s}) \quad (1)$$

where z = packet size used.

Δt_r = time interval between the first and last received packet

P_r = number of packets received

S upon receiving probe reply packets gathers the AB values sent by the D to decide whether to admit the connection or not.

3.3 Admission Control Using Bandwidth Sharing

3.3.1 Bandwidth Sharing Technique

The bandwidth sharing scheme is implemented which involves sharing the available bandwidth into two regions as follows.

1) Reserved region (αAB)

2) Shared region (βAB)

$$\therefore AB = \alpha AB + \beta AB \quad (2)$$

In this approach, the bandwidth is initially used in the reserved region and further in the shared region. The bandwidth is reserved either for voice or video flows or both if required. Other type of traffic like Best Effort utilize the bandwidth from the shared region.

3.3.2 Admission Control

During the process of admission control, the existing voice or video flows is guarded from new voice or video flows. The source node broadcasts AB value through the

beacon frames, and this value is shared among the voice and video flows. The mobile host (MH) determines an internal transmission limit per access category (AC) for each beacon interval depending on the transmission count during the previous beacon period and AB announced from S. The local video or voice flow's bandwidth limit per beacon interval may not exceed the internal transmission per AC. When AB is depleted, new flows will not be able to gain AB, while existing flows will not be able to increase AB per beacon interval, which is already in use. Thus, this mechanism protects the existing flows.

3.4 Bandwidth Allocation Policy

This approach describes a basic scheme for monitoring the actual time employed for transmission by each AC. In specific, each mobile node maintains two counters for each of the ACs that requires admission control

- 1) Used time T_u : It represents measurement of the medium occupancy the AC actually within the last 1 second duration.
- 2) Admitted time T_a : It is the aggregated time length within 1 second interval offered by the mobile node to all the admitted traffic flows of an AC that requires admission.

Initially both the counter values are set to zero. In specific, the mobile host updates T_u at every 1 second interval as:

$$T_u = \max(T_u - T_a, 0) \quad (3)$$

After each frame exchange, T_u becomes

$$T_u = T_u + T_e \quad (4)$$

Where T_e = exchange time. It is defined as time required for transmitting a nominal MAC service data unit (MSDU) packet including the inter-frame time-lengths and the transmission acknowledgments policy overheads.

In wireless networks, some transmission conditions may lead to an excessive number of packet retransmissions. Also, an admitted traffic flow may suddenly produce data bursts not accurately defined in the corresponding traffic specifications (TSPEC). Both situations cause T_u values to be greater than the T_a . This situation can be handled using the following conditions.

If $T_u > T_a$ Then

The additional bandwidth is obtained from reserved region of available bandwidth.

Else

The excess bandwidth is restored to the shared region of available bandwidth thus minimizing the reserved region.

End if

3.5 Overall Algorithm

Step 1

The source and destination node communicates to access the available bandwidth (AB) using bandwidth probing technique while implementing the admission control approach.

Step 2

The source node upon receiving probe reply packets gathers the AB values sent by the destination node to decide whether to admit the connection or not.

Step 3

The bandwidth sharing scheme is implemented which involves sharing the available bandwidth into two regions that includes the reserved and shared region.

Step 4

In bandwidth sharing scheme, the bandwidth is initially used in the reserved region and further in the shared region. The bandwidth is reserved for either both voice or video flows or both if required.

Step 5

The employed admission control scheme guards the existing voice or video flows guarded from new voice or video flows.

Step 6

The bandwidth allocation policy monitors the actual time employed for transmission by each AC.

Step 7

During the situation of excessive number of packet retransmissions and data bursts, the counter value maintained by mobile hosts exceeds i.e the used time T_u exceeds the admitted time T_a . During this condition, the additional bandwidth is obtained from reserved region of available bandwidth.

Step 8

If T_u is less than T_a , the excess bandwidth is restored to the shared region of available bandwidth thus minimizing the reserved region.

4 Simulation Results

4.1 Simulation Model and Parameters

To simulate the proposed algorithm, Network Simulator version-2 (NS2) [15] is used. In the simulation, 50 mobile nodes move in a 1000 meter x 1000 meter region for 50 seconds simulation time. Initial locations and movements of the nodes are obtained using the random waypoint (RWP) model of NS2. All nodes have the same transmission range of 250 meters. The node speed is 5 m/s. and pause time is 5 seconds. In the simulation, for class1 and class2 traffic CBR and FTP are used respectively. The IEEE 802.11e MAC protocol is used.

4.2 Performance Metrics

The proposed Optimal Bandwidth Allocation Technique (OBAT) is compared with the DACME [9]. The performance is mainly evaluated according to the following metrics: Packet delivery ratio, Number of packets received, Number of packets dropped, average end-to-end delay, aggregated bandwidth and fairness index.

The performance results are presented in the next section. 1.

4.3 Results

A. Effect of Varying Rates

In the initial experiment, the transmission rate is varied as 250,500,750 and 1000Kb. The results are given for class1 and class2 traffic.

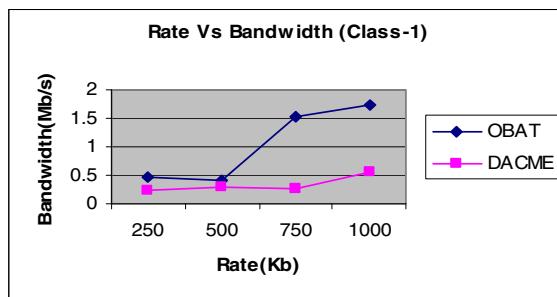


Fig. 1. Rate Vs Bandwidth

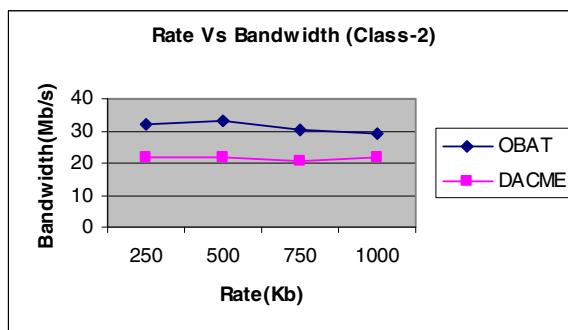


Fig. 2. Rate Vs Bandwidth

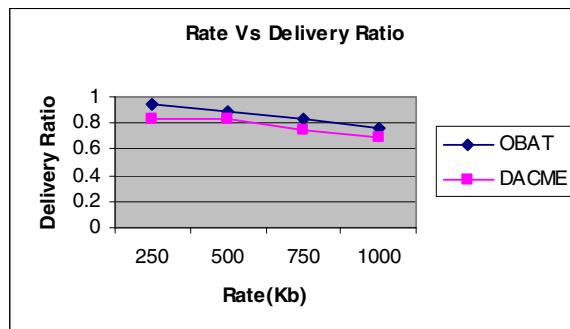


Fig. 3. Rate Vs Delivery Ratio

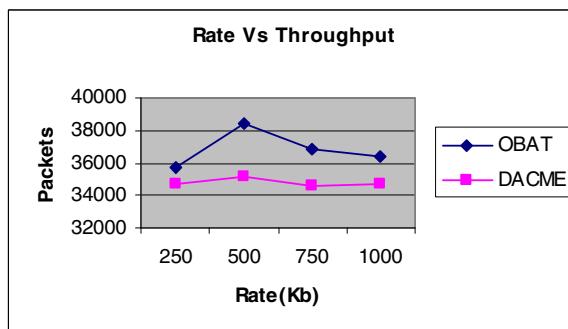


Fig. 4. Rate Vs Throughput

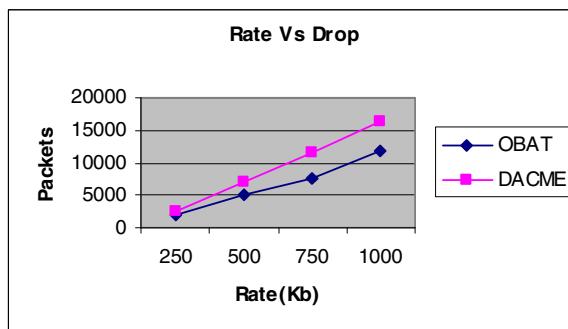


Fig. 5. Rate Vs Drop

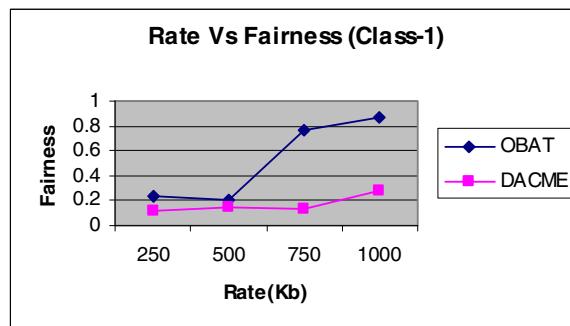


Fig. 6. Rate Vs Fairness

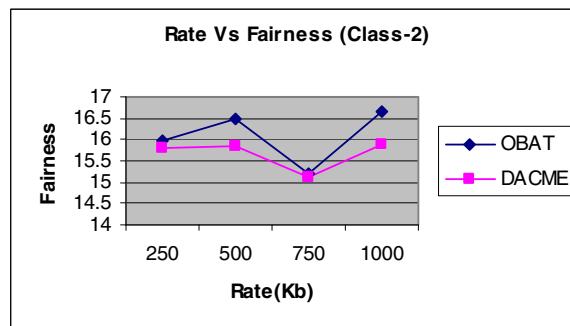


Fig. 7. Rate Vs Fairness

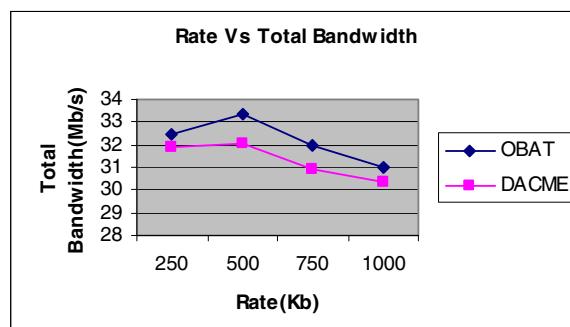


Fig. 8. Rate Vs Total Bandwidth

Fig: 1 and Fig: 2, gives the aggregated bandwidth for class1 and class2 traffic. From the figures, it can be seen that OBAT has received more bandwidth when compared with DACME.

Fig: 3 give the delivery ratio. From the figures, it can be seen that OBAT has received more delivery ratio when compared with DACME.

Fig: 4 give the throughput. From the figures, it can be seen that OBAT has achieved more throughput when compared with DACME

Fig: 5 give the packet drop. From the figures, it can be seen that OBAT has less packet drop when compared with DACME.

Fig: 6 and Fig: 7, gives the fairness for class1 and class2 traffic. From the figures, it can be seen that OBAT has more fairness when compared with DACME.

Fig: 8 give the Total bandwidth received. From the figures, it can be seen that OBAT has more bandwidth received ratio than the DACME.

5 Conclusion

In this paper, we have proposed an optimal bandwidth allocation technique in IEEE 802.11e MANET. In this approach, the available bandwidth is estimated using the bandwidth probing technique. The estimated available bandwidth is shared in reserved and shared region by the bandwidth sharing scheme for the real-time and non-real time flows respectively. The bandwidth allocation policy allocates additional bandwidth from the reserved region of available bandwidth, when the mobile host needs additional bandwidth for real-time flows. By simulation results, we have shown that the proposed optimal bandwidth allocation scheme offers higher bandwidth, fairness and throughput while minimizing the delay and packet drop. As a future work, the scheme can be extended for large number of networks.

References

1. Szott, S., Natkaniec, M., Canonico, R., Pach, A.R.: Impact of Contention Window Cheating on Single-hop IEEE 802.11e MANETs. In: IEEE Wireless Communications and Networking Conference, WCNC (2008)
2. Calafate, C.T., Manzoni, P., Malumbres, M.P.: On the interaction between IEEE 802.11e and routing protocols in Mobile Ad-hoc Networks. In: Proc. the 13th Euromicro Conference on Parallel, Distributed and Network-Based Processing, pp. 110–117 (2005)
3. Cabral, O., Segarra, A., Velez, F.J.: Implementation of Multi-service IEEE 802.11e Block Acknowledgement Policies. IAENG International Journal of Computer Science 36(1) (2009)
4. Friedman, R., Hay, D., Kliot, G.: Jittering Broadcast Transmissions in MANETs: Quantification and Implementation Strategies (2009)
5. Li, J., Li, Z., Mohapatra, P.: APHD: End-to-End Delay Assurance in 802.11e Based MANETs. In: 3rd Annual International Conference on Mobile and Ubiquitous Systems – Workshops, pp. 1–8 (2006)
6. Hamidian, A., Korner, U.: Providing QoS in Ad Hoc Networks with Distributed Resource Reservation. In: International Teletraffic Congress, pp. 309–320 (2007)

7. Yang, Y., Haverkort, B.R., Heijenk, G.J.: A centralized feedback control model for resource management in wireless networks. In: Eighth International Workshop on Performability Modeling of Computer and Communication Systems, PMCCS-8, Edinburgh, Scotland, September 20-21 (2007)
8. Floros, A.: Effective Bandwidth Control Policies for QoS enabled Wireless Networks. In: The Sixth Annual Mediterranean Ad Hoc Networking Workshop, Corfu, Greece, June 12-15 (2007)
9. Calafate, C.T., Malumbres, M.P., Oliver, J., Cano, J.C., Manzoni, P.: QoS Support in MANETs: A Modular Architecture Based on the IEEE 802.11e Technology. *IEEE Transactions on Circuits and Systems for Video Technology* 19(5) (2009)
10. Xiao, Y., Li, F.H., Li, B.: Bandwidth Sharing Schemes for Multimedia Traffic in the IEEE 802.11e Contention-Based WLANs. *IEEE Transactions on Mobile Computing* 6(7) (July 2007)
11. Wolf, J., Heckmuller, S., Wolfinger, B.E.: Dynamic Resource Reservation and QoS Management in IEEE 802.11e Networks. In: Proc. of the International Symposium on Performance Evaluation of Computer and Telecommunication Systems, SPECTS (2005)
12. Mynuddin Sulthani, R., Sreenivasa Rao, D.: "A Scalable and Reliable QoS Architecture (SRQA) for Mobile Ad-hoc Networks", International conference on computing, communication and networking (ICCCN), 2008.
13. Mynuddin Sulthani, R., Sreenivasa Rao, D.: A QoS Estimation Based Scheduler for Real-time Traffic in IEEE 802.11e Mobile Ad hoc Networks. *Information Security Journal* 20(6), 317–327 (2011)
14. Lal, S., Sousa, E.S.: Distributed Resource Allocation for DS-CDMA Based Multimedia ad hoc Wireless LAN's. *IEEE Journal on Selected Areas in Communications* 17(5) (May 1999)
15. Wolnger, B.E., Wolf, J., Le Grand, G.: Improving Node Behavior in a QoS Control Environment by Means of Load-dependent Resource Redistributions in LANs. *International Journal of Communication Systems* 18(4) (May 2005)

Trusted AODV for Trustworthy Routing in MANET

Sridhar Subramanian¹ and Baskaran Ramachandran²

¹ Dept. of Computer Applications, Easwari Engineering college, Chennai
ssridharmca@yahoo.co.in

² Dept. of Computer Science and Engineering, CEG, Anna University, Chennai
baskaran.ramachandran@gmail.com

Abstract. A Mobile ad-hoc network (MANET) is an extremely testing lively network. They are self configuring, autonomous, quickly deployable and operate without infrastructure. Mobile ad hoc networks consist of nodes that cooperate to provide connectivity and are free to move and organize randomly. Nodes can connect and depart the network at anytime and should be in position to relay traffic. These nodes are often vulnerable to failure thus making mobile ad hoc networks open to threats and attacks. Communication in MANET relies on mutual trust between the participating nodes but the features of MANET make this hard. Nodes sometimes fail to transmit and start dropping packets during the transmission. Such nodes are responsible for untrustworthy routing. A trust based scheme can be used to track this behavior of untrustworthy nodes and segregate them from routing, thus provide trustworthiness. In this paper a trust based AODV protocol is presented which assigns a trust value for each node. Nodes are allowed to participate in routing based on their trust values. A threshold value is assigned and if the nodes trust value is greater than this value its marked as trustworthy node and allowed to participate in routing else the node is marked untrustworthy. This scheme increases PDR and decreases delay thereby enhancing the trustworthiness in AODV based MANET routing. The work is implemented and simulated on NS-2. The simulation result shows the proposed protocol provides more reliable and consistent data transfer compared with general AODV in presence of unpredictable and unreliable nodes in MANET.

Keywords: Ad-hoc, MANET, AODV, Trust, Qos.

1 Introduction

Mobile ad hoc network is a standalone network capable of autonomous operation where nodes communicate with each other without the need of any existing infrastructure. Mobile Ad-Hoc network [1] is a system of wireless mobile nodes that self-organizes itself in dynamic and temporary network topologies. Every node is router or an end host, in general autonomous and should be capable of routing traffic as destination nodes sometimes might be out of range. Nodes are mobile since topology is very dynamic and they have limited energy and computing resources. The primary goal of MANET is to find an end to end path or route, minimizing overhead, loop free and route maintenance. A few challenges faced in mobile ad hoc networks are mobility, variable link quality, energy constrained nodes, heterogeneity and flat addressing.

Most traditional mobile ad hoc network routing protocols were designed focusing on the efficiency and performance of the network [2]. These protocols should meet some basic requirements like self starting, self organizing, loop free paths, dynamic topology maintenance, minimal traffic overhead etc to deal with the challenges involved in routing. Existing MANET routing protocols can be classified into mainly two types- proactive routing protocols and reactive routing protocols. Table driven (proactive) routing protocols such as dynamic Optimized Link State Routing (OLSR), Destination-Sequenced Distance-Vector routing (DSDV), Topology Broadcast based on Reverse Path Forwarding (TBRPF) and On-demand (reactive) routing protocols such as Ad hoc On demand Distance Vector (AODV), Signal Stability-based Adaptive routing (SSA), Dynamic Source Routing (DSR). Other categories are flooding based, cluster based, geographic and application specific. Proactive protocols are table driven protocols much similar to conventional routing, have little delay in route discovery and routing overhead is high. On-demand routing protocols are reactive protocols which obtain route information only when needed and the overhead is low since there is no periodic update of tables.

AODV is a reactive protocol where route discovery initiated when required only using route request (RREQ) and route reply (RREP) packets and stores only active routes in routing table. Explicit route error notification is done by using route error (RERR). Ad-hoc on demand Distance Vector (AODV) routing protocol [3] is an on demand routing protocol that focuses on discovering the shortest path between two nodes with no consideration of the reliability of a node. By broadcasting HELLO packets in a regular interval, local connectivity information is maintained by each node. However, the traditional AODV protocol seems less than satisfactory in terms of delivery reliability thereby affecting quality of service.

Due to the dynamic nature of Mobile Ad-Hoc Networks, there are many issues which need to be tackled and one of the areas for improvement is Quality of Service (QoS) routing. When it comes to QoS routing, the routing protocols have to ensure that the QoS requirements are met [4]. A few challenges faced in providing QoS are persistently changing environment, unrestricted mobility which causes recurrent path breaks and also make the link-specific and state-specific information in the nodes to be inaccurate.

This AODV protocol is to perform its task based on the trust based scheme where trust values calculated for each node and to decide whether the node can take part or to be isolated from routing. If nodes trust value is less than the threshold then the node is declared to be untrustworthy node and an alternate path is chosen. This trust based routing scheme facilitates in identifying and isolating untrustworthy nodes thus providing trustworthy routing in MANET and also improves the performance QoS parameters like PDR and delay.

2 Literature Survey

Mobile ad hoc networks are peer to peer wireless networks which operate without infrastructure and communicate without any centralized administration. MANETs have put on more significance in recent applications areas like security, routing, resource management, quality of service etc. The significance of routing protocols in

MANETs has anticipated for a lot of competent and inventive routing protocols. Continuous evaluation of node's performance and collection of neighbor node's opinion value about the node are used to calculate the trust relationship of this node with other nodes [5]. In this paper, existing AODV routing protocol has been modified in order to adapt the trust based communication feature and the proposed trust based routing protocol equally concentrates both in node trust and route trust.

RAODV (Reliant Ad hoc On demand Distance Vector Routing) [6] is a security-enhanced AODV routing protocol that uses a modified scheme called direct and recommendations trust model and then incorporating it inside AODV. This scheme assures that packets are not handed over to malicious nodes. Based on this trust value a node is selected to perform packet transfer. This protocol results in higher percentage of successful data delivery compared to AODV. A routing algorithm is proposed that adds a field in request packet which stores trust value indicating node trust on neighbor [7]. Based on level of trust factor, the routing information will be transmitted depending upon highest trust value among all that results not only in saving the node's power but also in terms of bandwidth. A trusted path irrespective of shortest or longest path is used communication in the network.

A routing protocol [8], that adds a field in request packet and also stores trust value indicating node trust on neighbor based on level of trust factor. This scheme avoids unnecessary transmit of control information thus efficiently utilizing channels and also saves nodes power. Route trust value is calculated based on the complete reply path, which can be utilized by source node for next forthcoming communication in the network that results in improvement in security level and also malicious node attacks are prevented. A trust based packet forwarding scheme [9] for detecting and isolating the malicious nodes using the routing layer information that uses trust values to favor packet forwarding by maintaining a trust counter for each node. A node will be punished or rewarded by decreasing or increasing the trust counter. If the trust counter value falls below a trust threshold, the corresponding intermediate node is marked as malicious.

A framework [10] for estimating the trust between nodes in an ad hoc network based on quality of service parameters is proposed based on Probabilities of transit time variation, deleted, multiplied and inserted packets, processing delays. It has been shown that only two end nodes need to be involved and thereby achieve reduced overhead. A Node-based Trust Management (NTM) scheme in MANET [11] is introduced based on the assumption that individual nodes are themselves responsible for their own trust level. Mathematical framework of trust in NTM is developed along with some new algorithms for trust formation in MANETs based on experience characteristics offered by nodes. The above listed works are spotlighting on reliability that is provided to the mobile ad hoc network by using trust schemes.

3 Proposed Work

It is tough to provide reliable routing in routing in mobile ad hoc networks because of its dynamic nature that keeps nodes moving and not stable. In spite of this nature nodes communicate with each other and exchange data among the nodes that are in its range on the network. But still there are nodes in the MANET which take part in

routing but drop packets while transmitting packets which affects the performance of the protocol. Hereby we introduce a scheme in the existing AODV routing which checks each node before involving it in the routing process. The structural design of the proposed work is presented in Fig. 1. In the MANET a observation is made on all nodes that transmit packets. The total packets they transmit, packets they receive and the packets they drop are taken in to account. Once a particular transmission is to be made the protocol decides the route and the nodes which are going to participate in routing are checked against their trust values which are calculated based on the total packets handled by each node. Based on this trust value a node is located if it is about to drop packets. Thus an alternate path is identified to carry on the routing effectively.

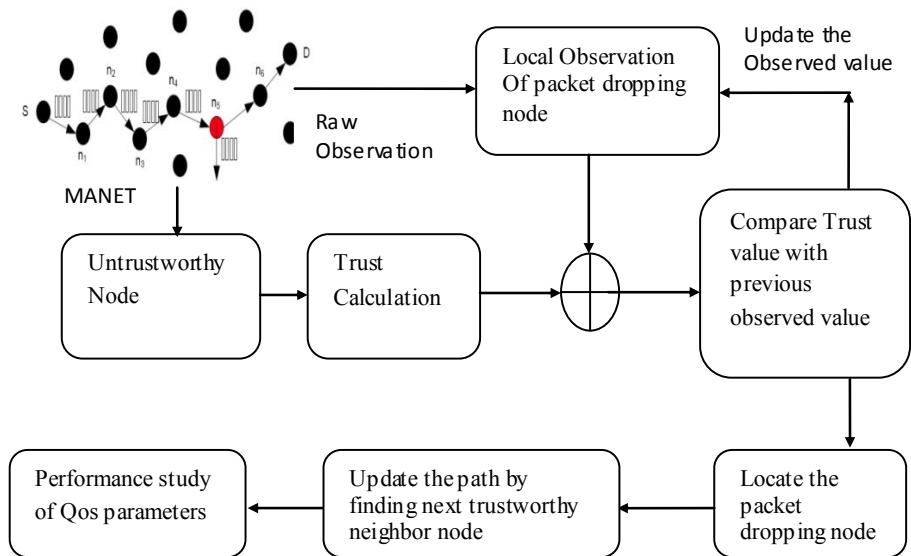


Fig. 1. Architecture of trust based AODV routing in MANET

The trust level value calculation [12] is based on the parameters shown in the table 1. The count field describes about two criteria success and failure which describes whether the transmission was a successful transmission or a failure.

Table 1. Trust value calculation parameters

Count Type	RREQ	RREP	Data
Success	Qrs	Qps	Qds
Failure	Qrf	Qpf	Qdf

RREQ and RREP are the route request and route reply respectively which are exchanged between nodes in the network. Data refers to the payload transmitted by the nodes.

The parameter q_{rs} is defined as the query request success rate which is calculated based on number of neighboring nodes who have successfully received (rreq) from the source node which has broadcasted it, q_{rf} defined as the query request failure rate which is calculated based on number of neighboring nodes which have not received the query request, q_{ps} is defines as the query reply success rate which is calculated as successful replies (rrep) received by the source node which has sent the rreq and q_{pf} is defined as the query reply failure rate which is calculated based on the number of neighboring nodes which have not sent the replies for the query request received. q_{ds} is defined as the data success rate calculated based on successfully transmitted data and q_{df} is defined as data failure rate calculated based on data which have failed to reach destination. However, it is known that for every network there will be minimum data loss due to various constraints

$$Qr = \frac{q_{rs} - q_{rf}}{q_{rs} + q_{rf}} \quad (1)$$

$$Qp = \frac{q_{ps} - q_{pf}}{q_{ps} + q_{pf}} \quad (2)$$

$$Qd = \frac{q_{ds} - q_{df}}{q_{ds} + q_{df}} \quad (3)$$

Where Qr , Qp and Qd are intermediate values that are used to calculate the nodes Request rate, Reply rate and Data transmission rate. The values of Qr , Qp , and Qd are normalized to fall in range of -1 to +1. If the values fall beyond the normalized range then it clearly shows that the failure rate of the node is high and denotes that the corresponding node may not be suitable for routing.

$$TL = T(RREQ) * Qr + T(RREP) * Qp + T(DATA) * Qd \quad (4)$$

Where, TL is the trust level value and $T(RREQ)$, $T(RREP)$ and $T(DATA)$ are time factorial at which route request , route reply and data are sent by the node respectively. Apart from the above mentioned normalised range, using the above formula the trust level value (TL) is calculated for each node during routing and is checked against the threshold value (assumed to be as 5). If lesser than threshold then there is a possibility for this node to drop packets for the current transmission and will not be suitable for routing and an alternate path is selected for routing. However, this node may be the best node for some other transmission between some other source and destination in the same network at different time interval. Therefore based on the above calculation the following two cases are derived based on the threshold value that is assumed to be 5. Case 1: The nodes trust value is checked with the threshold value and if the value is greater than the threshold value then the node is defined a

trustworthy node and are allowed to participate in routing thereby assuring a trustworthy routing in MANET. Case 2: If the nodes trust value is less than or equal to threshold value then the node is defined as untrustworthy node which cannot be allowed to participate in routing which causes packet dropping. In both cases the trust calculation is performed regularly to check the nodes performance and help it to be marked trustworthy or not.

For the sample network shown in figure 2, the path selected is $S \rightarrow F \rightarrow E \rightarrow G \rightarrow D$. For example, Node F has four neighbors and for this node the trust value calculation is to be done.

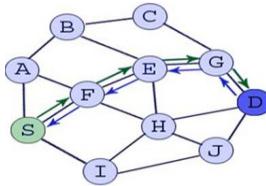


Fig. 2. Sample network

For node F the success and failure rate of route request, reply and data are calculated.

$$Qr = (4 - 0) / (4 + 0) = 1; Qp = (4 - 0) / (4 + 0) = 1; Qd = (900-100) / (900+100) = 0.8$$

The values of Qr , Qp , and Qd are falling within the normalized range fixed (i.e) -1 to +1. Thus the trust value is calculated for the node F.

$TL = 1*1!+1*2!+0.8*3! = 7.8$ (which is more than 5) thus making this node a reliable node for routing. This trust calculation is done for all nodes in the routing path to monitor nodes reliability. If the failure rate increases it automatically affects the Qr , Qp and Qd values thus making them fall beyond the normalized values thus resulting in trust value less than the threshold.

4 Evaluation Results

The proposed AODV protocol's performance is analyzed using NS-2 simulator. The network is planned and implemented using network simulator with maximum of 50 nodes and other parameters based on which the network is shaped are given in Table2. The simulator is applied with traditional AODV and with proposed trust based AODV and results are obtained for assessment. The proposed trust based AODV protocol has shown good progress over the Qos parameters like PDR & Delay. PDR is increased and delay is reduced compared to the traditional AODV and throughput is maintained in both cases. However there is a fraction of difference in throughput between general and proposed protocol which is rounded off as a whole value in result table. The performance of the proposed protocol is also represented graphically where it clearly shows the betterment of the Qos parameters.

Table 2. Simulation Parameter Values

Parameter	Value
Network size	1600 x 1600
Number of nodes	50
Movement speed	100 kbps
Transmission range	250 meters.
Packet size	5000
Traffic type	CBR
Simulation time	30 minutes.
Maximum speed	100 kbps
MAC layer protocol	IEEE 802.11
Time interval	0.01 sec.
Protocol	AODV
NS2 version	2.34

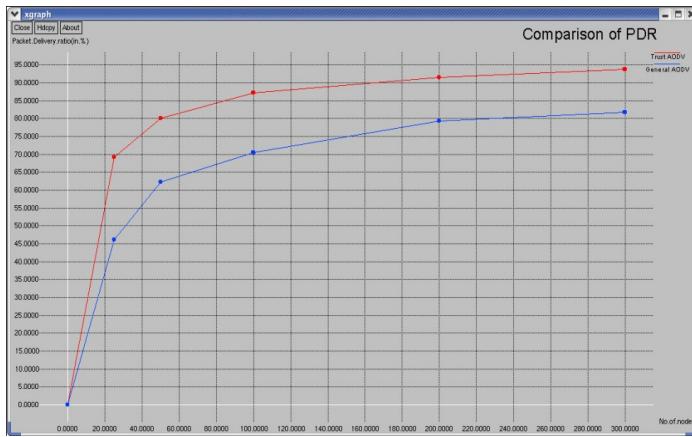
The values obtained using traditional AODV and proposed trust based AODV at different node sizes are listed in table 3. The traditional AODV doesn't provide reliable routing since the nodes present in the network drop packets while routing which degrades the performance of routing and results in reduced packet delivery ratio and increased delay.

Table 3. Result comparison with different node sizes

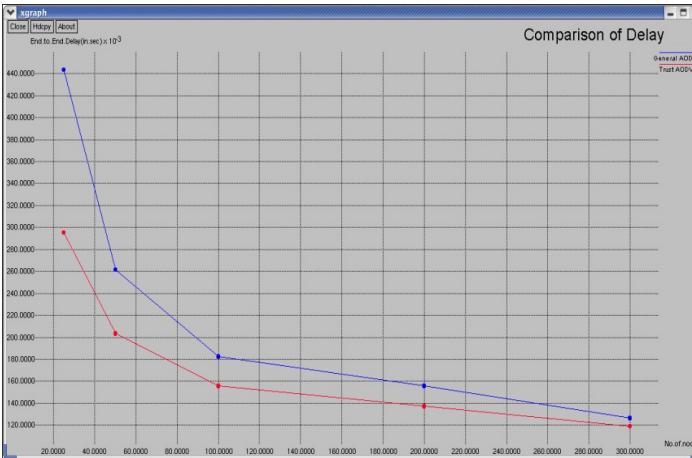
Node Size	General AODV			Proposed Trust based AODV		
	PDR	Delay	Throughput	PDR	Delay	Throughput
25	46.10	0.44306	757771.43	69.15	0.29538	757771.43
50	62.25	0.26151	120032.60	80.04	0.20340	120032.60
100	70.59	0.18225	115783.25	87.25	0.15595	115783.25
200	79.35	0.15584	113259.53	91.53	0.13759	113259.53
300	81.73	0.12635	110935.75	93.75	0.11925	110935.75

The Qos parameter values are showing better improvement when the routing takes place with the proposed AODV protocol which works using trust values that identifies untrustworthy nodes in the route and immediately take an alternate path to provide trustworthy and successfully routing. The results shown in the following table clearly shows the PDR and delay of the proposed AODV protocol are privileged compared to traditional AODV protocol at different node sizes.

Graph 1 specifies the increase in PDR by implementing the proposed trust based AODV protocol compared to the traditional AODV protocol. Graph 2 specifies the decrease in delay while using the proposed protocol compared to traditional AODV.



Graph 1. Comparison of general AODV PDR and Trust based AODV PDR



Graph 2. Comparison of general AODV Delay and Trust based AODV Delay

5 Conclusion and Future Enhancements

In this paper a trust based AODV protocol that identifies the nodes that drop packets while transmission is proposed where trust value for each node is calculated to spot the untrustworthy nodes in the path during routing. A node is declared as a trustworthy node if its trust value is greater than the threshold value thus resulting in a trustworthy MANET routing. This proposed scheme has shown a good development over Qos parameters like PDR and delay and has also provided trustworthy routing. The same scheme can also be implemented on other MANET routing protocols and check the performance with respect to the all Qos parameters. The future work will also provide reliability for the packets transmitted by the trustworthy nodes by introducing a secured scheme for checking whether the packets are tampered or not.

References

1. Kortuem, G., Schneider, J., Preuitt, D., Thompson, T.G.C., F'ickas, S., Segall, Z.: When Peer to-Peer comes Face-to-Face: Collaborative Peer-to-Peer Computing in Mobile Ad hoc Networks. In: 1st International Conference on Peer-to-Peer Computing, Linkoping, Sweden, pp. 75–91 (2001)
2. Narayan, P., Syrotiuk, V.R.: Evaluation of the AODV and DSR Routing Protocols Using the MERIT Tool. In: The Proceeding of ADHOC-NOW (2004)
3. Perkins, C.E., Belding Royer, E.M., Das, S.R.: Ad-hoc On-Demand Distance Vector (AODV) Routing. In: Mobile Adhoc Networking Working Group, Internet Draft (February 2003)
4. Jawhar, I., Wu, J.: Quality of Service Routing in Mobile Ad Hoc Networks. In: Cardei, M., Cardei, I., Du, D.Z. (eds.) Resource Management and Wireless Networking. Kluwer Academic Publishers
5. Pushpa, A.M.: Trust based secure routing in AODV routing protocol. In: IEEE International Conference (2009)
6. Jassim, H.S., Yussof, S.: A Routing Protocol based on Trusted and shortest Path selection for Mobile Ad hoc Network. In: IEEE 9th Malaysia International Conference on Communications (2009)
7. Mangrulkar, R.S., Atique, M.: Trust based secured adhoc On demand Distance Vector Routing protocol for mobile adhoc network. In: Sixth International Conference on Wireless Communication and Sensor Networks, WCSN (2010)
8. Mangrulkar, R.S., Atique, M.: Trust Based Secured Adhoc on Demand Distance Vector Routing Protocol for Mobile Adhoc Network (2010)
9. Sharma, S., Mishra, R., Kaur, I.: New trust based security approach for ad-hoc networks. In: 3rd IEEE International Conference on Computer Science and Information Technology, ICCSIT (2010)
10. Umuhzoza, D., Agbinya, J.I., Omlin, C.W.: Estimation of Trust Metrics for MANET Using QoS Parameter and Source Routing Algorithms. In: The 2nd International Conference on Wireless Broadband and Ultra Wideband Communications (2007)
11. Ferdous, R., Muthukumarasamy, V., Sattar, A.: Trust Management Scheme for Mobile Ad-Hoc Networks. In: IEEE 10th International Conference on Computer and Information Technology, CIT (2010)
12. Subramanian, S., Ramachandran, B.: Trust Based Scheme for QoS Assurance in Mobile Ad Hoc Networks. International Journal of Network Security & Its Applications (IJNSA) 4(1) (January 2012)

Dynamic Fuzzy Based Reputation Model for the Assurance of Node Security in AODV for Mobile Ad-Hoc Network

Arifa Azeez¹ and K.G. Preetha²

¹ Department of Computer Science

² Department of Information Technology

Rajagiri School of Engineering & Technology, Rajagiri valley, Cochin, India

arifa.azeez@gmail.com, preetha_kg@rajagiritech.ac.in

Abstract. Mobile ad hoc network is a self organized network with a collection of wireless nodes without a fixed topology and centralized administration among the nodes. There is an urgent need for trustworthiness in MANET, as the survival of the spontaneous network depends upon the trust and cooperation among different nodes. The nodes have unmitigated control over the data packets passed through them and so the malicious nodes may accomplish the control over the data packets, thus they may threat the normal nodes in the network. Here the need for establishing a trusted environment in MANET arises. Trusted environment increases the probability of a successful transaction and reducing the opportunities of being defrauded. Trust is used to determine the reliability and inter opinionative between nodes and their neighbors. Difference between actual and expected value of the node behavior determines the degree of trust of that node. In this paper first discuss the need for trusted environment, then the basic issues while designing trusted environment. Then survey some existing trusted models for MANET and point out some issues regarding with them. Based on these studies a solution suggested, dynamic fuzzy based trust model through a drop less route using AODV in MANET. Also advantages of the proposed model are discussed.

Keywords: MANET, AODV, Fuzzy systems, Reputation.

1 Introduction

Mobile ad hoc network is a collection of autonomous nodes and these nodes act as router to discover the dynamic routing topology and network connectivity. MANET uses their own algorithms to route through the network which can be a reactive or proactive or a hybrid routing technique. Based on wireless link quality, power limitation, mobility of nodes and multi user interference, routing may become difficult. MANET routing must assure security in routes and security in nodes. The security level can be internal security and external security. Many security schemes are proposed for MANET such as secured routing protocol, key based security solution, certificate based security solutions and trust and reputation based solutions.

Trust is a particular level of assurance reliance on the character, ability, strength or truth of someone or something. Trust management can be done by certificate based and reputation based. In the certificate based the trust is calculated by a certificate authority and the trust level is set as a fixed value throughout the network. While in case of reputation based model, the trust management depends on the behavior of each node in the network. This paper describes a fuzzy based reputation model. In this model the dynamic trust calculation is done based on the data packets shared between each node in the MANET.

The rest of this paper is organized as follows. Section 2 gives the brief idea of the need for trusted environment in MANET. Section 3 describes issues in designing the trusted environment. Related works are explained in section 4. A desired model is discussed in section 5. A conclusion is given in section 6.

2 Need for Trusted Environment in MANET

Traditional security systems are very effective but they are not applicable always, so trust and reputation based approaches are preferred to be a good alternative for security solutions in MANET. Basic attacks in MANET can be deceptive incrementing sequence number and deceptive decrementing hop count. From the perspectives of sociologists, economists, psychologists trust is stated as a duty imposed on faith or confidence in truth of relationships between someone. That is trust is context dependent, dynamic and monotonic and the basic actors of trusted environment are one, who is releasing trust information and the trustee who is being trusted. Reputation is a degree of opinion about a person about the other. Based on the reputation, trust level can be derived in a system. The entity which is providing trust information about the other, need to do more complex tasks to observe the other.

Trust can be of interpersonal, structural system dependent and temperamental make up which is independent of context and node behavior. In case of MANET, specific protocols have been proposed for cooperation among nodes. Assuming specific features to make a trusted node may degrade the performance due to low battery power, low network efficiency and vulnerability to intruders. There exists a need for providing a degree of trust to each node and when it misbehaves, penalty should be given so that data packets should be not transferred through it. MANET is a self configuring network, thus a self learning system by means of observing and collecting local neighborhood information and transfer is needed. Trust can be established and it can be cooperated among the nodes. Thus the security optimization can be achieved. In AODV, a number of data and control packet exchange is required between nodes. These information exchanges need to be secured enough so that no data packet should be gone through a malicious node.

The node inside the network can act as a malicious node itself, so a dynamic trust among the nodes need to be established. Thus a node who receives a data packet successfully can recommend its neighbor who had passed the data packets to them. For each successful data exchange and for each unsuccessful data exchange trust of each nodes should be updated.

3 Issues in Designing Trusted Environment

The basic issue in a huge network is a node can be malicious at any time. A node is malicious when it deviates from normalcy in operation. To ensure end to end security in MANET, both node and route should be trusted at certain level. Sometimes a selected route which is the shortest path route may be congested or may contain more selfish nodes [4]. When congestion arises, the degree of packet dropping also increases, therefore to resolve the problem of congestion a node is said to be trusted only if they are of having less packet dropping ratio. Trust is a software entity and using trust a tight security can be enhanced in the network [3]. Next issue is regarding unused calculated trust. Calculated trust need to expire if they are unused for a fixed time interval. Thus the lifetime for the trust value is considered for establishing an uncongested trusted environment. Trust should be time dependent and the trustworthiness should grow and decay according to the time. Next issue is how to implement the security in MANET. Secure routing protocol can be based on a trusted third party exchange and without a trusted authority that is by computing trust from inherent knowledge in the system.

4 Related Studies

In [15] authors proposed a CONFIDANT protocol based on node observation by its neighbors and reputation is calculated and updated according to Bayesian estimation. Positive and negative reputation values are assigned according to the behavior of the nodes in the network. An improved CONFIDANT approach is also proposed by the same authors which includes an adaptive Bayesian reputation and trust management system.

Michiardi and Molva are proposed a CORE system [16] which uses game theory based approaches. In that a good reputation is calculated by the weighted mean of observations. The system includes functional reputation to combine subjective and indirect reputation to obtain global reputation value.

TAODV protocol [17] is a modified protocol with node trust value requires the modifications such as trust request, trust reply and modified extended routing table with positive events, negative events and opinion. Using this approach, secured routing can be achieved depends on the TREQ packet and its reply. The paper does not deal with the delay in the data transfer and the rate of packet dropping by a node so that it does not achieve a complete dynamic behavior.

Trust based secured routing in [3] includes addition of one more field to the ACK packet of AODV so that it records how many data packets received by the destination since last acknowledgement. Author does not dealing with the issue of unused calculated trust and trust level computation depending upon the packet dropping ratio.

In [2], author proposed to establish trust in a pure ad hoc network in which trust derivation based on the trust table which includes success or fail of RREQ, RREP, ACK packets. In this the complexity of creating many trust tables based on the success or failure of gratuitous route replies, black list and salvaging is very high. As MANET is an emergency formed network, complexity of trust table creation and updating should be reduced for better performance of the network.

Paper [8] described a trusted routing based on a fixed value of trust assigned to the nodes. It has been demonstrated to use a fixed value of the trust through the program and when it reached successfully on the destination and acknowledgement back to the source, the trust value is changed. According to this value the nodes are treated as trusted or not. Here there is no consideration of dynamic behavior of the nodes in the network so the maximum trust is not ensured.

5 Desired Model for Fuzzy Based Reputation System

Trust and reputation are inter dependent terms of achieving internal security in a mobile ad hoc network. By definition trust is the degree of notion of a particular node's belief on other node which can be fixed. Reputation is based on observing node's previous behavior in the domain and thus a node may like to create a trust based bridge between the nodes. This paper proposes a dynamic model for calculating reputation of different nodes in the network. The proposed trust model has three components, trust level computation, trust updating, and route trust calculation. The desired model is depicted in the figure1.

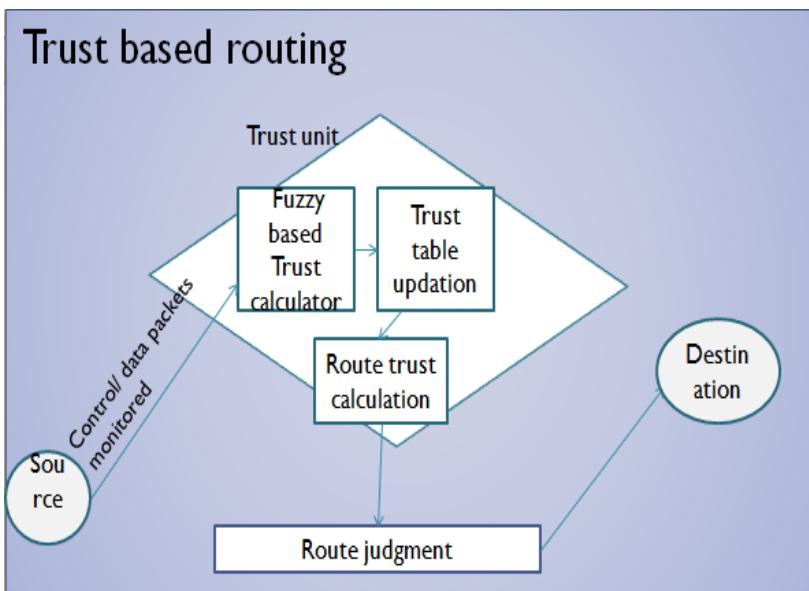


Fig. 1. Figure shows the Model for trust based routing

In this model when a source wants to send a data packet to a destination node, it first checks the trust table. The trust table stores the trust level determined by observing the node's past behavior. In addition to the trust value, packet delay and packet dropping value are also measured. Basic trust value computations are described in the subsections.

5.1 Trust Level Computation

A node's behavior is observed by its neighbors and trust level of a node is calculated from its previous experiences. A trust can be calculated at the regular intervals. The proposed model computes the trusting behavior based on the Mamdani fuzzy model. In the model, a ratio of the how many data packets it forwarded and how many data packet it needs to forward is calculated. Trust level is calculated based on this ratio is called trust membership. Degree of membership value of reputation based trust is in the fuzzy set [0, 1]. The node that sends all the desired packets will get 1 as the degree of membership function. Depends on the number of RREQ packets passed through the node, number of RREP packets, DATA packets and ACK packets the behaviour of the node can be analyzed. According to the membership function value, trust level of nodes can be treated as any of the following three categories. If the membership value is ≥ 0.7 are trusted nodes, if membership value $< 0.7 \text{ & } > 0.3$ are medium trusted nodes or normal node. If the membership value is less than or equal to 0.3 then they are treated as malicious nodes. This behavior is described in figure2.

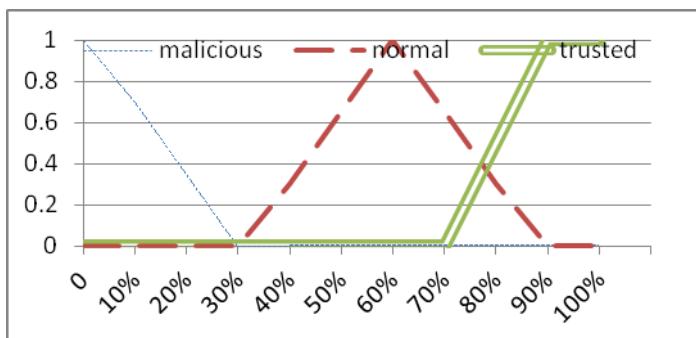


Fig. 2. Category of node based on the trust membership value

A node is having 30 % trust level is treated as malicious nodes, and the nodes between 30% and 70 % are are medium trusted nodes, and the nodes with trust level greater than 70 % are trusted nodes.

Each neighbor of the node is observing the node's behavior take the average of all neighbor's opinion for calculating the trust level of the particular node. Trust level of a node cannot be transitive, so an expected mean of all trusts on a particular node need to be computed to determine the trust level.

The fuzzy relation set for reputation includes

$$R = \{\text{trust value}/(x,y), \text{trust value}/(x,z), \dots\}$$

For each pair there exists a membership value for the trust between the node x and its neighbors y,z,...etc.

After calculating R then union of all reputation is calculated.

$$\text{Fuzzy union of membership values} = U [R_x^{nh1}, R_x^{nh2}, \dots, R_x^{nhn}]$$

where R_x^{nh1} denotes the reputation of node x .

$$\text{Expected mean of the trust level} = \mu_{R_x^{nh}} = f(R_x, n)$$

$$= U [R_x^{nh1}, R_x^{nh2}, \dots, R_x^{nhn}]$$

$$\frac{\sum}{n}$$

Where n is the number of neighbors of the node x.

5.2 Trust Table Computation

The dynamic fuzzy based trust model creates a trust table for each node. Contents of trust table include Node ID of the neighbor node, trust value according to the trust level ration, time stamp for denoting the time on which the particular node's trust value updated on the trust table.

Node ID	Trust level	Timestamp
X	R_x^{nh}	Time

Time stamp is added to each entry in the trust table denotes the life time of the trust value. The trust table needs to be update periodically. When a node has a particular level of distrust and even though this node may become the normal node still the node will be treated as malicious node till the next pause time.

5.3 Route Trust Calculation

The main components in the proposed model are the membership value of the trust level, delay through a path and packet delivery ratio of the nodes. At a regular interval, end to end delay is measured. If delay is greater than the threshold value, then a less priority is given to that route and avoids that delayed path for routing.. Also at regular intervals of time packet delivery ratio is calculated so that it is easy to know whether the node is dropping any data packet. Packet delivery ratio should be greater than the fixed threshold level of PDR. A node which is dropping less number of packets is given high priority.

$$\text{Node trust, } \delta = \Sigma [\Pi w_x(\alpha 1) R_x^{nh}, (1 - \Pi w_x(\alpha 2) D_x), (1 - \Pi w_x(\alpha 3) PDX)]$$

Where $\alpha 1$ denotes the category of nodes depends on the trust level value, $\alpha 2$ denotes the value of delay level, and $\alpha 3$ denote value of packet drop. $w_x(\alpha 1)$ is the weight of $\alpha 1$, $w_x(\alpha 2)$ is the weight of $\alpha 2$ and $w_x(\alpha 3)$ is the weight of $\alpha 3$. The range of these values is illustrated in the figure3.

Priority classification for each event can be given as

Prio (reputation based trust level) > prio (delay) > prio (packet loss)

Rule No	Trust level	Delay	PDR	Node trust
1	Max	Min	Max	Medium(normal)
2	Max	Min	Min	Trusted
3	Max	Max	Min	Trusted
4	Min	Min	Min	Medium(normal)
5	Min	Max	Min	Malicious
6	Min	Min	Max	Malicious
7	Min	Max	Max	Malicious
8	Max	Max	Max	Medium(normal)
9	Medium	Min	Max	Malicious
10	Medium	Max	Min	Medium(normal)
11	Medium	Min	Min	Medium(normal)
12	Medium	Max	Max	Malicious

Fig. 3. Fuzzy based trust computing table

From the figure we can interpret as if trust level is “max”, delay value “min” and PDR is “min”, then the node is a trusted node. The more the fuzzy rules the more the fuzzy compositions exist.

6 Conclusions

In this paper a dynamic fuzzy based reputation model for the assurance of node security for AODV in MANET is discussed. Internal security can be achieved by trust and reputation between nodes in the MANET. The need for trusted environment in mobile ad hoc network and the basic issues while designing the proposed model in AODV for MANET is also illustrated.

Reputation of the nodes is determined by observing the routing behavior of nodes in the past experiences. For this, set up a pause time and at regular intervals, the levels of trust is calculated. According to these levels the nodes are treated as either trusted node, normal node or malicious node. Trust value of the node is also depends on the delay through that node and the number of packets dropped by the node.

References

1. Martin Leo Manickam, J., Shanmugavel, S.: Fuzzy based trusted Ad hoc on demand distance vector routing protocol for MANET. In: 15th International Conference on Advanced Computing and Communications. IEEE (2007) 0-7695-3059-1/07
2. Pirzada, A.A., McDonald, C.: Establishing Trust In Pure Ad-hoc Networks. In: 27th Australasian Computer Science Conference, vol. 26. Australian Computer Society, The University of Otago, Dunedin, New Zealand (2004)
3. Menaka Pushpa, A.: Trust Based Secure Routing in AODV Routing Protocol. IEEE (2009) 978-1-4244-4793-0/09
4. Rashidi, R., Jamali, M.A.J., Salmasi, A., Tati, R.: Trust Routing Protocol based on Congestion control in MANET. IEEE (2009) 978-1-4244-4740-4/09
5. Rehmani, M.H., Doria, S., Senouci, M.R.: A Tutorial on the Implementation of Ad-hoc On Demand Distance Vector (AODV) Protocol in Network Simulator, NS-2 (June 2009)
6. Alfawaaer, Z.M., Al_zoubi, S.: A proposed Security subsystem for Ad Hoc Wireless Networks. IEEE (2009) 978-0-7695-3930-0/09
7. Veeraraghavan, P., Limaye, V.: Trust in Mobile Ad hoc Networks. In: Proceedings of the 2007 IEEE International Conference on Telecommunications and Malaysia International Conference on Communications, Malaysia. IEEE (2007) 1-4244-1094-0/07
8. Mangrulkar, R.S., Atique, M.: Trust Based Secured Adhoc on Demand Distance Vector Routing Protocol for Mobile Ad hoc Network. IEEE (2010) 978-1-4244-9730-0/10
9. Patnaik, G.K., Gore, M.M.: Trustworthy Path Discovery in MANET - A Message Oriented Cross-correlation Approach. In: 2011 Workshops of International Conference on Advanced Information Networking and Applications. IEEE (2011) 978-0-7695-4338-3/11
10. Chen, J.(T.), Boreli, R., Sivaraman, V.: TARo: Trusted Anonymous Routing for MANETs. IEEE (2010) 978-0-7695-4322-2/10
11. Inoue, S., Ishii, M., Sugaya, N., Yatagai, T., Sasase, I.: Trust Level Evaluation for Communication Paths in MANETs by Using Attribute Certificates. IEEE (2010) 978-1-4244-7057-0/10
12. Manoharan, R., Mohanalakshmie, S.: A Trust Based Gateway Selection Scheme for Integration of MANET with Internet. IEEE (2011) 978-1-4577-0590-8/11
13. Raza, I., Hussain, S.A.: A Trust based Security Framework for Pure AODV Network
14. Li, X., Lyu, M.R., Liu, J.: A Trust Model Based Routing Protocol for Secure Ad Hoc Networks. IEEE (2004) 0-7803-8155-6/04
15. Buchegger, S., Boudec, J.Y.L.: Performance analysis of the CONFIDANT protocol(Cooperation of Nodes-Fairness In Dynamic Ad-hoc NeTworks). In: The 3rd ACM International Symposium Mobile Ad-hoc Networking & Computing (2002)
16. Michiardi, P., Molva, R.: CORE: A Collaborative Reputation Mechanism to Enforce Node Cooperation in Mobile Ad-hoc Networks. In: The IFIP TC6/TC11 Sixth Joint Working Conference on Communications and Multimedia Security: Advanced Communications and Multimedia Security, Portoroz, Slovenia (2002)
17. Yunfang, F.: Adaptive Trust Management in MANETs. In: Proc. 2007 Int'l Conference on Computational Intelligence and Security, Harbin, China, December 15-19, pp. 804–808 (2007)
18. Al-Arayed, D., Pedro Sousa, J.: A Survey of Trust Management Systems

Policy Based Traffic in Video on Demand System

Soumen Kanrar

Vehere Interactive Pvt Ltd Calcutta 700053, India

Soumen.kanrar@veheretech.com

Abstract. Performance of the video on demand system highly depends upon the efficient traffic processing. Video stream data processing becomes a very interesting area over the decade for its exponential rapid growth market. Existence packet flow control mechanism can't handle the video streaming smoothly. Streaming video content sent in compressed form over the network and displayed by the viewer in real time. Stream data stream contains the audio and video data. A good policy is highly required to control the loss of video stream data over the high speed network. By efficient control mechanism of the video streaming to reduce the over burden load to the VOD system and increases the system performance of centralized as well as distributed VOD architecture. A good policy of the video stream control mechanism efficiently implements on the all types topologies. This paper presents the policy based traffic control mechanism for the video streaming in real time scenario.

Keywords: Video stream, Video on demand, Erlang, Policy.

1 Introduction

Over the year, researchers interested to develop the efficient and complex scheduling algorithm to handle the multiple request classes. The most of the research work is concentrated on the area of multiple video server design, process scheduling and admission control at storage server [2]. Some of works on the area of disk striping and video block placement. [3], [7], [10]. The papers [2] presents different types of models and server buffer management to transfer the video stream data to the user through the broadband network. In the paper [9][10] authors discussed about the real time net works, centralized or distributed analyzing network conditions for guaranteed services. Recent papers in the area of distributed VOD systems focus on the load balancing schemes. In the paper [4] authors focus on the movie placement and load balancing in a distributed VoD architecture. It follows a special placement method that each of the video clips or movies assigns to each level of connected node in the graph topology. So the expected demand increases the dynamic priority of the video clips or movie as the better utilization of the network resources. The bandwidth capacity is used to characterize the network in the load balancing scheme in [2]. This paper is different from [1] and [4] in that load balancing is not the primary focus. In this paper, explicitly presents an efficient policy model using a real-time network for general purpose request handling in a hierarchical VOD system. The distributed system use to enhance the performance of all types of multimedia services in the next generation network. The packet loss delay jitter has high impact over the performance in the distributed

VOD system performance for the end to end video stream delivery. An efficient stream handle policy is highly required for the next generation VOD network system. A policy is good means how efficiently it minimize the traffic load in the video on demand system to provide better utilization of the available bandwidth in the very race situation. The topic of this paper is the development of a unified traffic policy for a hierarchical Video-on-Demand (VoD) system. The policy integrates video server retrieval and video stream transfer mechanisms by relating buffer management at the server with rate-based scheduling at the network. The significance of this paper to development of a policy for designing and measuring the effectiveness of end-to-end video stream control and request handling policies. The aim of this paper to presents a unified policy model for a hierarchical VOD architecture consisting of clusters of client population subsets, cluster of storage servers and the interconnected nodes. Most of the previous research is focused on topics related to a single video server design such as disk striping, video block placement, and admission control at the level of disks and branch of disk handling [6], [7], [8]. The request comes from the client to the VoD server for the different types of request. Main categories are classified in two broad areas. One for the popular clips and other for the unpopular clips. The request are classified in two types in the VOD Network. The first one of the request for initializing or starting the video clips. Other type is the request for interactive service (e.g. stop/pause, jump forward, fast reverse etc) to be performed on the viewed clips. Since each of these request is independent from each other, and arrival requests come from large numbers of client set-up terminals, the arrival process of normal requests as well as of interactive requests to the video server can be modelled as a Poisson distribution with average rates λ_s (for steady session) and λ_i (for interactive session) respectively. With this assumption, the distribution of the sum of K of independent identically distributed random variables, representing the request inter arrival times (Which are exponential distributed mutually independent random variables) is then follow the Erlang distribution. Multicasting by the use of clustering concept significantly improves the VoD system performance greatly by reducing the required network bandwidth. So the overall network load reduces. In other way the multicasting alleviates the workload of the VoD server and improves the system throughput by batching requests. Multicasting offer excellent scalability which in turn, enables serving a large number of clients that provide excellent cost/performance benefits. The result presents in this paper explicitly, how a good traffic policy model makes control traffic inside the system. The paper is structured as follows. Section 1 introduces the major area related to Policy based traffic model requirement and the papers survey. The following section 2 discusses the Traffic flow policy model. The section 3 presents simulation parameter requirement for the experiment the Policy based traffic model. The section 4 presents the result and section 5 explains the further improvement of this policy model.

2 Traffic Flow Policy

Let the request comes from i^{th} class of population and served by the j^{th} partition block of the server where $1 < j$ and assume $A_j B_i B_{i+1} \dots B_{j-1}$ is the event that

the previous all partition from i to $j-1$ is blocked only j^{th} partition has at least one free port.

$B_i B_{i+1} \dots B_{j-1}$ represents the event that all the partition from i to $j-1$ is blocked.

We get,

$$\begin{aligned}
 & p(A_j / B_i B_{i+1} \dots B_{j-1}) \\
 &= \frac{p(A_j B_i B_{i+1} \dots B_{j-1})}{p(B_i B_{i+1} \dots B_{j-1})} \\
 &= \frac{p(A_j) p(B_i) p(B_{i+1}) \dots p(B_{j-1})}{p(B_i) \dots p(B_{j-1})} \\
 &= p(A_j) \\
 &= \left(1 - \frac{1}{k}\right)^{j-1} \cdot \frac{1}{k} \left(\frac{C_j - Q_i^{(j)}}{C_j}\right)
 \end{aligned} \tag{1}$$

Since $p_b(B_m \cap B_n) = \Phi$, $\forall (m, n) \in N$ and $m \neq n$ where $i \leq m \leq j-1, i \leq n \leq j-1$.

So we get

$$p_b(B_i B_{i+1} \dots B_{j-1}) = p_b(B_i) p_b(B_{i+1}) \dots p_b(B_{j-1}) \tag{2}$$

Each $p_b(B_i)$ for $i \in I^{\geq o}$ follows the Erlang distribution.

So the above expression can be written as

$$\begin{aligned}
 & p_b(B_i B_{i+1} \dots B_{j-1}) \\
 &= \left[\frac{\frac{E_i^{C_i}}{(c_i)!}}{\sum_{k=0}^{c_i} \frac{E_i^k}{(k)!}} \right] \cdot \left[\frac{\frac{E_{i+1}^{C_{i+1}}}{(c_{i+1})!}}{\sum_{k=0}^{c_{i+1}} \frac{E_{i+1}^k}{(k)!}} \right] \dots \dots \left[\frac{\frac{E_{j-1}^{C_{j-1}}}{(c_{j-1})!}}{\sum_{k=0}^{c_{j-1}} \frac{E_{j-1}^k}{(k)!}} \right]
 \end{aligned} \tag{3}$$

Now the request from the class a will be admitted to a partition b with probability

$$p^\#(a, b) = p_a^\# \tag{4}$$

Where, $\sum_{i=1}^k p_i^\# = 1$, A new request arrives with Poisson distribution for video stream.

$B_i B_{i+1} \dots B_{j-1}$ represents the event that all the partition from i to $j-1$ is blocked. Now by considering the expression (1) we get,

$$\begin{aligned}
& p(A_j / BB_i, BB_{i+1}, \dots, BB_{j-1}) \\
& = (p_a^\#) \frac{p(A_j | BB_i, BB_{i+1}, \dots, BB_{j-1})}{p(BB_i, BB_{i+1}, \dots, BB_{j-1})} \\
& = (p_a^\#) \frac{p(A_j) p(B_i) p(B_{i+1}) \dots p(B_{j-1})}{p(B_i) \dots p(B_{j-1})} \\
& = (p_a^\#) p(A_j)
\end{aligned} \tag{5}$$

3 Simulation Environment

The simulation environment is created according to the problem requirement.

For the simplicity, we present the simulation scenario as each sector of the server contained equal number of ports. The traffic arrived rate from different cluster of clients, started at 0.1 Mb/s and end at 20 Mb/s. The number of clusters of client varied between 10 to 50. The number of subsection in the server side is 20. Port access time for each client vary from 1sec to 100 seconds for each video clips. The simulation runs for 360 seconds i.e. 5 minute. Table 1 presents the overall simulation parameters for the VoD server.

Table 1. Simulation parameters.

The Simulation Parameters of VOD System	Parameters Value
Number of Clusters vary	10 to 50
Minimum traffic Arrival rate	0.1 Mb/sec
Maximum Traffic Arrival Rate	20 Mb/sec
Port access time minimum	1 sec
Port access time maximum	100 sec
Number Sub section	20
Simulation Time	360 seconds

4 Results and Performance Analysis

Figure 1 to 6 represents the performance of distributed video on demand system with respect to the policy based traffic. Fig.1 and Fig.2 represents the traffic scenario in the

distributed video on demand system without implementation of any traffic control policy. The request comes from the variable Clusters. The sample cluster client population varied from 10 to 50. The traffic rate of the video stream started from 0.1Mb/second and maximum the video stream rate 20 Mb/sec.

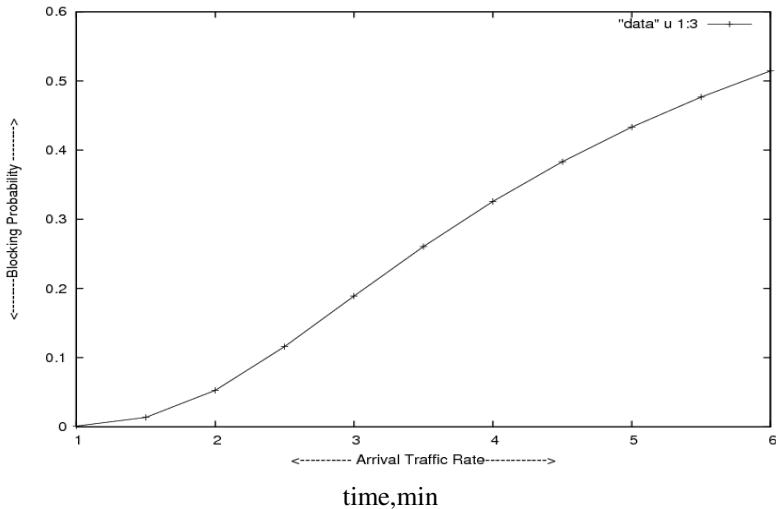


Fig. 1. Traffic arrival / blocking probability (with arrival traffic)

Figure 1 and figure 2 represents the traffic congestion inside the video on demand system with respect to the blockage probability. The diagram shown that if the arrival video stream is gradually increasing the blockage probability is gradually increasing that explicitly explains the congestion inside the VOD system.

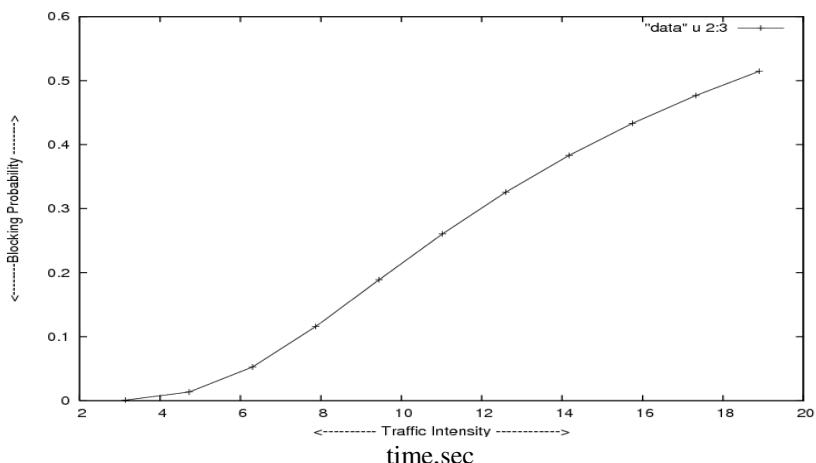


Fig. 2. Traffic arrival / blocking probability (with traffic intensity)

Figure 2 represents the scenario there is initially no blockage exist when the initial video stream traffic does not exist during the interval 1 second to 3 second. After that when the video stream traffic intensity increases the resultant blockage probability increase. The blockage probability directly indicates the high chances of video stream loss.

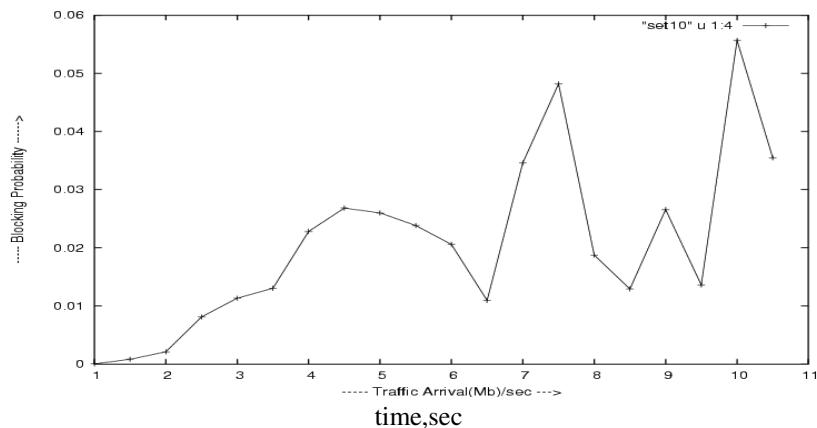


Fig. 3. Traffic arrival / blocking probability (with policy based traffic arrival)

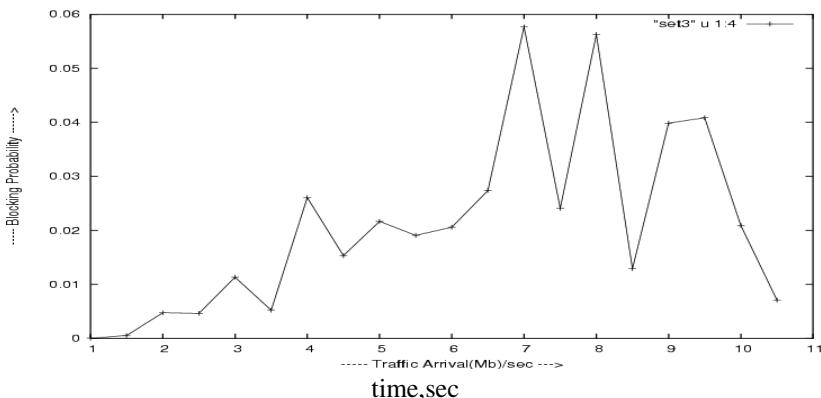


Fig. 4. Traffic arrival / blocking probability (with policy based traffic arrival)

The figures 3, 4, 5 and 6 represent the load of the traffic scenario with respect to policy based video stream model. Here the diagrams show the snapshot of the video on demand system with respect to the blockage probability. We consider the snapshot of the video on demand system for 11 seconds. During the 11 seconds, we see the blockage probability is well controlled in the probability interval [0,0.06].

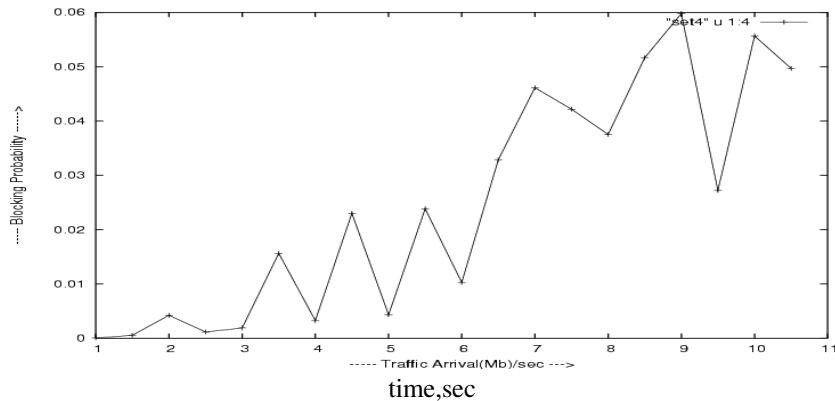


Fig. 5. Traffic arrival / blocking probability (with policy based traffic arrival)

The policy based traffic model clearly indicate that an efficient model well control the video stream inside the VOD system. So the packet loss, delay and the packet jitter are well controlled. So the system performance is increase efficiently.

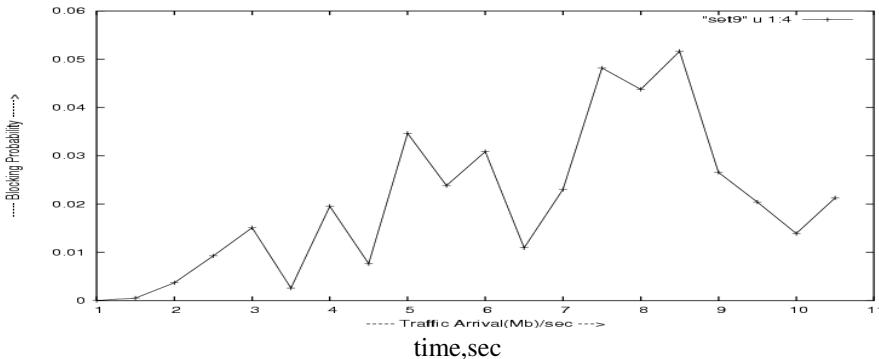


Fig. 6. Traffic arrival / Blocking probability (with policy based traffic arrival)

5 Conclusion and Future Research Section

The study is remaining to compare the result with real data. For the congestion control and to minimize the jitter of the VOD stream data, needs better algorithm to develop. This paper present an over views of the impact of the policy based traffic on the VOD system. The policy based traffic model implicitly represents the system performance.

References

- [1] Mundur, P., Simon, R., Sood, A.K.: Class Based Access Control for distributed Video on demand Systems. IEEE Transactions on Circuits and Systems for Video Technology 15(7) (July 2005) (March/April 2007)

- [2] Kanrar, S., Siraj, M.: Class Based Admission Control by Complete Partitioning Video on demand Server. IJCNC 2(1) (May 2010), doi:10.5121/ijcnc.2010.2308
- [3] Kanrar, S.: Efficient Traffic Control of VOD System. IJCNC 3(5) (September 2011), doi:10.5121/ijcnc.2011.3507
- [4] Bisdikian, C., Patel, B.: Issues on movie allocation in distributed Video-on-Demand systems. In: Proceedings of the IEEE ICC, pp. 250–255. IEEE Press (1995)
- [5] Brubeck, D.W., Rowe, L.A.: Hierarchical storage management in a distributed VoD system. IEEE Multimedia, 37–47 (Fall 1996)
- [6] Gemmell, D.J., Vin, H.M., Kandlur, D.D., Rangan, P.V., Rowe, L.A.: Multimedia storage servers: A tutorial. IEEE Computer 28(5), 40–49 (1995)
- [7] Wang, Y., Liu, J., Du, D., Hsieh, J.: Efficient video file allocation schemes for video-on-demand services. Multimedia Systems 5(5), 283–296 (1997)
- [8] Wolf, J., Yu, P., Shachnai, H.: Disk load balancing for Video-on-Demand systems. Multimedia Systems 5(5), 358–370 (1997)
- [9] Georgiadis, L., Guerin, R., Peris, V., Rajan, R.: Efficient support of delay and rate guarantees in an internet. In: ACM SIGCOMM, pp. 106–116 (August 1996)
- [10] Parekh, A., Gallager, R.: A generalized processor sharing approach to flow control - the multiple node case. IEEE/ACM Transactions on Networking 2(2), 137–150 (1994)
- [11] Kanrar, S., Siraj, M.: Performance of Multirate Multicast in Distributed Network. IJ CNS 3 (June 2010), doi:10.4236/ijcns.2010.36074
- [12] Annapu Reddy, S.: Exploring IP/VOD in P2P Swarming Systems. In: Proc. INFOCOM 2007, Anchorage, AK, pp. 2571–2575 (May 2007)
- [13] Agrawal, D., Beigi, M.S., Bisdikian, C., Lee, K.-W.: Planning and Managing the IPTV Service Deployment. In: 10th IFIP/IEEE International Symposium on Integrated Network Management, vol. 25(21), pp. 353–362 (May 2007)
- [14] Lee, G.M., Lee, C.S., Rhee, W.S., Choi, J.K.: Functional Architecture for NGN- Based Personalized IPTV services. IEEE Transaction on Broadcasting 55(2) (June 2009)
- [15] Deering, S., Cheriton, D.: Multicast routing in datagram internetworks and extended LANs. ACM Tran. on Computer Systems, gS-111 (May 1990)
- [16] Kanrar, S.: Efficient Traffic control of VoD System. IJCNC 3(5), 95–106 (September 2011), doi:10.5121/ijcnc.2011.3507
- [17] Souza, L., Ripoll, A., Yang, X.Y., Hernandez, P., Suppi, R., Luqu, E., Cores, F.: Designing a Video on Demand System for a Brazilian High Speed Network. In: Proceedings of the 26th IEEE International Conference on Distributed Computing Systems Workshops (2006)
- [18] Kanrar, S., Siraj, M.: Performance Measurement of the Heterogeneous Network. IJCSNS 9(8), 255–261 (2009)

Distance Aware Zone Routing Protocol for Less Delay Transmission and Efficient Bandwidth Utilization

Dhanya Sudarsan¹, P.R. Mahalingam¹, and G. Jisha²

¹ Department of Computer Science

² Department of Information Technology

Rajagiri School of Engineering & Technology, Rajagiri valley, Cochin, India

{dhanyasudarsan127, prmahalingam}@gmail.com, jishag@rajagiritech.ac.in

Abstract. Zone Routing protocol (ZRP) combines the best features of both proactive and reactive MANET routing protocol and hence comes under the category of MANET hybrid routing protocol. ZRP is based on the concept of zones where zone is defined based on the number of hops, the actual physical distance is not considered. The zone size is directly propositional to the distance between the sender and border nodes, so if the distance increases, the radio coverage of the sender node will not be able to reach the border nodes in the zone, so that the sender node will need to increase the number of broadcasts to find the border nodes in the zone, which will obviously increase the bandwidth utilization. Since effective utilization of bandwidth is one of the major issues faced by MANET routing protocols the paper proposes a modification for ZRP by considering physical distance of the nodes also as a factor in determining the Zone radius for the effective utilization of the bandwidth and an easy cost effective way to determine the distance between the nodes using triangulation. Moreover considering the actual distance leads to the selection of shortest path and thus decreases the delay in packet delivery.

Keywords: ZRP, DZRP, Ad-hoc network, Routing, MANET.

1 Introduction

Mobile Adhoc Network (MANET) is a promising area of research under wireless network since it is a self-configuring network of mobile devices connected by wireless links. Mobility of nodes is the main attraction of MANET [1]. Because of its fast and easy of deployment, robustness, and low cost MANET is supporting a wide range of applications. Many protocols are designed for MANET. On the basis of routing information update mechanism MANET protocols are mainly classified into proactive routing protocols (Table driven) and reactive routing protocols (On Demand)[2].In the proactive routing protocol every node maintains the routing information to every other node in the network. In the reactive approach topology information is not stored, whenever a node wants to communicate with the other node in the network a route discovery procedure is initiated.

The problem with the proactive routing protocols is that it uses excess bandwidth to maintain routing information since frequent topology change happens and routing

table of each node has to be updated periodically. This also results in increased traffic across the network. The problem with on demand approach is that it involves a long route acquisition delay since the path has to be computed each time when needed and is also inefficiently floods the entire network for route determination. All these lead to the hybrid MANET routing protocol which combines the best feature of both proactive and reactive approaches.

Zone routing protocol comes under the category of hybrid MANET Routing Protocol. Zone routing protocol as its name implies is based on the concept of zones. Based on the current network traffic an optimal zone radius is determined. The zone radius r expressed in hops. The zone thus includes the nodes, whose distance from the node is at most r hops. The zones of the neighboring nodes overlap[5].

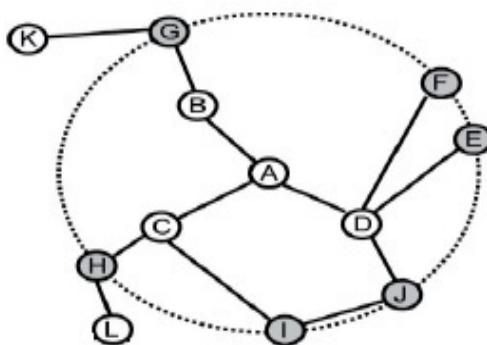


Fig. 1. Example routing zone with $r=2$

If for a node inside the zone, routes are immediately available, ie it uses a proactive approach. For the nodes whose destination is outside the zone, an on demand approach is used.

2 ZRP Routing Framework

The ZRP protocol suit consists of IARP (Intra Zone routing protocol)[6], IERP (Inter Zone Routing Protocol)[7], BRP (Border cast Routing Protocol) [8] and NDP (Neighborhood Discovery Protocol)[5].

If the destination node is within the zone, the packets are routed directly to destination since routing information is available immediately [5]. The IARP maintains routes for the nodes within the routing zone. The IERP is responsible for finding routes to destinations located beyond a node's routing zone. The IERP operation is initiated by checking whether the destination is within its zone. If so, no further route discovery processing is needed. If the destination is not within the source node's routing Zone, the source sends a route request to all its border nodes (nodes whose distance is equal to the zone radius).Now all the border nodes checks whether the destination is in its zone, if so route reply is sent back to source, else it will repeat

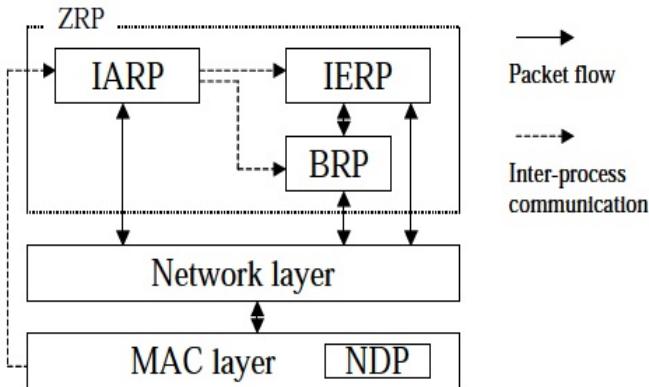


Fig. 2. ZRP Architecture[5]

the border casting process. The BRP is used in the ZRP to direct the route requests to the peripheral nodes(border nodes).ZRP relies on NDP provided by the Media Access Control (MAC) layer to check whether the neighboring nodes are alive or not.NDP transmits “HELLO” beacons at regular intervals say $T_{beacon}=0.25$ s.This short beacons contains only the source address. Upon receiving a beacon, the neighbour table is updated. If a new beacon fails to arrive within $2.T_{beacon}$ of the most recent beacon, a link failure is reported. Neighbours, for which no beacon has been received within this specified time, are removed from the table.

3 Motivation to Distance Aware ZRP

The main issue in ZRP is determining the optimal Zone radius for which hybrid min search/traffic adaptive zone radius estimator gives the optimal measurement [5]. But the ZRP protocol determines the zone size only based on the number of hops, the actual physical distance between the nodes is not taken into consideration. Since MANET uses radio waves for communication, the bandwidth of radio waves is limited. So if the radius of the zone gets increased the radio coverage of the sender node will not reach the border nodes in its zone, so the number of broadcasts needed to find the border nodes in the zone, which will obviously increase which result in the increased utilization of bandwidth [15]. Effective utilization of bandwidth is one of the major issues faced by MANET routing which motivates the proposal of distance aware ZRP.Speed of packet delivery also gets increased.

4 Proposed Modification

The proposed enhancement to ZRP for the effective utilization of bandwidth is to assign weight to edge where the edge weight is the actual physical distance between the nodes having links to each other. Other than number of hops, weight is also

considered while checking whether a node can be included within the zone. A threshold value for distance is set according to the available bandwidth. So the nodes which exceed the threshold are not included within the zone.

The reason why the physical distance is not considered for ZRP is that it is not feasible for MANET since high cost is incurred in measuring the physical distance between nodes due to its very high mobility. So the paper also proposes a novel cost effective approach to measure the physical distance.

5 Physical Distance Measurement

The concept of physical distance measurement has been available in MANET in a variety of proposals. But they had one inherent problem. They were totally based on the methods like:

1. Finding coordinates with respect to fixed point on the ground. The fixed points act like origin, and the position plotted like in a graph to find distance.
2. Using GPS to track respective positions, and find distance.

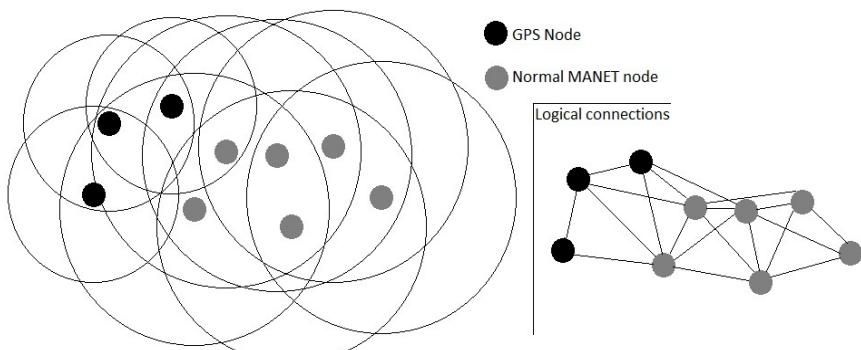
The first method is not very easy to implement since all nodes have to stay in range of the fixed coordinate nodes. So, it causes a lot of cluttered nodes, and restricts the mobility to a great degree.

GPS is a promising solution. But it needs special GPS trackers to be fixed to each individual node. Hardware cost may be considerable here.

Now, we propose a solution that incorporates some aspects of both the above methods.

Here, instead of equipping each device with its own GPS receivers, we equip 3 fixed nodes with GPS receivers, and a one-time computation. Once computed, it is stored in the node permanently. But even then, the first method cannot be applied since it cuts down the allowed range. So, instead of direct communication, we go for hops.

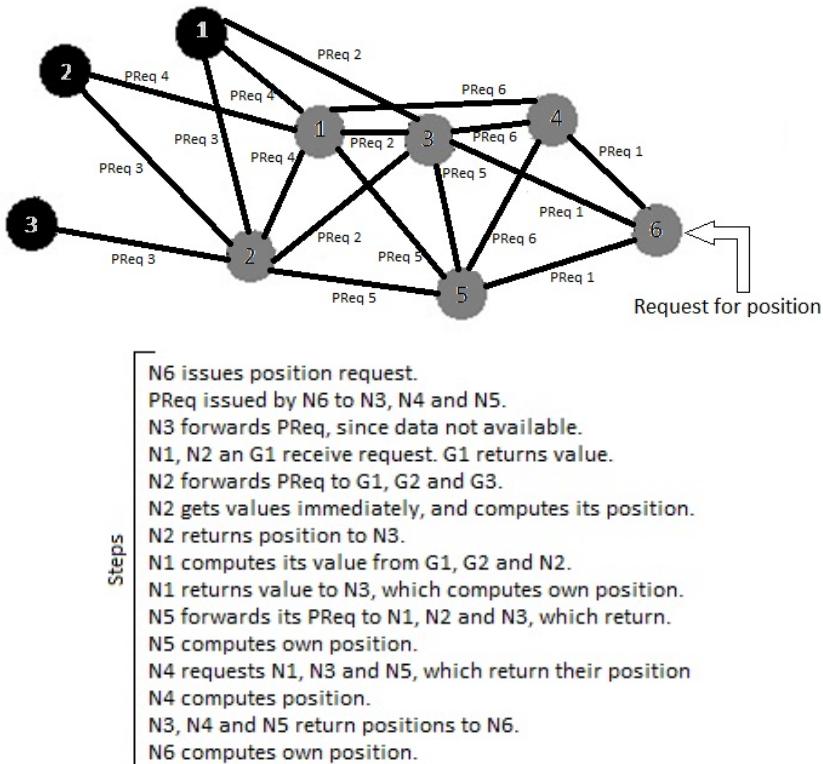
Consider the example here.



This is a sample MANET. The 3 black nodes are the fixed GPS receivers, while the gray nodes are normal MANET nodes.

When a node wants to know its position, it locates 3 one-hop neighbours and sends a *PReq* (Position Request) to all of them. If they have precomputed positions, they simply return it via a *Prep*. Else, they themselves perform the same procedure and get their positions.

An example is:



Position calculation: Here, position is calculated similar to the principle in GPS. ie, triangulation. Each node has to take a minimum of 3 available nodes and use them as a reference to get its own position. In GPS, we use 3 satellites, while here, we use either the GPS-node or precomputed MANET nodes.

Consider the case of a node N with position (x, y) to be computed. It has 3 reference nodes N0, N1 and N2, having coordinates (x_0, y_0) , (x_1, y_1) and (x_2, y_2) .

Now, once the *PReq* is received by a node, it immediately acknowledges it by a *PAck*. This is a high priority operation, and has minimal processing involved. So, it can give the RTT upto a high degree of accuracy, and this can give the distance from N to N0, N1 and N2.

Let distance from N to N0 = d_0 , N to N1 = d_1 and N to N2 = d_2 .

Then, we can calculate the coordinates (x,y) as:

$$x = \frac{c1(y1-y2)+c2(y1-y0)}{2[x0(y1-y2)+x1(y2-y0)+x2(y0-y1)]}$$

$$y = \frac{c1(x1-x2)+c2(x1-x0)}{2[y0(x1-x2)+y1(x2-x0)+y2(x0-x1)]}$$

where

$$c1 = x_0^2 + y_0^2 - x_1^2 - y_1^2 - d_0^2 + d_1^2$$

$$c2 = x_1^2 + y_1^2 - x_2^2 - y_2^2 - d_1^2 + d_2^2$$

6 Advantages

This saves a lot of hardware and helps in a lot of cases where positioning is vital. Also, this is not a very complex algorithm, and very less processing is needed at the nodes themselves.

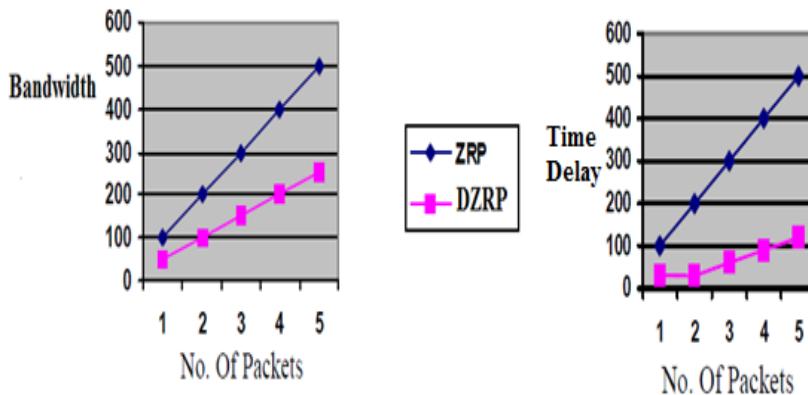
But at the same time, it can cause an increase in the traffic, upto 25%. But we can extend such that recomputation is done only when the node moves, cutting down the traffic to just above 10%.

7 Modified IERP and BRP

For the implementation of Distance aware Zone Routing Protocol the protocols IERP and BRP should be modified. Since the improvement comes into play only when the destination node is outside the zone where needs a bordercasting, the broadcast routing table of each table is be modified according to the actual distance from the source node to the border node. This threshold distance is set as constant based on the available bandwidth. The border node which exceeds the threshold distance from source node is removed from the list of border nodes in the broadcast routing table of the source node. This naturally improves the IERP and BRP protocol.

8 Result Obtained

The proposal is implemented using ns2.33.Result obtained is the graph showing the comparison of normal ZRP and the distance aware ZRP.Parameters considered in the awk script are number of nodes, number of broadcasting nodes, number of broadcast per node and Time delay.



According to the result obtained it is obvious that the bandwidth and time consumption is considerably reduced in DZRP.

Table 1. Simulation parameters

Parameter	Value
Number of Packet(Traffic) Sources	5
Topology Size	800x800 m
Transmission(radio) range	250 m
Traffic Type	Constant Bit Rate (CBR)
Packet size	512 Bytes
Standard packet sending rate	5 packets/sec [20Kbps]
Standard ad hoc host speed	20 m/s (Max)
Mobility Model	Random waypoint
Pause Time	5 sec
Simulation Time	900 sec
Wireless channel bandwidth	2 Mbps
ZRP periodic route update interval	10 sec
Zone radius	2 Hops

9 Applications

Since it imposes an additional overhead of distance measurement and increase in traffic it is most suitable for applications like military where delay is strictly intolerable .We can also use it in cases like Ham-Radio where positions have to be calculated.

10 Conclusion

Distance aware ZRP provides efficient bandwidth utilization and reduces the total time required by the packet to reach the intended destination in addition to all the advantages due to hybridization. More over triangulation based distance calculation reduces the cost of computing physical distance considerably.

11 Future Work

The proposal of considering the distance as a factor for determining the zone radius can be also used for efficient utilization of power [15] since according to Inverse Square Law, the power received by the receiving node is inversely proportional to square of the distance between the nodes. So the power required for delivery varies according to the node's distance.

References

1. Taneja, S., Kush, A.: A Survey of Routing Protocols in Mobile Ad Hoc Networks. International Journal of Innovation, Management and Technology 1(3) (August 2010)
2. VijayaKumar, Vasudeva Reddy, Y., Nagendra, M.: Current Research Work on Routing Protocols for MANET: A Literature Survey. International Journal on Computer Science and Engineering 02(03) (2010)
3. Royer, E., Toh, C.-K.: A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks. IEEE Personal Communications (April 1999)
4. Siva Ram Murthy, C., Manoj, B.S.: Ad Hoc Wireless Networks, pp. 213–248, 321–367. Pearson Education, Inc. (2004)
5. Pearlman, M.R., Haas, Z.J.: Determining The optimal configuration for Zone routing Protocol. IEEE Journal in Selected Areas of Communication 17(8) (August 1999)
6. Haas, Z.J., Pearlman, M.R., Samar: Intrazone Routing Protocol (IARP), IETF Internet Draft (June 2001)
7. Haas, Z.J., Pearlman, M.R., Samar: Interzone Routing Protocol (IERP), IETF Internet Draft (June 2001)
8. Haas, Z.J., Pearlman, M.R., Samar: The Bordercast Resolution Protocol (BRP) for AdHoc Networks, IETF Internet Draft (June 2001)
9. Li, H.J., Qiu, F.-Y., Liu, Y.-J.: Research on Mechanism Optimization of ZRP Cache Information Processing in Mobile Ad Hoc Network. IEEE (2007)
10. Samar, P., Pearlman, M.R., Haas, Z.J.: Independent zone Routing: An adaptive hybrid routing framework for adhoc wireless Network. IEEE/ACM Transaction on Networking 12(4) (August 2004)
11. Fahmy, I.M.A., Nassef, L., Saroit, I.A., Ahmed, S.H.: QoS Parameters Improvement for the Hybrid Zone-Based Routing Protocol in MANET. Institute of Statistical Studies and Research Computer Science & Information Department, Cairo University
12. Sinha, P., Krishnamurthy, S.V., Dao, S.: Scalable Unidirectional Routing with Zone Routing Protocol (ZRP) Extensions for Mobile Ad-Hoc Networks. IEEE (2000)

13. Aggelou, G., Tafazolli, R.: A bandwidth efficient routing protocol for mobile ad hoc Networks. In: ACM International Workshop on Wireless Mobile Multimedia (August 1999)
14. Subramaniam, A.: Power Management In Zone Routing Protocol (ZRP). University of Central England, Birmingham
15. Feeney, L.M.: An energy consumption model for performance analysis of routing protocols for mobile ad hoc networks. Mobile Networks and Applications (June 2001)

Mobile Data Offloading: Benefits, Issues, and Technological Solutions

Vishal Gupta¹ and Mukesh Kumar Rohil²

Department of Computer Science and Information Systems
Birla Institute of Technology and Science, Pilani
bestgupta@hotmail.com,
rohil@bits-pilani.ac.in

Abstract. Apart from voice services, data made its foray in cellular networks with 2.5G networks. Today, with 3G network already in place, the data requirements of mobile subscribers is very high. With the increasing demand for mobile internet and rich data services such as streaming media for audio and video, this data requirement is expected to multifold in near future. Correspondingly the existing network infrastructures will have to scale to satisfy this huge bandwidth demand in future. The simple solution for this is to build up new infrastructure. But huge investments are involved with it. So, network operators have started looking for the alternative ways of satisfying this data needs. Among many alternatives, mobile data offloading is a most promising one. This paper presents the extensive need, benefits, and technological solutions for Mobile data offloading.

Keywords: Wi-Fi, Mobile Data Offloading, 3G.

1 Introduction

Wireless cellular network now covers most of the populated-part of the world. According to the Information and Communication Technology (ICT) statistics of International Telecommunication Union (ITU), by the end of 2010 there were an estimated 5.3 billion mobile cellular subscriptions worldwide [1]. It also claims that access to mobile networks is now available to 90% of the world population and 80% of the population living in rural areas. In fact, mobile cellular growth has started slowing in developed countries and is slowly reaching saturation levels with an average 116 subscriptions per 100 inhabitants at the end of 2010, just a marginal growth of 1.6% from 2009-2010 [1].

For many years in the past, the cutting edge services provided by cellular network were voice and SMS. But mobile phones – especially smart phones and iphones - radically changed it and are now poised to take over the traditional information access devices as the dominant platform for accessing the information. The nature of data transformed from conventional and plain text to emails, multimedia and chats. Subscribers now have an easy access to streaming media for video and audio. So, now apart from voice services, it is the data services which govern the telecom industry. In May 2009, the packet data has put nine times more load than voice services in North

American networks [2]. So, to facilitate these, many countries have already started offering 3G services and people are moving rapidly from 2G to 3G platforms in both developed and developing countries. In fact, in 2010, there were 143 countries offering 3G services commercially, compared to 95 in 2007 [1]. Also, by the end of 2010, the number of 3G subscriptions, worldwide, was expected to reach the mark of 940 million [1].

It is also been speculated that in future no single existing wireless network technology can simultaneously provide low latency, high bandwidth, and wide-area data service to a large number of mobile users [3]. Among many alternatives, offloading the data onto some overlay network is a promising solution. In this paper we emphasize that Wi-Fi network is the best technology for Mobile Data Offloading.

2 Problems with 3G Network

In this section we show that why there is a problem and exactly where the problem lies.

2.1 Why There is a Problem

Following are the three basic reasons which contribute to why current 3G network infrastructure cannot sustain to the future bandwidth requirements:

1. Increasing Network Traffic

3G offers rich data services and the bandwidth consumption from this is not expected to slowdown. Let us see few statistics in support this point:

- Unwired Insight anticipates a 20-fold growth in 3G traffic to 2014 [2]. From present requirement of 10MB per month for audio, video, photos, software and email downloads, it is expected to grow up to 2GB within five years.
- Juniper Research estimates that the cost of delivering mobile data could rise sevenfold to \$370 billion by 2016 [4].
- According to network provider Cisco Systems, in 2015 mobile data traffic will be 26 times higher compared to 2010, mainly caused by the use of mobile video services [5].
- Mobile data traffic will grow at a compound annual growth rate (CAGR) of 92 percent from 2010 to 2015, reaching 6.3 hexabytes per month by 2015. The data capacity requirements are increasing 150 percent per year [6].

Thus, using current 3G network architectures, the cost of supporting the anticipated exponential traffic growth generated from mobile data services is unsustainable. This subsequent anticipated explosion of data traffic on 3G networks has caused an immediate need for carriers to seriously think of the alternatives so that both voice and data services can perform optimally.

2. Increasing usage of Smart Phones

Accelerated adoption of Smartphone by mobile phone subscribers, in combination with the much higher usage profile of Smartphone relative to basic handsets is the

major cause for the unexpected data surge. Operators are seeing increasing data traffic driven by the growth of Smartphone's and other connected devices that offer ubiquitous Internet access. Let us see again few statistics in support of this point:

- According to Juniper Research the amount of mobile data traffic generated by smart phones, feature phones and tablets will exceed 14,000 Petabytes by 2015, equivalent to almost 18 billion movie downloads or 3 trillion music tracks [7].
- Informa Telecoms and Media data indicates that the number of Smartphone in use grew by 32 percent during the year 2010 while it was anticipated as 22 percent [8].
- In 2011 over 85 percent of new handsets will be able to access the mobile Web [8].
- In developed countries the maximum of the mobile subscribers have an internet ready phone. In fact in US and Western Europe this number is about 90%.
- 98 percent of iPhone users use the data features of their phones. iPhone users are four times as likely to use the Internet as a typical subscriber, five times as likely to download an application, six times as likely to watch mobile video, and seven times as likely to use location based services [9].

Above all, out of total global handsets in use today, Smartphone represent only 13 percent but they represent over 78 percent of total global handset traffic [5]. Average Smartphone usage doubled in 2010 and this trend is expected to continue.

3. Spectrum is Costly and Scarce

Telecommunication systems all require a certain amount of electromagnetic bandwidth to operate. In different parts of the world, different organizations allot parts of the overall electromagnetic spectrum to different uses. Also, in many parts of the world, international agreements are required so that communications systems in neighboring countries are not interfering with each other. As the world becomes increasingly wireless (with cordless phones, cell phones, wireless internet, GPS devices, etc), allocation of the available spectrum to each technology becomes increasingly contentious. Each user community (usually manufacturers of the wireless equipment) wants more bandwidth in order to be able to sell and service more units.

So "Spectrum scarcity" is the apparent result. This requires that the means of allocation of radio communications resources to satisfy our future need for increasingly dense, fast, flexible mobile communications networks should be done judiciously. Because of this scarcity, the organizations allotting this spectrum to vendors charge them heavily.

2.1 Where Is the Problem?

Fig 1 shows exactly where the problem lies with respect to 3G network infrastructure and the possible solutions. There are two choke points, i.e. Radio and Backhaul (point 1 and 2 in Fig 1). These are the limiting factors for the amount of data traffic which can pass from Internet/Operator network to the end user. Increasing their capacity can alleviate the problem temporarily but certainly cannot provide a permanent solution.

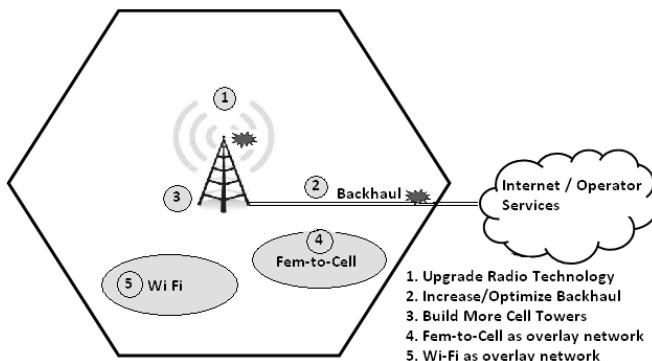


Fig. 1. Backhaul and Radio are two choke points

3 Possible Solutions

The possible techniques to solve the problem and the issues associated are as follows:

1. Scaling: It refers to building more cell towers and/or increasing the backhaul capacity. But this approach requires more infrastructures and thus more investment, which directly results in high cost/MB. Moreover, research reveals that top 3% of the smart phone users consume 40% of all smart phone data [2]; the main user gets a fractional benefit as the major consumers will continue to hog the increased bandwidth. So, scaling is not the solution, rather it is just masking the problem.

2. Optimization: It refers to optimizing the radio and backhaul usage. Starting from 1G to 3G many technological upgrades have happened for the optimized usage of radio [10], but there is certainly a limit on the number of bits which can be packed onto radio waves. Moreover, the rate of data consumption will continue to outpace technology upgrades as host of new services are introduced with any technological upgrade. Also, any technological upgrade results in new equipments to be installed, which again require more investment.

Backhaul usage can also be optimized and caching frequently used data items is one way to do it. It is a promising solution as it helps in flow control, but it poses two challenges:

1) Intensive packet inspection and correlation, which may slow down the network.

2) Privacy issues, as users do not like to be policed.

3. Mobile Data Offloading: It is the use of other (preferably “complementary”) network technologies for delivering data originally targeted for cellular users. It provides an alternative path of wireless delivery with a best performance capability. These networks can function with the macro-cellular network as an adjunct network either operating independently or as an overlay network. It is a new telecom industry buzzword, which is soon expected to become its separate revenue generating segment.

It has emerged as a promising solution. Though necessary, it is beneficial also. For the end users offloading contributes in higher bandwidth availability and reduced data services cost. For the operators its main contribution is in reducing the congestion of the cellular networks. Because a good amount of the telecom operators' revenue goes in paying for the exclusive usage of the spectrum, offloading also helps in reducing this cost without affecting the services.

4 Technologies for Mobile Data Offloading

The two candidate technologies for Mobile Data Offloading are:

1) **Fem-to-cell:** It is a small cellular base station, which connects to the service provider's network via broadband, and is typically designed for use in a small business or home. It allows service providers to extend their services where access would otherwise be limited or unavailable. It leverages on the licensed spectrum, offer better indoor coverage at low pressure and work with common single radio handsets.

Fem-to-cells are the natural extension of the main cellular network which allows them to support most of the services provided by mobile operator. Since wireless interface in Fem-to-cells is identical to cellular network, and control functions are also identical, this does not require handsets to have an additional Wi-Fi radio unit enabled. Moreover, Fem-to-cells allow for easier seamless roaming, and can provide managed quality of service, and provides improved battery life.

On the contrary, the main challenge with Fem-to-cells is that they have yet not reached widespread availability. They require expensive devices because the market is still small and since because of the issues of utilizing licensed spectrum, FCC rules etc, there deployment is more complicated. There may be spectrum conflicts between the macro network and fem-to-cell and between neighboring fem-to-cells.

2) **Wi-fi:** Based on IEEE 802.11, it is a name of a popular wireless networking technology that uses radio waves to provide wireless high-speed Internet and network connections. It leverages the unlicensed spectrum; and offer a much faster rate of service than the comparable 3G service.

4.1 Why Wi-Fi Is the Correct Technology

Though 3G-WLAN interworking should be built on top of harmonizing layer(s) (e.g. IP) and not limited to any specific WLAN technology, Wi-Fi has emerged as one of the primary candidate by the industry for offloading data. Following are the factors which contribute to the huge success of Wi-Fi.

1) **Vast unlicensed spectrum:** Wi-Fi operates in unlicensed ISM 2.4 GHz and 5 GHz frequency bands. The spectrum availability in the two respective bands is 83 MHz and 505 MHz. This means that regulator approval is not required for individual deployments, and Wi-Fi has a larger "free" spectrum available to cater to any size of network deployment.

2) High data rates and user experience: Though ITU has not provided a clear definition of the data rates which a user can expect from a 3G service provider, it is expected that IMT-2000 provides the minimum transmission data rates of 2 Mbit/s for stationary or walking users, and 384 Kbit/s in a moving vehicle [11]. Compared to this, IEEE 802.11n, which operates on both the 2.4 GHz and lesser used 5GHz band, can provide the data rates up to 600 Mbps [12]. So for consumers, all this it means is that in terms of downloading music or streaming video, or transferring a big file, Wi-Fi is a much better network solution.

3) Total ownership cost: Wi-Fi offers huge capital expenditure and operational expense benefits for operators. Over the last decade, since the launch of Wi-Fi technology, it has evolved and matured enough, thus bringing down the equipment cost significantly. In addition, with data rates of 600 Mbps and availability of more than 500MHz of unlicensed spectrum, Wi-Fi offers huge network capacity compared to 2G/3G, thus requiring less equipment to serve a given subscriber base. Also, without requiring large investment in channel planning and site surveys, the Wi-Fi networks can be easily and cost-effectively scaled.

4) Advanced Security and QoS: Since the introduction of the IEEE 802.11 WLAN standard in 1999, it has gone through a series of amendments to support quality of service (QoS) along with the standard-based business-grade security. With the most common wireless encryption standard, Wired Equivalent Privacy (WEP), been shown to be easily breakable, the advent of Wi-Fi protected access (WPA and WPA2) encryption aimed to solve the problem. WPA2 is based on IEEE 802.11i and it provides 128-bit AES-based encryption using Pre-Shared Key (PSK) or 802.1x RADIUS authentication, which is ideal for operators to provide Authentication, Authorization and Accounting (AAA) services. The Wi-Fi Multimedia enabled Wi-Fi networks offer a prioritized treatment to multimedia applications such as VoIp, interactive gaming, and video streaming, to support the jitter and latency requirements of these applications. As a result, QoS and security support in Wi-Fi is comparable to that of 2G/3G networks.

5) Increasing number of Wi-Fi hotspots: Wi-Fi operates in more than 220,000 public hotspots and in tens of millions of homes and corporate and university campuses worldwide [13]. There are about four million hotspots around the world and as more large venues and enterprises recognize the significant value of offering Wi-Fi for their customers, employees and operations, the number of hotspots is growing rapidly.

Also, AT&T's Q3 2009 hotspot connection numbers were 25.4million sessions, up from 15million the quarter before. Of these connections, 60% were from "integrated devices", meaning Smart phones [13]. It is estimated that in 2015 wireless hotspots will account for nearly 120 billion connect sessions [13].

6) Wi-Fi Complement to 3G: Mobile Data Offloading prefers complementary network technology to 3G for delivering data. Over the years Wi-Fi has proved to be complementary technology to 3G and there are several important ways in which Wi-Fi and 3G approach of offering wireless access services are substantially different.

First, the corresponding network deployment and business models are different. The basic business model of 3G is the telecommunication service model in which service providers own and manage the infrastructure and sell service on it. In contrast, Wi-Fi favors data communication industry (LANs). The basic business model is the equipment makers selling equipments to customers and services provided by the equipment are free to its customers. Second, 3G mobile technology use licensed spectrum, while Wi-Fi uses unlicensed shared spectrum. Thus there cost of service and quality of service are different. Third, the standardizing bodies of two are different. 3G is been standardized by 3GPP and is a relatively small family of internationally sanctioned standards. In contrast, Wi-Fi is one of the families of continuously evolving 802.11x wireless Ethernet standards. Finally, 3G offers communication in much broader geographical area with ubiquitous services, but at comparatively less speed. In contrast, Wi-Fi offers communication in smaller geographical area, but at very high speed.

5 Conclusion

With the expected increase in network traffic, increasing usage of smart phones, and considering the fact that spectrum is costly and scarce, Mobile Data Offloading is the new buzzword in the telecommunication industry. Today, it is not only the requirement and immediate need of the hour, but is also beneficial for operators as well as subscribers. In this paper, we have given several qualitative reasons and statistics in support of it. Also, we present the benefits and drawbacks of Fem-to-Cell and Wi-Fi, the two candidate technologies for Mobile Data Offloading. Based upon the comparative analysis we can conclude that a typical customer who don't want to pay for a dual mode smart phone or iphone, is a little technical savvy, want to get good coverage of network at home (which otherwise is bad), Fem-to-cell is a good solution. Rather, considering the continuous increase in usage of smart phones and iphones, projected increase in the data requirements of mobile subscribers in future, Wi-Fi is the clear winner.

References

1. The world in 2010 – The rise of 3G, technical report of ICT (2010),
<http://www.itu.int/ITU-D/ict/material/FactsFigures2010.pdf>
(accessed December 1, 2011)
2. Mobile data offload for 3G networks, Intellinet Technologies report (October 2009),
[http://www.intellinet-tech.com/Media/PagePDF/
Data%20Offload.pdf](http://www.intellinet-tech.com/Media/PagePDF/Data%20Offload.pdf) (accessed October 20, 2011)
3. Stemm, M., Katz, R.H.: Vertical handoffs in wireless overlay networks. *Mobile Networks and Application* 3(4), 335–350 (1998)
4. Juniper Research (August 2, 2011),
<http://juniperresearch.com/viewpressrelease.php?pr=254> (accessed November 15, 2011)

5. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010-2015, white paper,
http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf (accessed December 20, 2011)
6. Mobile Marketing: Insight & trends, Report, Quadlogix Technologies,
<http://www.slideshare.net/QuadLogix/quadlogix-technologies-mobile-marketing-an-overview> (accessed December 10, 2011)
7. Juniper Research (March 31, 2011),
<http://juniperresearch.com/viewpressrelease.php?pr=237> (accessed January 10, 2011)
8. Global Mobile Statistics 2012(January 2012),
<http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats> (accessed January 5, 2012)
9. Approaching shortages of mobile broadband spectrum threaten to limit broadband deployment and economic growth, Discussion Paper - ICC communication (October 2011), <http://www.iccwbo.org>
10. Lemstra, W., Hayes, V.: License-exempt: Wi-Fi complement to 3G. Journal of Telematics and Informatics 26(3), 227–239 (2009)
11. Cellular standards for the third generation, ITU (December 1, 2005)
12. IEEE standard 802.11n, Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications – amendment 5: Enhancements for higher throughput (2009)
13. Hotspot usage to reach 120 billion connects by 2015, Scottsdale (August 2011) (accessed November 13, 2011)

Performance Analysis of Gigabit Ethernet Standard for Various Physical Media Using Triple Speed Ethernet IP Core on FPGA

V.R. Gad, R.S. Gad, and G.M. Naik

Department of Electronics, Goa University,
SPO GU, Taleigao Plateau, Goa 403 206
vinaya_gad@rediffmail.com

Abstract. Gigabit Ethernet Standard provides 1 Gbps bandwidth and is backward compatible with 10 Mbps (Ethernet) and 100 Mbps (Fast Ethernet). It can also be installed with lower cost than other technologies having similar speed. The performance studies of Gigabit Ethernet is more complex than Ethernet or Fast Ethernet protocols. In this paper we have described the implementation of Gigabit Ethernet design on FPGA using Altera's Triple Speed Ethernet IP Core. The performance analysis of Gigabit Ethernet Standard has been studied using various physical media. This analysis includes performance measurements with different number of frames and frame lengths.

Keywords: MAC, Triple Speed Ethernet, Gigabit Ethernet, SFP.

1 Introduction

Software-based programmable network interfaces excel in their ability to implement various services.[1] These services can be added or removed in the network interface simply by upgrading the code in the system. However, programmable network interfaces suffer from instruction processing overhead. Programmable NICs must spend time executing instructions to run their software whereas ASIC (Application Specific Integrated Circuit) based network interfaces implement their functions directly in hardware. To address these issues, an intelligent, configurable network interface is an effective solution. A reconfigurable NIC (Network Interface Card) allows rapid prototyping of new system architectures for network interfaces. The architectures can be verified in real environment, and potential implementation bottlenecks can be identified. Architecturally, the platform must be processor-based and must be largely implemented using a configurable hardware. Thus, an FPGA (Field Programmable Gate Array) with an embedded processor can be the best platform to combine performance, efficiency and versatility. Dynamically reconfigurable platform will also reduce power consumption of the network device [2].Also, the reconfigurable NIC must have different memory interfaces including high capacity memory and high speed memory for adding new networking services.

2 Gigabit Ethernet Standard

Figure 1 presents a block diagram identifying the various components of IEEE Std 802.3z. The media access control (MAC) sublayer describes the algorithms used to control the transmission and reception of frames on an Ethernet network. IEEE Std 802.3z includes both the full duplex MAC and the carrier sense multiple access with collision detection (CSMA/CD) MAC [3]. The gigabit media-independent interface (GMII) allows any physical layer to be attached to the MAC and thus provide interoperability between different vendors. The GMII delivers 8-bit octets to the physical coding sublayer (PCS) on the transmit path, and accepts 8-bit octets from the PCS on the receive path at a rate of 125 million octets (1 billion bits) per second. The 10-bit symbols produced by the PCS are serialized by the physical medium attachment (PMA) sublayer. The PCS includes a function referred to as auto negotiation, which is a link startup and initialization procedure. Within IEEE Std 802.3z, auto negotiation is used to select between the CSMA/CD and full duplex operating modes, and to select whether the Pause flow control mechanism is enabled or disabled on a link-by-link basis.

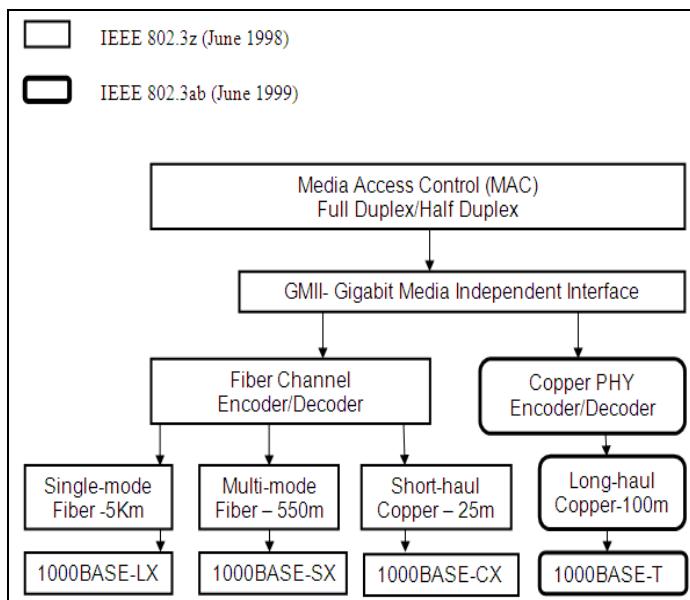


Fig. 1. Gigabit Ethernet Standard [3]

The transceiver specifications are given at the bottom of the diagram. The 1000BASE-SX specification for short-wavelength laser transceivers supports multi-mode fiber optic links at distances up to 275 m using 62.5 μ fiber, and 550 m using 50 μ fiber. 1000BASE-LX supports longer distances using higher-cost components, spanning 550 m on 62.5 μ or 50 μ fiber, and up to 5 km on single-mode fiber. The

1000BASE-LX laser transmitter is optimized for single-mode fiber, and requires a mode-conditioning patch cord to support multimode fiber optic cable. Both 1000BASE-SX and 1000BASE-LX specify the familiar duplex SC optical connector, eliminating the most common installation problem encountered in fiber optic networks, the misconnection of the transmitting and receiving fibers. IEEE Std 802.3z also includes a specification for a transceiver technology referred to as 1000BASE-CX, which supports shielded copper cables links spanning 25 m. The SerDes component which makes up the PMA sublayer is designed to drive this cable directly, which makes 1000BASE-CX an economically attractive choice for short-distance interconnections, for instance, between devices located within the same rack or within a computer room or telephone closet. 1000BASE-T supports 1000 Mb/s operation on four pairs of category 5 UTP cabling, at a maximum link distance of 100 m. A 1000BASE-T PHY transmits its signal on all four pairs of wire simultaneously, thus reducing the data rate on each pair to 250 Mbps. The use of a five-level pulse amplitude modulation scheme further reduces the signaling rate on each pair. Hybrids and digital echo cancellation are used to achieve full-duplex communication.

3 System Organization

Fig.2 shows a high-level block diagram of the Triple Speed Ethernet (TSE) design[4]. The design includes two Altera TSE MegaCore functions (MAC + PCS + PMA) and is downloaded on Altera's Stratix II GX PCI Express Development Kit. There are two SFP(Small Form-factor Pluggable) cages built onto the kit. This design interfaces the TSE MegaCore function[5] with a Copper or Optical Fibre SFP module via a 1.25 Gbps serial transceiver that enables all 10, 100, and 1000 Mbps Ethernet operations. The design sends stream of Ethernet packets to the TSE MegaCore function, which can be looped back using SFP modules with an Ethernet fibre optic cable, copper cable or a switch. The design can demonstrate the operation of the TSE MegaCore function in various modes with live traffic upto the maximum throughput rate and show the error rate in the receiver, if any.

The design is built using Altera's Quartus II software and SOPC (System On Programmable Chip) builder .The Nios II processor is used as a control plane component for setting up and configuring the system components. The Ethernet packets are generated and monitored by the processor. The on-chip memory is of block size 256 Kbytes, which is used for storage of software code. The parallel input/output (PIO) core provides easy I/O access to the 1000BASE-T Copper or Optical Fibre SFP module's PHY registers. The JTAG UART core transfers serial character streams between a Nios II processor and an SOPC Builder system. The phase-locked loop (PLL) core takes an input clock from a 100 MHz crystal on the development kit and generates an 83.33 MHz PLL output clock as a system-wide clock source for the SOPC Builder system. The TSE Megacore function transmits the Ethernet packets from the Avalon Streaming (Avalon-ST) interface to a 1.25 Gbps serial transceiver interface that is built in the Stratix II GX device and receives packets from the opposite direction. The Ethernet Packet Generator is an SOPC custom component, used to generate a stream of Ethernet packets. It drives the Transmit FIFO interface of TSE MegaCore

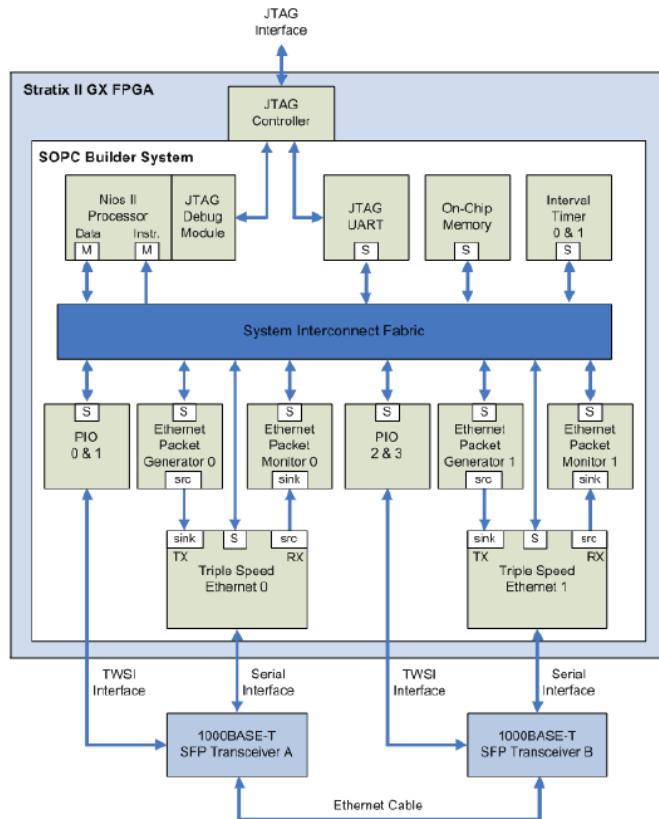


Fig. 2. Block Diagram of Triple Speed Ethernet Reference Design [4]

function. The Ethernet Packet Monitor block is an SOPC custom component created using the component editor. It has an Avalon-MM slave interface on one side for control purposes and an Avalon-ST sink interface on the other side for the data path. This block is fed a stream of Ethernet packets by the TSE MegaCore function Receive FIFO interface. The Ethernet Packet Monitor also verifies the accuracy of the received payload. The interval timer core is a 32-bit timer used by the Nios II processor system to calculate the performance and throughput rate of various Ethernet operations.

4 Implementation

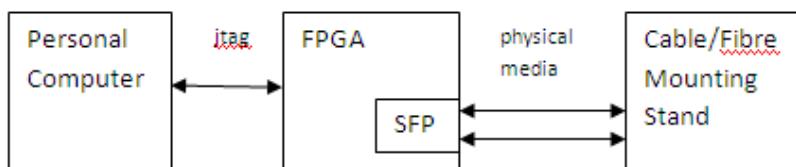
Altera's Triple Speed Ethernet Design has been used as a platform for studying the performance of Gigabit Ethernet Standards 1000Base-LX, 1000Base-SX and 1000Base-T. The design has been implemented on Altera's Stratix II GX device EP2SGX90FF1508C3. Table 1 summarizes the utilization result of this design.

Table 1. Resource Utilisation of Triple Speed Ethernet Design

Parameter	Value	Utilization(%)
Logic	15281	21
Combinational ALUTs	11209	15
Dedicated logic registers	10419	14
Total pins	32	4
Total block memory bits	390494	9
DSP block 9-bit elements	8	2
Total PLLs	1	13
Total GXB Receiver Channels	2	13
Total GXB Transmitter Channels	2	13

5 Performance Evaluation and Results

The network performance instrument with measuring ability of full line rate is an important component of the system[6].The Test System of the Ethernet design is given below in Fig.3. The Triple Speed Ethernet design is dumped onto the FPGA using Quartus II software and JTAG interface.

**Fig. 3.** Test System for Performance Evaluation

The performance of above design was studied for Gigabit Ethernet standards 1000Base-LX ,1000Base-SX and 1000Base-T using various physical media and corresponding SFP Transceivers as mentioned in Table 2.

Table 2. List of SFP Transceivers used for the different Gigabit Ethernet standards

Sr. No.	SFP Transceiver	Physical media	Gigabit Ethernet Standard
1	1000Base-LX 1310nm	Singlemode Fibre (SMF)	1000Base-LX
2	1000Base-LX Bi-Di 1310nm, 1550nm	Singlemode Fibre (SMF)	1000Base-LX
3	1000Base-SX 850nm	Multimode Fibre (MMF)	1000Base-SX
4	1000BaseT	Cat-5e	1000Base-T

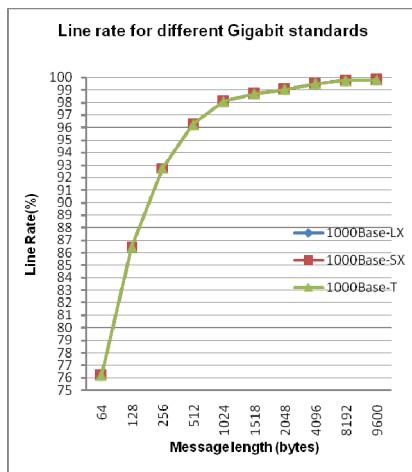
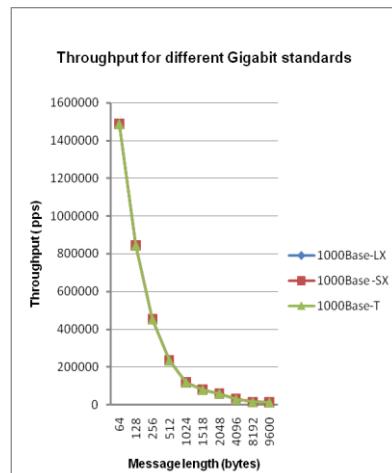
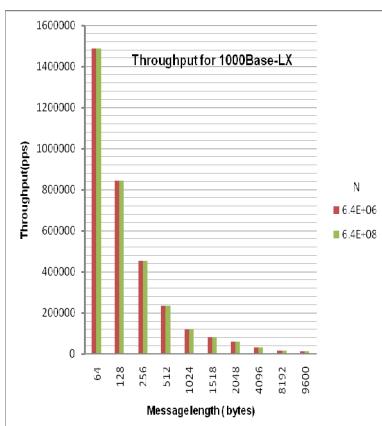
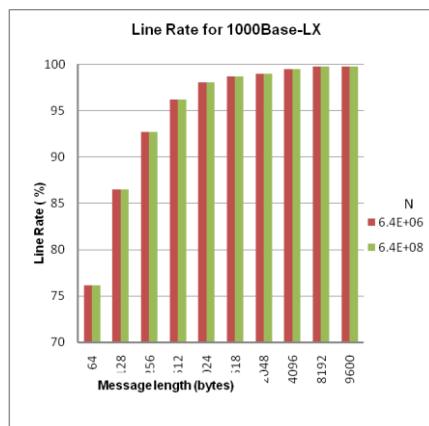
The system was tested by varying parameters such as message length and number of frames .The first test was performed by increasing the message length and keeping the number of frames fixed . The test was repeated for two different values of number of frames i.e. 10^5 and 10^7 . The results of the tests performed for all the Gigabit Ethernet Standards mentioned in Table 2 is given in Table 3. It is found that as the message length is increased from 64 bytes to 9600 bytes, the line rate increases and achieves 99.79% for 9600 bytes. The throughput is 1488115 packets per second(pps) for 64 bytes and 12993 pps for 9600 bytes.The results are found to be nearly the same for all the different Gigabit standards.

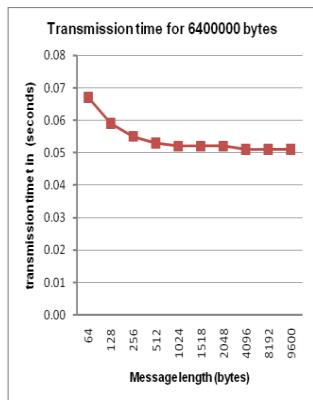
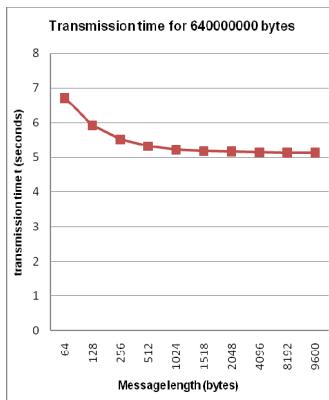
Table 3. Throughput of the Network

Length (bytes)	Line rate (%)	Throughput (pps)
64	76.19	1488115
128	86.48	844607
256	92.75	452905
512	96.24	234966
1024	98.08	119733
1518	98.70	81275
2048	99.03	60445
4096	99.51	30369
8192	99.75	15221
9600	99.79	12993

The second test was performed for the different Gigabit Ethernet standards by varying message length from 64 bytes to 9600 bytes and measuring Line rate and Throughput. From the Figure 4 and Figure 5, it can be seen that the curves for the 3 standards 1000Base-LX ,1000Base-SX and 1000Base-T almost overlap. Hence the performance of the 3 standards are found to be almost similar. This measurement was taken for fixed number of packets 10^5 and 10^7 .

The third test was performed by keeping the total number of bytes (N) sent constant.N =message length×number of packets .The message length was varied from 64 bytes to 9600 bytes and correspondingly number of packets was changed.The experiment was repeated for 2 different values of N i.e. 64×10^5 and 64×10^7 .Fig.6 shows that the throughput is the same for both values of N for a particular message length. Fig.7 shows that the line rate obtained is same for both values of N for a particular message length.Also the values of Line rate and Throughput are the same as in Table 3 (where number of frames is kept fixed). The Total transmission time(t) has reduced from .067s for 64 bytes message length to .051s for 9600 bytes when N= 64×10^5 as shown in Fig.8. Similarly, as the message length is increased , t is reduced from 6.72s to 5.13s for N= 64×10^7 as illustrated in Fig.9.

**Fig. 4.** Line rate vs. message length**Fig. 5.** Throughput vs. message length**Fig. 6.** Throughput vs. message length**Fig. 7.** Line rate vs. message length

**Fig. 8.** Transmission time vs. message length**Fig. 9.** Transmission time vs. message length

This test was performed only for 1000Base-LX standard and similar performance is expected of the other standards. The use of SOPC has given us the flexibility to choose from different software and hardware components and greatly reduce the system development cycle[7].

6 Conclusion and Future work

This paper has studied the performance analysis of Gigabit Ethernet Standards, implemented on Altera's FPGA. Quartus II software is used to synthesize and create .sof file .The design is downloaded to the FPGA chip using JTAG interface.Experimental results reveal that the line rate is 76.19% for minimum 64 bytes packet size and approaches 100% for 9600 bytes packet size.The throughput is lowest for 64 bytes packet size and increases with increase in packet size.for various packet lengths .For a particular packet size, the throughput and line rate remain almost the same for all the different Gigabit Ethernet standards. Also, as the packet length increases , the total transmission time is found to be decreasing.We have also developed an experimental platform to introduce errors into the network and future work includes Error Detection and Correction Analysis using the same platform.

Acknowledgement. The authors would like to acknowledge Altera Inc. USA for the MOU with Goa University and one of the authors V. R. Gad would like to thank University Grants Commission(UGC), New Delhi,India for providing FIP study leave under which this work is being carried out.

References

- [1] Mohsenin, T.: Design and Evaluation of FPGA-Based Gigabit-Ethernet/PCI Network Interface Card Thesis, Rice University (2004)
- [2] Kachris, C., et al.: Design and performance evaluation of an adaptive FPGA for network applications. Microelectron. J. (2008), doi:10.1016/j.mejo. 2008.05.011

- [3] Frazier, H.: The 802.3z Gigabit Ethernet Standard. *IEEE Network*, 6–7 (May/June 1998)
- [4] Triple Speed Ethernet Data Path Reference Design AN-483-June 2009 ver. 1.1 Altera Corporation (2009)
- [5] Triple-Speed Ethernet MegaCore Function User Guide © December 2010 Altera Corporation Application Note 483 (2010)
- [6] Duan, M., Han, H.: Research and Implementation of Gigabit Ethernet Full Line Rate. In: Cross Strait Quad-Regional Radio Science and Wireless Technology Conference, pp. 736–738 (2011)
- [7] Wang, Y., Zhang, C., et al.: Implementation of Gigabit Ethernet Network based on SOPC. In: Asia Pacific Conference on Wearable Computing Systems, pp. 341–343 (2010)

Assortment of Information from Mobile Phone Subscribers Using Chronological Model [IGCM]: Application and Management Perspective

Neeraj Kumar* and Raees A. Khan

Department of Information Technology, School for Information Science and Technology,
Babasaheb Bhimrao Ambedkar University (A Central University),
Lucknow 226 025, UP, India
neerajmtech@gmail.com, khanraees@yahoo.com

Abstract. In this study, proposed Information Gathering Chronological Model (IGCM) design offers a large-scale questionnaire, i.e. a good reference for evaluating the prevalence of symptoms and sensations from the usage of mobile phones or wireless devices. The IGCM was applied to assess the possible health effects from mobile phones on population. The study was carried out as a survey by posting the questionnaire both the online and manual among the randomly selected 307 Indians. Mobile subscribers were assessed for self reported symptoms and sensations and safety management adopted during calling. A good number of mobile phone users were found to be associated with symptoms i.e., headache, ringing delusion, forgetfulness, increase in the carelessness, dizziness, extreme irritation, speaking falteringly, neurophysiologic discomfort, cell phone side ear temperature increase and speaking falteringly when analyzed the output of IGCM. Study concluded that proposed model IGCM has significance in order to monitoring and evaluating frequently available holistic information especially health care management by mobile phone subscribers.

Keywords: Information Gathering Chronological Model (IGCM), Mobile Phone, Wireless Communication, Symptoms.

1 Introduction

Mobile communication technology has been emerged surprisingly over the last decade. Global System of Mobile Communication (GSM), Code Division Multiple Access (CDMA), Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA) are foremost among various cellular standards but GSM and CDMA cellular standards are most common of them and deployed in many parts of world. In present, mobile communication carries critical business, social and health information of many subscribers. With the introduction of mobile communication technologies

* Corresponding Author: Neeraj Kumar, RA-UPCST (Young Scientist Scheme), Department of Information Technology, School for Information Science and Technology, Babasaheb Bhimrao Ambedkar University (A Centre University), Lucknow – 226 025, UP, INDIA. Email: neerajmtech@gmail.com, Contact No. +91-522-2964765 (O); +91-9473594960 (M).

such complains related to cell phone, base stations, radar communicating devices, electronic gadgets etc. became more prominent. Since wireless technology has come into existence in the last decades but now the RFR exposure levels has amplified many folds because of the extensive use of RF devices and cell phone. Some individuals who experience the electromagnetic hypersensitivity (EHS) believe that their associations to EHS are caused by an increased exposure of EMRs[1]. Situation associated with these complaints are typically characterized by EMF exposure well below current reference values [2], but studies have not been able to show a reliable connection between EMF exposure symptoms [3]. Electromagnetic hypersensitivity (EHS) has come in common usage in recent years and considers particular in association with symptoms and sensations to subjects on exposure to EMR. A number of studies have investigated EHS symptoms and EMR exposure by wireless communication devices or cell phones or base stations.

In a prevocation study [4] of cell phone, twenty subjects reporting EHS symptoms. They reported on increase in symptoms during 30min exposure during RF exposure compared to sham exposure. In a Meta analysis[5], prevalence of EHS was reported to be 1.5% in Sweden [6], 3.2% in California [7], 5% in Switzerland [8]. But the study could not establish a causal link between exposures actually exists. Epidemiological studies conducted so far are very controversial and failed to point out a clear relationship between the use of mobile phones and the incident of diseases [9]. Contradictory results are parallel on previous findings [10-12].

The holistic available findings illustrated in literature tend us for better management of information. In present, mobile phones are playing a vital role in health care, industry, military and general communication services. As mobile phone utilization will increase the associated possibility of threats or symptoms to users will also increase. The requirement of wireless communicating services and devices are increasing globally and this concern tends us to establish a well defined security model to gather the information by the mobile phone subscribers to assess the risks or threats as subscribers are realized during the *calling* on Mobile Phone (MP) or just after *calling*. This study has four major phases of consideration and analysis; first phase is to gather information through online system from MP subscribers. A new "*Information Gathering Chronological Model*" (IGCM) was proposed and implemented in particular relevance to self reported threats or symptoms among urban and rural MP subscribers including male, female and children.

1.1 Problem Formulation

In the race to adopt new technologies, we often ignore the ill effects of the technology and do not realize about it. Presently people are adopting the wireless communication devices for varied purposes. They feel comfortable when they communicate to others through wireless communication devices in the air instead of wire. The common communicating device operates in the radiofrequency wave or microwave range. Microwave may cause thermal effects at high exposure levels. The radiofrequency (RF) fields from 10 MHz to 10 GHz penetrate exposed tissue and produce heating due to energy absorption in these tissues. The depth of penetration of RF field in to tissue depends on the frequency of the field and is

greater for lower frequencies. The Specific Absorption Rate (SAR) is the basic dosimeter quantity for the RF fields between about 1 MHz and 10 GHz.

Above 4 W/Kg SAR may generate an adverse health effects to people exposed by RF fields and such kind of energy are found around the tens of meter away from the powerful antennas. In India, two common communication services i.e. GSM (Global System of Mobile Communications) and CDMA (Code Division of Multiple Access) are using frequently by subscribers. The frequency bands for GSM ranged from 900 MHz - 1800 MHz and CDMA associated to 800 MHz band [13]. Now billions of mobile phone subscribers are using such devices in the world. Only in India, the numbers of mobile phone subscribers have reached over 82 million [14]. The close proximity to mobile phone, subscribers are more prone to electromagnetic hypersensitivity to cellular phone EMR. Some of the illustrated biomechanical concepts may describe well problem formulation of our concern regarding cellular communication. The present safety standards for radiofrequency radiation are largely based on preventing these effects from heating, the lower frequency (below to 10 MHz) may cause currents of biological significance. These are the basic reasons to raise the questions on mobile phone safety in relevance to exposure of radiofrequency radiation.

2 Objectives

As discussed in problem formulation section that in present scenario a number of threats are associated individually or within community. The complete information of threat associated to individuals and their relevant action, are an issue of deal in present scenario. In this paper we are being approached for a cost-effective and highly efficient independent assessment model to get basic reliable information for Health and Safety management. To validate the proposed model in respect to authenticity, risk analysis and security of the information gathering system is primary object of this study. Further, to analyze the composite threats associated to individuals in particular relevance to self reported symptoms and sensation by extensive usage of mobile phone among urban and rural population of India, is also application of the proposed model.

3 Advantages of Model - IGCM

This model has classical benefits in order to compile research findings in varied areas including information science, engineering, occupational health, biomedical, epidemiology and clinical research. The IGCM may-

- i. persuade focused, high quality research
- ii. able to incorporate research results frequently
- iii. facilitate the development of Internationally acceptable standards for mobile communication EMR
- iv. provide information on the management of mobile phone EMR risk perception, communication and management
- v. circulation on EMR health and environmental effects and actions needed
- vi. explore the safety measures and advices at regional and global level
- vii. develop a research and development networks in allied area.

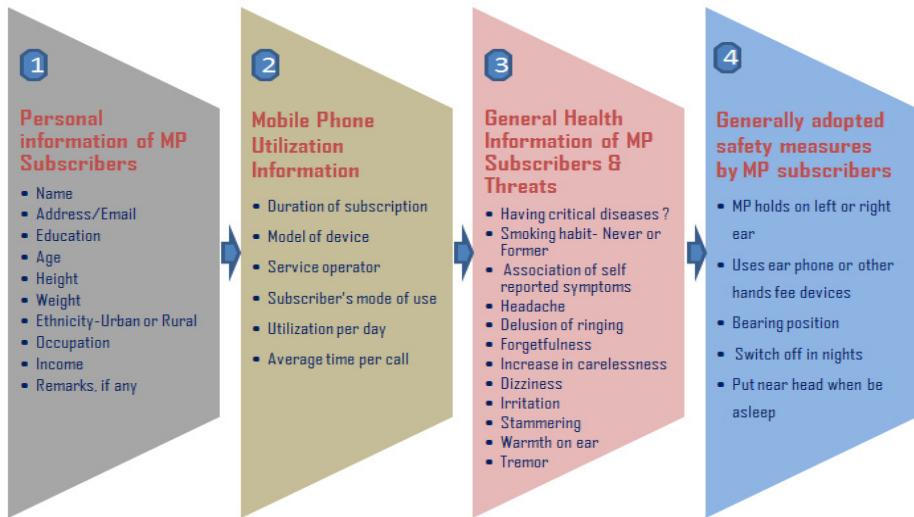


Fig. 1. Information Gathering Chronological Model [IGCM]

4 Methodology

Before being concerned on discussion on validation of the proposed model, primarily we have to discuss on input sources. To analyze the threats associated to individuals in particular relevance to self reported symptoms and sensations by extensive usage of mobile phone among urban and rural population of India, the modified methodology of Interphone study [9] was followed to design the basic format of the study. The information input key words regarding to symptoms and sensations with particular relevance to cellular phone usage to the subjects, was collected by the methodology of [15] with minor modifications.

4.1 Information Input Sources (IIS)

In this study three major input sources are proposed to gather online information, these input sources may be private networks (personal or organizational mails), public networks (social networks, open access links etc.) and manual questionnaire which may be used to send directly or by survey to concern authority or clinician.

4.2 Validation of IGCM

To validate the proposed model authenticity and risk management is most essential tools to get the well defined security frame work. Some of basic concepts are intended to apply in our proposed model to get the information secure and reliable from the Information Input Sources.

4.3 Information Security System of IGCM

In the Information Security System (ISS) of any proposed model, we have to discuss on four major issues i.e., (1) Information Security (2) Significance of Information (3) Information Security Management and (4) Risk Management. Before discussion on these major issues validation of information should be confirmed. If primarily output of IGCM is invalid, then individual may call further to gather the correct information as requirement. The looping of recalling should be for a second time whenever the information found invalid. To understand well the facts on information security of proposed IGCM, we followed three major components of security i.e., confidentiality, integrity, and availability.

Information Security and Management

Information security can understand well the activities which are related to the safety of information against the risks of loss, misuse, disclosure or damage. The Information security management describes controls that proposed model needs to implement to ensure that it is sensibly managing these risks.

Significance of Information

In order to be assuring the significance of Information of the IGCM, we have taken five major criteria including desirable, fast retrieval, integrity, confidentiality and availability. In this model we enquired the general profile of cell phone subscribers, health details, safety measures adopted during call and possibility to be associated with prescribed symptoms and sensations. First we checked the desirability on the basis of above variables, if individuals satisfied properly then gathered information is significant.

Risk Management

For the smooth assessment, the risks were controlled in the IGCM. The possible risk was managed in point of view on the controls, defense in depth, security classification, access controls for collection of information

5 Application of IGCM

We think, in Indian scenario this was initial stage effort which was taken to establish a link between the CP users to EHS as self reported symptoms and sensations. On the basis of demographic characteristics of CP users in India, this was first study which had been tried to generate baseline data of CP users and their fundamental link to self reported symptoms and sensations. Cell phones are used by majority of population throughout the world as one of the best and cost effective communication system. In India they are being used by the common man ranging from rag pickers, rickshaw pullers, farmers, academicians, industrialists and corporate personnel. The no of cellular phone subscribers in India have reached over 82 million till February, 2011

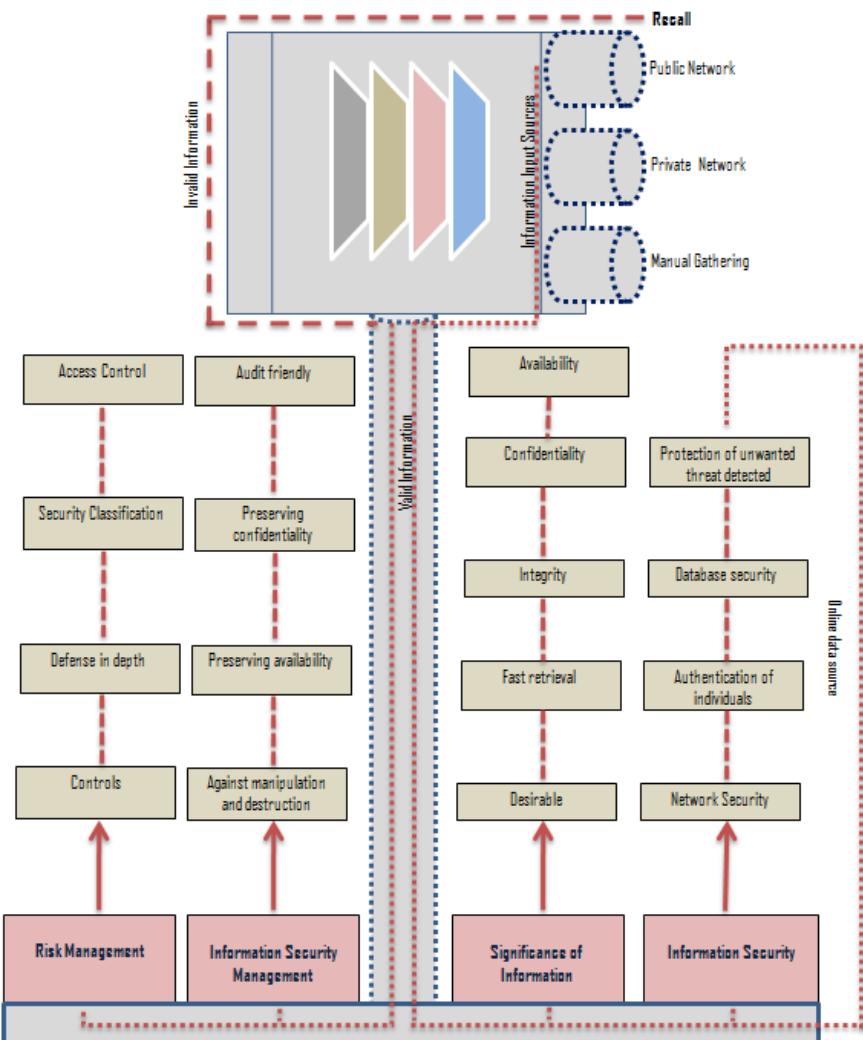


Fig. 2. Security and Management of IGCM

[14]. As much as number of cell phone users enlarges in a domain, the possibility of association with EMR, increase. This study is an attempt to develop a cofactor of possibilities of EHS sensitivity through the self reported symptoms and sensations by individuals in Indian scenario.

To evaluate the possibility of symptoms and sensation with extensive usage of mobile phone among the urban and rural population of India in particular relevance to self reported symptoms. Several symptoms were illustrated in well designed IGCM. The list of symptoms and sensations were collected with discussions of psychologists, clinicians and other biomedical experts and some of the published references [9, 15].

5.1 Designs for Requirement

Four segment of the IGCM was designed to gather the information among the mobile phone users. These segments of the IGCM were personnel details of mobile subscribers, mobile phone utilisation details, and general health information of mobile phone subscribers and associated threats. This segment of the IGCM model is very important in order to full fill our requirement. It can be say as optional segment and optional segment can be change as our desire or requirement of our study. In the segment of general health, self reported symptoms and sensations were mentioned as headache, ringing delusion, forgetfulness, increase in carelessness, dizziness, irritation, stammering, neurophysiologic discomfort, warmth on ear and tremor. Last segment of questionnaire was ‘personnel safety measures’, and before finalizing the assessment of EHS to individuals, a written consent was taken. The questionnaire was pre tested on some individuals and modifications done before final administration of the questionnaire to the subjects. The questionnaire was filled by a single experienced investigator so that there was no chance of intra observer error.

Subjects and Assessment of Variables

The Information of the individuals was collected by the variables as user’s name, address, occupation monthly income, education, age, height and weight. Second segment included variables as duration of cell phone use, mode of cell phone (ringing mode or vibration mode), received and dialed calls per day and number of adult and non adult cell phone users in individual’s family. Then investigator usually expressed the variables of EHS (symptoms and sensations) one by one to know symptom’s association with individuals. Individuals were informed that it is not compulsory to be associated with any symptoms and they were liberated to report “having no symptoms” prior express the variable of EHS. The variable used for the symptoms in this study were as *headache, ringing delusion, forgetfulness, increase in the carelessness, dizziness, extreme irritation, speaking falteringly, neurophysiologic discomfort, cell phone side ear temperature increase and speaking falteringly*. In the segment of safety measures individuals were asked three important questions as a variables. The safety variables were (i) generally you hold your cell phone in right or left ear during call? (ii) do you use your cell phone with headphone, speaker mode, other devices or not any one? (iii) do you keep your cell phone in bag, shirt, pant or hanging in neck? (iv) do you switched off your cell phone in night?, and (v) do you put cell phone near your head?.

Survey Strategy and Strength

At random 31 districts of India were selected to collect the information about the self reported symptoms in association with cell phone use. 307 CP users were participated from the both the urban and rural areas of the county. The male, female and children were included but the individuals they have no cell phone, were not included in this study. This survey was conducted during the year 2009-10.

Data Analysis

The details from individuals by questionnaire were transformed into Micro Soft Excel sheet and cross tabulated using EPI INFO soft ware. The significance prevalence of

signs and symptoms in relation to age, sex and duration of use of cell phones was tested using Chi Square Test. Fisher's exact test was used where expected cell frequencies were less than five. The P<0.05 was considered for significance in this study.

Results

Incomplete survey questionnaires were found 32 numbers. These numbers of individuals were not included in this study.

Demographic and Social Characteristics

The total, 307 CP users (age range 14 years to 62 years) were participated including 236(76.87%) males (mean age \pm SD: 28.95 ± 9.5) and 71(23.13%) females (mean age \pm SD: 25.34 ± 6.0). Study participants (mean age \pm SD: 28.95 ± 9.5) between age 14 and 62 years old were enquired for association to EHS. More than half (near 60%) individuals were highly educated up to Post graduate level and approx 75% participants were having formerly smoking or drinking habits. We get an immense contribution from the unemployed student community (near 50%) in this study and mostly they were. Only 24(7.82%) individuals of below 20 years (children) were participated but more than 36% CP user group were associated to (24-27) year's age group. The details of participation and distribution of individuals are summarized in Table-1.

Table 1. Participation and distribution of subjects (N=307)

Individual's	Group	No of Subject (%)
Age	≤ 20	24 (7.82)
	21-23	68 (22.15)
	24-27	111(36.16)
	28-30	51(16.61)
	≥ 31	53(17.26)
Sex	Female	71(23.13)
	Male	236 (76.87)
Education	Below Graduates	34(11.07)
	Graduates	86(28.01)
	Post Graduates	118(38.44)
	Doctorates	24(7.82)
	Professionals	45(14.66)
Smoking	Never	78(25.41)
	Former	229(74.59)
Safety	Head phone user	68(22.15)
	Switched off in night	41(13.36)
	Puts cell phone near the head during sleeping	118(38.44)
	Left ear side user	98(31.92)
	Right ear side user	209(68.08)
Cell phone Utilization	Exposure - Whole (hrs)	50(16.29)
	Group 1; HU (>5000)	
	Group 2; MU ($<1000 - 5000 \leq$)	95(30.94)
	Group 3; NU ($>500 - 1000 \leq$)	65(21.17)
	Group 4; LU (≤ 500)	97(31.60)

Adopted Safety Measures

Most of CP users (above 86%) were reported that generally they did not ‘switch off’ their cell phone in nights. They were keeping CP approx 3-4 fit distance from their bed during sleeping but near 41% individuals were continue exposed by cellular phone EMR at night because of having CP near the head at ‘switch on’ mode. Hence they were more prone to EMR. Only 68 individuals were use headphone. Right ear side CP users (68%) were larger than the left side (32%).

Cell Phone Exposure

We found four major groups (LU, NU, MU, HU) on the basis of CP usage in individuals whole life. Individuals were separated for the utilization of CP as (1) LU (low user range \leq 500 hours CP use in individual’s life); (2) NU (normal user range $>$ 500 – 1000 \leq hours CP use in individual’s life) (3) MU (moderate user range $>$ 1000-5000 \leq hours CP use in individual’s life) and (4) HU (heavy user range $>$ 5000 hours use in individual’s life). Moderate (MU-31.6%) and lower (LU-30.9%) group users were equally participated and participation of other groups were near 21% (NU-21.17%) and below (HU-16.29%) (Table-1)

Individual’s Association of Symptoms and Sensations

Approximately 12% individuals were not associated to EHS, they did not report any complain regarding the possibility to be associated with symptoms and sensation during *calling* or just after *calling* on CP but Incredible about 88% individuals were associated to symptoms or sensations (Fig.3).

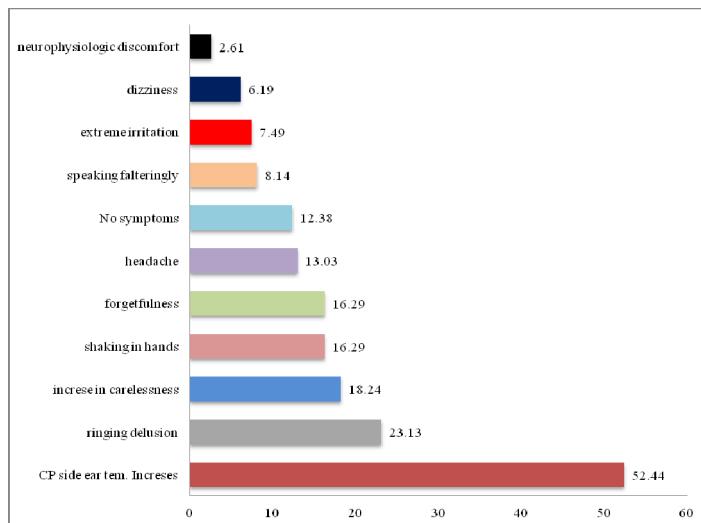


Fig. 3. Associations of self reported symptoms to CP users (in %)

The prevalence of overall link of EHS to individuals was found above 87%. They were associated to minimum one or more than one symptoms and sensations but above the half CP users were reported ‘ear temperature increase or warmth on ear. Second most frequent symptom was observed ‘ringing delusion’ (52.4%). The symptom ‘Ringing delusion’ is basically a confusion of ringing voice of cell phone realized by the CP user, though in reality there is no ringing voice at that moment. ‘Ringing delusion’ may consider as psychological sensitivity in which suddenly, a cellular phone user feels ringing voice of the device but in actually there is no ringing of the device. The associations of ‘increase in carelessness’ ‘forgetfulness’, and ‘headache’, symptoms was observed 18.2% 16.3%, and 13.0% respectively and rest of the symptoms like dizziness (6.19%), extreme irritation (7.49%) and neurophysiologic discomfort (2.61%) were minutely linked to individuals.

The possibilities of prevalence of EHS among the CP users were usually analogous to all age groups except the users of 31 years and above age. They were found in the reduced rate of possibility. Almost CP user’s link of EHS to CP usage did not support on the basis of participant’s gender but it was significant higher ($p<0.05$) in females (63.4%) than the male (49.2%) users. No significance link of EHS to CP users could be detected on the basis of literacy and smoking habits. EHS were found to be significantly higher ($p<0.001$) in individuals, they were keeping their device ‘switched on’ mode in the night (88%) than the ‘switched off’ mode. Particularly they were found significantly associated ($p<0.001$) to symptoms ‘ear temperature increase’.

6 Discussion

In the world wide a no of studies have been conducted over the possibilities of the symptoms and their association with usage of cell phone. In Indian scenario this is an initial level effort is being taken in particular relevance to subjective symptoms and sensation through usage of cell phone. Although, the prevalence of headache among handheld cellular phone users in Singapore was assessed in a community through a questionnaire [16], but this study could not conclude for association of significant increase of CNS related symptoms other than Headache. Another study was also conducted in Elazig, Turkey [15] among the cell phone users and cell phone non users. Study resulted as Headache-313 (72.1%), Dizziness-29 (55.8%), Extreme irritation-140 (71.8%), Shaking in hands-25 (61.0%), Speaking falteringly-11 (73.3%), Forgetfulness-75 (67.6%), Neuro-psychological discomfort-53 (81.5%), Increase in the carelessness-186 (86.9%), Decrease of the reflex -71 (97.3%), Clicking sound in the ears-80 (72.1%). Both the studies described above have shown an association of headache 44% and 72% likely but our study revealed an association of headache to CP users limited for approximately 13% only (Table-1).

7 Conclusion

The proposed model has played a significance role to gather the information in different ethnicities/regions cost effectively. In this study, we found a small number

of individuals were associated to EHS but significantly it could not be establish a link between the EHS to CP users. The vast statistics of participants may conclude better such a concern. The need-based cautious use of the scientific technologies must be accepted but the same technology may turn to be negative due to overuse. The Cell phones are required to be fabricated and to be used as per the guidelines of the regulatory organizations.

Acknowledgment. We acknowledge to Uttar Pradesh Council of Science and Technology for financial support under the Young Scientist Scheme - 2008.

References

1. Seitz, H., Stinner, D., Eikmann, T., Herr, C., Röösli, M.: Electromagnetic hypersensitivity (EHS) and subjective health complaints associated with electromagnetic fields of mobile phone communication-a literature review published between 2000 and 2004. *Science of The Total Environment* 349, 45–55 (2005)
2. Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz). International Commission on Non-Ionizing Radiation Protection. *Health Phys.* 74, 494–522 (1998)
3. Rubin, G.J., Das Munshi, J., Wessely, S.: Electromagnetic hypersensitivity: a systematic review of provocation studies. *Psychosom Med.* 67, 224–232 (2005)
4. Hietanen, M., Hamalainen, A.M., Husman, T.: Hypersensitivity symptoms associated with exposure to cellular telephones: no causal link. *Bioelectromagnetics* 23, 264–270 (2002)
5. Roosli, M.: Radiofrequency electromagnetic field exposure and non-specific symptoms of ill health: a systematic review. *Environ. Res.* 107, 277–287 (2008)
6. Hillert, L., Berglind, N., Arnetz, B.B., Bellander, T.: Prevalence of self-reported hypersensitivity to electric or magnetic fields in a population-based questionnaire survey. *Scand J. Work Environ. Health* 28, 33–41 (2002)
7. Levallois, P., Neutra, R., Lee, G., Hristova, L.: Study of self-reported hypersensitivity to electromagnetic fields in California. *Environ. Health Perspect* 110 (suppl. 4), 619–623 (2002)
8. Schreier, N., Huss, A., Roosli, M.: The prevalence of symptoms attributed to electromagnetic field exposure: a cross-sectional representative survey in Switzerland. *Soz. Präventivmed.* 51, 202–209 (2006)
9. Lahkola, A., Salminen, T., Raitanen, J., Heinavaara, S., Schoemaker, M.J., Christensen, H.C., Feychtig, M., Johansen, C., Klaeboe, L., Lonn, S., Swerdlow, A.J., Tynes, T., Auvinen, A.: Meningioma and mobile phone use—a collaborative case-control study in five North European countries. *Int. J. Epidemiol.* 37, 1304–1313 (2008)
10. Hardell, L., Carlberg, M., Mild, K.H.: Case-control study of the association between the use of cellular and cordless telephones and malignant brain tumors diagnosed during 2000–2003. *Environ. Res.* 100, 232–241 (2006)
11. Meo, S.A., Al-Drees, A.M.: Mobile phone related-hazards and subjective hearing and vision symptoms in the Saudi population. *Int. J. Occup. Med. Environ. Health* 18, 53–57 (2005)
12. Santini, R., Santini, P., Danze, J.M., Le Ruz, P., Seigne, M.: Symptoms experienced by people in vicinity of base stations: II/ Incidences of age, duration of exposure, location of subjects in relation to the antennas and other electromagnetic factors. *Pathol. Biol. (Paris)* 51, 412–415 (2003)

13. Issues in spectrum allocation and pricing in India (2002)
14. February report of Telecommunication Regulatory Authority of India Press Release No.29 /2011 (2011)
15. Balikci, K., Cem Ozcan, I., Turgut-Balik, D., Balik, H.H.: A survey study on some neurological symptoms and sensations experienced by long term users of mobile phones. *Pathol. Biol.(Paris)* 53, 30–34 (2005)
16. Chia, S.E., Chia, H.P., Tan, J.S.: Prevalence of headache among handheld cellular telephone users in Singapore: a community study. *Environ. Health Perspect* 108, 1059–1062 (2000)

Modeling Soft Handoffs' Performance in a Realistic CDMA Network

Moses E. Ekpenyong and Enobong Umana

Department of Computer Science

University of Uyo

PMB. 1017 520003 Uyo

Akwa Ibom State NIGERIA

mosesekpenyong@gmail.com, ekpenyong_moses@yahoo.com,
ennyin4u@yahoo.com

Abstract. The advent of code division multiple access (CDMA) technology has offered solution to the incessant termination of ongoing calls experienced in second generation networks. Seamless connectivity is now made possible as these networks provide an uninterrupted-transfer loop known as the overlap region, which must be effectively managed to ensure efficient handoffs. This contribution adopts a practical approach to the soft handoff problem, with concrete experimental solution that will benefit network operators and the wireless research community. We achieve this by studying an existing CDMA cellular network over a period of two months. Under ideal conditions, we adapt the existing data to the COST-231 Hata pathloss model and derive a soft handoff (SHO) probability model, peculiar to the study environment and generic to similar environments. In order to draw effective conclusions and advise on best practices, the model was simulated for various coverage areas and propagation exponents. Simulation results confirm that SHO thresholds should be carefully chosen in order to minimize network defects.

Keywords: Propagation pathloss, propagation exponent, soft handoff probability, soft handoff threshold.

1 Introduction

Cellular communication systems typically consist of fixed base stations (cells) that offer the transmission and reception of signals, to and from mobile units (users) within its communication area. Each base station is assigned a plurality of channels over which it allocates to the mobile units. A mobile unit within the range of the base station communicates with other mobile units through the base station (using these channels). Typically, the channels used by a base station are selected such that signals on the channel do not interfere with signals on other channels used by that base station. In order to allow mobile units to transmit and receive calls as the units travel over a wide geographical area, each cell or base station is usually positioned such that its coverage area is adjacent to and overlaps the areas of coverage of a number of other cells. When a mobile unit moves from its area of coverage to another coverage

area, communication is transferred (or handed off) to the nearest base station in an area where the coverage from different cells overlaps.

Handoff is therefore a process of transferring the support of a mobile unit from one base station to another. It happens when a cell (base station) establishes a ‘communication handshake’ with its neighboring cell as the mobile user approaches a new cell. This scenario is implemented in two ways: hard handoff (handover) and soft handoff. Hard handoff occurs when connection of the current cell is broken and a connection to the new cell is established. This class of handoff characterizes the second generation (2G) network or Frequency Division Multiple Access (FDMA) systems. Soft handoff enables an overlapping of the repeater coverage zones, such that every cell phone set is always within the range of at least one of the base stations. This process is achieved by switching and establishing connection with a neighboring (current) base station, before disconnecting from the previous base station in that network. Soft handoff is a characteristic of third generation (3G) or Code Division Multiple Access (CDMA) systems. CDMA systems implement a spread spectrum (Viterbi, Viterbi, Gilhousen and Zehavi, 1994) principle known as universal frequency reuse that allows all users to share the available radio channels. This principle provides seamless connectivity and that makes handoff connections relatively permanent. The connection becomes more stable compared to other cellular technologies.

Given the limited radio spectrum, soft handoff shows high promise, as it reduces the network signal load and data lose. This technique stabilizes transmission and enhances effective mobility handling. Apart from mobility, soft handoffs are also implemented in CDMA as power control and interference reduction mechanisms on both links (uplink and downlink).

2 Problem Statement

CDMA is interference limited in theory. However, in practice, the capacity is also restricted by channel elements (CEs) at the base stations. Hence, dealing with interference simultaneously, considering the capacity of the CEs becomes vital. Also, it is desirable that the difference between capacities determined by interference and the number of CEs should be small. But when there are unused CEs, the interference level becomes high and call admission control (CAC) is required to block the admission of new calls.

The problem of soft handoff arises in cellular communication systems, when a mobile unit seeks to communicate with multiple base stations simultaneously. The set of base stations with which the mobile unit communicates at a given time instance is called the active set. As the mobility and system traffic load conditions change, the active set requires an update to maintain an acceptable signal quality. This change in the active set is the soft handoff event and is governed by the soft handoff algorithm. Another problem lies in the ability to identify the start point level of a pilot signal referring to the forward CDMA channel. The strength of these pilot signals are reported by the mobile stations (to the base station) from time to time. As soon as the measured signal exceeds a predefined threshold value, the mobile unit automatically becomes a member of the active set.

3 Background Literature

In the active set, mobile stations can interact simultaneously with various base stations (BSs) within its neighborhood. A mobile station in the process of handoff eventually makes a definite decision to communicate with only one BS. Based on changes in the pilot signal strength of these BSs, the mobile is finally transferred to the BS with the strongest signal. Hard handoff occurs instantaneously, while soft handoff lasts over a period of time.

Fig. 1 illustrates the difference between hard and soft handoffs.

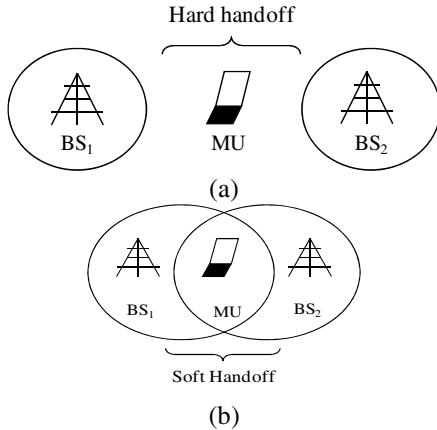


Fig. 1. Diagrammatic illustration of hard and soft handoffs

In hard handoff, the “ping-pong” effect (a phenomenon that describes the occurrence of frequent handoffs as a result of mobile users (MU) movement within cell boundaries) is mandatory. When this happens, the original traffic link with BS₁ is dropped before the setting up of a new link with BS₂ (Fig. 1(a)). Therefore, hard handoff is a process of “break before make”. This effect can however be reduced by the introduction of a hysteresis margin (Forkel, Schinnenburg and Wouters, 2003, Singh and Singh, 2003). In the case of soft handoff, as long as the soft handoff trigger condition is fulfilled, the MU enters a soft handoff state and a new link is established. Before BS₁ is dropped (handoff dropping condition is fulfilled), the mobile unit communicates with both BS₁ and BS₂ simultaneously (Fig. 1 (b)). Therefore, soft handoff is a process of “make before break”. So far, several algorithms have been proposed to support soft handoff and different criteria used for different algorithms.

Several researchers have studied the performance of soft handoff in CDMA systems. Their studies also attempt to solve problems using different models. Research on the tradeoffs and parameter settings are mostly in the form of simulation studies. Wang and Wang (1993) focus on how soft handoffs are made. They simulate a modified equation that includes factors directly or inversely proportional to the antenna gains along the direction from the particular interferer to its base station and

to the base station causing the interference. Analytical tools for soft handoff performance tradeoff analysis using basic framework are presented in Zhang and Holtzman (1998), Sheu and Hou (2005). They consider a mobile user at 180° to two base stations without interference and assume that new call generation rate per unit area is uniformly distributed over the service area. A comparative analysis using multiple base stations reveal more complicated traffic patterns and a prohibitively NP-complete case, when interference is considered. Aswa and Stark (1994) dwell on add and drop thresholds. They conclude that such thresholds are not good enough decision criteria for active set maintenance, hence, they formulate the problem as a reward/cost stochastic optimization problem, where the reward is associated with good signal and the cost with soft handoff overhead. Their model consists of N base stations and a user is permitted to access all the base stations in its active set.

Although these models reveal the prospects of soft handoff over hard handoff, especially in the area of uplink power control, they are however limited comparisons which do not adequately address the merits and demerits of soft handoff. When the soft handoff concept was formally introduced for IS-95 systems, a lot of researches followed. Most research publications came under three broad categories:

- (i) Evaluating soft handoff performance through the assessment of link level indicators such as the average E_0/I_0 and fade margin improvement for individual radio links (Viterbi, Viterbi, Gilhousen and Zehavi, 1994, Mihailescu, Lagrange and Godlewski, 1999, Kari, Mika, Jaana, and Achim, 1999).
- (ii) Using system level indicators for estimating the performance of soft handoff: QoS metric parameters such as outage probability, call blocking probability and handoff failure rate are useful indicators for estimating handoff performance. The level of performance however depends on the network load or related system optimization parameters such as capacity and coverage gain (Su, Chen and Huang, 1996, Lee, Un and Kim, 1996, Chen 1995, Homnan, Kunnsriksakul and Benjapolakul, 2000, Narrainen and Takawira, 2001, Choi and Kim, 2001, Fu and Thompson, 2002).
- (iii) Investigating soft handoff using resource efficiency indicators or adaptive techniques such as the average number of mobiles within the active set, active set update rate and handoff latency (Yang, Ghaheri-Niri and Tafazolli, 2000, Yang Ghaheri-Niri and Tafazolli, 2001, Chang and Sung, 2001, Wang, Sridsha and Green 2002).

Other researches dealt with the establishment of appropriate thresholds and solution to call overlap. In Rao and Mishia (2000), a report on the performance analysis of soft handoff algorithm proposed in the IS-95 CDMA standard is presented. They also present a simulation that enhanced the selection of appropriate handoff thresholds. A locally optimal handoff algorithm that improves the performance of the static threshold handoff algorithm is derived in Prakash and Veeravilli (2000). They report that ad-hoc dynamic threshold algorithm is a good approximation to the handoff algorithm. They also provide an analytical justification for the use of dynamic threshold algorithms. In Singh and Singh (2008), they show the effect of the characteristic parameters of the cellular environments such as pathloss exponent, standard deviation of shadow fading and correlation co-efficient of shadow fading on

soft handoff performance. They numerically prove that these propagation parameters have decisive effect on signal outage probability and hence on the overall performance of the soft handoff algorithm.

A focus on the effects of soft handoff applied to a realistically planned UMTS network is made in Forkel, Schinnenburg and Wouters (2005). They evaluate the interference mitigation and capacity loss tradeoffs using dynamic simulation. The effect of soft handoff techniques on cell coverage on reverse link capacity is investigated in Viterbi, Viterbi, Gilhousen and Zehavi (1994). The authors show that soft handoff significantly increases both parameters relatively to conventional handoffs.

Existing literature on the soft handoff problem constitutes a body of analytical evidence, which does not adapt proposed models to practical environments. The importance of this paper is not only to make data publicly available for future research, but to present a concise and working methodology that can be built on and enhance effective academia-industry collaborations.

4 Proposed System Model

One most important advantage offered by the CDMA technology is the ability of a mobile user to maintain seamless connectivity, which is made possible through soft handoff. This capability makes CDMA one of the viable technologies for future mobile communication systems. Mobile stations in soft handoff utilize multiple radio channels and receives signal from multiple base stations. In analyzing the performance of CDMA in soft handoff, several assumptions are required to simplify the handoff analysis. We start by defining a cell structure as a hexagonal formation, geometrically divided into three regions (Kim, 1999):

- (i) The inner cell region
- (ii) The soft handoff region
- (iii) The outer cell region

These regions are surrounded by inner and outer boundaries. A region surrounded by a cell boundary is called an ordinary cell. The overlap regions in the cell structure as shown in Fig. 2 are demarcated by thick lines. In this cell structure, considering cell 1,

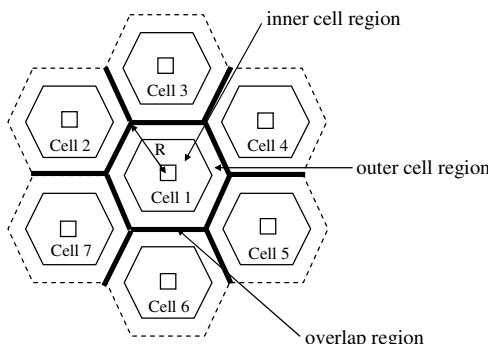


Fig. 2. A hexagonal cell structure

the soft handoff region (SHR) is a part of the six overlap regions. The region outside the SHR, of the ordinary cell is called the non-soft handoff region (NSHR).

4.1 Soft Handoff Traffic Model

We define the assumptions in analyzing the soft handoff problem as follows:

- (i) All mobile stations or users are uniformly distributed
- (ii) New call arrival follows a Poisson process, with rate λ_n
- (iii) Call holding time T_h is distributed exponentially with mean $1/\mu_h$

4.1.1 Handoff Call Attempt Rate

To compute the handoff call attempt rate for soft handoff, we consider the number of handoff call attempt rate (during a call holding time), \bar{N}_{CAR} , given as (Sheu and Hou, 2005):

$$\bar{N}_{CAR} = \frac{(1 - P_B)P_{NR}(P_{NS}P_I + P_S)}{(1 - (1 - P_{HF})P_{HR})^2} \quad (1)$$

where

P_B is the blocking probability

P_{HF} is the handoff failure probability

P_{NS} is the new call arrival probability in non-SR region

P_S is the new call arrival probability in SR region

P_I is the probability of new call leaving the inner cell.

P_{NR} is the probability of new call requesting handoff

P_{HR} is the probability of a handoff cell request.

The soft handoff call attempt rate λ_{CAR} can then be expressed as:

$$\lambda_{CAR} = \lambda_n \bar{N}_{CAR} \quad (2)$$

where λ_n is the new call arrival rate. In hard handoff, the old base station (BS) is released before connecting the new BS (no margin). Thus the hard handoff call attempt rate (λ_{CAR}) can be given as (Kim 1999):

$$\lambda_{CAR} = \lambda_n \frac{\mu_{cell}(1 - P_B)}{\mu_c + \mu_{cell}P_{HF}} \quad (3)$$

where

$1/\mu_{cell}$ is the residual time

$\frac{1}{\mu_c}$ is the mean call holding time.

4.1.2 Channel Holding Time

When a call terminates or the MU leaves the outer cell, channel occupancy is released. Hence, the channel holding time can be expressed as:

$$T_{ch} = \min(T_h, T_O) \quad (4)$$

where

T_h is the call holding time

T_O is the outer cell residual time.

T_{ch} is exponentially distributed with means $\frac{1}{\mu_h}$ and $\frac{1}{\mu_O}$ (Guerin 1987).

Therefore:

$$\frac{1}{\mu_{ch}} = \frac{1}{\mu_h + \mu_O} \quad (5)$$

4.1.3 Soft Handoff Probability

The probability that a user is in soft handoff mode (in the soft handoff region) is an important parameter for radio network planning. Due to the additional overhead caused by excessive soft handoffs on system resources, it becomes necessary to set the system parameters to appropriate levels that will optimize the network performance. In this paper, two important parameters are investigated:

- (i) The time the user is in soft handoff
- (ii) The total time simulated

We start by expressing the soft handoff probability as time ratios (Stjin, 2003):

$$P_{SHO} = \frac{\Delta t_{SHO}}{\Delta t_{total}} = \frac{\Delta d_R(SHO)/v}{\Delta d_R(total)/v} = \frac{\Delta d_R(SHO)}{\Delta d_R(total)} = \frac{w}{\sqrt{3}R} \quad (6)$$

where

Δt_{SHO} is the time the users is in soft handoff

Δt_{total} is the total time

$\Delta d_R(SHO)$ is the distance of trajectory crossed in handoff

$\Delta d_R(total)$ is the total length of trajectory.

v is the velocity of the mobile user

w is the width of soft handoff region

R is the cell radius

Suppose a base station is located at $d = 0$ (BS A) and the other at $d_R = \sqrt{3}R$ (BS B). Considering pathloss, the power of BS A at point x is:

$$P(x) = P(0) - Pathloss(x) \quad (7)$$

Using the COST-231 Hata pathloss model (Abhayawardhana, Wassell, Crosby, Sellars and Brown, 2005), given as:

$$PL = L_0 + n \log f - 13.82 \log h_b - C_H + [\sigma - 6.55 \log h_b] \log d + C \quad (8)$$

where

PL is the pathloss

n is the propagation exponent

f is the transmission frequency in MHz

h_b is the base station antenna height in m

d is the link distance in km

C_H is the mobile station antenna height correction factor

L_0 is the pathloss due to free space

σ is the standard deviation

$$C = \begin{cases} 0dB & \text{rural and suburban areas} \\ 3dB & \text{for urban areas} \end{cases}$$

we adapt this model to our study environment by substituting the empirical pathloss parameters obtained from the field (see Table 1)

Table 1. Empirical data for pathloss determination obtained from Starcomms Nigeria

Parameter	Value
Transmission frequency (f)	1900MHz
Base station antenna height (h_b)	45m
Mobile station antenna height correction factor (C_H)	1.4m
Pathloss due to free space (L_0)	46.30dB
Standard deviation (σ)	44.90dB

On substitution of these parameters, we arrive at:

$$\begin{aligned} PL &= 46.3 + 33.9 \log_{10}(1900) - 13.82 \log_{10}(45) - 1.4 + [44.90 - 6.55 \log_{10}(45)] \log_{10} d + 3 \\ PL &= 46.3 + 111.15 - 22.8474 - 1.4 + [44.90 - 10.8285] \log_{10} d + 3 \\ PL &= 136.20 + 34.07 \log_{10} d \text{ or } 34.07 \log_{10} d + 136.20 \end{aligned} \quad (9)$$

Hence, at $d = \frac{\sqrt{3}R}{2}$, the ratio of soft handoff threshold (SHO_TH) to the width of the soft handoff region (w) can be expressed as:

$$\frac{SHO_TH}{w} = \left| \frac{dPL}{dx} \right|_{d=\frac{\sqrt{3}R}{2}} = \frac{34.07}{\ln 10} \Big|_{d=\frac{\sqrt{3}R}{2}} = \frac{A}{R} \quad (10)$$

$$\therefore w = \frac{R}{A} SHO_TH \quad (11)$$

where

R is the cell radius

A is the coverage area

Therefore,

$$prob\ SHO = \frac{w}{\sqrt{3}R} = \frac{1}{\sqrt{3}A} SHO_TH \quad (12)$$

Introducing our case study parameters, we arrive at:

$$prob\ SHO = \frac{1}{2(\frac{34.07}{\ln 10})} SHO_TH \quad (13)$$

$$prob\ SHO = \frac{1}{29.59} SHO_TH \quad (14)$$

Soft handoff probability can also be expressed in relation to the overlap and cell regions. Thus, following similar derivation in equation (10), we obtain:

$$prob\ SHO = 1 - \frac{4}{\left(10^{\frac{SHO_TH}{34.07}} + 1 \right)^2} \quad (15)$$

Equation (15) is the soft handoff-propagation exponent model for our study environment. To allow for the simulation for different propagation environments, we replace the propagation exponent with a variable n , and arrive at:

$$prob\ SHO = 1 - \frac{4}{\left(10^{\frac{SHO_TH}{n}} + 1 \right)^2} \quad (16)$$

where n represents the propagation exponent of the different environments (urban, suburban, rural and free space).

5 Empirical SHO Data Analysis

In this section, we study realistic data that defines the propagation behaviour of the study environment. A survey of Starcomms, a CDMA telecommunication service

provider operating in Nigeria, was made. The survey captured relevant data from the various base stations in the Southeast region of Nigeria. We discovered that Starcomms network has seventy-five (75) sectorized base stations (cells) transmitting at a frequency of 1.9GHz. Their cells have an antenna height of 45m, radio frequency (RF) height ranging between 1.2m and 1.4m, with microwave diameter of 0.6m and average cell capacity of 2000 users. Soft handoffs ratio (%) measurements at the various base stations were also captured from the main control-switch room of Starcomms, Port Harcourt branch. These data were collated over a period of two months (September-October, 2010).

A SHO performance model for the area under study derived in this paper uses the mean of the observed data as model predictors to ensure prediction accuracy of the proposed model. The SHO control algorithm implementing the SHO process is shown in Fig. 3:

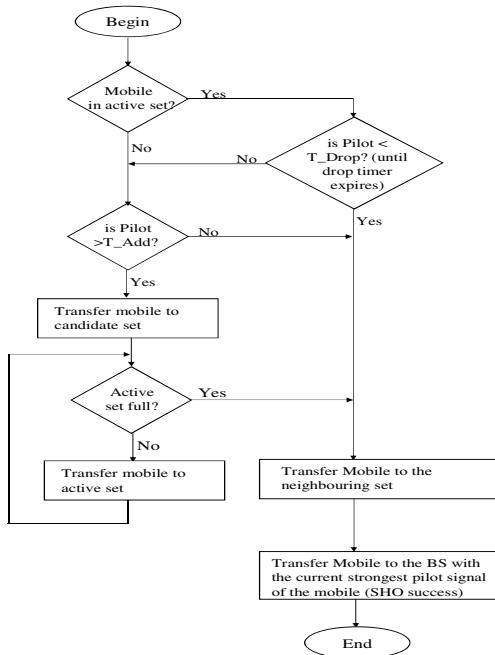


Fig. 3. Soft handoff algorithm

Initial investigation of the system (see Fig. 4), reveals that soft handoff (SHO) rate was not stable over the study period. This could be attributed to the following:

- (i) High co-channel interference due to the influx of mobile users during busy periods (e.g. festive seasons).
- (ii) Frequent handoff failures resulting from base station malfunctioning and lack of efficient call admission control scheme.

- (iii) Type of services mostly subscribed by customers. We observed that users subscribed more to data services such as internet access, than voice services. As such not much call handoffs took place.
- (iv) Impairments such as congestion, blocking, etc. were noticed, where the handoff ratio exceeded a certain threshold.

In Fig. 5, we probe further into the cause of this unstable system nature, by studying the monthly durations separately. A plot of the average daily data obtained from the field (see Appendix), shows that in the month of September, the SHO ratio increases slowly across the base stations and decreases in October. The decrease could be attributed to less traffic. In the month of October, there seemed to be heavy traffic as a result of more users' migration. The R^2 values ($R^2=0.2014$ and $R^2=0.0127$) reveal that the number of base stations do not significantly influence soft handoffs. Trend equations are fitted in the plots for the purpose of predicting new results.

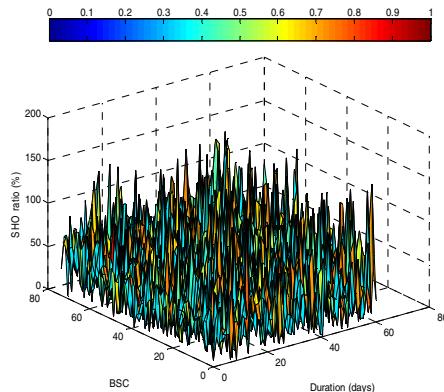


Fig. 4. Analysis of call SHOs across the various base stations for the period under study

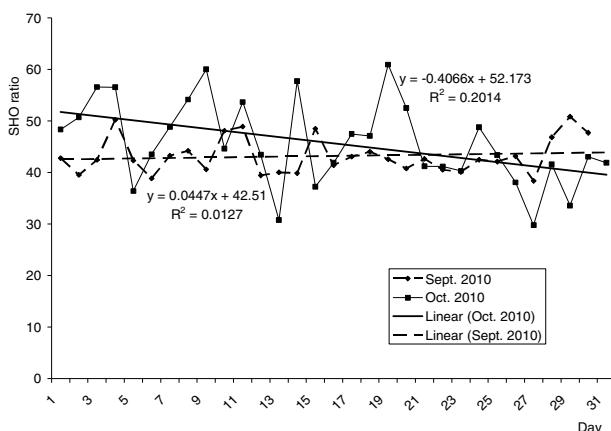


Fig. 5. Graph of average SHO ratio vs. number of days

6 Model Simulation and Discussion of Results

Table 2 shows the sample input parameters and their respective values for the various system environments under study:

Table 2. Simulation entries

Parameter	Value
Coverage Areas(A)	Urban-346m ² Suburban-573m ² Rural-968m ²
Propagation exponent (η)	Urban-34.07m ² Suburban-38.4m ² Rural-43.5m ²
Threshold values (SHO_TH)	Freespace-20m ² 2dB-20dB

Sample outputs were generated in the form of graphs, and are discussed as follows:

Fig. 6 shows the effect of SHO threshold on SHO probability in diverse propagation environments (urban, suburban and rural). We discover from the simulation that SHO threshold increases with the SHO probability, i.e., the higher the threshold value, the higher the SHO probability. Also, there is a proportional increase in growth across the respective environments. Therefore, network operators should maintain a minimum SHO threshold for different propagation environment (or deployments) to avert the detrimental effect of running into undesirable states that

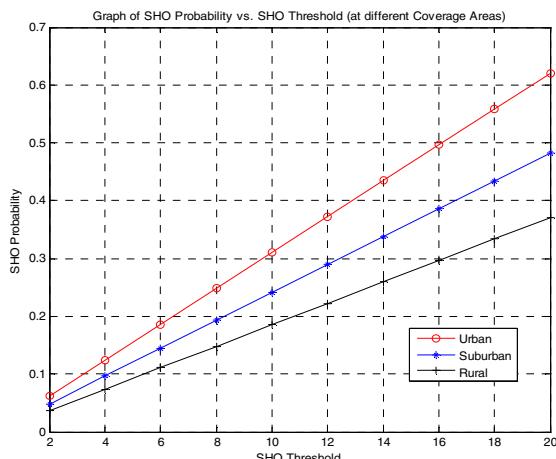


Fig. 6. Graph of SHO probability vs. SHO threshold for diverse environments

could cause system malfunctioning. From our investigations, we discovered that network site engineers were unfamiliar with the soft handover concept and how to deal with problems. This explains why the network degrades longer than expected without due attention. The neglect in turn affects the subscribers and negatively impact on the system's performance.

In Fig. 7, we study the effect of SHO threshold on SHO probability in different propagation environments, using the respective propagation exponents as varying parameters. The graph is a family of power curves, exhibiting rapid convergence of values to the optimum system utilization state (i.e. SHO probability = 1). However, as can be seen from the graph, free-space allows for faster convergence than other propagation environments.

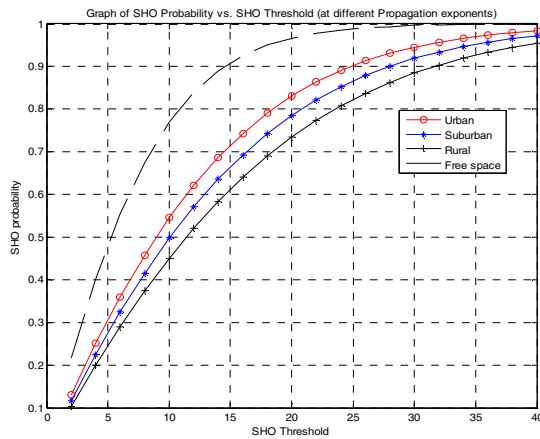


Fig. 7. Graph of SHO probability vs. SHO threshold for diverse propagation exponents

This shows that network systems could perform optimally in free-space than in other environments. The notion is true given the nature of impairments that tend to obstruct the line of site in other propagation environments. In practice, cellular network operators should not site base stations among obstacles, but use effective strategies to circumvent obstacles and observe a free-space broadcast. Furthermore, our model controls the likelihood of a system going into indeterminate states (rising above a SHO probability of 1), as experienced in the existing system (see Fig. 4).

7 Conclusion

Soft handoff is an intriguing technology. It promises better performance than hard handoff. In this paper, we have analyzed the performance of soft handoff in a realistic cellular network and revealed the problems associated with soft handoffs in telecommunication systems. We have also proposed a SHO model, implementing same using realistic data for the purpose of improving the performance of the system, taking into consideration different coverage areas and diverse propagation exponents.

The paper has helped establish a practical solution using propagation models that improves the performance of soft handoffs in CDMA networks. We have focused on the establishment of soft handoff threshold values that are appropriate for urban, suburban, rural, and free space communities, hence, have provided best practices to network operators.

Acknowledgements. We are grateful to Starcomms staff for assisting during the data collection phase of this research.

References

- Abhayawardhana, V.S., Wassell, I.J., Crosby, D., Sellars, M.P., Brown, M.G.: Comparison of Empirical Propagation Path Loss Models for fixed Wireless Access Systems. In: Proceedings of 61st IEEE Vehicular Technology Conference, VTC 2005-Spring, vol. 1, pp. 73–77 (2005)
- Aswa, M., Stark, W.E.: A Framework for Optimal Scheduling of Handoffs in Wireless Networks. In: Proceedings of IEEE Global Telecommunication Conference and Exhibition, San Francisco, CA, pp. 1828–1833 (1994)
- Chang, J.W., Sung, D.K.: Adaptive Channel Reservation Scheme for Soft Handoff in DS-CDMA Cellular Systems. *IEEE Transactions on Vehicular Technology* 50(2), 341–353 (2001)
- Chen, X.H.: Adaptive Traffic-load Shedding and its Capacity Gain in CDMA Cellular Systems. In: Proceedings of IEEE Conference on Communications, vol. 142(3), pp. 186–192 (1995)
- Choi, W., Kim, J.Y.: Forward-link capacity of a DS/CDMA System with Mixed Multirate Sources. *IEEE Transactions on Vehicular Technology* 50, 737–749 (2001)
- Forkel, I., Schinnenburg, M., Wouters, B.: Performance Evaluation of Soft Handover in a Realistic UMTS Network. In: 57th IEEE Semiannual Conference on Vehicular Technology (VTC 2003-Spring), vol. 3, pp. 1979–1983 (2003)
- Fu, H., Thompson, J.S.: Downlink Capacity Analysis in 3GPP WCDMA Networks System. In: Proceedings of Third International Conference on 3G Mobile Communication Technologies, pp. 534–538 (May 2002)
- Guerin, R.A.: Channel Occupancy Time Distribution in a Cellular Radio System. *IEEE Trans. Veh. Tech.* VT-35, 89–99 (1987)
- Homnan, B., Kunsriruksakul, V., Benjapolakul, W.: A Comparative Performance Evaluation of Soft Handoff between IS-95A and IS-95B/CDMA2000. In: IEEE APCCAS 2000, pp. 34–37 (2000)
- Kari, S., Mika, J., Jaana, L.S., Achim, W.: Soft Handover Gains in a Fast Power Controlled WCDMA Uplink. In: Proceedings of IEEE VTC 1999, pp. 1594–1598 (1999)
- Kim, D.K.: Characterization of Soft Handoff in CDMA System. *IEEE Transactions on Vehicular Technology* 48(4), 1195–1202 (1999)
- Lee, D.J., Un, C.K., Kim, B.C.: An Improved Soft Handover Initiation Algorithm in Microcellular Environment. In: Proceedings of 5th IEEE International Conference on Universal Personal Communication, vol. 1, pp. 310–314 (1996)
- Mihailescu, C., Lagrange, X., Godlewski, P.: Soft Handover Analysis in Downlink UMTS WCDMA System. In: Proceedings of IEEE Workshop on Communications, pp. 279–285 (1999)

- Narrainen, R.P., Takawira, F.: Performance Analysis of Soft Handoff in CDMA Cellular Networks. *IEEE Transactions on Vehicular Technology* 50(6), 1507–1517 (2001)
- Prakash, R., Veeravilli, V.V.: Locally Optimal Soft Handoff Algorithms. *IEEE Transactions on Vehicular Technology* 52(2), 347–356 (2000)
- Rao, K.B., Mishra, L.N.: Soft Handoff in CDMA Systems. *IEEE Transactions on Vehicular Technology* 47(2), 710–714 (2000)
- Sheu, T.-L., Huei Hou, J.-H.: An Analytical Model of Cell Coverage for Soft Handoffs in Cellular CDMA Systems. *GESTS Int'l Trans. Computer Science and Engineering* 18(1), 209–223 (2005)
- Singh, N.P., Singh, B.: Effect of Soft Handover Parameters on CDMA Cellular Networks. *Journal of Theoretical and Applied Information Technology*, 110–115 (2005–2010)
- Singh, N.P., Singh, B.: Performance of Soft Handover Algorithm in Varied Propagation Environments. *World Academy of Science, Engineering and Technology Journal* 45, 377–381 (2008)
- Stjin, N.P.: Study of soft handover in UMTS. Master Thesis, Technical University of Denmark, Denmark (2003)
- Su, S.L., Chen, J.Y., Huang, J.-H.: Performance Analysis of Soft Handoff in CDMA Cellular Networks. *IEEE Journal on Selected Areas in Communications* 14(9), 1762–1769 (1996)
- Viterbi, A.J., Viterbi, A.M., Gilhousen, K.S., Zehavi, E.: Soft Handoff Extends CDMA cell coverage and increase Reverse Link Capacity. *IEEE Journal on Selected Areas of Communications* 12(8), 1281–1288 (1994)
- Wang, S.S., Sridha, S., Green, M.: Adaptive Soft Handoff Method Using Location Information. In: *IEEE 55th Conference on Vehicular Technology, VTC Spring*, vol. 4, pp. 1936–1940 (2002)
- Wang, S.-W., Wang, I.: Effects of Soft Handoff, Frequency Reuse and Non-ideal Antenna Sectorization on CDMA System Capacity. In: *Proceedings of 43rd IEEE Vehicular Technology Conference, VTC Spring*, pp. 850–854 (1993)
- Yang, X., Ghaheri-Niri, S., Tafazolli, R.: Evaluation of Soft Handover Algorithms for UMTS. In: *Proceedings of 11th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 26, pp. 772–776 (2000)
- Yang, X., Ghaheri-Niri, S., Tafazolli, R.: Performance of Power Triggered and E_0/N_0 -Triggered Soft Handover Algorithms for UTRA. In: *3G Mobile Communication Technologies Conference Publication*, vol. 477, pp. 7–10 (2001)
- Zhang, N., Holtzman, J.M.: Analysis of a CDMA Soft-handoff Algorithm. *IEEE Transactions on Vehicular Technology* 47(2), 710–714 (1998)

Appendix: Average daily SHO ratio data collected from Starcomms Telecommunications for two base station controllers located in the south-east region of Nigeria.

Month: September, 2010

1	2	3	4	5	6	7	8	9	10
40.33	35.92	38.25	50.15	37.99	40.25	37.83	44.36	33.38	51.24
45.29	43.16	46.70	50.34	46.69	37.43	48.70	44.09	47.87	44.98
11	12	13	14	15	16	17	18	19	20
42.40	38.92	33.18	39.77	49.14	34.46	49.25	40.24	45.26	38.24
55.44	40.00	46.88	40.00	47.88	48.32	36.90	47.81	39.91	43.34
21	22	23	24	25	26	27	28	29	30
39.96	29.97	33.84	47.19	44.17	48.70	51.34	46.68	35.43	46.77
45.34	51.11	46.45	37.79	40.11	37.66	25.43	47.04	66.29	48.69

Month: October, 2010

A Security Approach for Mobile Agent Based Crawler

Vimal Upadhyay, Jai Balwan, Gori Shankar, and Amritpal

Department .of CS/EC, St. Margaret Engineering College, Neemrana (Alwar)(Rajasthan)
vimalupadhyay2002@gmail.com, balwan_smec@rediffmail.com,
gorishanker44@gmail.com, palamrit83@gmail.com

Abstract. Mobile agents are active objects that can autonomously migrate in a network to perform tasks on behalf of their owners. Though they offer an important new method of performing transactions and information retrieval in networks, mobile agents also raise several security issues related to the protection of host resources as well as the data carried by an agent itself. Mobile agent technology offers a new computing paradigm in which a program, in the form of a software agent, can suspend its execution on a host computer, transfer itself to another agent-enabled host on the network, and resume execution on the new host. Mobile Agent (MA) technology raises significant security concerns and requires a thorough security framework with a wide range of strategies and mechanisms for the protection of both agent platform and mobile agents against possibly malicious reciprocal behavior. The security infrastructure should have the ability to flexibly and dynamically offer different solutions to achieve different qualities of security service depending on application requirements. The protection of mobile agent systems continues to be an active area of research that will enable future applications to utilize this paradigm of computing. Agent systems and mobile applications must balance security requirements with available security mechanisms in order to meet application level security goals.

A security solution has been introduced, which protects both the mobile agent itself and the host resources that encrypt the data before passing it to mobile agent and decrypt it on the visited host sides i.e. it transfers the URL to the Mobile Agent System that will pass that encrypted URL to the server where it will be decrypted and used. The methods of Encryption/Decryption used are a Public-key Cipher System and a Symmetric Cipher System that focuses on submitting data to the server securely. The proposed approach solves the problem of malicious host that can harm mobile agent or the information it contain.

Keywords: Uniform resource locator (URL), Mobile agent (MA), Hyper Text Transfer Protocol (HTTP).

1 Introduction

Mobile agents are small threads of execution that are able to migrate from machine to machine, performing operations locally. An application area for mobile agents is internet computing. Mobile agents provide a very attractive paradigm for this area. The agents can be launched from a machine, navigate from web to web, collecting

information or performing transactions, finally returning home with the goods or results.

Mobile agents are an appealing alternative to the client-server architecture for many applications. MA's have extended the mobile-code concept to "mobile object" in which an object (code + data) are moved from one host to another. This approach extends the concept by moving code, data and state (thread) from one host to another as well. MAs run at one location, move with their state to another host, and continue execution at that host. Mobile code and mobile objects are normally moved by an external entity while MAs are usually migrate autonomously as shown in Fig. 1.

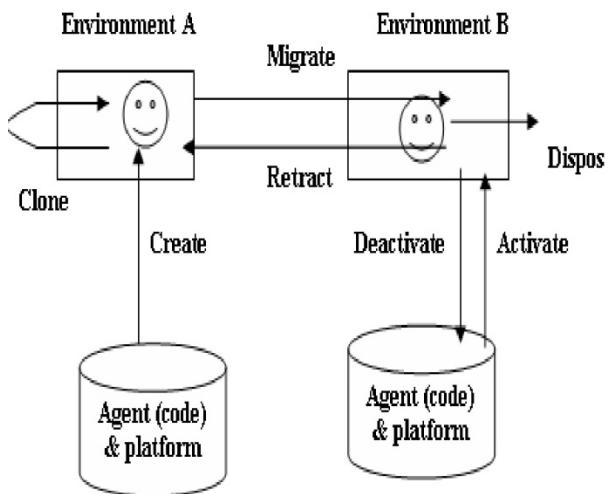


Fig. 1. Mobile Agent with Platform

The goals of Mobile Crawling System are:-

- To minimize network utilization.
- To keep up with document changes by performing on-site monitoring.
- To avoid unnecessary overloading of the web servers by employing time realization.
- To be upgradeable at run time.

2 Drawbacks of Mobile-Agent Based Crawler

Mobile agent based crawler has some limitations, primarily in the area of security. Current research efforts in the area of mobile agent security adopt two different points of view. Firstly, from the platform perspective, there is a requirement to protect the host from malicious mobile agents such as viruses and Trojan horses that are visiting it and consuming its resources. Secondly, from the mobile agent point of view, it

needs to protect the agent from malicious hosts. Therefore, security is a fundamental precondition for the acceptance of mobile agent applications. In other words, The system should have a program that actively protects itself against execution environment that possibly may divert the intended execution towards a malicious goal. Many approaches aim at protecting mobile agents. There are some problems, which have to be solved before these approaches can be used. The particular attacks that a malicious host, malicious agent can make can be summarized as follows.

- Observation of code, data and flow control,
- Manipulation of code, data and flow control, including manipulating the route of an agent,
- Incorrect execution of code – including re-execution,
- Denial of execution – either in part or whole,
- Masquerading as a different host,

Eavesdropping on agent communications,

Manipulation of agent communications,

- It is very difficult to protect mobile agent as it visit different node in network so security should be apply on the platform. This paper contain the application in which protection on data by encrypt it on platform so no other malicious node or malicious plate form access the data carry by mobile agent. The technology use for the protection is computing with cryptography.

3 Proposed Architecture of Security Based Mobile Agent Crawler

The method of Encryption/Decryption is used to ensure the security of the agent and the host resources. The mobile agent program is encrypted on the creator side before it is transferred to other hosts, and deciphered on the remote hosts side when it gets to the right hosts. The encryption needs not only a private key for the computation with cryptography method but also the private key of every destination host. After the destination hosts have received the encrypted agent, they decipher the cryptograph and recover the original mobile agent code using their corresponding private keys. The proposed security architecture is shown in Fig. 2, which starts after firing the query to internet. Then, the query is passed to crawler which works in three stages:

- Retrieval stage: In this, the crawler has to retrieve the resources which will be part of the index. It contacts a remote HTTP (Hypertext Transfer Protocol) server, requesting a web page specified by a URL (Uniform Resource Locator) address.

- Analysis stage: After a certain resource has been retrieved, the crawler will analyze the resource in a certain way depending on the particular crawling algorithm. For example, in case the retrieved resource is a Web page, the crawler will probably extract hyperlinks and keywords contained in the page.
- Decision state: Based on the results of the analysis stage, the crawler will make a decision how to proceed in the crawling process. After that, it passes the URL one by one to encryption layer. The secure URL given to mobile agent is encrypted before transferring to it. This encryption is done after the completion of crawling. This solves the problem regarding any malicious host that can harm mobile agent or any information.

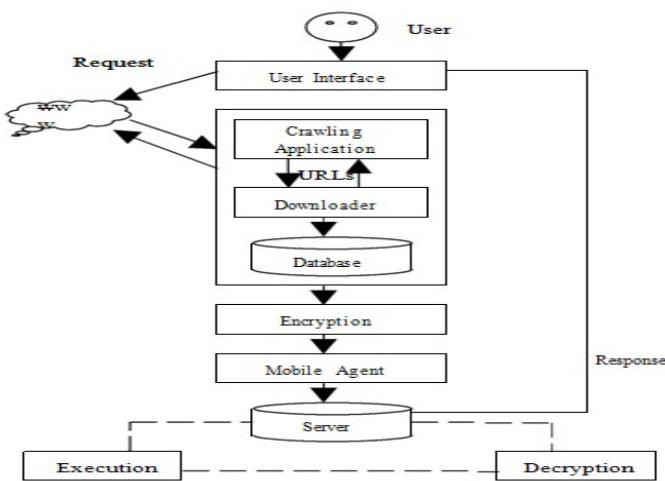


Fig. 2. Security Based mobile Agent Crawler

4 Conclusion

Mobile agents have gained a great deal of attention in research and industry in the recent past. Although mobile agents are a promising technology, the large-scale deployment of agents and the existence of hosts running agencies will not happen until proper security mechanisms are well understood and implemented. Mobile agents are plain enough for malicious parties to read and to analyze. A malicious host may read or alter the content of the agent, or analyze the accumulated information carried by the mobile agent. Another program or agent running on the same host as the agent is another source of threat for the agent. It is more difficult to ensure security in the mobile agent paradigm than in some other technologies where hardware solutions are practical. Problems in security have been seen as an obstacle in the way of success of mobile agent technology.

References

1. Anderson, J.P.: Computer Security Threat Monitoring and Surveillance. Technical Report, James P. Anderson Co., Fort Washington, PA (April 1980)
2. Asaka, M., Okazawa, S., Taguchi, A., Goto, S.: A Method of Tracing Intruders by Use of Mobile Agents. In: INET 1999 Conference (June 1999)
3. Balasubramaniyan, J., Garcia-Fernandez, J.O., Isacoff, D., Spafford, E.H., Zamboni, D.: An Architecture for Intrusion Detection using Autonomous Agents. Department of Computer Sciences, Purdue University, Coast TR 98-05 (1998)
4. Boudaoud, K., Labiod, H.: MA-NID: A Multi-Agent System for Network Intrusion Detection. In: Eighth International Conference on Intelligent Systems (June 1999)
5. Cabri, G., Leonardi, L., Zambonelli, F.: The Impact of the Coordination Model in the Design of Mobile Agent Applications. In: Twenty-Second Computer Software and Applications Conference, COMPSAC (August 1998)
6. Jansen, W., Karygiannis, T.: Privilege Management Mobile Agents. In: Twenty-Third National Information Systems Security Conference, pp. 362–370 (October 2000)
7. Karjoth, G., Asokan, N., Gülcü, C.: Protecting the Computation Results of Free-Roaming Agents. In: Second International Workshop on Mobile Agents, Stuttgart, Germany (September 1998)
8. Lange, D., Oshima, M.: Programming and Deploying Java Mobile Agents with Aglets. Addison-Wesley (1998) ISBN:0-201-32582-9
9. Martino, S.: A Mobile Agent Approach to Intrusion Detection. In: Joint Research Centre-Institute for Systems, Informatics and Safety, Italy (June 1999)
10. Yee, B.S.: A Sanctuary for Mobile Agents. Technical Report CS97-537, University of California in San Diego (April 28, 1997)

Prospects and Limitations of Organic Thin Film Transistors (OTFTs)

B.K. Kaushik¹, Brijesh Kumar², Y.S. Negi², and Poornima Mittal³

¹ Department of Electronics and Computer Engineering,
Indian Institute of Technology, Roorkee, 247667, India

² Polymer Science and Technology Program, DPT,
Indian Institute of Technology, Roorkee, 247667, India
Department of Electronics and Communication Engineering,

Graphic Era University, Dehradun, 248001, India
{bkk23fec, bk228dpt, ynegifpt}@iitr.ernet.in,
poornima2822@gmail.com

Abstract. Organic Transistor (OT) modeling, fabrication and applicability has undergone remarkable progress during last ten years. Organic Thin Film Transistors (OTFTs) have received significant attention recently because of their considerable utility. They can be fabricated at lower temperature and significantly reduced cost as compared to Hydrogenated Amorphous Silicon Thin Film Transistors (a-Si: H TFTs). Fabrication of OTFTs at low temperature allows utilization of wide range of substrates, thereby permitting usage of organic transistors as future candidate for many low-cost electronics applications that require flexible polymeric substrates such as RFID tags, smart cards, electronic paper, and active matrix flat panel displays. This paper provides detailed insight of OTFTs, their operating principles, device materials and various structures such as Top Gate Top Contact (TGTC), Top Gate Bottom Contact (TGBC), Bottom Gate Top Contact (BGTC) and Bottom Gate Bottom Contact (BGBC). Although OTFTs find tremendous and widespread applications, but is marred by few limitations related to factors such as speed, compatibility, stability, degradability and variability. This paper comprehensively discusses the performance and limitations of OTFTs.

Keywords: Operating principle, Organic thin film transistor, Organic transistors and Pentacene.

1 Introduction

During the last few years, research in OTFTs is being actively pursued because of their potential for low cost fabrication of large area circuits. OTFTs possible applications range from Active-matrix displays, E-paper, E-book, Signage, Advertisement, Storage devices, Electronic display cards (Smart cards, Gate pass, and Game cards), Touch screen mobile phones, Automobile, Wearable cloths, RFID tags, Price/Inventory tags, Flexible integrated circuits to Sensors and other novel products [1-7]. Due to their unique material properties, organic semiconductors have several

advantages over their inorganic counterparts. The most important advantage is in the processing techniques, where organic devices can be fabricated on large area [2], at much lower temperature and at considerably lower cost thereby allowing the use of flexible substrates [6, 7]. Therefore, researchers have actively considered fabricating OTFTs and other organic devices on a variety of unconventional substrates like glass, plastic [4], paper [5] and fibers. Furthermore, organic materials supports novel and unconventional fabrication techniques, such as spin coating process, nano imprinting techniques which enables rigorous downscaling of the devices to nano level size in a more cost effective way than for inorganic devices [7, 10, 12].

Performance of OTFTs can be expressed in terms of conventional parameters such as on-off current ratio (I_{ON}/I_{OFF}), field effect mobility (μ), threshold voltage (V_T) and transconductance (g_m). The performance of the best OTFTs now rivals that of commercial a-Si- H TFTs, which are commonly employed as the pixel switching elements in active matrix flat panel displays [8, 9]. A number of industrial laboratories are working hard to develop low-cost, large-area plastic electronics materials making use of TFTs and diodes based on organic semiconductors [9]. Inspite of impressive progress in their performance, a relevant physical description or model is still not available that can adequately predict or interpret observed phenomena of OTFTs [9, 10]. However, there are still many issues remaining with OTFTs that are to be resolved for their successful and reliable future use. Although, OTFT technology is on the verge of commercialization, but still faces a number of challenges that require fundamental understanding at the fabrication, materials chemistry and device physics to make significant progress [11-13].

This paper contains five sections. The present section 1 introduces the contents of the paper. Section 2 describes various top and bottom gate structures, material and fabrication of OTFTs. Compact model and operating principle are explained in section 3. Performance parameters of various structures are analyzed and discussed in section 4. Finally, section 5 discusses some of the limitations of OTFTs that needs to be overcome in future.

2 OTFT Structure, Material and Fabrication

An OTFT is a transistor made up of thin film current carrying organic semiconductor (OSC), an insulator layer and three electrodes. Two of the electrodes, source (S) and drain (D) are in direct contact with organic semiconductor and the third, gate (G) electrode is isolated from semiconductor by dielectric insulator. The structure of device is stated not only by its operating mode but also by issues and limitations arising from its fabrication process [7]. Firstly, the main difference between the geometry of conventional MOSFETs and OTFTs is that latter does not have a fourth terminal that is body, thus making these transistors free of body effect. Secondly, in MOSFETs the conducting channel is formed by an inversion layer while in OTFTs, it is because of accumulation layer [22]. Classification of different OTFTs structure based on top or bottom gate is shown in Fig.1.

The development of different structures had motivation to improve electrostatic control of gate (G) over organic conducting channel (such as pentacene) as well as to reduce contact resistances at source and drain regions [8]. Based upon the relative

position of S, D and G contacts with respect to OSC layer different structures can be made for OTFTs [12]. Certain merits and demerits are associated with each of the OTFT structures.

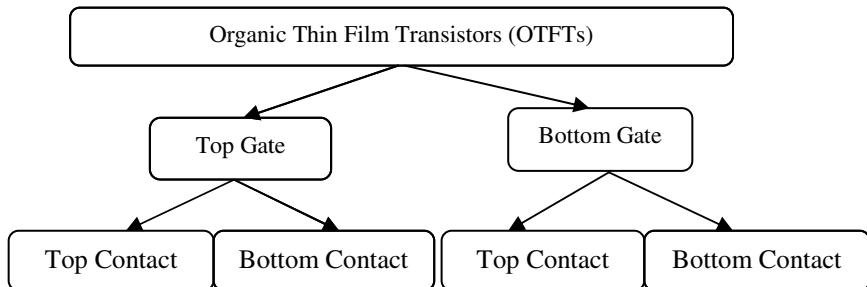


Fig. 1. Various structures of Organic Thin Film Transistors (OTFTs)

2.1 Top Gate Structures

The structure of top gate OTFT is like conventional MOSFET, where gate is placed on the top of organic semiconductor layer [7, 8]. Further it has two different configurations, i.e. top contact and bottom contact structures. These structures are based on relative position of contacts (S and D) with respect to semiconductor layer. In top contact structure, the contacts are deposited above OSC layer through shadow mask, whereas in bottom contact structure, the contacts are placed below OSC layer using microlithography technique. Based on the placement of contacts, the two configurations of top gate structures are named as Top Gate Top Contact (TGTC) and Top Gate Bottom Contact (TGBC). As shown in Fig. 2 (a) for TGTC structure, gate is defined on the top of organic semiconductor layer and source and drain contacts are patterned between gate dielectric and organic semiconductor film. However, in TGBC structure source and drain contacts are patterned on a substrate rather than on an insulator layer or organic semiconductor layer as shown in Fig. 2 (b). The performance of top gate structure may get severely degraded when the underlying active organic semiconductor layer is affected during deposition of gate [12]. Thus top gate structures require precise and accurate fabrication procedure.

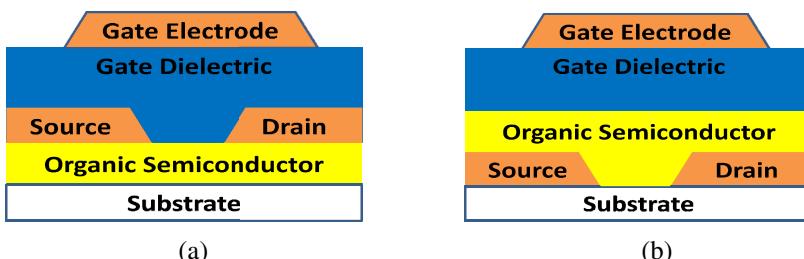


Fig. 2. Top gate OTFTs structures (a) Top Gate Top Contact (TGTC) structure (b) Top Gate Bottom Contact (TGBC) structure

2.2 Top Gate Structures

Bottom gate structures are built in majority for existing OTFTs since deposition of organic semiconductor on insulator is much easier than the reverse due to fragile nature of organic semiconductors. Furthermore, bottom gate structure can be categorized to two different configurations *viz.* Bottom Gate Bottom Contact (BGBC) and Bottom Gate Top Contact (BGTC). As shown in Fig. 3 (a) for BGBC structure, the gate is at the bottom, above which the insulator (dielectric) layer is deposited. S and D contacts are formed above the dielectric layer and finally organic semiconductor layer is deposited on the top of the structure [7]. BGBC structure is advantageous because the methods involving solvents and/or thermal treatments can be safely employed to prepare the gate dielectric and contacts without harming the semiconductor layer [8]. Moreover, bottom contact structures have an advantage of utilizing standard lithographic techniques straightforwardly for obtaining short channel length devices [12]. However, these devices usually suffer from higher contact resistances because of smaller effective area for charge carrier injection [22].

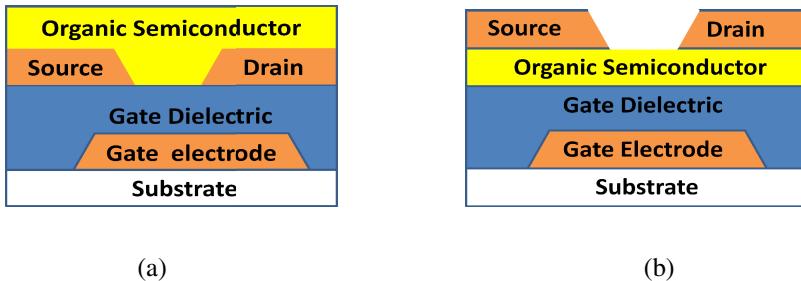


Fig. 3. Bottom gate OTFTs structures (a) Bottom Gate Bottom Contact (BGBC) structure (a) Bottom Gate Top Contact (BGTC) structure

In BGTC structure, contacts on top are deposited through shadow mask where as for bottom gate formation microlithography technique is used [7]. The top contact OTFT fabricated using shadow mask technique results in relatively large channel length devices [10]. However, for same semiconductor and dielectric materials, the contact resistance and mobility are better in top contact devices. The better field effect mobility for BGTC OTFT is due to less contact resistance than that of BGBC [12]. OTFTs mobility in BGBC device structure is generally observed to be lower by two orders of magnitude than to BGTC device. The reasons for this difference is often explained by the large metal-semiconductor contact resistance due to interface contact barrier and irregular deposition or poor morphology of semiconductor film around the already patterned S and D contacts [8]. Organic semiconductor films grown on metal often have inferior properties as compared to those grown on a dielectric [11]. In BGTC, lower contact resistance and larger injection area enables higher currents for the same applied voltages in comparison to BGBC structure [21].

2.3 Organic Thin Film Transistor Materials

The important factors for high performance OTFTs are dielectric/semiconductor interface, metal/semiconductor interface, ordered molecular structure of active layer, efficient injection at contacts and stability. A wide variety of materials for all the layers of OTFT are explored in various combinations to optimize their performance. Organic materials offer strong assurance in terms of properties, processing and cost effectiveness.

2.3.1 Organic Semiconductor Thin Film

Organic semiconductor materials are divided in two groups such as polymers and small molecules. The mobility of polymers is lower than small molecules because of higher molecular weight [10]. To obtain higher mobility in OTFTs, active semiconductor layer should have large grain size. There are various polymers and small molecule semiconductor materials that can be used as active OSC material such as pentacene, poly (3-hexylthiophene) (P3HT) [14], poly (3-alkylthiophene) (P3AT) and poly (3-octylthiophene) (P3OT) [15]. Pentacene is the most extensively used small molecule material, because of its highest mobility among all organic semiconductor materials. Pentacene on SiO_2 has yielded higher carrier mobility. Pentacene has other advantages such as good chemical stability in adverse environmental conditions, very high mobility in thin film form and stable interface with commonly used electrode metals such as Al and Au [8, 22].

2.3.2 Electrodes

Electrodes can be fabricated using either inorganic or organic materials. The contact metal for S and D should have low interface barrier with active semiconductor layer so that large number of carriers can be injected. Moreover, these contacts should have small contact resistance [18, 19]. Prominent inorganic contact materials include aluminum (Al), gold (Au), titanium (Ti), copper (Cu), calcium (Ca), nickel (Ni), platinum (Pt), magnesium (Mg) and chromium (Cr) [8, 22]. New classes of organic materials known as conducting polymers are also available for S/D contacts. Poly-3, 4-ethylenedioxythiophene : styrene sulfonic acid (PEDOT: PSS), polyaniline doped with camphorsulphonic acid (PANI-CSA) [14] deposited by spin coating are commonly used conducting polymers for S/D electrodes. Among all inorganic contact materials, Au is the most commonly used contact metal because of its higher work function *i.e.* 5.1 eV, which enhances the injection of holes into active semiconductor layer. The material for gate electrode should have several characteristics, like good patterning capabilities, and good adhesion with substrate and gate dielectric. The gate metal work function should be comparable to active semiconductor layer to attain low threshold voltage. Such materials include heavily doped silicon, Al, Au, indium tin oxide (ITO) [12, 22].

2.3.3 Gate Dielectric

Nowadays, researchers are working to enhance the properties of dielectric materials, since, they are extremely crucial to achieve reliable and high-performance OTFTs.

Apart from compatibility with organic semiconductors, the dielectric materials should exhibit high insulation. High resistivity reduces the interface trap density between OSC and gate dielectric, which, in turn, prevents leakage between gate metal and OSC [14]. Further, the dielectric material should have high dielectric constant to ensure enough capacitance for channel current flow. This lowers the required drive voltage for a given channel current [12, 17]. Moreover, the thickness of gate dielectric layer is kept low to not only operate at low voltages, but also to reduce short channel effects in submicron devices. Besides this, the dielectric materials should have ability to form thin, pinhole-free films with high breakdown voltages and long-term stability [7-10]. These requirements are met by various dielectric materials such as SiO_2 , Al_2O_3 , polyvinyl phenol (PVP), propylene, poly (4-vinylphenol) (P4VP), poly-methyl methacrylate (PMMA) [16], barium zirconate titanate (BZT) and polyvinylidene fluoride (PVDF).

2.3.4 Substrate

Most OTFTs have been fabricated on heavily doped silicon substrate due to availability of good dielectric in the form of SiO_2 [16, 17]. Si is often used in electronics not only because of the intrinsic properties but also because of nearly ideal interface it forms with its thermally grown oxide [25]. Novel substrate materials on which working OTFTs have been fabricated include polyethylene naphthalate (PEN), poly-ethyleneterephthalate (PET), polyimide, and polyethylene, glass, plastic, paper and fibers [4-6]. These organic substrate materials have shown their remarkable performance in term of flexibility which motivates researchers for its utilization in design of flexible electronic devices and circuits.

2.4 OTFTs Fabrication

Performance of an OTFT strongly depends on the way it is fabricated. There are various fabrication approaches for OTFTs using different materials for substrates, electrodes, active semiconducting layers and gate insulators [8]. The devices are prepared either on top of highly n-doped silicon (Si) substrate or on a plastic foil. This section emphasizes on fabrication of bottom gate OTFTs with top and bottom contacts. Top and bottom contact are generally fabricated on a highly doped n-Si wafer, which acts as substrate as well as gate electrode [12]. A thermally grown SiO_2 on the Si substrate acts as gate insulator for both top and bottom contact devices. First step is thermal oxidation of Si that exposes wafer to an oxidizing environment of oxygen at an elevated temperature in quartz tube to grow SiO_2 . Thereafter, substrates are cleaned by RCA (Radio Co-operation of America) method, followed by photolithography and lift-off process to pattern S and D electrodes [12]. Subsequently, pentacene is deposited as active semiconductor layer using spin coating technique. Finally, source, drain and gate contacts are made from gold. Major OTFT's fabrication process steps are depicted in Fig. 4. After fabrication of complete OTFT device, various characterization methods are applied to extract its electrical properties and parameters [18, 22].

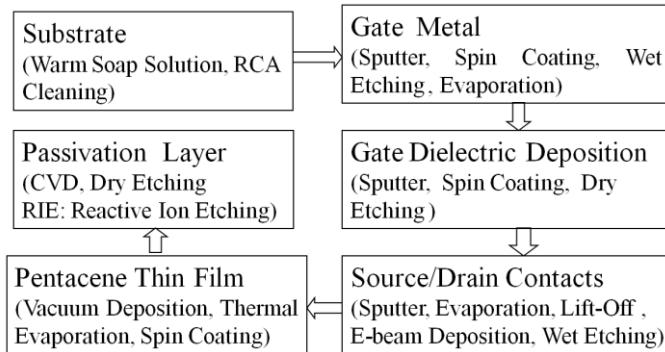


Fig. 4. Major steps used in fabrication of organic thin film transistors

3 OTFT Models and Operating Principle

Analytical models are often incorporated in simulations to forecast and optimize performance of organic integrated circuits. These models should be essentially upgradable, reducible, easily implementable and modifiable [13]. The model should also simultaneously allow for separation of characterization techniques. Researchers have recently proposed some analytical models which are discussed in following subsection along with their operating principle.

3.1 OTFTs Compact DC Models

The electrical characteristics of OTFTs are mostly similar to those of a-Si: H TFTs [24]. Some of the existing models of a-Si: H TFTs have been modified in order to represent the characteristics of OTFTs [6]. These OTFT models are referred to as dc compact models and behave similar to crystalline field-effect transistors (c-FETs) and a-Si: H TFTs [13]. Several models and features have been proposed to reflect one or the other specific charge transport mechanism in OTFTs. However, dc modeling involves the challenge of taking into account the variations in measurement results [13]. The electrical current-voltage characteristics of OTFTs are often incorrectly modeled similar to MOSFETs until the first dc compact model of OTFT was proposed by Xie *et al.* in 1992 [23]. This compact model for OTFT included both contact and leakage resistances. A generic analytical dc model, proposed by Marinov *et al.* [13], claims to be fully symmetrical, and valid for all regimes of TFT operation, *i.e.*, linear and saturation above threshold, subthreshold, and reverse biasing.

Although researchers have made serious efforts to describe the physical characteristics and properties of OTFTs by introducing analytical compact models [13, 15], but due to incomplete information about nature of carrier transport mechanism in organic semiconductors, these rare models are still far from representing OTFT compact dc model in usual circuit analysis. Furthermore, these

models are not fully suitable due to their incompleteness and/or their complexity, which leads to convergence issues and long computation time.

3.2 Operating Principle

The operating principle of OTFT in this paper is explained using the schematic diagram of BGBC structure, as shown in Fig. 5 (a). The device consists of pentacene organic semiconductor material with source/drain and gate electrodes made of Au and Al, respectively. The Fermi level of gold and the HOMO: LUMO (Highest Occupied Molecular Orbital: Lowest Unoccupied Molecular Orbital) levels of pentacene are shown in Fig. 5 (b).

Fundamentally, an OTFT operates like a capacitor. When voltage is applied between gate and source (V_{GS}), charge carriers get accumulated at insulator-OSC interface. For instance, when positive voltage is applied to the gate, negative charges are induced at the interface. However, the electron injection is poor in this scenario because the Fermi level of gold is far away from the LUMO level. Accordingly, no current passes through the channel. However, a small current exists, essentially, due to leakages through the insulating layer.

On the other hand, when negative gate voltage is applied, holes get injected from source to semiconductor. This charge forms conducting channel, since Fermi level of gold is now closer to HOMO level of pentacene, as depicted in Fig. 5 (b). Threshold voltage V_T describes activation potential for channel formation. On applying drain-source voltage (V_{DS}), current I_{DS} flow from source to drain.

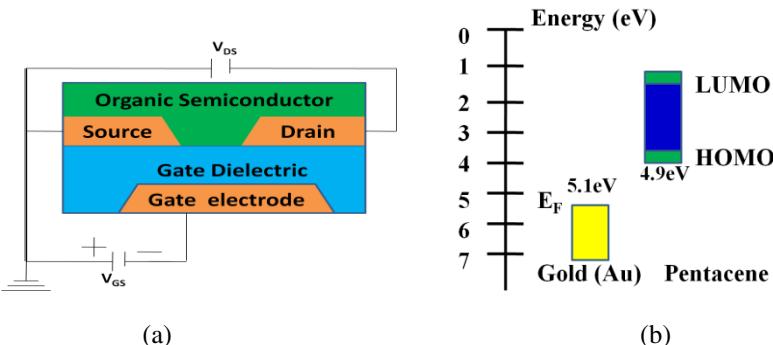


Fig. 5. (a) Schematic of BGBC structure operation, pentacene as active semiconductor layer, with S/D gold contact electrodes. (b) Energy band level of gold and pentacene interface.

Apart from voltages, critical device dimensional parameters that affect current flow include channel length L , width W , thickness of dielectric t_{ox} and organic semiconductor film thickness t_{osc} . Due to majority of holes in the conducting channel, pentacene is said to be a *p*-type semiconductor. Similarly, an organic semiconductor will be considered

n-type when S and D electrodes can inject electrons in its LUMO level. However, it is necessary to point out that this concept differs from that of doping in conventional semiconductors, which can be made either *n*-type or *p*-type.

4 Performance Analysis of OTFT

In general, OTFT performance is analyzed through its output and transfer characteristics. These characteristics are used to extract various performance parameters such as field effect mobility, on/off current ratio, threshold voltage, subthreshold slope and transconductance.

4.1 Characteristics of Different OTFT Structures

This section analyzes the functioning of various OTFT structures. For carrying out the analysis the parameter values for channel length (L), channel width (W), aluminum oxide (Al_2O_3) insulator thickness (t_{ox}), source/drain length (t_S/t_D), and pentacene active layer thickness (t_P) are taken as 10 μm , 100 μm , 5.7nm, 20nm and 30nm, respectively [7]. The gate electrode is considered to be made of aluminum with 20nm thickness. The properties of pentacene organic semiconductor material used in simulation include 2.2eV energy gap, 2.8eV electron affinity, electron density of state of 2.0×10^{21} per cm^3 in valance band, 1.7×10^{21} per cm^3 in conduction band and permittivity of 4.0 [12]. Top and bottom gate analyses have been performed for bottom and top contact OTFT structures using Silvaco ATLAS two-dimensional numerical device simulator setup. Out of these four structures, two structures are referred as top gate and the other two as bottom gate, as shown in Fig. 2 and Fig. 3, respectively. The mobility in OTFT structure is described using Poole-Frenkel model given by

$$\mu(E) = \mu_0 \exp \left[-\frac{\Delta}{kT} + \left(\frac{\beta}{kT} - \gamma \right) \sqrt{E} \right] \quad (1)$$

where $\mu(E)$ is the field dependent mobility, μ_0 is the zero field mobility, E is the electric field, Δ is the zero field activation energy, β is the electron Pool-Frenkel factor, γ is the fitting parameter, k is the Boltzmann constant and T is the temperature. Poole-Frenkel conduction is due to field enhanced thermal excitation of trapped charge carriers. The drain current reduces in the low field region, due to charge carriers being localized around the traps. This implies that the device drain current can be appreciably enhanced through release of these trapped charge carriers. Increase in the electric field is one of the ways to ensure such desirable phenomenon.

Various OTFT structures are simulated using organic TFT display module. The transistor sizes for all the proposed structures are kept same, for valid comparison among them. Simulation uses proper boundary condition and device physics to analyze the OTFT structures. The resulting output and transfer characteristic plots for TGTC, TGBC, BGBC and BGTC OTFTs structures are shown in Fig. 6 to Fig. 9, respectively.

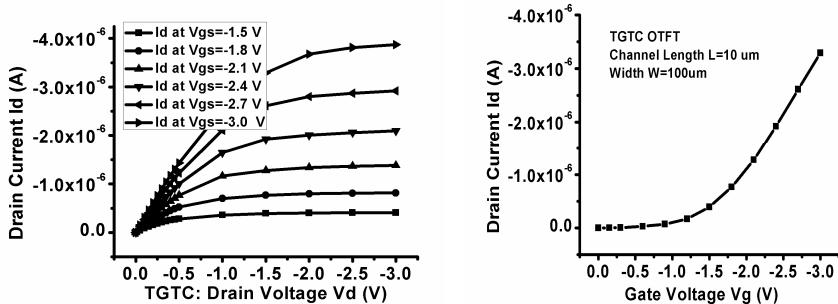


Fig. 6. Output and Transfer (at $V_{ds} = -1.5$ V) characteristics of TGTC OTFT

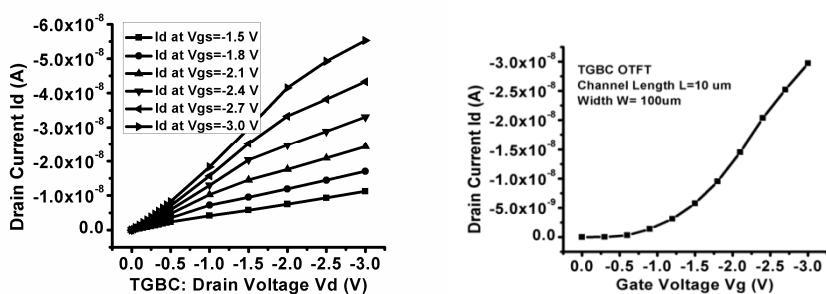


Fig. 7. Output and Transfer (at $V_{ds} = -1.5$ V) characteristics of TGBC OTFT

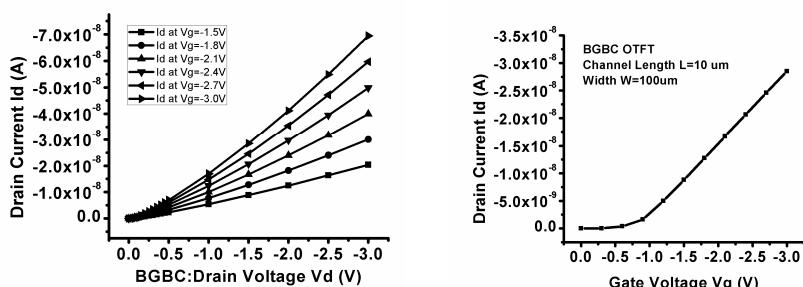


Fig. 8. Output and Transfer (at $V_{ds} = -1.5$ V) characteristics of BGBC OTFT

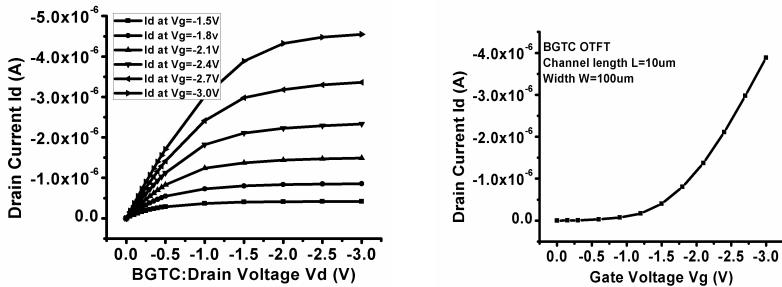


Fig. 9. Output and Transfer (at $V_{ds} = -1.5$ V) characteristics of BGTC OTFT

4.2 Performance Parameters

The performance parameters such as drive current, mobility, threshold voltage, subthreshold slope, transconductance, on/off current ratio are extracted from characteristic plots of different OTFT structures. The resulting parameter values are presented in Table 1.

Table 1. Extracted parameters for various OTFT structures

Parameters	TGBC	TGTC	BGTC	BGBC
Mobility (μ) $\text{cm}^2/\text{V.s}$	0.0016	0.2081	0.2512	0.0029
Threshold Voltage (V_T) V	-0.86	-0.97	-1.06	-0.63
On/Off Current Ratio (I_{ON}/I_{OFF})	10^6	10^5	10^5	10^6
Subthreshold Slope (SS) V/decade	0.163	0.119	0.130	0.147
Transconductance (g_m) S/m	0.019	2.30	3.00	0.013
Drive Current (I_{DSmax}) μA at $V_{GS} = V_{DS} = -3\text{V}$	-0.055	-2.32	-4.60	-0.069

It is observed that OTFT structures with top contact demonstrate superior performance as compared to those with bottom contact. These improvements in performance are achieved in terms of mobility, transconductance and maximum drain current. Higher injection area and reduced drain and source contact resistances are main reasons for better behavior of top contact based OTFT architectures. Among all OTFT structures, BGTC structure exhibit best operating functionality in terms of all the afore-mentioned parameters.

4.2.1 Mobility

Mobility of organic semiconductor layer is receiving significant attention in order to improve transistor characteristics. In fact, reliable operation of transistor requires larger mobility [8, 22]. Earlier, OTFTs exhibited poor performance due to their low field effect mobility. This motivated most of the research efforts to improve mobility through changing processing conditions. Consequently, mobility of *p*-type OTFTs has

already exceeded $1\text{cm}^2/\text{V.s}$ for pentacene and $0.1\text{cm}^2/\text{V.s}$ for poly-3-hexylthiophene (P3HT) [15]. Pentacene is the most widely used OSC for organic transistors due to its higher hole mobility. However, its electron mobility is substantially lower as compared to hole mobility. As a result, a significant effort is being devoted to improve its electron mobility [22], so as to realize complementary OTFT logic circuits using pentacene as active material.

The analysis in section 4.1 shows that a larger mobility is obtained for top contact rather than the bottom one. The results, tabulated in Table 1, indicate that TGBC and TGTC have much lower mobility than BGBC and BGTC structures, respectively. Moreover, simulation results indicate that device structure by itself is not the only factor that affects mobility devices. Other factors, which may be responsible for their performance, depend on the device fabrication process and material properties.

4.2.2 Threshold Voltage (V_T)

Threshold voltage is the minimum gate voltage at which an OTFT begins to conduct. In other words, it is the minimum gate voltage at which accumulation of holes takes place at the OTFT insulator-semiconductor interface. As indicated in table 1, threshold voltage (V_T) of TGBC, TGTC, BGTC and BGBC are -0.86V , -0.97V , -1.06V , and -0.63V , respectively, at gate and drain voltages of -3V each. Hence, the analysis reveals that bottom contacts show lower threshold voltages as compared to top contacts. Threshold voltage depends on dielectric capacitance and, hence, thickness of insulator layer. Dielectric thickness is smaller for bottom contact in comparison to top contact for constant dimensions of each layer of OTFT. This leads to lower threshold voltages for bottom contact structures.

4.2.3 On/Off Current Ratio (I_{ON}/I_{OFF})

I_{ON}/I_{OFF} is ratio of the current in the accumulation mode to the current in the depletion mode. I_{ON} is drain current above threshold voltage at which saturation takes place and I_{OFF} is the drain current below threshold voltage. The on/off current ratio of OTFTs are higher for thinner semiconducting layer. I_{ON}/I_{OFF} should be more than 10^6 for memory and display devices [8]. Short channel devices have higher on/off current ratio than long channel devices [22]. On/Off current ratio is evaluated from transfer characteristics of various OTFT structures.

The extracted on/off current ratios for TGBC, TGTC, BGTC and BGBC structures are 10^6 , 10^5 , 10^5 and 10^6 , respectively. This shows that bottom contact devices exhibit higher on/off current ratio. The primary reason behind this is that the device exhibits negligible current at zero drive voltage. This also means that leakage current is extremely small for bottom contact devices thus making them suitable for memory and display applications.

4.2.4 Sub-threshold Slope (SS)

Sub-threshold slope is ratio of change in gate voltage to the change in drain current on log scale at constant drain voltage. Steeper curve correlates to better switching behavior of device.

$$SS = \frac{\partial V_G}{\partial \log_{10}(I_{DS})} \quad (2)$$

Sub-threshold slope is evaluated using equation (2) from $\log_{10}(I_{DS})$ versus V_{GS} curve. It is an important parameter that allows efficient usage of transistor as a switch. Sub threshold slope is a measure of switching behavior of OTFT. Table 1 indicates that for given dimensions the channel formation between source and drain contacts depends on the biasing voltages and, hence, switching behavior of a device. The simulated results of subthreshold slopes for TGBC, TGTC, BGTC and BGBC structures are 0.163, 0.119, 0.130, and 0.147V/decade, respectively. Therefore, bottom contact OTFTs demonstrates higher sub-threshold slopes as compared to top contacts with same gate structures. This can be attributed to different carrier injection density from the source electrode into the OSC thin film.

5 Limitations

Almost all electronic devices used in daily life are based on inorganic semiconductors, like, silicon, gallium arsenide, gallium nitride, indium phosphate, silicon-germanium, etc. Inorganic semiconductor based devices are dominant in electronic market because of their high operating speed, environmentally stable and everlasting performance. For instance, silicon based integrated circuits have been integral part of satellite communication systems for over last 40 years.

On the other hand, significant progress in organic semiconductors technology has provided designers with an alternative to inorganic materials. However, still numbers of challenges have to be met out in order to make organic material based devices, practically viable. Currently, organic devices show various constraints at fabrication, material and device physics level. Firstly, the organic semiconductor based devices possess complex structure. Besides this, they degrade with time and can deform easily. It is well-known that the characteristics of organic materials change with environmental conditions after long duration. Therefore, stability of these devices has to be worked out and modeled properly to better comprehend the process of degradation. Researchers agree that most of the instability comes from the chemical structure of the compound and, hence, are trying to find ways to synthesize more stable organic compounds.

Furthermore, OTFT current-voltage characteristics degrade at higher temperatures. In addition, the noise level increases considerably at low frequencies. Academia and industry throughout globe intend to develop organic semiconductor with high mobility and fast switching time. Rapid development of organic electronics with improvements in compatibility of organic transistors and micro-fluidics opens wide horizons for use of OTFTs in compact sensing systems or biochips. However, several issues are still open, particularly, those related to long-term stability and batch-to-batch, roll-to-roll or even device-to-device variability that will determine whether this technology will move beyond the laboratory stage. Currently, OTFTs are not suitable for very high switching speed applications due to low mobility. Consequently, high voltage and reverse recovery time are required to drive OTFTs devices.

6 Conclusions

Low cost and considerably lower temperature fabrication of flexible organic transistors make them suitable candidates for several futuristic low-cost, flexible devices and circuits. This has attracted researchers from all over the globe to the exciting domain of organic electronics. Consistent research work in organic devices has brought about steady improvement in the transistor's performance.

Many technical issues related to OTFT structures, materials, compact dc models, operating principle, fabrication and performance parameters have been discussed. Comparative analyses of various OTFT structures, *viz.* TGBC, TGTC, BGTC and BGBC, have been done based on simulated results. OTFT performance has been characterized in terms of field effect mobility, drive current, on/off current ratio, threshold voltage, sub-threshold slope and transconductance.

Observations reveal that OTFT structures with top contact exhibit superior performance as compared to those with bottom contact in terms of mobility, transconductance and maximum drive current. Moreover, BGTC structure exhibit best operating functionality among all OTFT structures. The analysis further indicates that not only device structure, but also, material properties and fabrication process affects device performance, particularly, operating speed and signal sensitivity.

References

1. Rogers, J.A., Bao, Z., Katz, H.E., Dodabalapur, A.: Thin-Film Transistors, 377 (2003)
2. Dimitrakopoulos, C.D., Malenfant, P.R.L.: Organic Thin Film Transistors for Large Area Electronics. *Adv. Mater.* 14, 99–117 (2002)
3. Klauk, H., Halik, M., Zschieschang, U., Eder, F., Schmid, G.: Pentacene Organic Transistors and Ring Oscillators on Glass and on Flexible Polymeric Substrates. *Appl. Phys. Lett.* 82(23), 4175 (2003)
4. Moore, S.K.: Just One Word: Plastics. *IEEE Spectrum* 39(9), 55 (2002)
5. Kim, Y.H., Moon, D.G., Han, J.I.: Organic TFT Array on a Paper Substrate. *IEEE Electron Dev. Lett.* 25, 702 (2004)
6. Baude, P.F., Enter, D.A., Haase, M.A., Kelley, T.W., Muyres, D.V., Thesis, S.D.: Pentacene-Based Radio-Frequency Identification Circuitry. *Appl. Phys. Lett.* 82(22), 3964 (2003)
7. Klauk, H.: Organic Thin Film Transistor. *Chem. Soc. Rev.* 39, 2643–2666 (2010)
8. Mittal, P., Kumar, B., Kaushik, B.K., Negi, Y.S.: Organic Thin Film Transistor Architecture, Parameters and their Applications. In: Proc. IEEE Int. Conf. on Communication Systems and Network Technologies (CSNT 2011), Katra, pp. 436–440 (2011)
9. Klauk, H., Halik, M., Zschieschang, U., Schmid, G., Radik, W.: High-Mobility Polymer Gate Dielectric Pentacene Thin Film Transistors. *J. Appl. Phys.* 92(9), 5259–5263 (2002)
10. Horowitz, G.: Organic Thin Film Transistors: From Theory to Real Devices. *J. Mater. Res.* 19(7), 1946–1962 (2004)
11. Kymissis, I., Dimitrakopoulos, C.D., Puroshothoman, S.: High Performance Bottom Electrode Organic Thin Film Transistors. *IEEE Trans. Electron Devices* 48, 1060 (2001)

12. Gupta, D., Katiyar, M., Gupta, D.: An Analysis of the Difference in Behavior of Top and Bottom Contact Organic Thin Film Transistors using Device Simulation. *Organic Electronics* 10, 775–784 (2009)
13. Marinov, O., Deen, M.J., Zschieschang, U., Klauk, H.: Organic Thin Film Transistors: Part I. Compact DC Modeling. *IEEE Trans. Electron Devices* 56(12), 2952–2961 (2009)
14. Chander Shekar, B., Lee, T., Rhee, S.W.: Organic Thin Film Transistors, Material, Processes and Devices. *Korean J. Chem. Engg.* 21(1), 267–287 (2004)
15. Deen, M.J., Kazemeini, M.H., Holdcroft, S.: Contact Effects and Extraction of Intrinsic Parameters in Poly (3-Alkylthiophene) P3AT Thin-Film Field-Effect Transistors. *J. Appl. Phys.* 103(12), 124509–124516 (2008)
16. Puuigdollers, J., Voz, C., Fonrodona, M., Martin, I., Orpella, A., Vetter, M., Alcubilla, R.: Flexible Pentacene/PMMA Thin Film Transistors Fabricated on Aluminium Foil Substrates. In: Materials Research Society Symposium Proceedings, vol. 871, pp. 323–328 (2005)
17. Puuigdollers, J., Voz, C., Orpella, A., Quidant, R., Martin, I., Alcubilla, R.: Pantacene Thin Film Transistors with Polymeric Gate Dielectric. *Organic Electronics* 5, 67–71 (2004)
18. Maeda, T., Kato, H., Haruo Kawakami, A.: Organic Field Effect Transistors with Reduced Contact Resistance. *Appl. Phys. Lett.* 89, 123508 (2006)
19. Bettinger, J., Bao, Z.: Organic Thin-Film Transistors Fabricated on Resorbable Biomaterial Substrates. *Adv. Mater.* 22, 651–655 (2010)
20. Marinov, O., Deen, M.J., Datars, R.: Compact Modeling of Charge Mobility in Organic Thin-Film Transistors. *J. Appl. Phys.* 106(6), 064501–064501-13 (2009)
21. Street, R.A., Salleo, A.: Contact Effects in Polymer Transistors. *Appl. Phys. Lett.* 81(15), 2887 (2002)
22. Kumar, B., Kaushik, B.K., Negi, Y.S., Mittal, P.: Characteristics and Applications of Polymeric Thin Film Transistor: Prospects and Challenges. In: Proc. IEEE Int. Conf. on Electrical and Computer Technology (ICETECT 2011), Nagercoil, pp. 23–24 (2011)
23. Xie, Z., Abdou, M., Lu, A., Deen, M.J., Holdcroft, S.: Electrical Characteristics of Poly (3-Hexylthiophene) Thin Film MISFETs. *Canadian J. of Physics* 70(10/11), 1171–1177 (1992)
24. Marinov, O., Deen, M.J., Iniguez, B.: Charge Transport on Organic and Polymer Thin Film Transistors: Recent issues. In: Proc. Int. Elect. Eng. Circuit Devices Syst., vol. 152(3), pp. 189–209 (2005)
25. Klauk, H.: *Organic Electronics: Materials, Manufacturing and Applications*. Wiley-VCH Verlag GmbH & Co., KGaA (2006)

Active Learning with Bagging for NLP Tasks

Ruy Luiz Milidiú, Daniel Schwabe, and Eduardo Motta

Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro
Rua Marquês de São Vicente, 225
Rio de Janeiro, Brazil, 22451-900
{milidiu,dschwabe,emotta}@inf.puc-rio.br

Abstract. Supervised classifiers are limited by the annotated corpora available. Active learning is a way to circumvent this bottleneck, reducing the number of annotated examples required. In this paper, we analyze the benefits of active learning combined with bagging applied to Quotation Start, Noun Phrase Chunking and Text Chunking tasks. We employ query-by-committee as query strategy to actively select examples to be annotated. By using these techniques, we achieve reductions up to 62.50% on the annotation effort depending on the task to obtain the same quality as in passive supervised learning.

1 Introduction

Active learning has been successfully applied to several natural language processing tasks like part-of-speech-tagging, named entity recognition, word sense disambiguation, among others. On the other hand, Olsson reports that active learning is not widely spread among the NLP community [8]. To the best of our knowledge quotation start, noun phrase and text chunking tasks in general are not yet approached using active learning techniques. Thus, in this work we explore active learning combined with bagging [1] to solve these tasks. Active learning can significantly reduce annotation effort for these tasks. In table 1, we show the effort reduction obtained for these tasks.

Table 1. Annotation effort reduction

Task	Effort Reduction
quote start	62.50%
noun phrase chunking	31.25%
NP+VP chunking	6.25%
NP+VP+PP chunking	18.75%
NP+VP+PP+ADJP+ADVP chunking	18.75%

The remainder of this work is organized as follows. In Section 2, we discuss active learning aspects. In section 3 we describe the natural language processing tasks we tackle. In Section 4, we present the characteristics of corpora used. In Section 5, we describe the experiments we perform and the corresponding results. Finally, in Section 6, we present our conclusions.

2 Active Learning

In active learning, instead of randomly getting examples from an annotated corpus, we actively select examples to be annotated by an oracle, typically a human domain expert. Example selection is driven by informativeness and representativeness in order to maximize the benefit of including the new annotated examples.

In this work, we use vote entropy [2] as query strategy. Vote entropy is calculated for a sentence S as

$$VE(S) = -\frac{1}{|S|} \sum_{t \in S} \sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C}. \quad (1)$$

where y_i is each possible label for a token, $V(y_i)$ is the number of votes a label receives, C is the number of members in the committee and $|S|$ is the sentence length in tokens. We use the average vote entropy of tokens, since all tasks are based on token classification and the examples correspond to whole sentences. After evaluating $VE(S)$ for each unlabeled sentence, the sentences that have the highest vote entropy are selected and submitted to the oracle. Then the annotated sentences are included in the model during the next active learning iteration.

3 Tasks

3.1 Quotation Start

Quotation start identification is a subtask of identifying quotations from a text and associating them to their authors. Quotation start is a binary token classification that tags each token as either a quotation start or not. In this task, the classes are highly unbalanced, since 99.5% of tokens are not quotation starts.

3.2 Chunking

Text chunking belongs to a broader concept called shallow parsing, which includes other tasks whose objective is to recover only a limited amount of information from natural language sentences [6].

Chunking is defined as dividing a text into “phrases” in such a way that syntactically related words become members of the same phrase. These phrases are non-overlapping, which means that one word can only be a member of one of them at most [9].

Shallow parsing in general, and the text chunking task in particular, are considered relevant. Not all applications require a complete syntactic analysis, and often a full parse provides more information than needed. One example is Information Retrieval, for which it may be enough to find simple noun phrases and verb phrases [6]. Text chunking usually provides enough syntactic information for several such applications.

In this work, text chunking task has four variations, depending on the chunk types addressed. The first one includes only Noun Phrase (NP). The second task adds Verb

Phrase (VP). The third adds Prepositional Phrases (PP) and the last also considers Adverbial Phrases (ADVP) and Adjectival Phrases (ADJP).

4 Corpora

For *Quotation Start* task we use GLOBO.COM corpus, with golden annotation of named entities, coreferences, quotations and associations between quotations and authors. This corpus is composed by Brazilian Portuguese news from the GLOBO.COM portal, dated from August, 2007 to August, 2008 and contains 685 news [3].

For the *Noun Phrase Chunking* task we use SNR-CLIC, a Portuguese noun phrase chunking corpus [4].

For the other *Text Chunking* tasks we use the Bosque corpus, which is part of the Floresta Sintá(c)tica Project [5]. It consists in a syntactic treebank of European and Brazilian Portuguese texts.

Table 2 summarizes the corpora statistics.

Table 2. Corpora statistics

Task	Classes	Training		Test	
		Sentences	Tokens	Sentences	Tokens
quotation start	2	1,104	174,415	266	41,613
noun phrase chunking	3	3,504	83,086	878	20,798
NP+VP chunking	5	1,200	33,213	1,405	34,169
NP+VP+PP chunking	7	1,200	33,213	1,405	34,169
NP+VP+PP+ADVP+ADJP chunking	11	1,200	33,213	1,405	34,169

5 Experiments

For each task we perform two experiments. The first one is the baseline passive supervised learning, where the examples are randomly selected from the available pool of unlabeled instances. The second experiment is active supervised learning where examples are actively selected from the pool of unlabeled instances using vote entropy as a measure of disagreement among committee members.

Every experiment starts with a set of 25% examples randomly selected from the training corpus to train the initial classifier. During each active learning iteration a pool of unlabeled examples to be annotated by the oracle is selected and the process repeats until there is no more unlabeled examples available. All the remainder examples are evaluated using the current classifier and ranked according to vote entropy.

All classifiers are based on Entropy Guided Transformation Learning algorithm [7] and committees have 11 members.

In Figures 1 to 5 we draw the F1-measure as a function of the number of annotated examples for passive and active with vote entropy experiments for the five tasks.

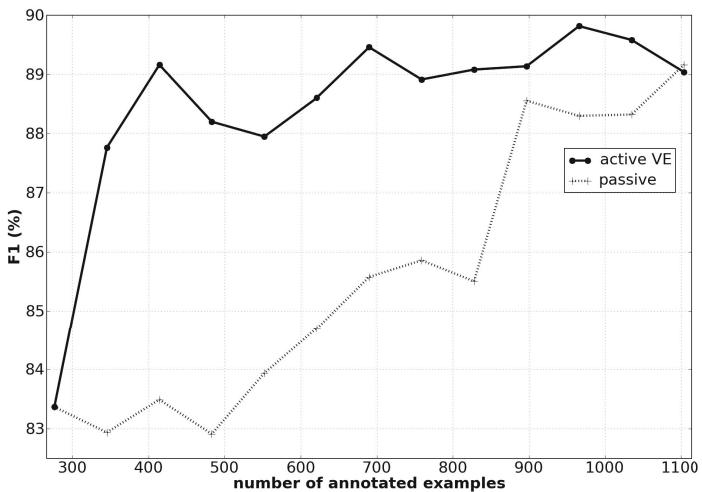


Fig. 1. Comparison between passive and active learning for quote start task

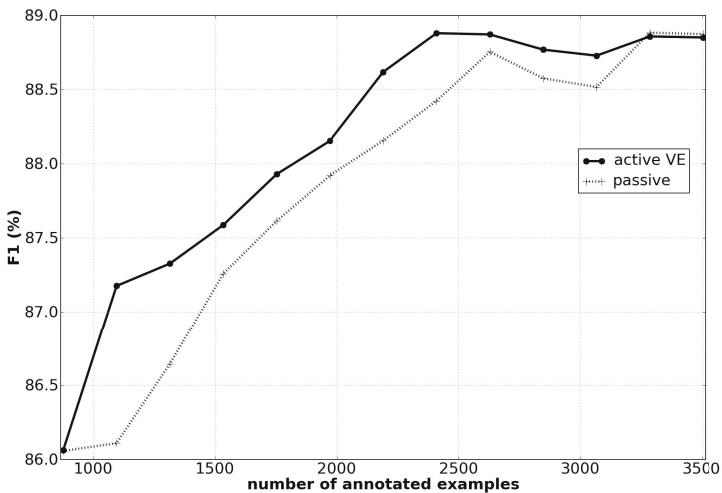


Fig. 2. Comparison between passive and active learning for noun phrase chunking task

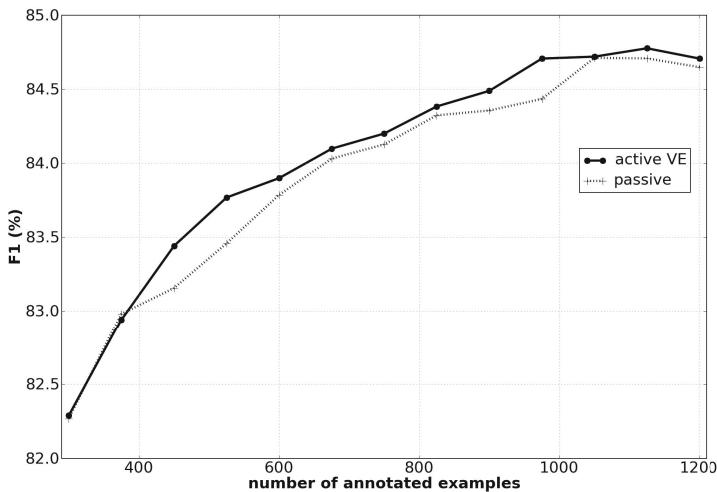


Fig. 3. Comparison between passive and active learning for NP+VP chunking task

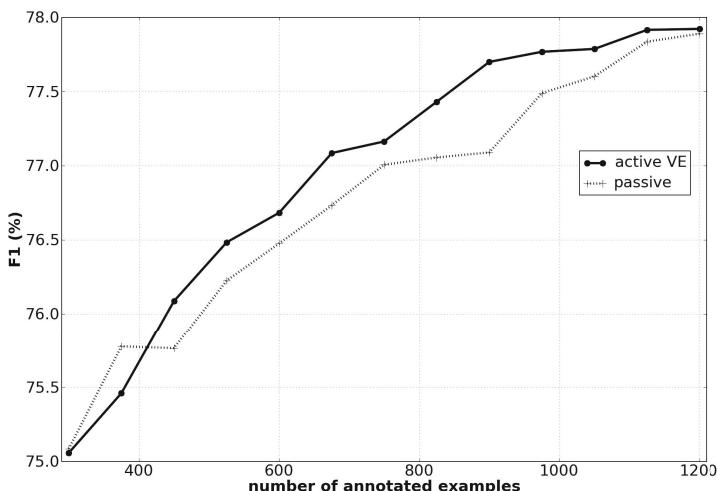


Fig. 4. Comparison between passive and active learning for NP+VP+PP chunking task

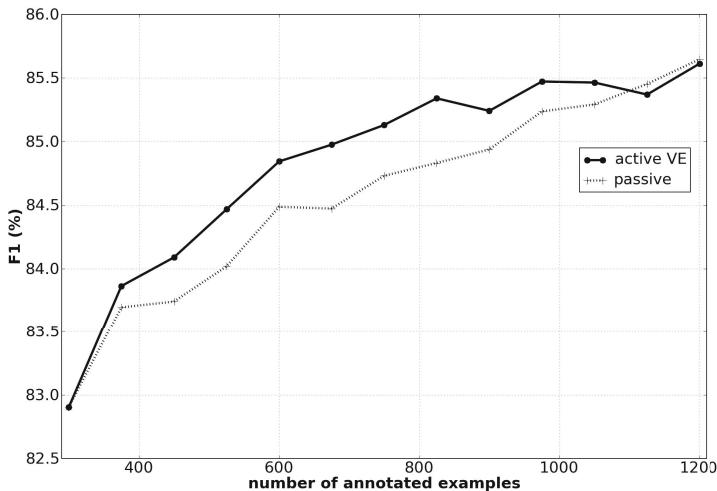


Fig. 5. Comparison between passive and active learning for NP+VP+PP+ADJP+ADVP chunking task

6 Conclusions

In this work we apply active learning techniques combined with bagging to a set of natural language processing tasks. We observe a significant reduction in the amount of annotated examples needed to achieve a given classification quality, ranging from 6.25% for the VP+NP text chunking task to 62.50% for quotation start task.

References

1. Abe, N., Mamitsuka, H.: Query learning strategies using boosting and bagging. In: Proceedings of the Fifteenth International Conference on Machine Learning, ICML 1998, pp. 1–9. Morgan Kaufmann Publishers Inc., San Francisco (1998), <http://dl.acm.org/citation.cfm?id=645527.657478>
2. Dagan, I., Engelson, S.P.: Committee-based sampling for training probabilistic classifiers. In: ICML 1995, pp. 150–157 (1995)
3. Fernandes, W.P.D., Motta, E., Milidiú, R.L.: Quotation extraction for portuguese. In: Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL 2011), Cuiabá, pp. 204–208 (2011)
4. Freitas, M.C., Garrao, M., Oliveira, C., dos Santos, C.N., Silveira, M.: A anotação de um corpus para o aprendizado supervisionado de um modelo de sn. In: Proceedings of the III TIL/XXV Congresso da SBC, São Leopoldo - RS - Brasil (2005)
5. Freitas, C., Rocha, P., Bick, E.: Floresta Sintá(c)tica: Bigger, Thicker and Easier. In: Teixeira, A., de Lima, V.L.S., de Oliveira, L.C., Quaresma, P. (eds.) PROPOR 2008. LNCS (LNAI), vol. 5190, pp. 216–219. Springer, Heidelberg (2008)

6. Hammerton, J.: Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing. *Journal of Machine Learning Research* 19(2), 313–558 (2002), doi:10.1162/153244302320884533
7. Milidiú, R.L., Santos, C.N., Duarte, J.C.: Phrase chunking using entropy guided transformation. In: Proc. of ACL 2008: HLT, pp. 647–655 (2008)
8. Olsson, F.: A literature survey of active machine learning in the context of natural language processing. Tech. Rep. 06, Box 1263, SE-164 29 Kista, Sweden(2009), <http://soda.swedish-ict.se/3600/1/SICCS-T2009-06--SE.pdf>
9. Sang, E.F.T.K., Buchholz, S.: Introduction to the conll-2000 shared task: Chunking. In: Proceedings of CoNLL 2000 and LLL 2000, Lisbon, Portugal, pp. 127–132 (2000)

Mining Queries for Constructing Materialized Views in a Data Warehouse

T.V. Vijay Kumar¹, Archana Singh², and Gaurav Dubey³

¹ School of Computer and Systems Sciences, Jawaharlal Nehru University,
New Delhi-110067, India

² Amity Institute of Information Technology, Amity Campus,
Sector 125, Noida, UP-201301, India

³ Amity School of Computer Sciences, Amity Campus,
Sector 44, Noida, UP-201303, India

Abstract. A data warehouse stores historical information, continuously being generated over time, to support decision making. The queries posed for decision making are usually exploratory, long, complex and analytical in nature. These queries, when posed against a large and continuously growing data warehouse, consume a lot of time for processing and thereby resulting in high response times. This problem of high response time can be addressed by constructing materialized views on the data warehouse. These views, which store data along with its definition, cannot be arbitrarily constructed as they need to contain relevant and required information for answering most future queries. The approach proposed in this paper attempts to identify such information, from previously posed queries on a data warehouse, using clustering and association rule mining techniques. The information identified using the approach is likely to answer most future queries in reduced query response times. As a result, the decision making would become more efficient.

1 Introduction

A large amount of data that is continuously being generated by disparate data sources spread across the globe. Various organizations are aiming to exploit this data for decision making purposes in order to gain a competitive edge vis-à-vis other organizations. This data can be accessed using two approaches namely, the Lazy (on-demand) approach or, the eager (in-advance) approach [28]. In the former approach, the data is accumulated based on the user query. The latter approach enables the data to be accumulated in advance and queries are posed against this accumulated data. Data warehousing is based on the latter approach [28].

Data warehousing is concerned with the extraction of relevant data from the disparate data sources, its transformation and subsequent loading into a central repository, called a data warehouse [10]. The resultant data warehouse contains subject oriented, integrated, time variant and non-volatile information meant to support decision making [10]. Decision making queries posed on a data warehouse are usually analytical in nature. These

queries are long, complex and exploratory and, when posed against a large data warehouse, consume a lot of time for processing. As a result, the query response time is high. One way to address this problem is by materializing views in a data warehouse with the aim of improving the query response time. These materialized views, unlike virtual views, store relevant and required data separately from the data warehouse. They are significantly smaller in size, when compared with the data warehouse, and constructed with the purpose of providing answers to most future queries in improved query response times. This requires that materialized views contain relevant data that provide answers to future queries. This problem of selecting materialized views with relevant data is referred to as the view selection problem [5]. View selection deals with selecting, or constructing materialized views, which improve the query response time while conforming to resource constraints like storage space, memory etc [5, 6].

One way to construct materialized views is by materializing all possible views. This may not be feasible as the number of views grows exponentially with the number of dimensions thus leading to exceeding the available space for materialization. Also, optimal selection of subsets of views, from amongst all possible views, is shown to be an NP-Complete problem [8]. Alternatively, views can be selected heuristically or empirically. Heuristically views are selected using greedy approaches [6, 7, 8, 19, 20, 21, 24, 25, 27] or evolutionary approaches [9, 11].

In empirical based selection, past query patterns are monitored and assessed on factors like frequency or size of data and this information is used to construct materialized views [12, 15]. This paper focuses on constructing materialized views empirically based on previously posed queries on the data warehouse. Most existing query based approaches [1, 3, 4, 14, 17, 29, 30] for constructing materialized views are workload driven, with the rationale that new queries are likely to be closer to the queries posed in the past and thus the materialized views constructed using them are likely to provide answers to future queries. In this paper, an approach is presented that selects queries, from among all the queries posed in the past, using clustering and association rule mining techniques. The selected queries contain frequently accessed information, which has high likelihood of providing answers to future queries. As a result, the query response time would be reduced and would thereby facilitate the decision making process.

The paper is organized as follows: The approach for selecting frequent queries is given in section 2 and an example based on it is given in section 3. Section 4 is the conclusion.

2 Approach

The approach aims to identify relevant information, from previously posed queries on the data warehouse, for constructing materialized views. This identification of relevant information is carried out in two phases namely Subject Area Identification and Frequent Query Selection. The approach is similar to the approach given in [22, 23, 26]. The two phases used in the approach are discussed next.

2.1 Subject Area Identification

It is usually observed that queries on a data warehouse are subject specific i.e. focused around a particular subject. Therefore, it becomes more appropriate to identify the various subject areas based on queries posed in the past. The queries that access similar data are likely to belong to the same subject and therefore should be part of the subject area. On the other hand, queries that access totally dissimilar data are likely to belong to different subject areas. This would necessitate the need to group closely related queries to form clusters. The approach uses density based clustering algorithm OPTICS(Ordering Points to identify Clustering Structure)[2] to group closely related queries. The closeness of the queries is computed using the similarity measure Overlap Coefficient [18], where similarity between a pair of queries Q_i and Q_j , i.e. $\text{Sim}(Q_i, Q_j)$ is given by

$$\text{Sim}(Q_i, Q_j) = \frac{|R(Q_i) \cap R(Q_j)|}{\text{Min}(|R(Q_i)|, |R(Q_j)|)}$$

where $R(Q_i)$ and $R(Q_j)$ are the relations accessed by queries Q_i and Q_j respectively.

The similarity between pairs of queries is computed and a similarity matrix is constructed using them. This matrix is then used to create clusters using algorithm OPTICS [2]. The subject area identification algorithm, based on OPTICS[2], is given in Fig. 1. This algorithm takes the previously posed queries, the similarity matrix,

```

ALGORITHM SubjectAreaIdentification
Inputs:   QP : Previously posed Queries queries,
            ε : Minimum query similarity threshold,
            SimMat : Similarity Matrix showing similarity between queries
            MinQ : Minimum Query Threshold
Output:  Cluster of Queries CQ
Method:
For each Q in QP
    IF Q is not processed
        Insert (Q, UNDEFINED) into QueryList
        WHILE QueryList not empty
            Select first Query (Q, rQsim) from QueryList
            Retrieve ε–neighborhood of Q, i.e. Nε(Q), using SimMat
            If |Nε(Q)| ≥ MinQ
                Set cQsim=csim(Q) //csim(Q) is the core similarity of query Q
                Set Q as processed
                Write (Q, rQsim, cQsim) to file
                If Q is core query at any similarity greater than equal to ε
                    For each Q' in Nε(Q) not yet processed
                        Determine rQsimQ' = rQsim(Q', Q)
                        If Q' is not in QueryList
                            Insert (Q', rQsimQ') in QueryList
                        Else If (Q', rQ'sim) is in QueryList and rQ'sim< >rQsimQ'
                            Update (Q', rQsimQ') in QueryList
                        End If
                    End For
                End If
            End While
        End If
    End For
CQ is the number of valleys in the query reachability similarity plot due to the QueryList

```

Fig. 1. Algorithm SubjectAreaIdentification based on OPTICS[2]

showing similarity between queries, and a minimum query similarity threshold as input and produces the cluster of queries as output. The algorithm is used to determine meaningful clusters of queries of varying density. These queries are ordered in a manner whereby queries that are similar, as per the minimum query similarity threshold, are neighbors in the ordering. In order for the queries to belong to the same cluster, reachability query similarity (rqsim), which represents the acceptable density of queries defined by minimum query threshold, is computed. A reachability query similarity graph, showing core query similarity (cqsim) versus the query neighbors, is plotted. These graphs show the formation of clusters. The number of valleys in the plot is used to identify the cluster of queries. These identified clusters specify the various subject areas. Each such subject area consists of queries that are similar with respect to the data accessed by them.

It is possible to have large numbers of queries in each subject area. All these queries may not be of equal importance. The data that is accessed by most queries is considered important and the queries accessing them need to be selected, from amongst all queries, in a subject area. The selection of such frequent queries is discussed next.

2.2 Frequent Query Selection

As mentioned above, materialized views are required to contain relevant and required information for answering future queries. This information cannot be arbitrarily identified as it may result in the materialized views becoming unnecessary bottlenecks. The approach attempts to identify such information by selecting queries, from amongst all the queries in a subject area, that access frequently accessed information. These queries, referred to as frequent queries, contain information that is capable of answering most queries likely to be posed in future. The approach selects such frequent queries using the association rule mining technique DHP(Dynamic Hashing and Pruning) [13]. The frequent queries selection algorithm, based on DHP [13], is given in Fig. 2. This algorithm takes queries in a subject area, minimum query support threshold as input and produces a set of frequent queries in the corresponding subject area as output.

This algorithm is used to select frequent queries in each subject area. It considers the relations accessed by each query in a subject area to determine the frequent relation sets for a pre-specified minimum query support threshold. The algorithm prunes the number of queries and the relations accessed by them, after each scan of the set of queries in a subject area. For this, it uses a Hash table where, for any pair of relations, it computes the bucket value based on a hash function. If the bucket value is less than the minimum query support threshold, the candidate relation set is removed from the hash table. This process is repeated until all frequent relation sets have been identified. The queries containing any of the frequent relation sets are then selected as frequent queries.

```

ALGORITHM FrequentQuerySelection
Input : QS = set of queries in a subject area
           MinSupp = Minimum Query support threshold
Output : FQS = Frequent Queries set
Method :
/*Initialization and Identification of frequent itemsets */
  Scan each query in the query set QS
  Collect Rs, the set of Relations along with their support count.
  Now Generate L1 from C1
/* Ck is generated by joining Lk-1with itself */
  set k = 1
  Suppose Ck Qk : CandidateRelationSet(RkQk) of size k
  Lk Qk : frequent RelationSet of size k
  L1 = {frequent RelationSet};
  For (k = 1; Lk Qk != ∅; k++)
    Ck+1 = candidates generated from Lk Qk;
/* In k-iteration, hash all k+1 CandidateRelationSets in a hash table, */
  1. Scan each QuerySet from (k+1)-CandidateRelationSet Generation,
  2. Take (K+1)-CandidateRelationSet in QuerySet and Compute Hash Function
  3. Consider (K+1)-CandidateRelationSet ,in a Query(Qk), where k=1,2,3,...n.
     For any pair of CandidateRelationSet (RkQk, Rk+1Qk), the bucket # value according to
     Hash function using Rs
     h((RkQk, Rk+1Qk)) = ((support count of RkQk)*10+(support count of Rk+1Qk)) % N
     where N is any arbitrary number.
  4. Count all the CandidateRelationSets in each bucket.
/*Pruning from the QuerySet: */
  If bucket value of a CandidateRelationSet is less than MinSupp then remove the
  candidate relation set from the hash table
  Repeat the process until all frequent relations sets are identified
  Store the frequent relations sets into FRS
  Compute FQS as the queries in QS that contains the atleast one frequent relation set in FRS.

```

Fig. 2. Algorithm FrequentQuerySelection based on DHP[13]

3 Example

Consider the relations accessed by the previously posed queries on the data warehouse given in Fig. 3.

Q1	Doctor, Patient, Ward	Q6	Booking, Train, Class	Q11	Patient, Bill, Nurse	Q16	Doctor, Admission, Patient
Q2	Train, Ticket, Booking	Q7	Train, Ticket, Booking	Q12	Ticket, Class, Booking	Q17	Patient, Bill, Nurse
Q3	Train, Fare, Ticket	Q8	Patient, Ward, Admissions	Q13	Booking, Train, Class	Q18	Train, Class, Tickets
Q4	Ticket, Train, Class	Q9	Doctor, Admissions, Bill	Q14	Seat, Booking, Train	Q19	Bill, Admissions, Doctor
Q5	Patient, Ward, Admissions	Q10	Ticket, Fare, Train	Q15	Bill, Doctor, Patient	Q20	Doctor, Admission, Patient

Fig. 3. Relations accessed by Previously Posed Queries on a Data Warehouse

The similarity between the queries in Fig. 3 is computed using the Overlap Coefficient[18]. Using these computed similarities, a similarity matrix is constructed and is given in Fig. 4.

Using the similarity matrix in Fig. 4, the previously posed queries in Fig. 3, minimum query threshold MinQ=3 and minimum query similarity threshold $\varepsilon=0.5$, the subject areas are identified as given in Fig. 5. The plots depicting the formation of the first cluster is shown in Fig. 6.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	
Q1	0	0	0	0	0.665	0	0	0.333	0.333	0	0.333	0	0	0	0.665	0.665	0.33	0	0.333	0.66	
Q2	0	0	0.665	0.665	0	0.33	0.6	0	0.33	0.333	0.33	0.333	0.333	0.66	0.66	0	0.33	0.333	0.33	0	
Q3	0	0.665	0.33	0.665	0	0.33	0.66	0	0	1	0	0	0.333	0	0	0	0	0.665	0	0	
Q4	0	0.665	0.665	0	0.33	0.665	0	0	0	0	0.333	0	0.33	0	0.333	0	0	0	0.66	0	
Q5	0.665	0	0	0	1	0	0	0.665	0.665	0	0	0	0	0.33	0.33	0.66	0	0	0.665	0.665	
Q6	0	0.665	0.33	0.665	0	1	0.665	0	0.33	0.333	0.33	0.333	0.665	0.66	0.33	0	0.33	0.665	0.33	0	
Q7	0	1	0.665	0.665	0	0.665	1	0	0.33	0.665	0.33	0.333	0.333	0.33	0.33	0	0.33	0.333	0.33	0	
Q8	0.665	0	0	0	0.665	0	0	1	0.333	0	0.33	0	0	0	0.333	0.665	0.333	0	0.333	0.665	
Q9	0.333	0	0	0	0.665	0	0	0.333	1	0.333	0.333	0	0	0.333	0.665	0.665	0.665	0	1	0.665	
Q10	0	0.665	0.665	0.665	0	0.333	0.665	0	0.333	1	0	0.333	0	0.333	0.333	0	0.333	0.665	0.333	0	
Q11	0.333	0	0	0	0.333	0	0	0	0.333	0.333	1	0	0	0.333	0.333	0.333	0	0.665	0.333		
Q12	0	0.665	0.333	0.665	0	0.665	0.665	0	0	0.333	0	1	0.333	0	0	0	0	0.333	0	0	
Q13	0	0.333	0.333	0.333	0	0.665	0.333	0	0	0	0	0.665	1	0	0	0	0	0.665	0	0	
Q14	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0	0.333	0.333	0.333	0	0	1	0	0.333	0	0.333	0.333	0.333	
Q15	0.665	0	0	0	0.333	0	0	0.665	0.333	0.333	0.333	0	0	0	1	0.665	0.333	0	0.333	0.333	
Q16	0.665	0	0	0	0.665	0	0	0.665	0.665	0	0.333	0	0	0	0.333	0.333	0.665	0.333	0	0.333	0.665
Q17	0	0	0	0	0.333	0	0	0.333	0.333	0.333	0.665	0	0	0	0	0.333	0.333	1	0	0.665	
Q18	0	0.665	0.665	1	0	0.665	0.333	0	0	0.665	0	0.665	0.333	0.333	0	0	0	1	0	0	
Q19	0.333	0	0	0	0.665	0	0	0.333	1	0.333	0.665	0	0	0	0.333	0.333	0.665	0.665	0	1	0.333
Q20	0.665	0	0	0	0.665	0	0	0.665	0.333	0	0.333	0	0	0	0.333	0.333	0.665	0.333	0	0.665	1

Fig. 4. Similarity Matrix showing similarity between previously posed queries Q1 . . Q20

```

Retrieve ε-neighborhood points from Q1 using similarity Matrix.
Insert(Q1, UNDEFINED) into QueryList
ε-neighbor of Q1 are Q5, Q15, Q16, Q20
Sim(Q1, Q5)=0.66, Sim(Q1,Q15)=0.66, Sim(Q1,Q16)=0.66, Sim(Q1,Q20)=0.66,
|Nε(Q1)| >= 3, cqsim(Q1) = 0.66, Processed = true
Write Q1 to file and remove it from the QueryList
Nε(Q1)={Q5, Q15, Q16, Q20}
Similarly the Neighboring Queries Q15, Q8, Q5, Q9, Q16, Q20, Q19, Q17 are processed
Update rqsim in QueryList
Q11, Q17 are identified as Noise Point
Updated File = {Q1, Q5, Q8, Q15, Q16, Q19, Q20, Q9}
Let the core query = Q2,
Processing of Query CQ=Q2
ε-neighbor of Q2 are Q3, Q4, Q7, Q14, Q15
Sim(Q2, Q3)=0.66, Sim(Q2,Q4)=0.66, Sim(Q2,Q7) = 0.66, Sim(Q2,Q14) = 0.66, Sim(Q2,Q15)=0.66
|Nε(Q2)| >=3, cqsim(Q2)= 0.66, Processed = true,
Write Q2 to file and remove it from the QueryList
Nε(Q2)={Q3, Q4, Q7, Q14, Q15}
Similarly the Neighboring Queries Q4, Q3, Q7, Q6, Q13, Q12, Q10, Q18 are processed
Update rqsim in QueryList
Q14 is identified as Noise Point
Updated File = {Q2, Q4, Q3, Q7, Q6, Q12, Q10, Q13, Q18}
Two clusters of queries specifying the two subject areas S1 and S2 are as given below:
S1 = {Q1, Q5, Q8, Q9, Q15, Q16, Q19, Q20}
S2 = {Q2, Q3, Q4, Q6, Q7, Q10, Q12, Q13, Q18}
Noise Points: {Q11, Q14, Q17}

```

Fig. 5. Subject Area Identification using previously posed queries Q1 . . Q20

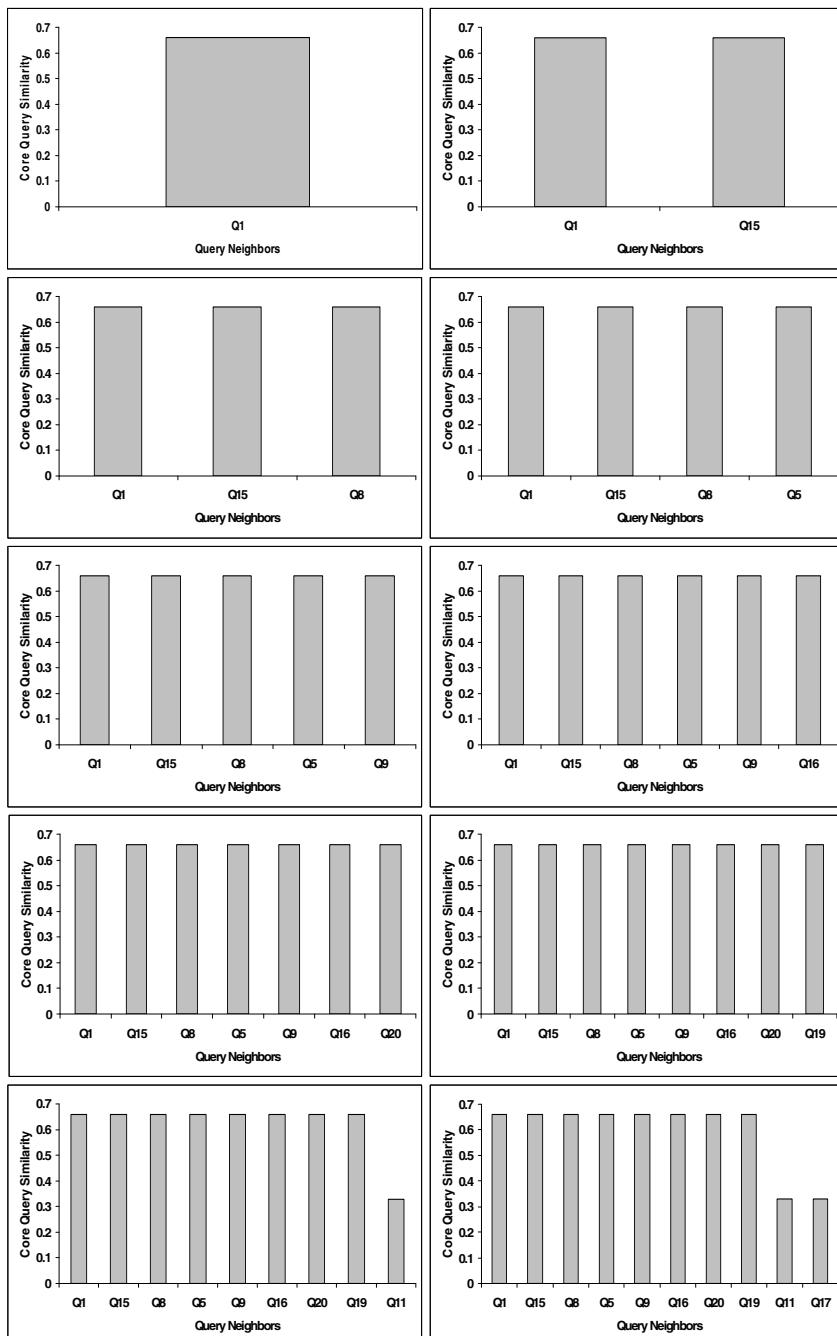


Fig. 6. Reachability Query Similarity Plot depicting formation of subject area S1

As can be observed from the reachability query similarity plot, the cluster representing the first subject area S1 contains queries Q1, Q5, Q8, Q9, Q15, Q16, Q19, Q20. The queries Q11 and Q17 are identified as noise queries. Similarly, the formation of the second cluster representing subject area S2 is identified and contains queries Q2, Q3, Q4, Q6, Q7, Q10, Q12, Q13, Q18. Query Q14 is identified as noise.

Next, the frequent queries are selected from the queries in subject areas S1 and S2. The selection of frequent queries in S1 for minimum query support threshold $q_s=0.5$ is given in Fig. 7.

$S_1 = \{Q_1, Q_5, Q_8, Q_9, Q_{15}, Q_{16}, Q_{19}, Q_{20}\}$	$FR = \{\text{Doctor}, 6\}, \{\text{Admissions}, 6\}, \{\text{Patient}, 5\}, \{\text{Ward}, 4\}$																																																			
$SFRlist = \{\text{Doctor}, 6\}, \{\text{Admissions}, 6\}, \{\text{Patient}, 5\}, \{\text{Ward}, 4\}$																																																				
For $k=1$,	$FR = \{\text{Doctor}, \text{Patient}, \text{Ward}\}$ $R_1Q_1 = \{\text{Doctor}\}, R_2Q_1 = \{\text{Patient}\}$ $\text{Since, } R_1Q_{(1)} \times R_2Q_1 \text{ where, } R_1Q_{(1)} < R_2Q_1$ $\Rightarrow R_1Q_{(1)} = R_2Q_1;$ $\text{So, } C_1Q_1 = \{\text{Doctor, Patient}\}$																																																			
For $k=2$,	$FR = \{\text{Doctor, Patient, Ward}\}$ $R_1Q_1 = \{\text{Doctor}\}, R_3Q_1 = \{\text{Ward}\}$ $\text{Since, } R_1Q_1 \times R_3Q_1 \text{ where, } R_1Q_1 < R_3Q_1$ $\Rightarrow R_1Q_1 = R_3Q_1;$ $\text{So, } C_2Q_1 = \{\text{Doctor, Ward}\}$																																																			
For $k=3$,	$FR = \{\text{Doctor, Patient, Ward}\}$ $R_2Q_1 = \{\text{Patient}\}, R_3Q_1 = \{\text{Ward}\}$ $\text{Since, } R_2Q_1 \times R_3Q_1 \text{ where, } R_2Q_1 < R_3Q_1$ $\Rightarrow R_2Q_1 = R_3Q_1;$ $\text{So, } C_3Q_1 = \{\text{Patient, Ward}\}$																																																			
Similarly applying the same method (DHP) for Queries Q5, Q8, Q9, Q15, Q16, Q19, Q20 in subject area , the frequent relation set generated $FR = \{\text{Doctor, Admission}\}$. The 2-candidate relation sets, hash table created from the 2-candidate relation set and the hash table after pruning are shown below:																																																				
<table border="1"> <thead> <tr> <th>Quer y</th> <th>2-RelationSets</th> <th>Bucket Address</th> <th>0</th> <th>4</th> <th>2</th> <th>5</th> <th>4</th> <th>6</th> <th>$RQ(1)*RQ(1)$</th> <th># in the bucket</th> </tr> <tr> <th>Bucket Count</th> <td>1</td> <td>2</td> <td>3</td> <td>3</td> <td>3</td> <td>5</td> <td></td> <td></td> <td></td> </tr> <tr> <th>Bucket Contents</th> <td>$\{\text{P W}\}$</td> <td>$\{\text{D W}\}$ $\{\text{D W}\}$</td> <td>$\{\text{P D}\}$ $\{\text{P D}\}$</td> <td>$\{\text{Ad P}\}$ $\{\text{Ad P}\}$</td> <td>$\{\text{Ad W}\}$ $\{\text{Ad W}\}$</td> <td>$\{\text{D Ad}\}$ $\{\text{D Ad}\}$</td> <td></td> <td></td> <td></td> </tr> </thead> <tbody> <tr> <td>Queries= Q5, Q9, Q16, Q19 and Q20 contains Doctor, Admission</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>FQS = {Q5, Q9, Q16, Q19 and Q20} are frequent queries in Subject Area 1</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table>		Quer y	2-RelationSets	Bucket Address	0	4	2	5	4	6	$RQ(1)*RQ(1)$	# in the bucket	Bucket Count	1	2	3	3	3	5				Bucket Contents	$\{\text{P W}\}$	$\{\text{D W}\}$ $\{\text{D W}\}$	$\{\text{P D}\}$ $\{\text{P D}\}$	$\{\text{Ad P}\}$ $\{\text{Ad P}\}$	$\{\text{Ad W}\}$ $\{\text{Ad W}\}$	$\{\text{D Ad}\}$ $\{\text{D Ad}\}$				Queries= Q5, Q9, Q16, Q19 and Q20 contains Doctor, Admission										FQS = {Q5, Q9, Q16, Q19 and Q20} are frequent queries in Subject Area 1									
Quer y	2-RelationSets	Bucket Address	0	4	2	5	4	6	$RQ(1)*RQ(1)$	# in the bucket																																										
Bucket Count	1	2	3	3	3	5																																														
Bucket Contents	$\{\text{P W}\}$	$\{\text{D W}\}$ $\{\text{D W}\}$	$\{\text{P D}\}$ $\{\text{P D}\}$	$\{\text{Ad P}\}$ $\{\text{Ad P}\}$	$\{\text{Ad W}\}$ $\{\text{Ad W}\}$	$\{\text{D Ad}\}$ $\{\text{D Ad}\}$																																														
Queries= Q5, Q9, Q16, Q19 and Q20 contains Doctor, Admission																																																				
FQS = {Q5, Q9, Q16, Q19 and Q20} are frequent queries in Subject Area 1																																																				

Fig. 7. Frequent Queries Selection in subject area S1

From Fig. 7, the frequent queries in subject area S1 are Q5, Q9, Q16, Q19 and Q20. Similarly, the frequent relation set in subject area S2 is identified as {Train, Ticket}. The queries Q2, Q3, Q4, Q7, Q10, Q18 contain Train and Ticket and thus are selected as frequent queries in subject area S2.

The selected frequent queries would thereafter be used to construct a materialized view for the corresponding subject area.

4 Conclusion

In this paper, an approach that mines data warehouse queries for constructing materialized views is proposed. The approach groups the previously posed queries on a data *warehouse*, based on their similarity, to form clusters of queries. Each identified cluster specifies a subject area. The approach then identifies frequent queries, from among all the queries, in a subject area. The selected frequent queries indicate data that has been accessed frequently in the past. It is usually seen that queries posed in the past serve as indicators to queries likely to be posed in future. Accordingly the selected frequent queries are considered appropriate for constructing materialized views for the respective subject area. The resultant materialized views would be able to answer most future queries in improved query response times. Further, since most of the queries posed on the data warehouse and the resultant materialized views, are subject specific, it would require fewer numbers of materialized views to answer most of the future queries. This would further improve the query response time. Consequently, it would facilitate the decision making process.

References

1. Agrawal, S., Chaudhari, S., Narasayya, V.: Automated Selection of Materialized Views and Indexes in SQL databases. In: 26th International Conference on Very Large Data Bases (VLDB 2000), Cairo, Egypt, pp. 495–505 (2000)
2. Ankerst, Breunig, Kriegel, Sander: OPTICS: Ordering Points to Identify the Clustering Structure. ACM SIGMOD Record Archive 28(2) (June 1999)
3. Aouiche, K., Darmont, J.: Data mining-based materialized view and index selection in data warehouse. Journal of Intelligent Information Systems, 65–93 (2009)
4. Baralis, E., Paraboschi, S., Teniente, E.: Materialized View Selection in a Multidimensional Database. In: 23rd International Conference on Very Large Data Bases (VLDB 1997), Athens, Greece, pp. 156–165 (1997)
5. Chirkova, R., Halevy, A.Y., Suciu, D.: A Formal Perspective on the View Selection Problem. In: Proceedings of VLDB, pp. 59–68 (2001)
6. Gupta, H., Mumick, I.S.: Selection of Views to Materialize in a Data warehouse. IEEE Transactions on Knowledge & Data Engineering 17(1), 24–43 (2005)
7. Gupta, H., Harinarayan, V., Rajaraman, V., Ullman, J.: Index Selection for OLAP. In: Proceedings of the 13th International Conference on Data Engineering, ICDE 1997, Birmingham, UK (1997)
8. Harinarayan, V., Rajaraman, A., Ullman, J.D.: Implementing Data Cubes Efficiently. In: ACM SIGMOD, Montreal, Canada, pp. 205–216 (1996)
9. Horng, J.T., Chang, Y.J., Liu, B.J., Kao C. Y.: Materialized View Selection Using Genetic Algorithms in a Data warehouse System. In: Proceedings of the 1999 Congress on Evolutionary Computation, Washington DC, USA, vol. 3 (1999)
10. Inmon, W.H.: Building the Data Warehouse, 3rd edn. Wiley Dreamtech India Pvt. Ltd. (2003)
11. Lawrence, M.: Multiobjective Genetic Algorithms for Materialized View Selection in OLAP Data Warehouses. In: GECCO 2006, Seattle, Washington, USA, July 8–12 (2006)

12. Lehner, W., Ruf, T., Teschke, M.: Improving Query Response Time in Scientific Databases Using Data Aggregation. In: Proceedings of 7th International Conference and Workshop on Database and Expert Systems Applications, DEXA 1996 (1996)
13. Park, J.S., Chen, M., Yu, P.S.: An effective hash based algorithm for mining association rules. In: ACM SIGMOD International Conference on Management of Data (May 1995)
14. Rizzi, S., Saltarelli, E.: View Materialization vs. Indexing: Balancing Space Constraints in Data Warehouse Design. In: Eder, J., Missikoff, M. (eds.) CAiSE 2003. LNCS, vol. 2681, pp. 502–519. Springer, Heidelberg (2003)
15. Teschke, M., Ulbrich, A.: Using Materialized Views to Speed Up Data Warehousing. Technical Report, IMMD 6, Universität Erlangen-Nürnberg (1997)
16. Theodoratos, D., Sellis, T.: Data Warehouse Configuration. In: Proceeding of VLDB, Athens, Greece, pp. 126–135 (1997)
17. Theodoratos, D., Xu, W.: Constructing Search Spaces for Materialized View Selection. In: 7th ACM Internatioanl Workshop on Data Warehousing and OLAP (DOLAP 2004), Washington, USA (2004)
18. Clemonsa, T.E., Bradley Jr., E.L.: A nonparametric measure of the overlapping coefficient. Published in Journal “Computational Statistics & Data Analysis” 34(1) (July 28, 2000)
19. Vijay Kumar, T.V., Ghoshal, A.: A reduced lattice greedy algorithm for selecting materialized views. In: Prasad, S.K., Routray, S., Khurana, R., Sahni, S. (eds.) ICISTM 2009. Communications in Computer and Information Science, vol. 31, pp. 6–18. Springer, Heidelberg (2009)
20. Vijay Kumar, T.V., Haider, M., Kumar, S.: Proposing candidate views for materialization. In: Prasad, S.K., Vin, H.M., Sahni, S., Jaiswal, M.P., Thipakorn, B. (eds.) ICISTM 2010. Communications in Computer and Information Science, vol. 54, pp. 89–98. Springer, Heidelberg (2010)
21. Vijay Kumar, T.V., Haider, M.: A Query Answering Greedy Algorithm for Selecting Materialized Views. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010. LNCS, vol. 6422, pp. 153–162. Springer, Heidelberg (2010)
22. Vijay Kumar, T.V., Jain, N.: Selection of Frequent Queries for Constructing Materialized Views in Data Warehouse. The IUP Journal of Systems Management 8(2), 46–64 (2010)
23. Vijay Kumar, T.V., Goel, A., Jain, N.: Mining Information for Constructing Materialised Views. International Journal of Information and Communication Technology 2(4), 386–405 (2010)
24. Vijay Kumar, T.V., Haider, M.: Greedy views selection using size and query frequency. In: Unnikrishnan, S., Surve, S., Bhoir, D. (eds.) ICAC3 2011. Communications in Computer and Information Science, vol. 125, pp. 11–17. Springer, Heidelberg (2011)
25. Vijay Kumar, T.V., Haider, M., Kumar, S.: A view recommendation greedy algorithm for materialized views selection. In: Dua, S., Sahni, S., Goyal, D.P. (eds.) ICISTM 2011. Communications in Computer and Information Science, vol. 141, pp. 61–70. Springer, Heidelberg (2011)
26. Vijay Kumar, T.V., Devi, K.: Frequent Queries Identification for Constructing Materialized Views. In: The proceedings of the International Conference on Electronics Computer Technology (ICECT 2011), April 8-10, vol. 6, pp. 177–181. IEEE, Kanyakumari (2011)
27. Kumar, T.V.V., Haider, M.: Selection of views for materialization using size and query frequency. In: Das, V.V., Thomas, G., Lumban Gaol, F. (eds.) AIM 2011. Communications in Computer and Information Science, vol. 147, pp. 150–155. Springer, Heidelberg (2011)

28. Widom, J.: Research Problems in Data Warehousing. In: 4th International Conference on Information and Knowledge Management, Baltimore, Maryland, pp. 25–30 (1995)
29. Yang, J., Karlapalem, K., Li, Q.: Algorithms for Materialized View Design in Data Warehousing Environment. *The Very Large Databases (VLDB) Journal*, 136–145 (1997)
30. Zhou, J., Larson, P., Goldstein, J., Ding, L.: Dynamic Materialized Views. In: IEEE 23rd International Conference on Data Engineering, Istanbul, pp. 526–535 (2007)

Similarity Based Cluster Analysis on Engineering Materials Data Sets

Doreswamy and K.S. Hemanth

Department of Post-Graduate Studies and Research in Computer Science
Mangalore University, Mangalagangotri-574 199, Karnataka, India
doreswamy@yahoo.com, reachhemanthmca@gmail.com

Abstract. Nowadays with rapidly growing databases in manufacturing industries it's really an unmanageable timing problem to analyze them and to make decision from them. Studying this type of problem using data mining techniques leads more clarification for manufacture and also for better research work. Here in this paper a similarity based cluster technique is proposed on engineering materials database and implemented using c sharp .net.

Keywords: Clustering, Engineering materials, K-mean.

1 Introduction

Today growing engineering materials database quickly, which is more complex and complicated to handle each and every values of database. It is arduous to make decision for manufacturing industry on such kind of databases. Where as in the design stage engineer confronting more difficulty and confused to choose the suitable materials for their applications [2]. First designer had better empathize the materials to be capable to match the product. There is so many materials and so much of information, and so however to mine from them are the challenging task. There are many techniques from computation aspects to mine. Data mining adds such form of techniques to mine information such as classification prediction , association, cluster etc., Using clustering techniques for grouping the materials makes designer to select the materials. There are several works been done with this technique. Simply implementing about engineering materials database during design stage depict a Modern approach. Clustering is a data mining technique applied to place data components into associated groups without advance knowledge by the group definitions[5].Clustering could be believed the most significant unsupervised learning problem; so, as ever other problem of these kind, it deals on finding a structure in a collection of untagged data. Accumulating and compounding several information almost defined clusters, contributes to qualitative decision making about best cluster number and elements that constitute them. Accordingly, to find qualitative outcomes as conceivable, and to alleviate cluster interpretation, analysts had better aggregate unlike tools in the process of data clustering[6].

From many years data clustering is an important technique used for explorative data analysis. It has shown to be valuable in many practical application areas such as data classification and image processing. Newly, there is a developing emphasis on searching analysis of very large datasets to discover useful models and/or correlations among attributes. This is called data mining, and data clustering is considered as a

particular branch. Still existing data clustering techniques do not adequately handle the problem of processing heavy datasets on a limited amount of resources. So as the dataset size increases, they do not scale up well in terms of memory demand, running time, and result quality.

In classification the objects are imputed to predefined classes, whereas in clustering the classes are also to be specified. The clustering problem has been directed extensively in statistics and machine learning. These techniques are idle for big amounts of data and as well assume that all the data to be clustered can be held in main memory concurrently. Data produced from practical application areas such as spatial databases, banking sector, supermarkets, information recovery, insurance, and text mining are identical heavy and high multidimensional. This has led to data clustering in data mining, which handles with efficient and effective clustering techniques[9][10].

The paper is organized as follows section 1 clearly says the introduction part of clustering and engineering materials. Methodology well be discuss in section 2 and section 3 present the experimental results.

2 Methodology

Data clustering is a process of grouping objects into groups so that clusters objects within one cluster are similar to each other and objects in different clusters are dissimilar to each other. It is referred to as unsupervised learning because it does not use any data labels for grouping of data. A similarity function or dissimilarity function is used to measure the similarity/dissimilarity between objects[3]. Clustering techniques can be useful tools for exploring the underlying structure of a given data set and are being applied in a wide variety of engineering disciplines[1].

2.1 Similarity Measures

Similarity measures depend on the type of data used. If the variables are continuous measurements, the dissimilarity or similarity between the objects described by these variables is computed based on the distance between them, the most popular distance measure is Euclidean distance and is defined as

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{in} - x_{jn})^2}$$

where $i = (x_{i1}, x_{i2}, \dots, x_{in})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jn})$ are two n-dimensional data objects.

2.2 K-Mean Clustering

The K-Means algorithm is to cluster n objects based on attributes into K partitions, $K < n$. It is alike to the expectation-maximization algorithm for mixtures of Gaussians in that they both set about to find the centers of natural clusters in the data. It assumes that the object attributes form a vector space. The objective it tries to achieve is to minimize total intra-cluster variance, or, the squared error function [4].

$$v = \sum_{i=1}^k \sum_{x_j \in s_i} (x_j - u_i)^2$$

where there are K clusters S_i , $i = 1, 2, \dots, K$, and u_i is the centroid or mean point of all the Data x_j in S_i . The most common form of the algorithm uses an iterative refinement heuristic known as Lloyd's algorithm. Lloyd's algorithm begins by partitioning the input data into K initial sets, either at random, or applying a few heuristic data. It then calculates the mean point, or centroid, of each set. It builds a new partition by consorting each point with the closest centroid. Then, the centroids are recalculated as the new clusters, and algorithm repeated by alternate application by this two steps until convergence, which is found as the points no longer alternate clusters (or alternatively, the centroids are no longer changed).unfortunately.

The K-Means Partitioning Algorithm

The k-mean algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input : K:the number of clusters, D: a data set containing n objects.

Output : A set of k clusters.

Method:

1. arbitrarily choose k objects from D as the initial cluster centers;
2. repeat
3. (re) assign each objects to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4. update the cluster means, i.e. calculate the mean value of the objects for each cluster; until no change;

2.3 Clustering Algorithm

Step 1: Start

Step 2: Database and with number of cluster K

Step 3: Initializing the value K list.

Step 4:Database.Datacount (No. of Attribute(row) and No. of Object(col))

```
Cnt1=0;
for ( Datacount = 0; Datacount <= col ; Datacount++)
{
    Cnt1[col]= Datacount[Row,col+1];
    return(0);
}
```

Step 5: Find Centroids Cnt1,Cnt2;

Step 6:Find distance (point Pt1, point Pt2)

```
return(0);
```

Step 7: Repeat 5 & 6 till distance lies between (0 < distance < 1);

Step 8: Group based on minimum distance;

Step 9:Display the cluster elements with their ID;

Step 10: Stop;

3 Experimental Results on Cluster Analysis

As pre day by day technology improves and new technology emerging back at the older one. Coding with new technologies bring more easier and much more readable to user. This makes a way of studying data mining techniques with advances technologies for large databases. A prototype software system as implemented for K-Means clustering algorithm on engineering materials data set using modern language C sharp .Net technology which is more readable and clear to user with the results after clustering as shown in fig.1. Experiment is conducted on engineering materials database consisting of 5097 data with 25 attributes, and for various cases of (K : I to V). Cluster analysis are shown in the table representations.

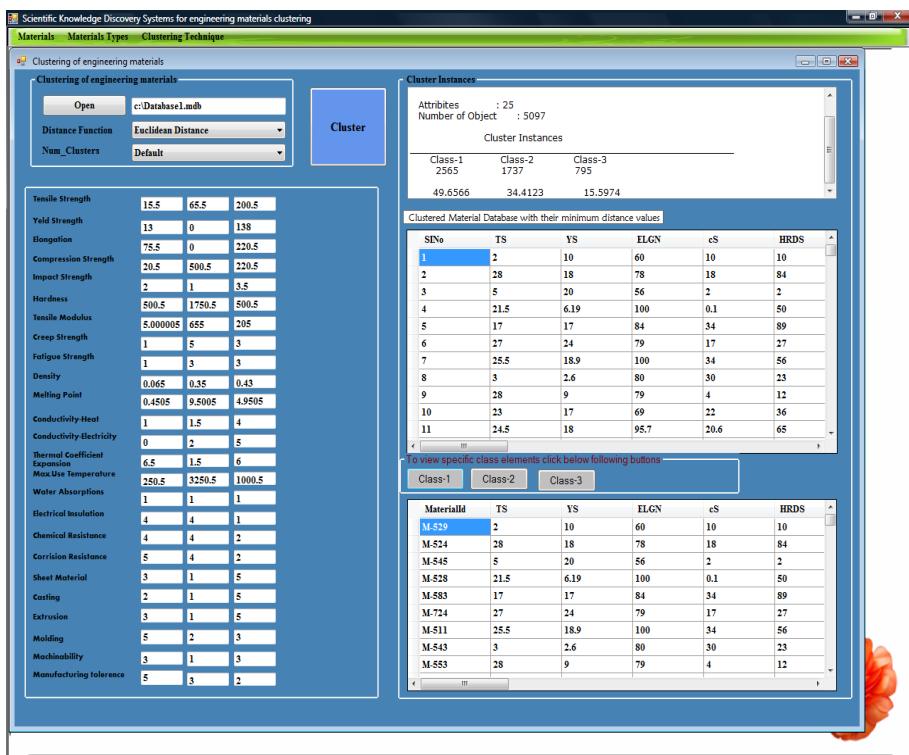


Fig. 1. A prototype software module for clustering engineering materials database on input design requirements.

Case-I

In this case, a known individual classes such as Polymer, Ceramic and Metal data sets with their centroid values respectively 37.10064, 250.714 and 101.7752 are considered for experimentation. At a time a single class is employed for experimentation, the data

clustered for individual classes, Polymer ,Ceramic and Metal are 2655, 795 and 1737 respectively, which is shown in the table1.

Table 1.

Number of Clusters	Individual Classes	Centroid Values	Number of Data Clusters
K = 1	Polymer	37.10064	2565
	Ceramic	250.714	795
	Metal	101.7752	1737

Case-II

In this case, K-means algorithm is executed for two known classes(K = 2). Known classes such as Polymer, Ceramic and Metal data sets with their centroid values respectively 37.10064, 250.714 and 101.7752 are considered for experimentation. At a time two classes are employed for experimentation, the data clustered for two classes are shown in the table2.

Table 2.

Experiments	Number of Clusters	Individual Classes	Centroid Values	Number of Data Clusters
1	K = 2	Polymer	37.10064	1754
		Ceramic	250.714	812
2	K = 2	Ceramic	250.714	812
		Metal	105.6752	1031
3	K = 2	Polymer	37.10064	1754
		Metal	101.7752	1737

Case-III

In this case, known three classes (K = 3) data sets are considered for validating K-mean algorithm on engineering materials data sets. Known classes such as Polymer, Ceramic and Metal data sets with their centroid values respectively 37.10064,

250.714 and 101.7752 are considered for experimentation. At a time three classes are employed for experimentation, the data clustered for three classes are shown in the table 3.

Table 3.

Experiments	Number of Clusters	Individual Classes	Centroid Values	Number of Data Clusters
1	K = 3	Polymer	37.10064	2565
		Ceramic	250.714	1737
		Metal	101.7752	795

Case- IV

In order to validate the algorithm, unknown data sets are considered. Four classes Class-1, Class-2 ,Class-3 and Class-4 with their centroid values 23.24, 35.44, 45.28 and 18.20 are considered randomly. At a time four classes are employed for experimentation, the data clustered for four classes are shown in the .table 4.

Table 4.

Experiments	Number of Clusters	Individual Classes	Centroid Values	Number of Data Clusters
1	K = 4	Class1	23.24	1011
		Class2	35.44	2553
		Class3	45.28	860
		Class4	18.2	673

Case- V

The maximum possible clusters in the proposed software package is 5. In order to fulfill the clustering of engineering materials into 5 classes, unknown five classes data sets are employed for testing the algorithm. Five classes Class-1, Class-2 ,Class-3, Class-4 and Class-5 with their centroid values 17.20, 15.40, 16.12 ,12.4 and 30.48 are considered randomly. At a time five classes are employed for experimentation, the data clustered for five classes are shown in the Table 5.

Table 5.

Experiments	Number of Clusters	Individual Classes	Centroid Values	Number of Data Clusters
1	K = 5	Class1	17.2	1629
		Class2	15.4	937
		Class3	16.12	890
		Class4	12.4	971
		Class5	30.48	670

4 Conclusion and Future Scope

In this experiment, K-Means algorithm is employed for clustering of engineering materials into predefined classes. Clustering of engineering materials is done based on the similarity of material's characteristics. Similarity of materials is computed by Euclidian distance function. Clustering results obtained by K-Means algorithm is verified with the known data sets. For instance, the k -means clustering algorithm has a tendency to discover well separated spherical clusters. Further, clustering results is validate with unknown class data sets.

Further, the proposed method can be extended to evaluate K-mean algorithm with various similarity measure functions for determining the better similarity function that results the best clustering of engineering materials.

Further, the scope of this work is focused on dissimilarity or outliers analysis of engineering materials data sets.

Acknowledgments. This work has been supported by the University Grant Commission(UGC), India under Major Research Project entitled “Scientific Knowledge Discovery Systems (SKDS) For Advanced Engineering Materials Design Applications” vide reference F.No. 34-99\2008 (SR), 30th December 2008. The authors gratefully acknowledge the support and thank the authors

References

1. Pham, D.T., Afify, A.: Clustering techniques and their applications in engineering. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 221, 1445–1459 (2007)
2. Doreswamy, Sharma, S.C.: An Expert Decision Support System for Engineering Materials Selections And Their Performance Classifications on Design Parameters. International Journal of Computing and Applications (ICJA) 1(1), 17–34 (2006)
3. Doreswamy: Similarity measuring approach based engineering materials selection. International Journal of Computational Intelligence Systems (IJCIS) 3, 115–122 (2010)
4. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn.

5. Teknomo, K.: K-Means Clustering Tutorials,
<http://people.revoledu.com/kardi/tutorial/kMean/>
6. Grlejvic, O., Bošnjak, Z.: Combining different Clustering Techniques for Improved Knowledge Discovery. In: Proceedings of the 20th Central European Conference on Information and Intelligent Systems, pp. 287–292 (September 2009)
7. Santhi, P., Murali Bhaskaran, V.: Performance of Clustering Algorithms in Healthcare Database. International Journal for Advances in Computer Science 2(1), 26–31 (2010) ISSN - 2218-6638
8. Chauhan, R., Kaur, H., Afshar Alam, M.: Data Clustering Method For Discovering Clusters In Spatial Cancer Databases. International Journal of Computer Applications (0975 – 8887) 10(6), 9–14 (2010)
9. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7) (July 2002)
10. Kumar, V., Rathee, N.: Knowledge discovery from database Using an integration of clustering and classification. International Journal of Advanced Computer Science and Applications 2(3), 29–33 (2011)

A Chaotic Encryption Algorithm: Robustness against Brute-Force Attack

Mina Mishra¹ and V.H. Mankar²

¹ Electronics & Telecommunication
Nagpur University, Nagpur
Maharashtra, India

minamishraetc@gmail.com

² Department of Electronics Engineering
Government Polytechnic, Nagpur
Maharashtra, India

vhmankar@gmail.com

Abstract. An encryption method is proposed that uses self-invertible matrix, modular function, Non-Linear shift register and the chaotic map known as Logistic. Parameter of Logistic map act as secret key. As chaotic system used in this algorithm is 1-D system, the key space is lesser than 2^{100} which shows that the method is weak against Brute-force attack but identifiability property of the selected key from key space assures its strength against the attack. Key sensitivity and plaintext sensitivity of the key chosen from key space for the algorithm is analyzed and its strength against known-plaintext attack is also tested and conclusions are derived.

keywords: Non-Linear Shift Register, Logistic map, Brute-force attack, Identifiability.

I Introduction

Chaos has been used in cryptography since 1992. It is a concept which presents a behaviors of non-linear systems that lies somewhere in between perfect order and a complete disorder or unbounded stage where the system becomes in deterministic (uncontrollable). The signals resulting from chaotic systems [1] [2] are broadband, noise like, unpredictable and have highly random behavior. There exists an interesting relationship between chaos and cryptography. The properties of chaotic cryptosystems [3] such as, sensitivity to initial conditions/system parameters, mixing property, deterministic dynamics and structural complexity found to be similar to the confusion and diffusion with small change in plaintext/secret key, deterministic pseudo randomness and algorithmic complexity properties of traditional cryptosystems.

Thus chaos provides a promising approach for cryptography. Field of Chaos Cryptography is very young and has a great potential for research. Much fundamental work as well as practical problems needs to be addressed before high performance perfect chaotic cryptosystem can be designed.

The communication schemes discussed in [4] [5] involving a chaotic transmitter system and a receiver system is illustrated in fig 1.

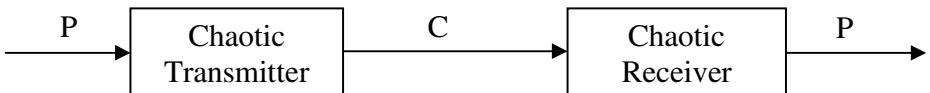


Fig. 1. Communication Scheme based on Chaos

P is the information to be encrypted i.e. the plaintext;

C is the encrypted information conveyed to the receiver i.e. the cipher text;

One of the most promising chaotic cryptosystem scheme named message-embedding [6] is illustrated in fig 2.

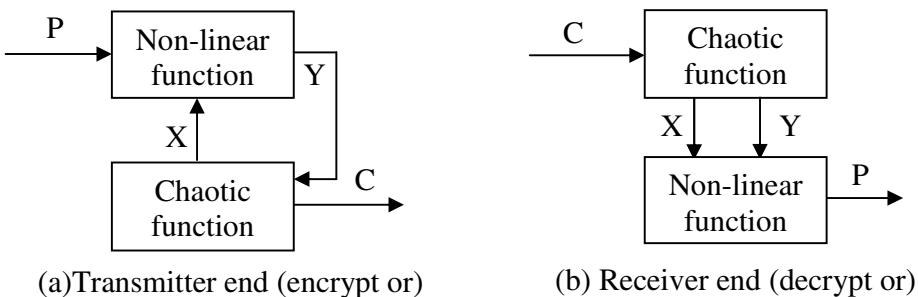


Fig. 2. Message-embedded chaotic cryptosystem

P : plaintext; C : cipher text; X : state of chaotic function; Y : Intermediate encrypted plaintext;

(i) Transmitter and Encryption: The plain text is encrypted by an encryption rule which uses non-linear function and the state generated by the chaotic system in the transmitter. The scrambled output is inputted further to the chaotic system such that the chaotic dynamics is changed continuously in a very complex way. Then another state variable of the chaotic system in the transmitter is transmitted through the channel.

(ii) Receiver and Decryption: Recovery of the plaintext is done by decrypting the input (cipher text) using reverse process of encryption, as used in the transmitter.

In this scheme state X is not directly transmitted through the noise interference channel but quantity Y is available at the output of transmitter. In the proposed scheme, plaintext is encrypted with a traditional method known as Hill cipher and then the cipher text is further used as plaintext of message-embedded scheme as discussed in fig.2.

An essential issue for the validation of any cryptosystems is the cryptanalysis that is the study of attacks against cryptographic schemes in order to reveal their possible weakness. A fundamental assumption in cryptanalysis [7], first stated by A. Kerkhoff

in 1883, is that the adversary knows all the details of the cryptosystem, including the algorithm and its implementation, except the secret key, on which the security of the cryptosystem must be entirely based.

The various cryptanalytic attacks are:

- (i) **Cipher text-only attack:** The attacker possesses a string of cipher text.
- (ii) **Known plain text:** The attacker possesses a string of plain text, P, and the corresponding cipher text, C.
- (iii) **Chosen plain text:** The attacker has obtained temporary access to the encryption machinery. Hence he/she can choose a plain text string, P, and construct the corresponding cipher text string, C.
- (iv) **Chosen cipher text:** The attacker has obtained temporary access to the decryption machinery. Hence he/she can choose a cipher text string, C, and construct the corresponding plain text string, P.
- (v) **Brute Force Attack:** A brute force attack is the method of breaking a cipher by trying every possible key. The brute force attack is the most expensive one, owing to the exhaustive search.

There are some other specialized attacks, like, differential and linear attacks. Differential cryptanalysis is a kind of chosen-plaintext attack aimed at finding the secret key in a cipher. It analyzes the effect of particular differences in chosen plaintext pairs on the differences of the resultant cipher text pairs. These differences can be used to assign probabilities to the possible keys and to locate the most probable key. Linear cryptanalysis [8] is a type of known-plaintext attack, whose purpose is to construct a linear approximate expression of the cipher under study. It is a method of finding a linear approximation expression or linear path between plaintext and cipher text bits and then extends it to the entire algorithm and finally reaches a linear approximate expression without intermediate value.

From the crypto graphical point of view [9], the size of the key space should not be smaller than 2^{100} to provide a high level security so that it can resist all kind of Brute force attack. To get a number of keys with 2^{100} (approx. 10^{30}), in chaotic system the resolution must be 10^{-15} , but, it may be possible that thousands of keys would become equivalent with that resolution, unless, there is a strong sensitivity to parameter mismatch. The quicker the brute force attack, the weaker the cipher. Whether brute force attacks gets succeeded or not depends on the key space size of the cipher and on the amount of computational power available to the attacker. However, this requirement might be very difficult to meet by proposed cipher because the key space does not allow for such a big number of different strong keys. The brute force attack is the most expensive one, owing to the exhaustive search. A fundamental issue of all kinds of cryptosystem is the key. No matter how strong and how well designed the encryption algorithm might be, if the key is poorly chosen or the key space is too small, the cryptosystem will be easily broken. Unfortunately, proposed chaotic cryptosystem has a small key space region and it is non-linear because all the keys are not equally strong.

A cryptanalytic method, known as output equality based on the identifiability concept cited in [10], is the solution to the problem of less key space in chaotic ciphers. It is found that in chaotic ciphers, there exists a unique solution for a particular input for certain domain of values of parameters. The response of any system to a particular input is the solution of that particular system and it contains all the information about the parameters of system. This type of analysis is also known as parametric analysis. Identifiability concept fulfills the necessary condition but not sufficient as the developed cryptosystems must be tested for sensitivity and other statistical tests to result in a robust cipher.

The aim of this work is to present an algorithm for text encryption using Logistic chaotic map which can provide security against Brute-Force attack and improve plaintext sensitivity and key sensitivity compared to message-embedded scheme using the same chaotic system and Non-Linear function. The analysis result concludes that the proposed encryption algorithm shows improvement in plaintext sensitivity and key sensitivity property. Method is found to resist known plaintext attack for almost all the selected keys as shown in the analysis table for available first five characters of plaintext string. If available character string of plaintext, which may be in any number are not the starting characters of plaintext then in such situation method proves to resist the attack for all the selected keys. Conclusion about the identifiability of almost all chosen key is derived which concludes that this algorithm provides security against Brute-force attack and the selected identifiable key can play role of secret key against Brute-force attack.

The paper is organized as follows. In Section 2, a brief description of the approaches involved in the analysis of the proposed algorithm and section 3, discusses the overview of Non-Linear Shift Register and Logistic map used for developing the algorithm. Section 4, an algorithm of the proposed encryption method is presented. In Section 5, simulated analysis results are cited in tabulated form. Then in Section 6, the conclusion derived from simulated analysis is discussed and it is shown that this method provides security against Brute-force attack.

2 Cryptanalytic Procedures Used for the Analysis

A. Output Equality: The output equality describes that - For the same inputs and initial condition, transmitter system is parameterized at different values of parameter taken from the existing domain of parameter space, if the output response of the system obtained after some value of iteration, parameterized at a particular value coincides with the output response of the same system parameterized at some other value of parameter within the domain for the same number of iteration, then both the parameters are said to be equal and identifiable. The system is said to possess unique solution at that particular value of parameter and the system is said to be structurally identifiable.

There exists a connection between uniqueness in the secret parameters (acting as key) and identifiability concept, which, reduces the probability of finding actual parameter by the eavesdropper. If parameter of the transmitter is identifiable, it is more difficult for the eavesdropper to find it by a brute force attack. Consequently, this parameter may be a good candidate to play the role of the secret key against a brute force attack. If the parameter is not identifiable, the eavesdropper has a higher favorable chance to find it by a brute force attack. Thus, this parameter vector is a bad candidate to play the role of the secret key against brute force attack.

B. Plaintext sensitivity Test: It is the percentage of change in bits of cipher text obtained after encryption of plaintext, which is derived by changing single bit from the original plaintext from the bits of cipher text obtained after encryption of original plaintext. With the change in single bit of plaintext, there, must be ideally 50% change in bits of cipher text to resist differential cryptanalysis (chosen-plaintext attack) and statistical analysis, corresponds to plaintext sensitivity test.

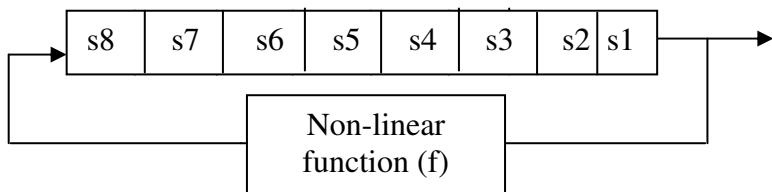
C. Key sensitivity Test: Key sensitivity is the percentage of change in bits of cipher text obtained after encryption of plaintext using key, which is flipped by single bit from the original key, from bits of cipher text obtained after encryption of plaintext using original key, which requires ideally 50% change in cipher text bits to resist linear and statistical attacks.

To resist common attacks, the designed cryptosystem should have the confusion property. To achieve the confusion property, statistical properties of the cipher text, such as distribution, correlation and differential probability of the cipher text should be independent of the exact value of the key and of the plaintext. With the increase of randomness of cipher text, confusion strengthens. No pattern should involve in the cipher text for secured cryptosystem. Confusion is intended to make the relationship between cipher text and plaintext statistically independent.

Plaintext sensitivity and key sensitivity describes the diffusion property of the system. Diffusion refers spreading out of the influence of single plaintext digit over many cipher text digits so as to hide the statistical structure of the plaintext. These methods together are also known as Avalanche effect.

3 Functions Used In Designing Cryptosystem

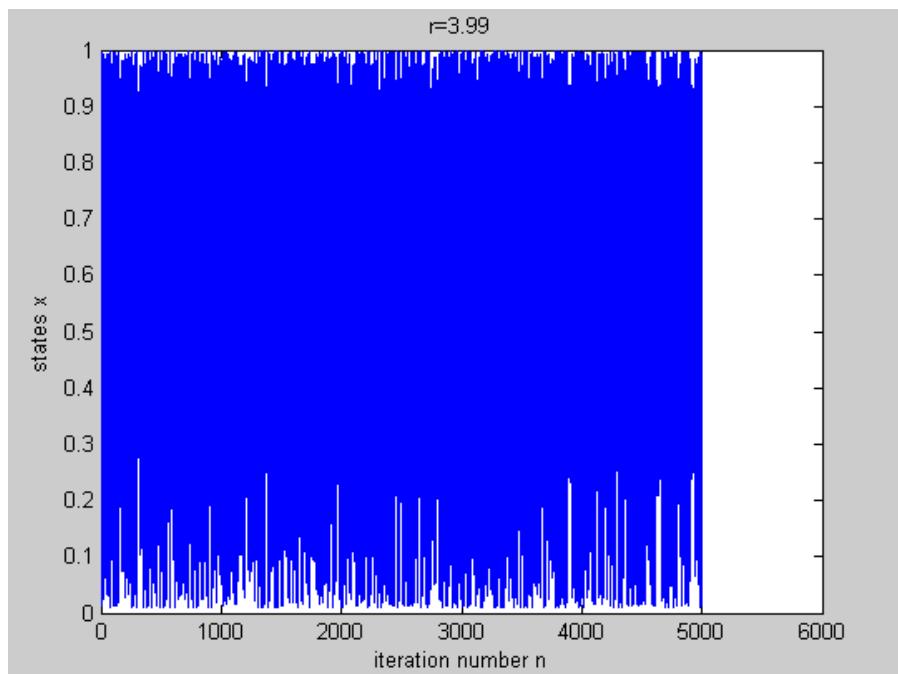
A. Non-linear Feedback Register: NLFSR (Non-Linear Feedback Shift-register) is a common component in modern stream ciphers, especially in RFID and smartcard applications. NLFSRs are known to be more resistant to cryptanalytic attacks than Linear Feedback Shift Registers (LFSR's), although construction of large NLFSRs with guaranteed long periods remains an open problem. A NLFSR is a shift register whose current state is a non-linear function of its previous state. The NLFSR used in this paper is shown in fig 3.

**Fig. 3.** NLFSR using 8-bit shift register

B. Logistic map: The logistic map is a polynomial mapping of degree 2, it takes a point, in a plane and maps it to a new point using following expressions:

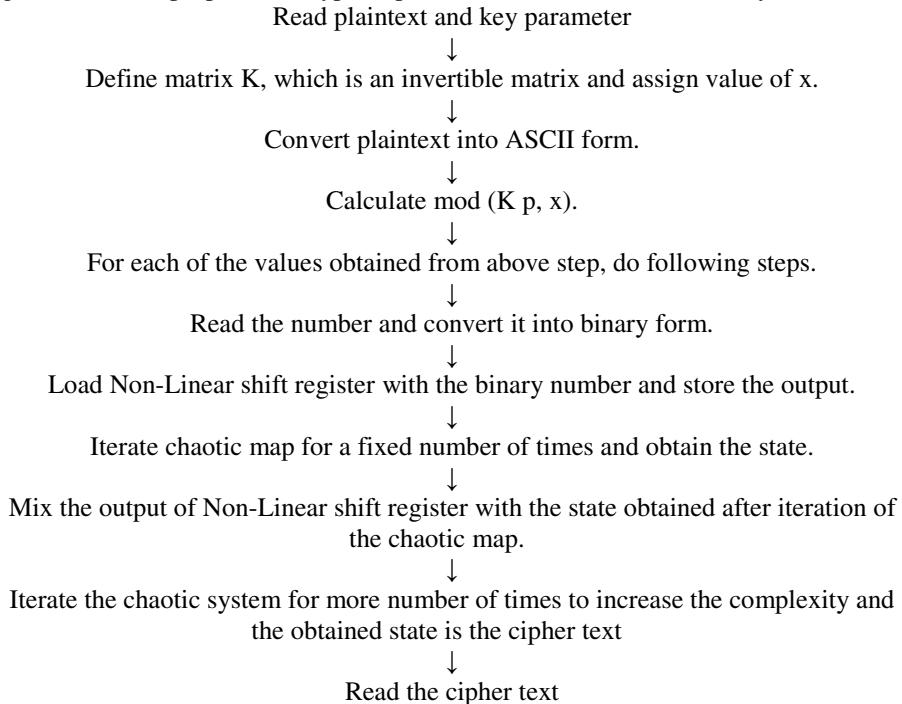
$$x(k+1) = rx(k)[1 - x(k)]$$

Where, map depends on the parameter r. From $r = 3.57$ to $r = 4$, the map exhibits chaotic behavior which is shown in fig 4.

**Fig. 4.** Plot of Logistic map for $r = 3.99$, $x(0) = 0.99$, $n = 5000$.

4 Proposed Encryption Method

Algorithm for the proposed encryption procedure is discussed in brief way.



Decryption algorithm is inverse of encryption algorithm which is used at the receiving end so as to recover the original information.

5 Analysis Result

Providing a high level of security is dramatically increasing. Cryptanalysis plays most important role in the development and advancement of cryptography. In this section ten different texts have been chosen as plaintext to the developed method and are analyzed for plaintext sensitivity, key sensitivity, Identifiability and known plaintext attack for different keys selected from domain of key space. The analysis results are presented in tabular form obtained from simulation of the cryptanalytic procedures. As we know that logistic map is chaotic for the value of parameter ' r ' = 3.57 to 4.0, hence key space of the method is also limited to chaotic region, which is very less compared to 2^{100} . The chaotic system is sensitive to a small change in initial condition or parameters, which shows that if value of key is incremented up to the precision of 10^{-9} , then also the system shows a drastic change in its output behavior, which enlarges the key space still not sufficient to resist brute-force attack because to resist this attack it is required to have precision of 10^{-15} . The key space for this algorithm is determined approx. to be 47×10^7 .

Table 1. Analysis result

Sl. No .	Plaintext	Key value	Cipher text	Plaintext sensitivity (in %)	Key sensitivity (in %)	Domain for key With increment =0.0001	Identifiability of key for iteration value =1 or 2	Robustness against known plaintext attack.	Whether key can act as secret key against Brute Force attack?
1.	What is your name?	3.6424	§À~ä¶X<Ö~semÄ	9.2105	16.4474	(3.57,3.77)	I	R for p=[p1 p2 ... p5]	YES
2.	I am going to market.	3.7328	v!^(')~'66yr2&	10.7955	17.0455	(3.57,3.77)	I	R for p=[p1 p2 ... p19]	YES
3.	My college name is s.s.c.e.t .	3.7455	,Üü,ç&~ñØ&Z@ÉÖ:Ö~pÙp2&	5.4167	13.3333	(3.57,3.77)	I	R for p=[p1 p2 ... p27]	YES
4.	Hello! how are you?	3.7694	¶ç>Rí~Ú^É\$ôjÄ	7.8947	9.8684	(3.57,3.77)	I	R for p=[p1 p2 ... p15]	YES
5.	Sita is singing very well.	3.8544	BÜ½~çÄ=BRflädÖùF i;d*p	6.4815	12.9630	(3.66,3.86)	I	R for p=[p1 p2 ... p19]	YES
6.	Ram scored 98 marks in Maths.	3.8551	L½½jdSÜ*•éæ%<Ac&;SleÅ<±&	5.8333	14.5833	(3.76,3.96)	I	R for p=[p1 p2 ... p19]	YES
7.	Jaycee publication.	3.8529	=txý®&Ó`Fåå#	14.3750	23.1250	(3.77,3.97)	I	R for p=[p1 p2 ... p15]	YES
8.	Thank you,sir.	3.8641	ñÖæN;ÉP½œé:o	10.8333	20	(3.78,3.98)	I	R for p=[p1 p2 ... p11]	YES
9.	The match was very exciting.	3.9065	ÝílWýÜ*)j=ÉEOPá?l×4•f6K	6.4655	22.8448	(3.785,3.985)	I	R for p=[p1 p2 ... p19]	YES
10.	I will be leaving at 9p.m.	3.9189	O~`Ø,§§-üÉ"0 Lfo	6.4815	12.9630	(3.79,3.99)	I	R for p=[p1 p2 ... p19]	YES

NI – Non-Identifiable; I – Identifiable; R – Robust; p [p1 p2... p n] – First ‘n’ characters of available plaintext string.

It is observed from the results cited in the given table 1 that plaintext sensitivity of the method ranges from 5% to 15% and key sensitivity is ranging from 12 % to 23%, which concludes that the method may or may not be weak against linear, statistical and differential attacks. For almost all the selected keys, method resist known plaintext attack for available first five characters of plaintext string. If available character string of plaintext, which may be in any number are not the starting characters of plaintext then in such situation method proves to resist the attack for all the selected keys. Different texts analyzed for known-plaintext attack for different keys requires different numbers of available characters of plaintext to find out the secret key as mentioned in the table ,which concludes that strength of the algorithm

against the attack depends on the length of text and strength of key. Conclusion about the identifiability of the all chosen key is derived for the given specifications in the above table which concludes that method can provide security against Brute-force attack and the selected identifiable key can play role of secret key against Brute-force attack. It is seen that length of cipher text is equal to the length of plaintext. Time taken is less and tolerable. Memory space required during encryption is also optimum. The space between the characters in plaintext is encrypted. In encrypted text, it is seen that the letter of plaintext, that repeats is not encrypted as same encrypted text, which avoids any pattern to exist in cipher text, thus shows randomness of cipher text.

A complete file as shown in fig 5 is encrypted using the procedure as shown in fig 6, which concludes that the method is capable of encrypting and decrypting a complete file successfully.

```
C:\Documents and Settings\ussr\My Documents\MATLAB\plaintext.txt*
1 Branch and subject code-CSVTU
2 Course Name: ME (PT), ME - M Tech
3 Sr. No. Branch Code Branch name Subject Code      Subject
4 MACHINE DESIGN
5 1 Design 48 548111 (37) Tribology
6 2 Design 48 548112 (37) Mechanical Vibration
7 3 Design 48 548113 (37) Advanced Dynamics of Machine
8 4 Design 48 548114 (37) Theory of Elasticity & Plasticity
9 5 Design 48 548131 (37) Optimization Techniques (Elective I)
10 6 Design 48 548211 (37) Finite Element Methods
11 7 Design 48 548212 (37) CAD/CAM Application
12 8 Design 48 548213 (37) Advanced Machine Tool Design
13 9 Design 48 548214 (37) Advanced Mechanism
14 10 Design 48 548231 (37) Experimental Stress Analysis (Elective II)
15 11 Design 48 548311 (37) Robotics
16 12 Design 48 548312 (37) Fatigue & Creep
17 PRODUCTION ENGINEERING
18 1 Production Engg.    42 542111 (37) Advanced Manufacturing Engineering
19 2 Production Engg.    42 542112 (37) CAD/CAM Applications
20 3 Production Engg.    42 542113 (37) Production & Materials Management
21 4 Production Engg.    42 542114 (37) Maintenance Engineering
22 5 Production Engg.    42 542131 (37) Applied Fuzzy Logic & Fuzzy Sets
23 6 Production Engg.    42 542211 (37) Machine Tools Engg.
24 7 Production Engg.    42 542212 (37) Robotics
25 8 Production Engg.    42 542213 (37) Quality Control & Reliability
26 9 Production Engg.    42 542214 (37) Measurement System Analysis
27 10 Production Engg.   42 542232 (37) Productivity Management
28
```

Fig. 5. Plaintext file

```

1   ]
2 zA0bJM) S10f~om $m| (000(D~2z\bT+000LE) Q00/0 (@eQ00pw_pi0I06)c200] : [0(>000
3 &9 ?0P9=cOz
4 9 ?0Fp m1] (!00I0<0) -] 00J0f~o3&1h00<k[j0`qT^ rgxX(0h
5 n00BQ00`k@Q0PCGD3>^~04Ua-
6 .H) (0F0y/kW/0%0- 0) 3740E000@lbaumu/0c s00A"~S17GV/. `2000~OR00 UF403700
7 00i0HiSiII)?g0000#0[000000c2V]+Jy~0M
8 [0-0] 003740?2H0@1bkL0z+tI~)0<+0=3vbP"c>S000E;E;m,z1
9 YU0Nj;>(*Dy' (4'00 (03m2R00g00*v0`4G\001.N*0WZ,]5000:001C6U<0)000vzsQ00`k0Q
10 0PCGDnY004Uatt0T0z s:i0I)Q,~3 (0o0(s0s20Q00`k@Q0PCGD9dG04UaeE3m7t0)6Y&J
11 ~5cV/. `) &=g~OR00 UF4037~0gDi0HiSiII)?g0000zV+[j000] TNz002L~tAqb0-0 00
12 ) 3740oL5Yc@1b0y"0X_00p$+s?09 Y?e>30;Q00`k@Q0PCGD(00004Ua<<PItSz_)Q,~0fB0
13 ("(%0Nq8z;00 g.0F[f0hGn;06V0&=Rr~OR00 UF4037J0G0 0nON0?uNuWEuCx-p5Q00
14 `k@Q0PCGDte_b04Ua003de$~0~000^K0d0H0Q0g0yVi (0+0PK0W0
15 u\0PL>Uj0dJAj] gZ0-*0"=Hh=n0) q5r,hH(U@1b0y"0X_00Dx;z;0ex) sf) `0d0z80^D>##J0s
16 >0[30M0GfY6I0z00P+fF(1R0Q0.h)e9#[0i?_
17 P9Zv0^wz0o00R,E000L:000h0GfY6I0z00P+fF(1R0Q0.h@ei0Hi8.nc0GfY6I0z.DVD0b;.4Y
18 0b0;b>=6h0)Q,~E|:y%kB0(Re$0~000mmBy:t9q.h%0m0 00g0jF~-h;V/B!'X80^D>##J0s
19 >00M 00GfY6I0z00P+fF(1R0Q0D>;A 0n0SiI1o0o00;0!Sbb_DH10C1hz!s`04q&e00=
20 Dzvzs%kB0(Re$0~000mmBy:t9<0SSS 800g0zV+[j000] T(Bq,Or' $0 0T0GfY6I0z00P+
21 fF(1R0Q0Ke0"9#[0N0?uNuWEuC=gzoGfY6I0z00P+fF(1R0Q0Ke0"i0Hi00y0J.o0dshQ4Q
22 00vMI;{1
23 ?h000
24 A"qT0000dJAj] gZ0-*0"=Hh=n0) q5r,oL5Yc@1bK0_S0.E>)Q,~0R`00+D+20,G505[RGD0dJ
25 Aj] gZ0-*0"=Hh=n0) q5r,+0k0c@1b0dJAj] gZCjCd4 p0=6h0)Q,~0(0!

```

Fig. 6. Encrypted file

6 Conclusion

This paper aims at presentation of an encryption method self-invertible matrix, modular function, Non-Linear shift register and 1-D chaotic map known as Logistic whose parameter act as secret key. The key space of the method is lesser than 2^{100} which shows that it is weak against Brute-force attack but identifiability property of the selected key from key space assures its strength against the attack. Key sensitivity and plaintext sensitivity of the key chosen from key space for the algorithm is analyzed and its strength against known-plaintext attack is also tested and conclusions are derived.

All the keys selected from the domain of key space derives the conclusion of identifiability hence, it is concluded that the chaotic encryption algorithm can resist Brute-force attack, which is the most basic attack. Avalanche effect of the proposed method is average.

References

- [1] Kocarev, L.: Chaos-based cryptography: A brief overview. *IEEE Circuits Syst. Mag.* 1(3), 6–21 (2001)
- [2] Jakimoski, G., Kocarev, L.: Chaos and Cryptography: Block Encryption Ciphers Based on Chaotic Maps. *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications* 48(2), 163–169 (2001)
- [3] Dachselt, F., Schwarz, W.: Chaos and cryptography. *IEEE Trans. Circuits and Syst. I* 48(6), 1498–1509 (2001)
- [4] de Oliveira, L.P., Sobottka, M.: Cryptography with chaotic mixing. *Chaos, Solutions and Fractals* 35(3), 466–471 (2008)
- [5] Alvarez, G., Montoya, F., Romera, M., Pastor, G.: Chaotic cryptosystems. In: Sanson, L.D. (ed.) *Proc. 33rd Annual 1999 International Carnahan Conference on Security Technology*, pp. 332–338. IEEE (1999)
- [6] Millérioux, G., Hernandez, A., Amigó, J.: Conventional cryptography and message-embedding. In: *Proc. Int. Symp. Nonlinear Theory and its Applications*, Bruges, vol. 35, pp. 469–472 (2005)
- [7] Guo, X., Zhang, J., Guo, X.: An Efficient Cryptanalysis of a Chaotic Cryptosystem and Its Improvement. In: *IEEE Conference on Information Theory and Information Security*, China, pp. 578–581 (2010)
- [8] Yin, R., Yuan, J., Yang, Q., Shan, X., Wang, X.: Linear cryptanalysis for a chaos-based stream cipher. *World Academy of Science, Engineering and Technology* 60, 799–804 (2009)
- [9] Alvarez, G., Li, S.: Some basic cryptographic requirements for chaos-based cryptosystems. *Int. J. Bifurc. Chaos* 16(8), 2129–2151 (2006)
- [10] Anstett, F., Milleroux, G., Bloch, G.: Message-embedded cryptosystems: Cryptanalysis and identifiability. In: *Proc. 44th IEEE Conf. Decision and Control*, vol. 44(3), pp. 2548–2553 (2005)

Introducing Session Relevancy Inspection in Web Page

Sutirtha Kumar Guha^{1,3}, Anirban Kundu^{2,3}, and Rana Dattagupta⁴

¹ Seacom Engineering College, Howrah, West Bengal - 711302, India
sutirthaguha@gmail.com

² Kuang-Chi Institute of Advanced Technology, Shenzhen - 518057, P.R. China
³ Innovation Research Lab, West Bengal - 711103, India
anirban.kundu@kuang-chi.org, anik76in@gmail.com

⁴ Jadavpur University, Kolkata - 700032, India
rdattagupta@cse.jdvu.ac.in

Abstract. In this paper, we propose a new technique for checking the relevancy of sessions created by visitors on a web-page for measuring the web-page ranking, since session of a web-page is considered as an important parameter for web-page ranking calculation in a search engine. It is assumed that session on a web-page depends on the relevancy of the web-page contents with respect to the requirement. A longer session on a web-page may not yield high relevancy of the web-page, hence a threshold value (THV) is considered for individual web-page based on the contents to avoid the probable noise. The threshold value (THV) is calculated by Keyword Matching Index (Kindex) and Data Transfer Speed of the client-server. The Kindex is measured by implementing fuzzy logic on Pattern Matching of requirement and web-page contents. Field Matching information is fetched through hierarchical database.

Keywords: Session, Threshold value (THV), Field Matching, Pattern Matching, Keyword Matching Index (Kindex).

1 Introduction

In our paper relevancy of session on a web-page is examined by a predefined threshold value (T_{HV}) based on the web-page contents. A threshold value (T_{HV}) is considered as the feasible range of the session on a web-page at a particular time instance. T_{HV} is calculated by implementing fuzzy logic and Hierarchical database.

In this paper the decision making procedures are implemented with fuzzy logic. Fuzzy logic is used in our paper since it is easy to implement, fast to react, reliable to use, less sensitive to external noise and it provides sensible transition between the entities [1][2]. Since our work is dealing with the real time scenario in web-sphere implementation of fuzzy logic may yield better result [3].

Several hypothetical proposals have already been published to ensure realistic and better ranking of web-pages. Minimum rank matrix solution has been proposed to minimize the rank of a web-page for large scale applications [4]. Query based ranking approach has been proposed to generate precise ranking as user requirement often varies for different perspectives [5]. Perceptron with margin based ranking method

results better performance as an online ranking algorithm [6]. A new approach has been proposed in the form of ontology ranking based on semantic web to calculate web-page rank based on the relevancy [7]. Inbound and outbound links of a web-page have been considered as attributes to calculate the web-page rank [8]. Machine learning approach has already been started by the researchers to develop better approach to rank the web-pages [9].

In the proposed paper the concept of ontology based semantic web is implemented in hierarchical database that is used for field matching. An ontology is a set of concepts within a domain and their inter relationship. Semantic web is a modified intelligent version of World Wide Web (WWW) where a web-page understands meaning of the contents, since the web-page satisfies the requirement of users [10]. Semantic web comprises of a set of design principles and templates, its wide range of predefined knowledgebase makes the task more intelligent.

Proposed system framework has been mentioned in Section 2 combined with field study. Section 3 shows experimental results. Section 4 concludes the paper.

2 Proposed System Design

It is assumed in this paper that in typical case user's requirement in a Search Engine is specified as 'keyword'. In this work visitor's session on a web-page is measured by a predefined threshold value. It is assumed that introduction of intruder or external noise or intentional longer session on a web-page cause erroneous increment of web-page importance. The value of ($T_H V$) is defined as shown in Equation 1

$$T_H V = K_{\text{index}} / \text{Data Transfer Speed}. \quad (1)$$

K_{index} is calculated based on Field matching and Pattern matching between the requirement and web-page contents as shown in Equation 2. Data Transfer Speed is a dynamic value assigned depends on the speed of the client server on that time instance and measured as bps. Hence dynamic value assignment to $T_H V$ yields more better and feasible result.

$$K_{\text{index}} = (\text{Field Matching Value} + \text{Pattern Matching Value})/2. \quad (2)$$

K_{index} is measured by implementing fuzzy logic on the parameters Field and Pattern. The pattern matching is measured by implementing hierarchical database. These implementations are elaborated in analytical study part.

The session on a web-page is measured by a predefined threshold value ($T_H V$) as shown in Equation 1. K_{index} indicates the degree of keyword matching of the requirement with the web-page. Data Transfer Speed is considered to make the $T_H V$ calculation more realistic, that value is a dynamic entity, and hence the $T_H V$ is a dynamic value depends on the data transfer speed of the client server. It is assumed that session on a web-page depends on the relevancy of the web-page and data transfer speed of the client server.

In our paper T_HV works as a validation checker for any session, if session value exceeds the T_HV then presence of unwanted noise is assumed.

2.1 Analytical Study

Field Matching

Let client requirement at any particular instance is “web traffic”. Field matching information is acquired by hierarchical database as shown in Fig. 1. It is assumed that in typical search engine database web-page information are kept in a hierarchical manner. The database is traversed in a depth first manner and the goal node is achieved after getting the pattern matching. It is assumed in our proposed paper the database is traversed along the left most path for the client given pattern string. The traversing would be considered successful if the goal node is achieved, otherwise the searching would be continued through the database. The traversing procedure and corresponding field matching value (F_M) calculation is depicted in Algorithm1.

Algorithm 1. Field_Match

Input: User submitted string

Output: Field Matching Value (F_M)

Step1: Start

Step2: User given String is taken as input from End user.

Step3: User given String would be matched with the Search Engine predefined database.

Step4: Left most wings of the predefined hierarchical database would be traversed first.

Step5: If matching keyword is found then,

$$F_M = 1/\text{Total Path Length}. \quad (2.1)$$

Otherwise,

$$F_M = 1/(\text{Total Path Length} - \text{distance between goal node and common parent node of successful path and last visited path}). \quad (2.1)$$

Step6: End

Analytical Result of Field Matching

Let, the client given string “web traffic” is situated as left child of a node at level 6 in the hierarchical database as shown in Fig.1. The left wing of the database would be traversed first and an unsuccessful searching would be returned. The search controller would be moved backward and a new traversal would be generated from the node ‘Internet’. The new traversal would return a successful search result. Hence, it is found that Total Path Length would be SIX(6), Distance between goal node(“web traffic”) and common parent node of successful path and last visited path would be ONE(1) as depicted in Fig.1.

Pattern Matching

Pattern matching between the requirement and the web-page is measured by implementing fuzzy logic. It is assumed in our paper that pattern of the requirement and the web-page may be identical if any one of the two conditions satisfies:

Condition 1: 100% string matching between the requirement and any keyword of the web-page

Condition 2: Partial string matching with maximum number of keywords.

It is assumed that in any of the above stated cases pattern may match, hence the decision can not be predefined, and fuzzy decision making procedure may yield better and feasible result.

In the proposed work it is assumed that final fuzzified decision can be either ‘Similar Pattern’ or ‘Moderate Similar Pattern’ or ‘Different Pattern’, hence different weightage value (F_W) is assigned for different decisions as depicted in Table 1 to reflect the appropriate pattern matching value. The input to the fuzzy system will check the two conditions stated above, based on the checking result the control will flow according to. The detailed control flow of the fuzzy system is shown in Fig. 2.

Table 1. Fuzzified Decision with weightage value

Final Fuzzified Decision	F_W
Similar Pattern	1
Moderate Similar Pattern	0.5
Different Pattern	0.3

Final field matching value calculation is represented in Equation 2.2.

$$P_M = (F_D * F_W) . \quad (2.2)$$

Where, P_M = Pattern Matching Value

F_D = Final Decision value according to fuzzy Set theory

F_W = Pattern Factor ($F_W = n; 1 \leq n < 0$)

Analytical Result of Pattern Matching

Let user requirement at any time instance is ‘Web Browser’ and the keywords present on a web-page are: ‘Web’, ‘Browsing’, ‘Web Traffic’, ‘Internet’, ‘Web Sphere’.

According to our proposed system, inputs are as follows:

- Maximum string matching with any keyword;
- Percentage of full or partial matching with the keywords;

In the above stated hypothetical case study most synchronized keyword of the web-page is ‘Browsing’ since five characters (BROWS) are matched out of eleven characters (45% matching) and 80% full or partial matching with the available keywords of the web-page with the requirement ‘Web Browser’. Hence the crisp inputs can be considered as follows:

0.45 as input ‘String Matching’ and 0.80 as input ‘Keyword Matching’.

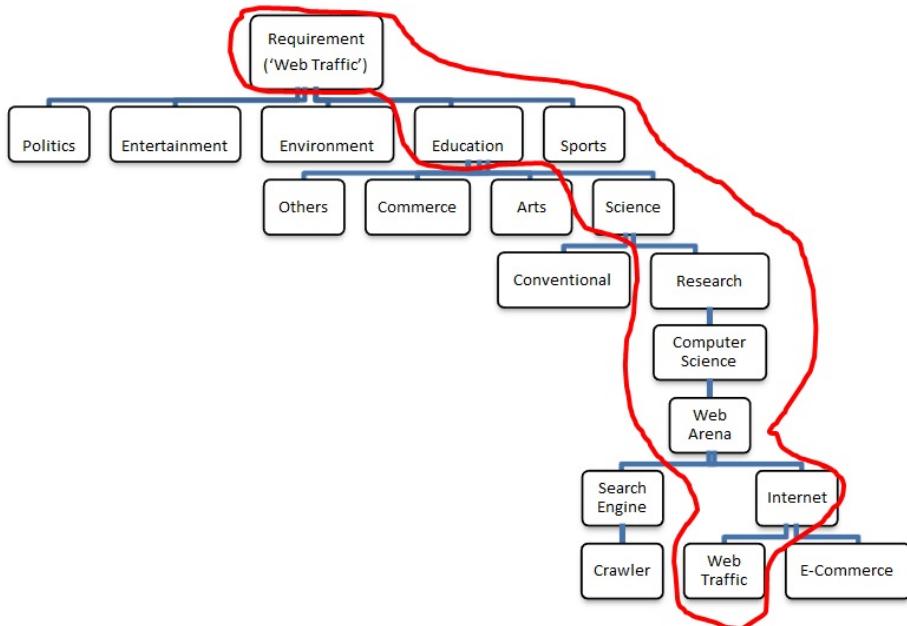


Fig. 1. Example of hierarchical database

The range of crisp input values are shown in Table 2.

Table 2. Input Specification Range

Crisp Input Value	Range	Specification
String Matching	0-0.4	Mismatch
	0.3-0.8	Moderate Match
	0.7-1.0	Complete Match
Keyword Matching	0.2-0.4	No Match
	0.3-0.8	Moderate Match
	0.7-1.0	Total Match

According to the input specification range illustrated in Table1 String Matching input resides in ‘Moderate Match’ specification range whereas Keyword Matching input resides in ‘Moderate Match’ and ‘Total Match’ ranges. Let the corresponding membership function values are as follows:

Moderate Match (String Matching): 0.24

Moderate Match (Keyword Matching): 0.0

Total Match: 0.77

It is assumed that final decision would be as depicted in Table 3.

Table 3. Initial Rulebase

	Mismatch	Moderate Match	Complete Match
No Match	Different Pattern	Different Pattern	Moderate Similar Pattern
Moderate Match	Different Pattern	Moderate Similar Pattern	Similar Pattern
Total Match	Different Pattern	Similar Pattern	Similar Pattern

Corresponding field value of the constructed rulebase is replaced by the membership function value of the field. Reconstructed rulebase is shown in Table 4.

Table 4. Rulebase with Fuzzy calculated Values

	Mismatch	0.24	Complete Match
No Match	Different Pattern	Different Pattern	Moderate Similar Pattern
0.0	Different Pattern	0.0	Similar Pattern
0.77	Different Pattern	0.24	Similar Pattern

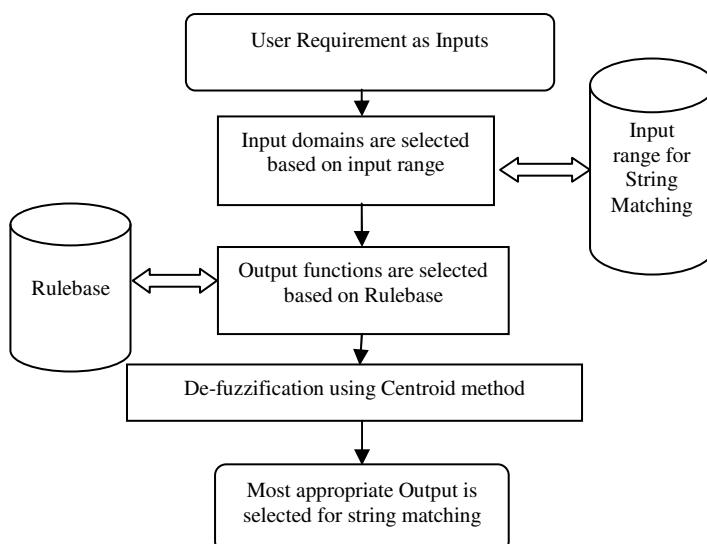
Corresponding values of “Moderate Similar Pattern” and “Similar Pattern” are 0.0 and 0.24 respectively using “AND” operation in Table 4.

According to the “MAX” method, final decision would be web-page of “Similar Pattern”. The field “Similar Pattern” contains higher numeric value.

According to “Centroid” method, final decision would be calculated based on a mathematical formula: $FD = (\sum \mu * D) / \sum \mu$.

According to the above stated case study, final decision calculated by Centroid Method is “Similar Pattern”.

Hence the fuzzy logic yields better and accurate result for pattern checking on a web-page.

**Fig. 2.** Control flow of the fuzzy system

Hence the calculated K_{index} is a combination of field and pattern matching. $T_H V$ is measured based on the K_{index} and Data Transfer Speed which is a dynamic value, each time the above stated procedure executes data transfer speed can be different, so the $T_H V$.

3 Experimental Results

Let the search result of a typical search engine for a searching string ‘Page Ranking’ at any time instance yields the list of web-page link that contains the related information of searching string as shown in Fig. 3. The available web-pages are displayed according to their ranking in the search engine.

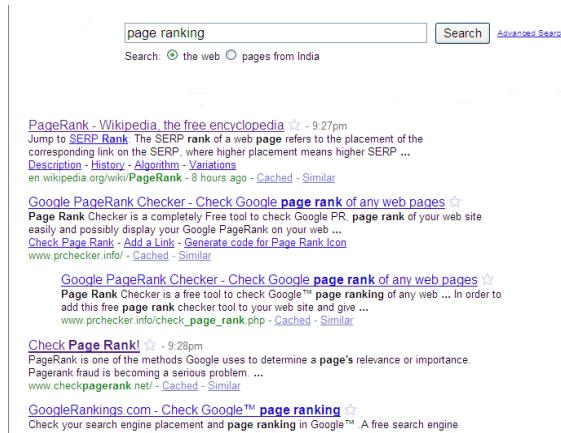


Fig. 3. Searching Result of a Typical Search Engine for string ‘Page Ranking’

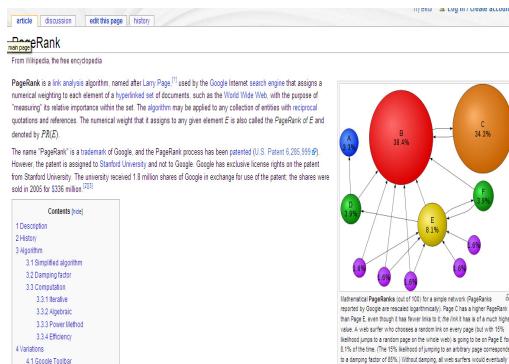


Fig. 4. R2: Rank 3 Web Page according to the most widely used Search Engine Web Page Ranking

It is depicted from the real time data, acquired from a typical search engine, that web-page ranking procedure followed in typical search engine may not yield the most desired result. It is obtained from our collected data that Rank 5 web-page contains more relevant and useful information about the searching string ‘Page Ranking’ rather than Rank 3 web-page that contains some marketing promotional advertisements as the web-page has higher session value. Hence as a consequence of inefficient session value calculation, the search result of a typical search engine may not yield most enviable result. It is observed that in the search result of a typical search engine many web-pages contain commercial promotional demonstration of the searching string rather than containing informative knowledge.

Notes:
Not all links are counted by Google. For instance, they filter out links from known link farms. Some links can cause a site to be penalized by Google. They rightly figure that webmasters cannot control which sites link to their sites, but they can control which sites they link out to. For this reason, links into a site cannot harm the site, but links from a site can be harmful if they link to penalized sites. So be careful which sites you link to. If a site has PR0, it is usually a penalty, and it would be unwise to link to it.

[Top]

How is PageRank calculated?

To calculate the PageRank for a page, all of its inbound links are taken into account. These are links from within the site and links from outside the site.

$$PR(A) = \frac{1-d}{N} + d(PR(1)/C(1) + \dots + PR(n)/C(n))$$

That's the equation that calculates a page's PageRank. It's the original one that was published when PageRank was being developed, and it is probable that Google uses a variation of it but they aren't telling us what it is. It doesn't matter though, as this equation is good enough.

In the equation d - d are pages linking to page A, ' C ' is the number of outbound links that a page has and ' d ' is a damping factor, usually set to 0.85.

We can think of it in a simpler way:
a page's PageRank = $0.15 + 0.85 * (\text{"share"} \text{ of the PageRank of every page that links to it})$

"share" = the linking page's PageRank divided by the number of outbound links on the page.

A page "votes" an amount of PageRank onto each page that it links to. The amount of PageRank that it has to vote with is a little less than its own PageRank (its own value $\cdot 0.85$). This value is shared equally

Website Design Los Angeles
UK Holidays search database of UK Holiday Accommodation
UK accommodation directory

Fig. 5. R3: Rank 5 Web Page according to the most widely used Search Engine Web Page Ranking

How to Increase Page Rank?

Five Steps to Increase your Page Rank:

- Update your website every day by adding more unique content.** If your site has some information for a visitor then it is 100% chance for him to come back to your site again.
- Provide a link to your website.** For example you can provide a link of your previous and next articles on an article page. Or you can provide a link of your website so that a visitor can remain a long time on your website.
- Create sitemap** for your website (XML based file). You can upload your sitemap to submit sitemap to Google and submit sitemap to Yahoo. Get a Google webmaster account and sign up for Google Analytics. It will also tell you if your site is indexed or not, and your page rank in Google, click here to [Get Started](#).
- And the most important tip: Trade your link with other web owners.** Put their link on your website and put your link on their website. This is for free and the very fast way to improve your visibility in search engines. ongsono.com is here to help you to trade or exchange your link.

Free Submit Your Website Now

Highest Referral

- Simea Bilgaya (2535)
- Your daily begin here (2346)
- Art Cinta (2223)
- DRG & Associates URL (2135)
- Place gain knowledge (2054)
- Islam Portal I Portal Ads (2009)

[More Highest Referral](#)

Latest Referral

- Download free Indonesia (1567)
- Education & Learning equipments (98)
- Simea Bilgaya (2535)
- Web Hosting Forum in Pakistan (1375)
- My-SocialNetwork.com - make friends online (140)
- girl Photo Collection (47)

[More Latest Referral](#)

Latest Website Submit

- putting the peanutbutter in Random (0)
- Car and Limo service NJ NY (0)
- FBI Criminal Justice (0)
- FSEEE E-SOCIOZ (0)
- Bag Organizer (0)
- Pakistan's #1 Learning Forum (0)

[More Latest Website Submit](#)

Fig. 6. R4: Rank 8 Web Page according to the most widely used Search Engine Web Page Ranking

A real time case analysis results the following information as shown in Table 5.

Table 5. Comparative Study of Session Relevancy for Sample Web Pages

	Average Visitor Session (min.)	Relative Web Page Ranking in Typical Search Engine among the selected Web Pages	Ideal Relative Web Page Ranking
R1	$(10+8+3+5+9+9+6+6)=7$	1	1
R2	$(12+3+3)=6$	2	4
R3	$(7+9+2+4+3)=5$	3	2
R4	$(6+1+2)=3$	4	3

It is typically assumed that a visitor can wait on a particular web-page only if the web-page is found to be useful, hence less waiting time on the web-page indicates less information regarding visitor query. Introduction of mal-function and spamming web-pages causes deliberate longer session on the web-page. Hence the authenticity of session could not be trustworthy.

In our proposed work relevancy of session would be examined by a threshold value, combination of pattern and field matching data of the web-page as depicted below.

It is assumed that fetched web-pages regarding the searching query ‘page ranking’ reside in the sub domain Search Engine as shown in Fig. 7. Hence the values of F_M are calculated based on Equation 2.1 as shown in Table 5.

It is assumed that, according to the hierarchical database shown in Fig. 7, common parent node of web-pages R1, R2, R3 and R4 are ‘Search Engine’, ‘Computer Science’, ‘Requirement’ and ‘Web Arena’ respectively. Henceforth, F_M value for web-pages R1, R2, R3 and R4 would be 1, 1, 0.7 and 1 respectively as shown in Table 6.

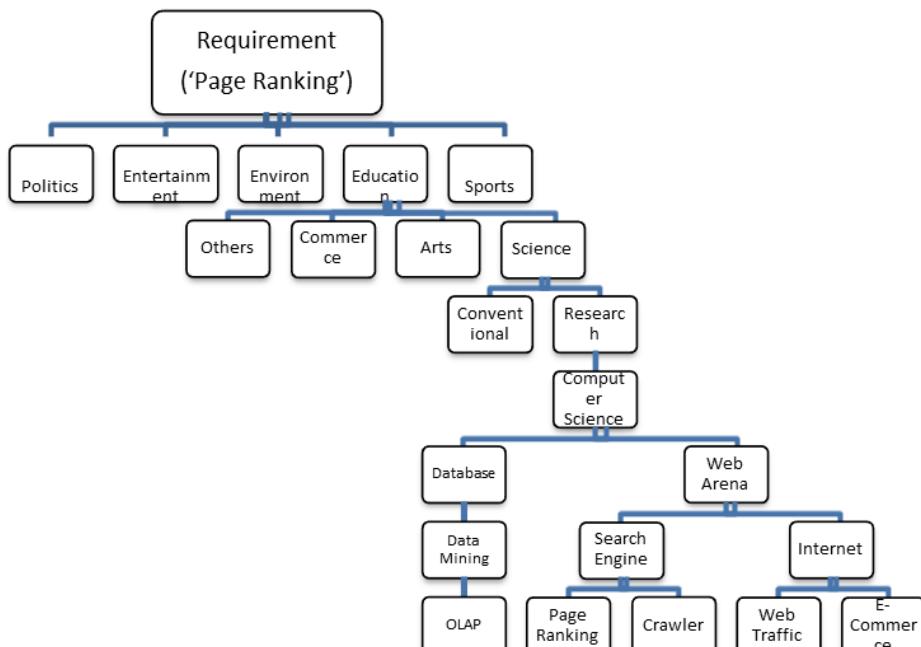


Fig. 7. Hierarchical database

Table 6. Field Matching value for selected Web Pages

Web Page	Common Parent Node	Goal Node	Field Matching Value (F_M)
R1	Search Engine	Page rank	1
R2	Computer Science	Page rank	1
R3	Requirement	Page rank	0.7
R4	Web Arena	Page rank	1

Let, a real time case analysis yields the crisp input values for “String Matching” and “Keyword Matching” for the selected web-pages as shown in Table 7. In this section brief analysis based on real time conquered data is discussed. Details of fuzzy set theory implementation in our proposed work are discussed in Section 2.1.

Table 7. Crisp Input Values for Inputs “String Matching” and “Keyword Matching” for Selected Web Pages

Web Page	String Matching Value	Keyword Matching Value
R1	0.68	0.8
R2	0.56	0.6
R3	0.5	0.74
R4	0.4	0.9

Constructed rulebase for crisp inputs of selected web-pages R1, R2, R3 and R4 would be as shown in Table 8.

Table 8. Rulebase with value for selected Web Pages R1, R2, R3, R4

Web Page R1				Web Page R2			
	Mismatch	0.65	Complete Match		Mismatch	1	Complete Match
No Match	Different Pattern	Different Pattern	Moderate Similar Pattern	No Match	Different Pattern	Different Pattern	Moderate Similar Pattern
Moderate Match	Different Pattern	Moderate Similar Pattern	Similar Pattern	1	Different Pattern	1	Similar Pattern
0.77	Different Pattern	0.65	Similar Pattern	Total Match	Different Pattern	Similar Pattern	Similar Pattern

Web Page R3				Web Page R4			
	Mismatch	1	Complete Match		Mismatch	1	Complete Match
No Match	Different Pattern	Different Pattern	Moderate Similar Pattern	No Match	Different Pattern	Different Pattern	Moderate Similar Pattern
0.25	Different Pattern	0.25	Similar Pattern	Moderate Match	Different Pattern	Moderate Similar Pattern	Similar Pattern
0.35	Different Pattern	0.35	Similar Pattern	1	Different Pattern	1	Similar Pattern

Finally, final decision for the selected web-pages would be measured by defuzzification using ‘Centroid’ method and pattern matching value P_M would be calculated according to Equation 1. Calculated Pattern Matching value is shown in Table 9. Hence, calculated K_{index} according to Equation 2 for the web-pages R1, R2, R3 and R4 are shown in Table 10. It is depicted that the proposed methodology yields more feasible and accurate result.

Table 9. Calculated P_M value for R1, R2, R3, R4

Web Page	Final Decision	Pattern Matching Value (P_M)
R1	100% in Similar Pattern	1
R2	50% in Moderate Similar Pattern	0.5
R3	75% in Moderate Similar Pattern	0.75
R4	60% in Similar Pattern	0.6

Table 10. Calculated Web Page Rank of the Web Pages

Web Page	P_M	F_M	K_{index}	Web Page Rank
R1	1	1	1.0	1
R2	0.5	1	0.5	4
R3	1	0.7	0.75	2
R4	0.6	1	0.6	3

4 Conclusion

In this paper a new approach has been described to measure the relevancy of session on a web-page. A threshold value is calculated based on some real time parameters, such as, comparison between the requirement and the web-page, data transfer speed at any particular time instance. Comparison is measured by field and pattern matching between the requirement and the web-page. Fuzzy logic is implemented for pattern checking whereas field checking is depicted through hierarchical database. Data transfer speed is considered to get more realistic and better threshold value. The session is compared with threshold value of the web-page at any specific time instance for checking the presence of intruder or unwanted noise. Overall the session relevancy checking for a web-page yields error-free and unambiguous session value. The calculation of pragmatic session capitulate the major impact in specified areas since session acts as a major part on any web-page based calculation.

References

- [1] <http://www.iau.dtu.dk/~jj/pubs/logic.pdf>
- [2] <http://www.fuzzy-logic.com/Ch1.html>
- [3] Kundu, A., Guha, S.K., Pal, A.R., Sarkar, T., Mandal, S., Duttagupta, R., Mukhopadhyay, D.: Fuzzy Based Multi Agent-System Offering Cost Effective Corporate Environment. The Open Automation and Control System Journal 1, 65–81 (2008)

- [4] Do, T.T., Chen, Y., Nguyen, N., Gan, L., Tran, T.D.: A Fast and Efficient Heuristic Nuclear-Norm Algorithm for Affine Rank Minimization. In: ICASSP (2009)
- [5] Ni, W., Huang, Y., Xie, M.: A Query Dependent Approach to Learning to Rank for Information Retrieval. In: The Ninth International Conference on Web-Age Information Management (2008)
- [6] Ni, W., Huang, Y.: Online Ranking Algorithm based on Perceptron with Margins. In: The Seventh World Congress on Intelligent Control and Automation, Chongqing, China (2008)
- [7] Rajapaksha, S.K., Kodagoda, N.: Internal Structure and Semantic Web Link Structure Based Ontology Ranking. In: ICIAF 2008 (2008)
- [8] Yates, R.B., Davis, E.: Web Page Ranking using Link Attributes. In: The Thirteenth International World Wide Web Conference, NewYork, USA (2008)
- [9] He, C., Wang, C., Zhong, Y.-X., Li, R.-F.: A Survey on Learning to Rank. In: Seventh International Conference on Machine Learning and Cybernetics, Kunming (2008)
- [10] Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American Magazine (May 17, 2001), <http://www.sciam.com/article.cfm?id=the-semantic-web&print=true> (retrieved March 26, 2008)

Way Directing Node Routing Protocol for Mobile Ad Hoc Networks

M. Neelakantappa¹, A. Damodaram², and B. Satyanarayana³

¹ Brindavan Inst. of Tech & Science, Kurnool-518002, AP, India
m_neelakanta@yahoo.com

² SC&DE, J.N.T. University Hyderabad, AP, India

³ Prof. & Head S.K. University, Anantapur, AP, India

Abstract. A Mobile Ad hoc Network (MANET) is a collection of wireless nodes, forming a temporary network without using any fixed architecture. Communications among the nodes of a MANET are accomplished by forwarding data packets for each other, on a hop by hop basis. The research in this area is commonly simulation based as not many ad hoc networks are currently deployed. Based on routing topology, routing protocols in MANETS are categorized as flat routing and hierarchical routing. In flat routing protocols like DSR and AODV, every node has assigned uniform functionalities. But their performance degrades as the network size grows. Hierarchical routing protocols like CGSR and HSR maintain hierarchy on entire network. But overhead may be high to maintain hierarchy in the mobile network environment. In this paper we propose an On Demand Hierarchical Routing protocol, in which certain number of intermediate nodes present on the route-path are selected as way-nodes and the entire route-path is partitioned into segments by these way-nodes. We call this protocol as Way Directing Node Routing (WDNR), in which the source and destination nodes run a high level inter-segment routing approach. Within each segment it runs a low-level intra-segment routing protocol. The main advantage of this protocol is, when a link on a route-path fails due to node mobility, instead of discarding entire route and rediscovering the fresh route between source and destination, the broken link can be repaired locally. Our model is light weight compared to basic hierarchical routing, as the selection of way-nodes is made only for active routes, on an on-demand basis. Our WDNR protocol uses AODV as intra-segment routing protocol and DSR as inter-segment routing protocol. It mainly solves the scalability problem of flat routing and overhead problem of hierarchical routing. WDNR protocol also exhibits the functionality required to scale large networks and reduce the overhead in hierarchy maintenance. Simulations are carried out in GloMoSim. The simulation results show that WDNR scales better for larger networks with higher than 800 nodes, incurring about 50 to 70 percent less overhead than AODV protocol, while other performance metrics are comparable to basic DSR and AODV.

Keywords: Routing protocols, mobile ad hoc networks, flat routing, hierarchical, scalability, overhead, DSR, AODV.

1 Introduction

Recent advances in technology have provided portable computers with wireless interfaces that allow networked communication among mobile users. The resulting computing environment, which is often referred to as mobile computing, no longer requires users to maintain a fixed and universally known position in the network. And enables almost unrestricted mobility. A Mobile Ad hoc NETwork (MANET) [1,2] is a special type of wireless mobile network in which a collection of mobile hosts with wireless network interface may form a temporary network, without aid of any established infrastructure or centralized administration. The application ranges from civilian to disaster recovery and military. Routing in the MANETs is one of the major challenges and this becomes more difficult when the size of the networks grows. Many routing protocols [1] have been proposed for MANETs and these protocols can be classified into various categories based on different criteria. Based on the manner in which they react to the changing network topology, routing protocols can be classified into two groups: proactive (or table driven) and reactive (or on-demand). Based on the organization and role of the nodes of network, routing protocols can be categorized into flat routing and hierarchical routing protocols.

Proactive routing protocols finds the routes continuously by periodic propagation of information, while the reactive routing protocols find routes only on need basis. Simulation results and performance analysis show that the reactive protocols outperform proactive protocols in all performance metrics like packet delivery ratio, delay and energy efficiency. Hence the research interests have been focused on reactive routing protocols. Ad hoc On-Demand Distance Vector [5] and Dynamic Source Routing [3] are two popular on-demand routing protocols.

In flat routing protocols like DSR and AODV, all nodes are assigned same functionalities. These protocols work well for networks of size within few hundred nodes; but because of extensive routing overhead, their performance deteriorates rapidly as the network grows in size. To solve this scalability problem, hierarchical routing protocols were developed. In the hierarchical routing protocols, the network is divided in to regions and some of the nodes are assigned specific functionalities, for coordinating the routing. In Cluster-head Gateway Switch Routing protocol (CGSR) [1], which is a popular hierarchical routing protocol, the network is divided into clusters. Each cluster is a circular region, with predefined hop count as radius. In this environment, local route maintenance activities which are performed periodically, will affect only a few neighboring clusters. During these activities the other clusters remain untouched. Hence scalability can be easily achieved. But the overhead due to periodic (dynamically) maintenance of the hierarchy will be high in high mobility environments. For larger networks, routes between source nodes and destination nodes are longer. In these networks, if a route breaks due to node crash or node mobility, flat routing protocols like AODV or DSR has to discard the entire original route and re-initiate yet another expensive route discovery process for establishing a new path between source and destination. Normally when a route fails (breaks), only a few hops are declined, but other links are still intact. There fore these protocols wastes the precious knowledge of original route path and causes heavy routing overhead due to global route discoveries. Local repair to AODV is proposed in [5], but it is suitable for only those cases, in which the link failures happen near the destination node. This is because in AODV,

intermediate nodes maintain the routing tables through which they can only know the destination and the next hop to reach that destination. But the target node of the local repair process has to be the destination node. Therefore in most of the cases, it would be better to rediscover a new route directly, when the link failure occurs far from the destination node. Hence, the main objective of our work can be stated as follows: “*Routes are established and maintained hierarchically through segmentation of active route paths. During the route maintenance phase, any broken route can locally be repaired, and thus limiting the number of global route discoveries by the source node*”. Therefore our new routing protocol exhibits better scalability and performance with less routing overhead.

In this chapter, we illustrate a scalable routing model for ad hoc networks, namely On-Demand Hierarchical Routing (ODHR), which maintains a hierarchy on-demand, i.e., only for active routes. In ODHR, along a route, certain number of intermediate nodes are chosen as *way-nodes*. And these way-nodes, divides the active route path into segments. The end points of a route path, which are source and destination, are also represented as way-nodes. With this technique, the main advantage of our model is that, when a node which is on a route path crashes or moves out, instead of discarding the entire route and rediscovering a new route path between source node and destination node, only the two way-nodes of the declined segment have to be reconnected by finding a new segment. This mechanism will have a performance advantage like low end-to-end delay and low routing overhead. The model requires operating two routing protocols, one which run on way-nodes, as a high-level inter-segment routing protocol and the other one which run on nodes of each segment as a low-level intra-segment routing protocol.

Unlike regular hierarchical routing protocols like Zone Routing Protocol (ZRP) [1] and CGSR, the ODHR is a light weight technique because of two reasons. First, nodes on active routes are only involved in the hierarchy of ODHR, where as in CGSR and ZRP, the hierarchy involves all the nodes exists in the network. Another reason is the maintenance of hierarchy in ODHR is easy as it built in one dimension. The maintenance of hierarchy in ZRP and CGSR is complex, as it is built on two dimensions, by dividing the whole network is divided into clusters as depicted in Fig. 1. But in our approach, only the active route paths are linearly divided into various segments as depicted in Fig. 2.

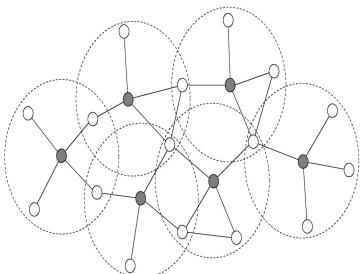


Fig. 1. A two-dimension hierarchy in CGSR/ZRP

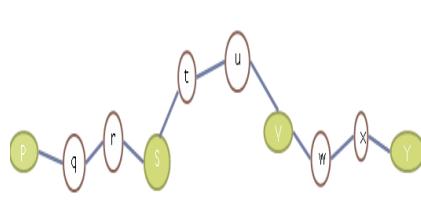


Fig. 2. Division of routes into segments

In case of CGSR and ZRP, the hierarchy is built in two dimensions as network is divided into clusters as shown in Fig.1. Also, the nodes in ODHR work uniformly, as assigning the duty of way-node or non-way-node is specific for a route-path. So, a node may act as way-node in one route-path and non-way-node in another route-path. Therefore uniform consumption of resources takes place at each node in MANETS. But in hierarchical routing protocols like CGSR, nodes have to work non-uniformly and a centralize node may frequently become a traffic center and drain all its battery power early. This results in performance degradation of the protocol.

Using ODHR approach, we designed a new protocol referred to as Way Directing Node Routing (WDNR). This protocol requires two flat routing protocols for its operation. We selected DSR and AODV as the inter-segment protocol and the intra-segment protocol respectively. Here DSR and AODV two popular routing protocols for MANETS are combined hierarchically. Actually, these two protocols become two specific cases of our mechanism: when the length of segment is set to 1, WDNR works purely as DSR because each link is a segment, and the inter-segment protocol dominates; when the length of segment is set to big number (nearly equal to path length), WDNR works as AODV, because entire route-path is to be considered as one segment and thus intra-segment routing protocol dominates.

2 The on Demand Hierarchy Routing

To make flat routing protocol like DSR scalable to large networks, our model On Demand Hierarchical Routing (ODHR) evolves from features of both DSR and AODV protocols. Normally in a flat routing protocol like DSR or AODV, when a route breaks due to node mobility or node failure, the broken route is removed and a new route is discovered from the source node to the destination node. Hence they have scalability problems in larger networks. In an optimization to AODV [7], states that the local routes repair mechanism is suitable only for few cases, where link failures happen near to the destination node. Also as specified in an Internet Draft of DSR [3], the DSR protocol is best suitable for ad hoc networks of size up to near by 200 nodes.

The above problems which are inherent in flat routing protocol are common because, in these protocols, a route is maintained as one unit in terms of single whole path from source node to destination node. This approach is suitable only for smaller networks, because of shorter routes. But for larger networks, routes become lengthier which may break more often. New routes have to be discovered and discarded frequently, resulting enormous increase in the routing overhead.

2.1 The On-Demand Hierarchy Routing Approach

In this approach, the active routes are maintained by using a hierarchy of nodes, which are on the routes. In this hierarchy, a route-path is partitioned into segments, where each segment represents a sub-route. Our model uses a two-level routing hierarchy: the high-level global inter-segment routing and the low-level local intra-segment routing. A global inter-segment routing protocol works at a higher level from the source node to destination node. A local intra-segment routing protocol works at a

lower level through the nodes within each segment. When ever it is needed, a multiple hierarchy with greater than two levels can be obtained by further partitioning the segments. But in our model we use only the two-level hierarchy. With this the route maintenance activities can be done locally within a segment.

We divide a route-path into several segments and some nodes on the route-path are chosen as way-nodes. Other nodes on the route-path are called as forwarding nodes. As we are categorizing the nodes into two-level hierarchy, which are dynamically assigned a role; we call our approach as On Demand Hierarchical Routing (ODHR). Every segment on a route contains two specific nodes extreme to the segment: start-node and end-node. Each segment begins with a way-node called start-node and ends with a way-node called end-node. These start-node and end-node of a segment are connected by a number of nodes called as forwarding-nodes. The adjacent segments share a common way-node, which act as start-node for the down-stream segment and as end-node for the up-stream segment.

The ODHR approach has a significant number of advantages: First, since routes are maintained as segment, a declined route can be repaired locally within a segment. Repairing a broken route at the level of a segment extends the life time of that route-path and prevents time-consuming, expensive global route discoveries. There fore ODHR will significantly reduce routing overhead and improves its performance. Second, the ODHR approach allows lengths of a segment on a route-path can be different. Also different route-paths can have different segment length. Thus ODHR is an adaptive hierarchy routing scheme, which is essential for MANETS of various network scenarios. For stable networks, where nodes move slowly, segments of longer length can be used for overhead reduction in maintaining the hierarchy. For unstable networks, where nodes move faster, segments of shorter length can be used for provision of route repairs. Finally, in ODHR unlike other hierarchical routing techniques, the hierarchy is built with nodes on active routes only, so that the other nodes save their resources.

Figure 2 shows an example of how the route-path from node P to node Y is divided into segments, with size 4. In this example, nodes S and V are chosen as way-nodes. Source P and destination Y are also referred to as way-nodes. The way-nodes S and V divide the route into three segments as P-q-r-S, S-t-u-V, and V-w-x-Y. The nodes which are in between the start-node and end-node of a segment are designated as forwarding-nodes. In a simple way, a segment can be represented by its start-node and end-nodes. For instance, the segment S-t-u-V can be represented as segment SV.

2.2 Using AODV and DSR in ODHR

Many existing routing protocols can be chosen as inter-segment and intra-segment routing protocols in our ODHR model. In our model, we have chosen AODV as local intra-segment routing and DSR as global inter-segment routing. Here DSR works on way-nodes between source and destination and AODV works on forwarding-nodes between way-nodes of a segment. Hence this protocol can be called as Way Directing Node Routing (WDNR) protocol. The performance comparison of AODV and DSR was specified in [4]. The reasons behind selection of AODV and DSR protocols are illustrated as follows.

- By designing WDNR, it can be made possible for DSR and AODV routing to exist in the same network. Actually, DSR and AODV are two special cases of WDNR. When the length of segment is set to big number, WDNR becomes AODV and when the segment length is set to 1, WDNR becomes DSR.
- Because of its efficient performance and its mere ability to run on larger network, AODV is chosen as intra-segment routing. Also, it allows WDNR to have longer segments. This feature makes the partition of segments in WDNR more flexible. If required, longer segments can be used so that the number of way-nodes on a route-path can be reduced.
- By combining AODV and DSR hierarchically. We can inherit the inherent strengths of both the protocols in our technique. The simulation results proves that this combination not only greatly improves the scalability of DSR but also significantly reduces the routing overhead of AODV.

3 Operation of WDNR Routing Protocol

In our instantiation of On-Demand Hierarchy Routing ODHR, DSR runs at inter-segment level, which uses the source route operation and therefore only way-nodes are listed in header of data packets. As AODV runs at the intra-segment level, overhead of it can be limited to a local range i.e., within a segment. By the combination of DSR and AODV in hierarchy, we can make DSR work for larger network. Also we can limit the routing overhead within the segment by using a separate protocol AODV in it. WDNR works with a mechanism of an on-demand basis. Even though the promiscuous mode in DSR has its own advantages, we prefer not to use the promiscuous mode in our simulation as it increases the processing overhead and battery power consumption. For illustrations the operation of WDNR, we focus on its three functions: route discovery, route maintenance and loop handling.

We define four types of control messages, route request (RRQ), route reply (RRP), route error (RER) and route activate (RACT). Route activate message is used to find a reverse path from end-node to source-node of a segment. To differentiate messages used for inter-segment and intra-segment routing, we add subscript *inter* or *intra* to messages like RRP_{inter}, RRP_{intra}.

3.1 Route Discovery

Similar to route discovery step of DSR, in WDNR the source-node discovers the route to destination, whenever it requires in an on-demand basis. Consider the route process in two steps as follows.

3.1.1 Inter-segment Route Request

The source node broadcasts an RRQ_{inter} can be uniquely identified by its contents which consist of the pair (source address, source's broadcast ID number). Like DSR protocol, the RRQ_{inter} message records the list of traversed nodes. Once the intermediate node receives the broadcast RRQ_{inter} message, it first finds whether this request is a fresh or duplicate by looking up its request-seen-table. The duplicate requests will be discarded. If the RRQ_{inter} is a new one and its TTL is not zero the

intermediate node augments its address to the route-path recorded in the RRQ_{inter} and forwards the request to all its neighbors. Intermediate node route-reply option of DSR is not enabled in WDNR, because most of the routes obtained by intermediate nodes were identified as outdated. Finally when this request reaches the destination-node, it replies to the request.

3.1.2 Inter-segment Route Reply

In DSR mechanism, for one route request sent by source node, there is no practical limit on the number of RRP_s that the destination-node can respond to the source. But in most of the paths found by RRQ messages have more number of common nodes. Generally along those paths, only the last few links are different. Based on this fact, for a single inter-segment route discovery, a limit on the number of RRP_s that the destination can send is maintained. We indicate this limit using the parameter MAX-REPLY-TIMES, whose value is set to 2 for our simulation. If the path length recorded in RRQ_{inter} message exceeds a threshold value (DEFAULT-SEG-LEN), the destination node divides the path into smaller segments by choosing way-nodes along the path. Many ways exist to select way-nodes for segmentation of routing path. We have chosen way-nodes such that, they divide the path into segments of same size. That means the hop count of each segment is approximately equal to DEFAULT-SEG-LEN. Consider for the above example shown in Figure 2, the path from a source node P to destination node Y is

$$P - Q - R - S - T - U - V - W - X - Y - Z$$

And for the value of the parameter DEFAULT-SEG-LEN is 3, a possible path division into segments will be

$$P - q - r - S - t - u - V - w - x - Y$$

This contains three segments: PS, SV and VY. Here the segment length is uniform that is 3. Other criteria can also be used for selecting way-nodes like security or speed concern. Using nodes of high secure as way-nodes makes the routes of high secure and using low-speed nodes, more stable routes can be obtained improving the performance of the routing protocol. In our approach, the parameter DEFAULT-SEG-LEN decides the number of segments a route-path is divided into.

Once the route-path has been divided, the destination constructs a RRP_{inter} message and sends back to the source. The path which includes both way-nodes and forwarding-nodes is included into the RRQ_{inter} and a strict source routing is followed by all intermediate nodes. In our example, $q - r - S - t - u - V - w - x - Y$ is placed in the RRP_{inter}. The source address P, need not be included as it appears as the destination-node in the RRP_{inter} message.

In the intra-segment routing, we need to add additional information into the RRP_{inter} message. Similar to normal AODV protocol, the reply message carries sequence number and the distance (hop-count) to the end-node of this current segment. These two fields are represented as end-seq-no and end-hop-count. They have to be used for the purpose of loop prevention, which is described in Section 4.3.3. For our example, it can be observed that the current-segment is the end segment

and the end-node is the destination-node. Therefore destination sets end-seq-no field to its own sequence number and also sets the end-hop-count field to 1.

Similar to DSR, in WDNR also, RRP_{inter} is unicast message to the source-node along the exact reverse path present in the RRP_{inter} message. Since two protocols DSR and AODV co-exist in the same network, all the nodes have to maintain two tables: routing table and route cache. Among these two, intra-segment AODV uses routing table and inter-segment DSR routing uses the route cache table.

Each entry in the route cache uses the structure (destination: source-route), similar to the way in the DSR protocol. Here the source route is the path between the current node and destination. But in this WDNR protocol, source route contains only way-nodes not all the nodes along the route-path. Therefore, two adjacent nodes on a source route-path are exactly one segment distance from each other, unlike one link away in normal DSR. Specifically, one segment will contain multiple hops, in our example it is 3.

Similar to AODV protocol, each entry present in the routing table uses the structure (end-node: next-hop). In AODV end-node specifies the destination node. But, the routing table in our WDNR stores the field next-hop to the end-node of the current segment, where the current-node is either a start-node or forwarding-node, of this segment. The main difference between routing tables of AODV and WDNR is that in AODV, the entry of destination field represents the final destination, but in WDNR it is the end-node of the current-segment but is not always necessarily be the final destination. The intermediate node, after receiving the RRP_{inter} message, it checks whether it is a way-node or forwarding-node on the reply path. If it is founds as a forwarding-node, it updates only the entries of its routing table but if it is a way-node it updates the entries in both its routing table and route cache.

For updating its route-cache, this node places source routes to the destination-node and also to downstream segment way-nodes into its route-cache. For instance in the above network, the source route is P – q – r – S – t – u – V – w – x – Y . When the nodes receives RRP_{inter}, the node V adds <Y: Y> into its route-cache, node S adds <V: V> and <Y: VY> to its route-cache. For updates the routing table, the node adds or updates an entry in its routing-table, if the intrasegment route (end-node: next-hop) in the RRP_{inter}, is fresh or better compared to existing entry. For this updated or inserted entry, the end-node is the last node of the current segment and the next-hop is the node from which the RRP_{inter} message relayed to the current node. In addition to normal fields, every entry in the routing table contains a Boolean-type field called start-node. This field is set to 1, if the current node is the start-node of current segment; else it is set to 0. This start-node field decides whether to forward or not an intra-segment route error message further.

For the above illustrated example network, where the replied route from the destination is P – q – r – S – t – u – V – w – x – Y; when the node w receives the RRP_{inter}, it checks for the route to the end node Y in its routing table. In case the route is not found or the existing route is old, then a new route to Y will be added or updated, along with next hop entry x and the hop count 2. When the RRP_{inter} receives at an intermediate way-node, which illustrates that the RRP_{inter} is leaving a segment, for which the current-node is the start-node and also that the RRP_{inter} message is entering an upstream-segment, in which the current node is end-node. Hence, this

node sets the value of end-seq in the RRP_{inter} message with its own sequence number and also resets the end-hop-count in the RRP_{inter} to 1.

Similar to AODV, at intra-segment level of WDNR, a reverse path between end-node and the start-node of a segment is required. In every segment the start-node sends a route-activate (RACTT) message to its extreme end-node, so that end-node and all other forwarding nodes of the segment will know the sequence number of and hop-count to the start-node. The message RACTT consists of two parameters represented as start-seq and start-hop-count. The start-seq field is set with the sequence number of the start-node. The initialization value of start-hop-count is set to 1 and will be increased by forwarding-nodes, during the relay of RACTT message. The RACTT message is needed because when RRQ_{inter}s were broadcasted, even though down-stream nodes can learn about list of up-stream nodes carried by RRQ_{inter}s, the route partition into segment might have not been done yet. When the RRP_{inter} reached back to source node, which initiated inter-segment route discovery, this source-node upgrades both its routing table and route-cache as illustrated above. With this, the source-node is ready to transmitting data packets along the path, just it has learned.

3.2 Route Maintenance

After the discovery of the route by the source node, data transmission takes place between source and destination nodes. Data packets from the source node reach the destination by both inter-segment routing and intra-segment routing. The data packet uses source route option, consists of a field called current-seg, which indicates index of current segment. The value of the parameter current-seg will be incremented by 1, when the data packet moves from an upstream segment to a destination segment.

Links on a route may fail, because of instability of the nodes in the ad hoc networks. To handle these breaks route maintenance mechanism is used in our protocol. A node can confirm the receipt of a packet to the neighboring node by any one of the three ways of acknowledgements: link-level, network-layer-level and passive (listening to the neighbor node forwarding). Normally in a flat routing protocol, when a link fails, error message is sent to the source node, which erases the route from its cache. The source node again rediscovers the path to destination, causing high route request overhead. But in WDNR, a route contains segments. So, a broken route-link can be repaired locally with in a segment or few segments near to the broken segment.

Route repairs in WDNR are defined in two cases: First is intra-segment route repair and the other one is inter-segment route repair. When a link is failed, first intra-segment route repair is tried and it operates within the range of the nodes in one segment. If this repair succeeds, way-nodes on the route-path need not be changed; and hence the source-node does not require notifying the change. If this repair fails, next inter-segment route repair is tried which operates over multiple segments along broken segment and its few adjacent segments towards the destination-node. If this repair along multiple segments succeeds, the new repaired source route is sent to the source-node, which uses this repaired route to send data packets from there onwards. If this repair also fails, a message of inter-segment route error is sent to the source-node, to initiate global route discovery again. Both of the inter-segment and

intra-segment repairs are to be initiated by the start-node of the broken-segment. Hence this node is designated as *initiator* and the node this initiator searches for is designated as *target*.

3.2.1 Intra-segment Route Repair

As AODV protocol is used for intra-segment routing, this route repair operates in AODV mode. If a node observes the next-hop node is not reachable; it sends an error message $\text{RER}_{\text{intra}}$ to all its predecessor nodes, which use this node as next-hop node for some segments. If multiple segments use this broken link as one of the hop, then all these segments will be broken. This $\text{RER}_{\text{intra}}$ message consists of the broken link and also the end node of these broken segments. Soon after receiving of the $\text{RER}_{\text{intra}}$, the predecessor nodes, sets the broken segments to invalidated state in their routing tables. The $\text{RER}_{\text{intra}}$ message relayed by intermediate nodes until it reaches to the start-node of all these broken segments. Upon receiving $\text{RER}_{\text{intra}}$ by start-node, it tries to repair the broken segment, with in a current segment and it becomes route repair initiator. Similar to AODV route discover process, the initiator broadcasts a route request $\text{RRQ}_{\text{intra}}$ looking for the end-node (target) of this segment and the TTL is set to MAX-SEG _LEN. Upon receiving $\text{RRQ}_{\text{intra}}$, the intermediate node forwards it, if it is not duplicate and also its TTL is greater than zero; otherwise it simply discard the message. When the end-node receives the first $\text{RRQ}_{\text{intra}}$ message, it replies $\text{RRP}_{\text{intra}}$ message back to the initiator (start-node). This reply message is uni-cast along the reverse path on its arrived path. The local route repair is said to be successful if the initiator gets the reply message.

Consider the example shown in figure 3, where original path for the segment SV is $S - t - u - V$ and the link between t and u fails. The new path for this broken segment SV, after the route repair process is $S - a - b - V$. The initiator has to buffer the data packets which are received during the route repair phase. After the success of the intra-segment route repair, buffered packets are transmitted by the initiator. In case of failure of this route repair, the initiator tries inter-segment repair and continue to keep non transmitted data packets in its buffer.

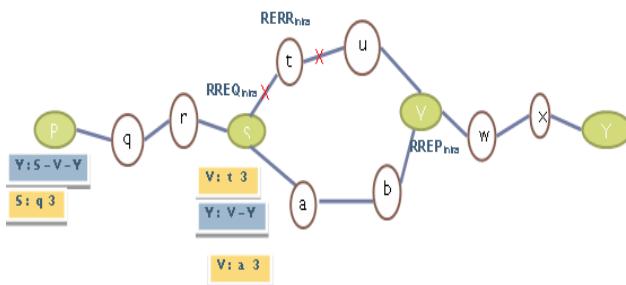


Fig. 3. Intra-segment route repair

3.2.2 Inter-segment Route Repair

When the initiator identifies, the failure of intra- segment route repair, it has to starts the inter-segment route-repair. If the timer expires and no path exists, alternate in the

broken segment, the initiator can detect the failure of intra-segment route repair and therefore begins the inter-segment rout-repair. This repair process includes discovery of routes to down-stream way-nodes along the route-path, repairing the broken route-link, and sending the repair result back to the source-node.

For this repair process and loop detection process, we use inter-segment route error (RER_{inter}) message in our WDNR protocol. Different types of message like whether route repair is successful or not, are needed by upstream way-nodes. Mainly we use RER_{inter} message to carry 3 types of information as specified below.

REPAIR-MSG: This message indicates that the existing route was broken, and inter- segment repair of route succeeded. This message carries repaired route.

BROKEN-MSG: It specifies that the route was broken and the inter-segment repair of route failed. In this case the source node will have to initiate yet another global route discovery.

LOOP-MSG: This message removes a loop existing on the route. This kind of message is used in loop detection process of loop detection.

We represent these three kinds of RER_{inter} messages as RER_{inter}^{repair} , RER_{inter}^{broken} and RER_{inter}^{loop} respectively. In inter-segment router repair, the initiator begins a localized inter-segment route discovery by broadcasting RRQ_{inter} message. This discovery process is very much similar to global inter-segment route discovery illustrated in section 4.3.1. As per the procedure, the intra-segment route repair finds a path to end-node of the broken-segment of range one segment (MAX-SEG-LEN); where as the inter-segment route repair find a path to end-node of the segment next to the broken segment of range two segments ($2 \times MAX-SEG-LEN$). For the case where broken-segment is the final segment of the route, the route discovery target will be the destination.

After receiving RRQ_{inter} message, the intermediate nodes process the message as specified in section 4.3.1. The RRQ_{inter} is forwarded if it is not a duplicate and its time Time-To-Live (TTL) field is greater than zero. After the target node receives RRQ_{inter} , it partitions the path recorded in RRQ_{inter} , provided the path is longer to DEFAULT-SEG-LEN hops. After this the target node replies an inter-segment route reply message (RRP_{inter}) to the initiator node. Once this RRP_{inter} message is received by the initiator node, it repairs that route by substituting the new segments in place of broken segments followed by few old downstream segments. The initiator node modifies its route cache and sends RER_{inter}^{repair} (repair) message, which contains the repaired route to source node through upstream way-nodes. Also, the initiator sends the buffered date messages (packets) by using the repaired route. All the upstream way-nodes, which receives the RER_{inter}^{repair} message have to update their route cache based on the repaired route path.

Using the example network shown above, we illustrate the method of inter- segment route repair the original route between source P and destination Y is

$$P - q - r - S - t - u - V - w - x - Y$$

In inter-segment level, it can be represented as a P-S-V-Y. Assume the segment SV is broken, denoted as P-S-V-Y. Here node S will be the initiator for the route repair

process. Practically intra-segment route repair will be attempted to find path for the segment SV and assume it will fail. Then the node S will try the inter-segment route repair to find the path to end node Y of the next segment VY .The node S starts localized inter-segment route discover with target Y. Assume the route discovery is success and the new path between S and Y is

$$S - a - b - C - d - e - Y$$

In this new path node C is chosen as way-node because the hop count between S and Y is longer then DEFAULT_SEG_LEN, which is 3 in our case. So, the path from S to Y is partitioned into two segments. The initiator node S repairs the old route and the new repaired route is

$$P - q - r - S - a - b - C - d - e - Y$$

At inter-segment level, using only way-nodes it can be written as P – S – C – Y. The node S modifies its route cache, by removing the routes S – V and S – V – Y, and adds the new routes S-C and S-C-Y. Afterwards node S send RER_{inter}^{repair}, repair message to the source node P, to inform the success of inter-segment route repair with repaired route P-S-C-Y. Similarly, the upstream way-nodes modify their route caches. Now, the buffered packets and hence forty new data packets will be sending through repaired route.

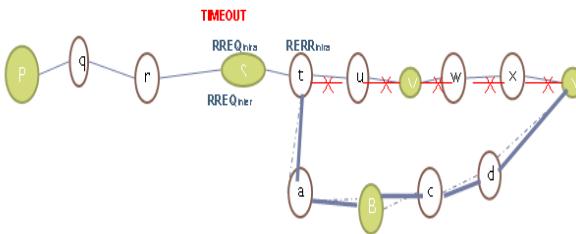


Fig. 4. Inter-segment route repair

At the start of inter-segment route repair process the initiator sets the times called Inter-Reply-Cheek. If no reply RRP_{inter} is received before the expiry of time, the initiator decides the failure of route pair. In this case, the initiator node sends RER_{inter}^{broken} message to the source node through upstream way-nodes, to inform that the route is broken and repair of route (both intra-segment and inter-segment) are failed. All the upstream-nodes remove the routes of broken segment from their individual route caches. All the buffered data packets stored at way-nodes during route repair process are deleted. In this case, it is necessary for the source node to begin another global inter-segment route discovery process to establish the route to that destination.

4 Performance Evaluation of WDNR

The performance of WDNR protocol is evaluated by conducting extensive simulation on GloMoSIM 2.3 [14], which is a scalable wireless network system simulation

environment. The simulator uses parallel discrete event capability in simulation supported by PARSEC [15].

4.1 Simulation Model and Scenarios

The main goal of our simulation is to analyze the performance of our protocol under various networks size, to evaluate its scalability. The Simulation runs by IEEE 802.11 MAC layer protocol, which uses control packets namely Request To Send (RTS) and clear to send Clear to send (CTS) for uni-casting transmissions. This protocol uses the logic of Carrier Sense Multiple Access Protocol with collision Avoidance (CSMA/CA) [11] at MAC level. The radio modal uses the two ray propagation, with radio band width of 2 Mb/Sec, in the 250m radio range. The traffic for load used in this simulation is Constant Bit Rate (CBR). The source and destination for CBR flow were selected randomly, which will not alter until the life time of a simulation scenario. Each source sends four data packets of size 512 byte per each second. The random way point (RWP) mobility model is used with speed spreads between 0 and 20 m/sec with a pause time of 30 seconds. These are default parameter values.

For DSR simulation, we disabled the option of route replies from any intermediate nodes, to improve its performance. The simulations were conducted for evaluating the performance of WDNR mainly to check its scalability. The scalability of WDNR is evaluated in networks with 100 to 1500 nodes and compared it with DSR and AODV. In our simulation, we studied four performance metrics namely, packet-delivery-ratio (PDR), average end-to-end delay, control packets overhead and number of broken links. The first three primary metrics were evaluated in every simulation set. The protocol parameter values assigned for this simulation study are as follows.

```
MAXM-REPLY-TIMES = 2
DEFAULT-SEG-LEN = 3
MAXM-SEG-LEN = 4
```

4.2 Results and Analysis

In the simulation, WDNR performance is evaluated with scalability criteria and compared it with basic protocols DSR and AODV. Appropriate network areas are used for network sizes ranging form 100 to 1500 nodes. Suitable network area as shown in Table 1 is selected to maintain approximately constant network densities, which properly forms the base for comparing scalability of these routing protocols. Four performance metrics are used for comparing WDNR with DSR and ADV and the results are shown in Fig. 5.

Table 1. Network Sizes and Terrestrial Areas

Size	Area (m x m)	Size	Area (m x m)
100	1500 x 1500	1000	4000 x 4000
250	2000 x 2000	1250	4500 x 4500
500	3000 x 3000	1500	5000 x 5000
750	3500 x 3500		

Packet delivering Ratio: Fig. 5(a) shows the DSR comparing of WDNR, DSR and AODV. It is clear from the graph that DSR is not scalable beyond 200 nodes. But WDNR and AODV exhibits high PDR even for all networks of size greater then 1000 nodes. We observe that for all cases, WDNR consistently maintain about 3% more PDR than AODV. Routes are maintained hierarchically in WDNR and broken routes are repaired locally. So, in this routings protocol, active routes last longer and hence more data packets will be delivered. AODV performance of PDR is comparable to WDNR, buy it has high control packets overhead (Fig.5b). DSR is not scalable because it generates high packet header overhead and it maintains routing information entirely in non-distributed approach.

Control over head: The main objective of using hierarchy levels in ad hoc networks is to reduce overhead of routing. Fig. 5(b) depicts the control packets overhead comparing of WDNR, DSR and AODV. It is clear that as network size increases WDNR and DSR have less overhead compared to AODV. For small network of 100 nodes, three protocol exhibits equal quantity of control overhead. For network of size 500, 1200 nodes, WDNR approximately saves 60% and 80% control packets respectively compared to AODV. DSR exhibits slightly less control packets overhead than WDNR, but its packets delivery ratio is very poor. The main reason for reduced control packet overhead in WDNR is the use of on-demand routing hierarchy. Route maintenance process can be localized in WDNR, resulting over all less over head which is critical for a protocol to scale for larger networks.

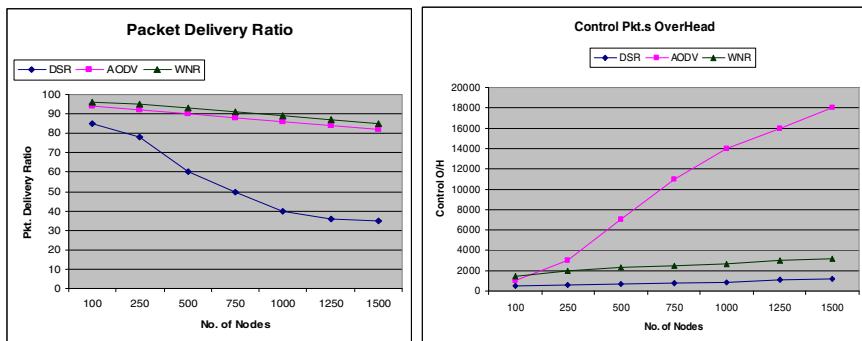
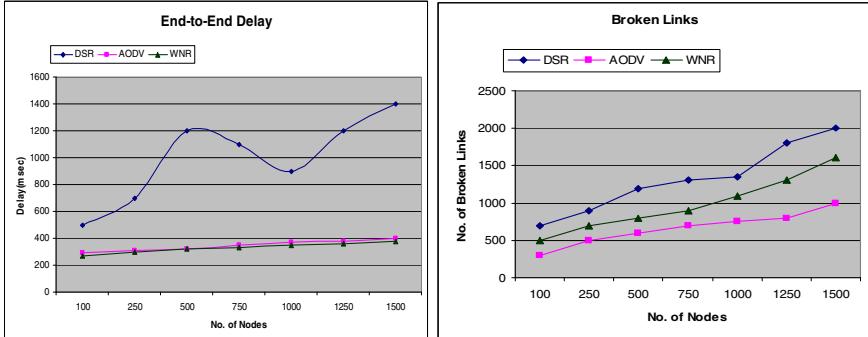


Fig. 5. Performance Comparison. (a) Packet delivery ratio (b) Control Packets Overhead

Average end-to-end delay: Fig. 5 (c) illustrates the end-to-end delay of WDNR, DSR and AODV. In many cases WDNR shows lowest average end-to-end delay. This performance metric is nearly same for both WDNR and AODV, but DSR has high time delay compared to other two protocols. In WDNR, the time delay is lower because, the broken routes are repaired locally, so that data packets need not wait for another route discovery for their transmission. The end-to-end delay does net increase as the network size increases. The main cause for this is, we used same number of CBR sources. In small networks, the length of routes on average is short but networks are congested causing the high delays. In longer networks of equal node density, the length of routes on average is long; however the networks are rarely congested.

Number of broken links: Fig. 5(d) depicts the Number of broken links comparison of WDNR, DSR and AODV. Broken links increases as the size of the network grows. This is because, the route-path length increases as network grows, causing instability of the path. It is clear that as network size increases WDNR and AODV have less number of broken links compared to DSR. In DSR, as there is no scope for route repair, the number of broken links will be more compare to other two protocols.



(c) Average End-to-End delay

(d) Number of Broken Links

5 Conclusion

In this paper we propose an On Demand Hierarchical Routing protocol, called as Way Directing Node Routing (WDNR), in which the source and destination nodes run a high level inter-segment routing approach. WDNR works in on-demand basis and maintains active routes hierarchically. WDNR partitions the routes in to segments, by selecting some of the nodes as way-nodes. In this protocol, inter-segment routing protocol works across segments globally, and intra-segments work locally among the nodes within a network. Hence, route maintenance is easy and light weight as it is done locally within a segment or a few neighboring segments. The main advantage of this protocol is, when a link on a route-path fails due to node mobility, instead of discarding entire route and rediscovering the fresh route between source and destination, the broken link can be repaired locally. This technique incurs a clear advantage in performance of the protocol in terms of all performance metrics like delivery ratio, end-to-end-delay and overhead compared to both flat routing and hierarchical routing.

We designed WDNR, using On Demand Hierarchical Routing approach by taking two well known on demand routing protocols namely DSR and AODV. Our WDNR protocol uses AODV as intra-segment routing protocol and DSR as inter-segment routing protocol. It mainly solves the scalability problem of flat routing and overhead problem of hierarchical routing. Extensive simulations were conducted in GloMoSIM to evaluate the performance of WDNR compared to DSR and AODV. The simulation results show that WDNR scales better for larger networks with higher than 800 nodes,

incurring about 50 to 70 percent less overhead than AODV protocol, while other performance metrics are comparable to basic DSR and AODV.

Our future work includes using heuristic approaches to select way-nodes like selecting nodes having stability; using common way-nodes for various active routes to facilitate maintenance of hierarchy; securing WDNR routing approach and selecting alternate routing protocols for inter-segment and intra-segment routing in WDNR.

References

- [1] Murthy, C.S.R., Manoj, B.S.: Ad hoc Wireless Networks Architectures and Protocols. Pearson education (2009)
- [2] Schiller, J.: Mobile Communications. Pearson Education (2004)
- [3] Johnson, D.B., Maltz, D.A., Hau, Y.C.: The dynamic source routing protocol for mobile ad hoc networks, IETF Internet Draft (April 2003),
<http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-03.txt>
- [4] Perkins, C.E.: Performance Comparison of two On Demand Routing Protocols for Ad hoc Networks. IEEE Personal Communications (February 2001)
- [5] Perkins, C.E., Royer, E.M.B., Chakeres, I.D.: Ad hoc On Demand Distance Vector (AODV) Routing, Internet Draft (October 2003),
<http://www.ietf.org/internet-drafts/draft-ietf-manet-aodv-03.txt>
- [6] Tarique, M., Tape, K.E.: Minimum Energy hierarchical Dynamic Source Routing for MANETs. Ad Hoc Networks 7, 1125–1135 (2009)
- [7] Hidehisa, N., Sathoshi, K., Abbas, J., Yoshiak, N., Kato, N.: A dynamic anomaly detection scheme for AODV-based MANETs. IEEE Transactions on Vehicular Technologies 58(5) (June 2009)
- [8] Ho, Y.H., Ho, A.H., Hua, K.A.: Routing protocols for inter-vehicular networks: A comparative study in high-mobility and large obstacle environments. Elsevier, Science Direct, Computer Communs. 31, 2767–2780 (2008)
- [9] Guo, S., Yang, O., Shu, Y.: Improving source Routing Reliability in MANETs. IEEE Transactions on Parallel & Distributed Systems 16(4) (April 2005)
- [10] Bai, F., Sadagopan, N., Helmy, A.: The IMPORTANT framework for analyzing the Impact of Mobility on Performance of Routing protocols for MANETs. Ad Hoc Networks Journal 1(4), 383–403 (2003)
- [11] Kumar, A., To, K.A., Pal, S., Du, S., Johnson, D.B.: Design & development of PRAN: A system for Physical Implementation of Ad hoc Network Routing protocols. IEEE Trans. On Mobile Computing 6(4) (2007)
- [12] Goodman, D.J.: Wireless Personal Communications Systems. Addison Wesley (2002)
- [13] Frodigh, M., Parkvakk, S.: Future Generation Wireless Network. IEEE INFORMATIONS COM (2000)
- [14] GloMoSim Manual Version 2.03,
<http://pcl.cs.ucla.edu/projects/glomosim/GloMoSimManual.html>
- [15] Bajaj, L., Takai, M., Tang, K., Bagrodia, R., Gerla, M.: GloMoSim: A Scalable Network Simulation Environment
- [16] Dube, R., Rais, C.D., Wang, K.-Y.: Signal Stability-Based Adaptive Routing (SSA) for Ad Hoc Mobile Networks. IEEE Personal Comm. (February 1997)

Authors:

Prof.M.Neelakantappa, B.E,M.S.,M.Tech.[Ph.D]

Professor & Head, Computer Science & Engg Department, B.I.T.S,
KURNOOL, A.P, INDIA

M.Neelakantappa received his B.E(ECE) from Osmania University, Hyderabad, M.S(SS) from BITS,Pilani and M.Tech(CSE) from JNT University Hyderabad in 1997,1999 &2005 respectively. He is working as Professor & Head in CSE Dept. of Brindavan Institute of Technology & Science College, Kurnool,AP,India.He has 14 Years of Teaching experience . He is Research scholar in Faculty of CSE in JNT University, Hyderanad. His Current Research Interest includes Computer Networks, Mobile Computing, Digital Communications, Software Engineering and Network Security.



Dr. A. Damodaram, B.Tech (C.S.E.), M.Tech. (C.S.E.), Ph.D.(C.S.E).

Professor of CSE & Director, School of C&DE, JNTU Hyderabad, AP, INDIA

Dr. A.Damodharam, received his M.Tech & Ph.D.from JNT University, Hyderabad,AP,India. He is currently working as Director of School of C&DE in JNT University, Hyderabad,AP,India. He joined as Faculty of CSE in 1989 at JNTU, Hyderabad. During his 21 years of service, he has been the Head of the Department, Vice-Principal and presently is the Director of UGC Academic Staff College of JNT University Hyderabad. Also he is been, Nominated as UGC member and NBA (AICTE) sartorial committee and a member in various academic councils of the university. He is an active participant in various social/welfare activities. And he has acted as Secretary General and Chairman for the AP State Federation of University Teachers Associations, and Vice president for All India Federation of University Teachers Associations. Presently He is the Vice President for the All India Peace and Solidarity Organization from Andhra Pradesh. His Current Research Interest includes Image Processing, Pattern recognition, and Computer Networks.



Dr. B. Satya Narayana, B.Sc,M.C.A.,Ph.D(C.S)

Professor & Chairman, Board of Studies, Computer Science, S.K. University, Anantapur, A.P, INDIA

Dr. B.Satyanarayana received his B.Sc. Degree in Mathematics, Economics and Statistics from Madras University, India, in 1985; Master of Computer Applications from Madurai Kamraj University in 1988. He did his Ph.D in Computer Networks from S.K. University, Anantapur, A.P., India. He has over 22 years of Teaching and Research experience. His Current Research Interest includes Computer Networks, Network Security and Intrusion Detection.

Web-Page Prediction for Domain Specific Web-Search Using Boolean Bit Mask

Sukanta Sinha^{1,4}, Rana Dattagupta², and Debajyoti Mukhopadhyay^{3,4}

¹ Tata Consultancy Services, Victoria Park Building, Kolkata 700091, India

² Computer Science Department, Jadavpur University, Kolkata 700032, India

³ Maharashtra Institute of Technology, Pune 411038, India

⁴ WIDiCoReL, Green Tower C- 9/1, Golf Green, Kolkata 700095, India

sukantasinha2003@gmail.com, rdattagupta@cse.jdvu.ac.in,

debajyoti.mukhopadhyay@gmail.com

Abstract. Search Engine is a Web-page retrieval tool. Nowadays Web searchers utilize their time using an efficient search engine. To improve the performance of the search engine, we are introducing a unique mechanism which will give Web searchers more prominent search results. In this paper, we are going to discuss a domain specific Web search prototype which will generate the predicted Web-page list for user given search string using Boolean bit mask.

Keywords: Search engine, Ontology, Ontology Based Search, Relevance Value, Domain Specific Search, Boolean algebra, Web-page Prediction and Index Based Acyclic Graph.

1 Introduction

Search Engine is an information retrieving system of World Wide Web (WWW) [1]. The features of the search engine have become very complex. Web page prediction is an important feature of search engine, which produces search result more accurately for a user given search string.

In this paper, we will discuss the basic idea of domain specific search from Index Based Acyclic Graph (IBAG) using Boolean bit mask. We will also describe a design and development methodology for generation of bit pattern for all the Web-pages existing in IBAG and dynamic Web-page prediction list.

This paper discusses Web-page prediction in section 2. Section 3 tells about the existing model of Relevant Page Graph (RPaG) model and IBAG model. Section 4 depicts the proposed approach. Section 5 shows some performance analysis. Finally, section 6 concludes the paper.

Definition 1. Seed URL – It is a set of base URL from where the crawler starts to crawl down the Web pages from Internet.

Definition 2. Weight Table - This table contains two columns; first column denotes Ontology terms and second column denotes weight value of that Ontology term. Weight value must be within ‘0’ and ‘1’.

Definition 3. Syntable - This table contains two columns; first column denotes Ontology terms and second column denotes synonym of that ontology term. For a particular ontology term, if more than one synonym exists then it should be kept using comma (,) separator.

2 Web-Page Prediction

Web-page prediction implies predicting proper Web-page based on the given search string. The exponential proliferation of Web usage has dramatically increased the volume of Internet traffic and has caused serious performance degradation in terms of user latency and bandwidth on the Internet [2]. Web-page prediction [3] that involves personalizing the Web users' browsing experiences assist Web masters in the improvement of the website structure and helps Web users in navigating the site and accessing the information they need. Various attempts have been exploited to achieve Web-page access prediction by pre-processing Web server log files and analyzing Web users' navigational patterns. The most widely used approach for this purpose is Web usage mining that entails many techniques like Markov model, association rules, clustering, etc. [4]. In this paper, we are going to introduce a new method of domain specific Web-page prediction using Boolean bit mask. In our approach, we are mainly using bit wise exclusive-OR operation for finding predicted Web-page list [5].

3 Existing Models

In this section, we will describe two existing models; RPaG model, IBAG model.

Definition 4. Relevance Value – It is a numeric value for each Web-page; which is generated on the basis of the term Weight value, term Synonyms, Number of occurrence of Ontology terms are existing in that Web-page.

Definition 5. Relevance Limit – It is a predefined static relevance cut-off value to recognize whether a Web-page is domain specific or not.

Definition 6. Term Relevance Value – It is a numeric value for each Ontology Term; which is generated on the basis of the term Weight value, term Synonyms, Number of occurrence of that Ontology term in the considered Web-page.

Definition 7. Term Relevance Limit – It is a predefined static relevance cut-off value for each Ontology Term.

3.1 Relevance Page Graph Model

In this section, RPaG [6] is described along with the concept of its generation procedure. Every crawler needs some seed URLs to retrieve Web-pages from World Wide Web (WWW). All Ontologies, weight tables and syntables [7] are needed for retrieval of relevant Web-pages. RPaG is generated only considering relevant Web-pages. Each node in RPaG holds Web-page information. In RPaG, each node

contains Page Identifier (P_ID), Unified Resource Locator (URL), four Parent Page Identifiers (PP_IDs), Ontology relevance value (ONT_1_REL_VAL, ONT_2_REL_VAL, ONT_3_REL_VAL), Ontology relevance flag (ONT_1_F, ONT_2_F and ONT_3_F) and Ontology terms relevance value (ONT_1_TERM_REL_VAL, ONT_2_TERM_REL_VAL and ONT_3_TERM_REL_VAL) fields information. “Ontology Relevance Value” contains calculated relevance value if these value grater than “Relevance Limit Value” of their respective domains. Otherwise, these fields contain “Zero (0)”.

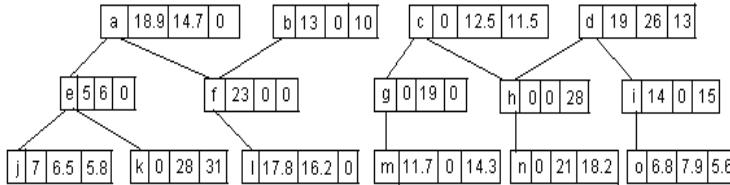
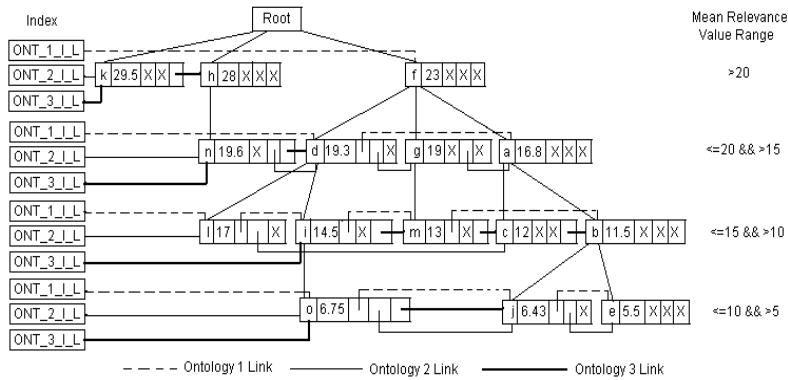


Fig. 1. Arbitrary example of Relevance Page Graph

If page P supports ‘Ontology 1’; i.e., relevance value grater than relevance limit; then ‘Ontology 1’ flag (ONT_1_F) must be ‘Y’. Same way we also define ‘Ontology 2’ flag (ONT_2_F) ‘Ontology 3’ flag (ONT_3_F). ‘Ontology 1’ each term relevance value (ONT_1_TERM_REL_VAL) of Page P is generated according to the ‘Ontology 1’. Similarly, ‘Ontology 2’ each term relevance value (ONT_2_TERM_REL_VAL) and ‘Ontology 3’ each term relevance value (ONT_3_TERM_REL_VAL) are generated according to the ‘Ontology 2’ and ‘Ontology 3’. A sample RPaG is shown in Fig. 1. Each node in this figure of RPaG contains four fields; i.e., Web-page URL, ONT_1_REL_VAL, ONT_2_REL_VAL and ONT_3_REL_VAL.

3.2 IBAG Model

An acyclic graph is a graph having no graph cycles. A connected acyclic graph is known as a tree. IBAG [8] means an indexed tree. IBAG is typically generated from RPaG. In Fig. 2, a sample IBAG is shown. RPaG pages are related in some Ontologies and the IBAG generated from this specific RPaG is also related to the same Ontologies. Each node in the figure (refer Fig. 2) of IBAG contains Page Identifier (P_ID), Unified Resource Locator (URL), Parent Page Identifier (PP_ID), Mean Relevance value (MEAN_REL_VAL), Ontology 1 link (ONT_1_L), Ontology 2 link (ONT_2_L) and Ontology 3 link (ONT_3_L) fields. Along with those fields we also transfer ONT_1_TERM_REL_VAL, ONT_2_TERM_REL_VAL and ONT_3_TERM_REL_VAL field information while generating IBAG from RPaG. Page Identifier (P_ID) is selected from RPaG page repository. Each URL has a unique P_ID and the same P_ID of the corresponding URL is mentioned into IBAG page repository.

**Fig. 2.** IBAG Model

Consider, one page supports ‘Ontology 1’ and ‘Ontology 2’; then we calculate MEAN_REL_VAL as $(\text{ONT_1_REL_VAL} + \text{ONT_2_REL_VAL})/2$. If one page supports ‘Ontology 1’, ‘Ontology 2’ and ‘Ontology 3’; then we calculate MEAN_REL_VAL as $(\text{ONT_1_REL_VAL} + \text{ONT_2_REL_VAL} + \text{ONT_3_REL_VAL})/3$. ‘Ontology 1 link’ (ONT_1_L) points to the next ‘Ontology 1’ supported page. Similarly, ‘Ontology 2 link’ (ONT_2_L) points to the next ‘Ontology 2’ supported page. ‘Ontology 3 link’ (ONT_3_L) points to the next ‘Ontology 3’ supported page. In Fig. 2, we have shown only five fields; i.e., Web-page URL, MEAN_REL_VAL, ONT_1_L, ONT_2_L and ONT_3_L. In each level, all the Web-pages’ “Mean Relevance Value” are kept in a sorted order and all the indexes which track that domain related pages are also stored.

Definition 8. Ontology – It is a set of domain related key information, which is kept in an organized way based on their importance.

4 Proposed Approach

In our approach, we have generated bit pattern for all Web-pages existing in IBAG and further have searched the Web-pages from IBAG for a given “Search String” to generate predicted Web-page list. Initially, we have generated RPAG which was constructed from typical crawling technique and then we have generated IBAG from that RPAG. After construction of IBAG we have generated bit pattern of all Web pages. Finally, a search string is given as input on the Graphical User Interface (GUI) along with other details; and as a result, corresponding list of predicted Web-page URLs is produced.

4.1 Bit Pattern Generation Algorithm

Bit pattern generation is a one time job. This job generates Web-page bit pattern for all Web-pages existing in IBAG. IBAG, Ontology terms relevance limit (OT_{limt}) and number of Ontology term (t) for the taken Ontology are considered as input for this

job. Each Ontology term holds a fixed position in the generated Web-page bit pattern. Those positions are absolutely predefined.

Method 1.1: Web-page Bit Pattern Generation
genWebpageBitPatrn (Web-page, Ontology)

1. generate dummy bit pattern with t number of 0's;
2. for each Ontology term perform 3-6;
3. fetch Ontology term limit (OT_{lmt}) for the selected Ontology Term;
4. Calculate Ontology term relevance value for the considered Web-page;
5. check Ontology term relevance value $> OT_{lmt}$ then perform 6 else goto 2;
6. change bit value 0 to 1 in dummy bit pattern for the selected Ontology term position;
7. store dummy bit pattern of selected Web-page for the corresponding Ontology;

Algorithm 1: IBAG Model Web-page Bit Pattern Generation
genIBAGWebpageBitPatrn (IBAG Model Web-page, Ontologies)

1. for each Web-page perform 2-3;
2. for each Ontology term perform 3;
3. call **genWebpageBitPatrn**(Web-page, Ontology);

For example, take one Web-page 'P' from IBAG. Now based on our Algorithm 1, first we have to generate a bit pattern of the Web-page 'P' which contains ' t ' number of bits, where ' t ' denotes number of ontology terms taken for the considered domain. Then check each ontology term relevance value with predefined ontology term relevance limit. Suppose, for Web-page 'P' 2nd and 5th ontology term exceeds the term relevance limit value. Hence the bit pattern of Web-page 'P' becomes like (0100100... t times).

4.2 Find Predicted Web-Page List

Predicted Web-page list is generated at runtime. This operation is performed for each search action. Initially we select the Web-pages from IBAG for the user given relevance range and selected Ontology. Then apply mask bit pattern to their respective Ontology to find whether the selected page belongs to predicted page list or not. Finally, predicted Web-page URLs are shown on the particular Web-page as the search result. Search string, relevance range, selected Ontology, number of search result, IBAG, Web-page bit pattern and number of Ontology terms are considered as input for this procedure

Method 2.1: Mask Bit Pattern Generation
getMaskBitPattern (Search String, Ontology)

1. extract Ontology terms in search string;
2. create a Mask Bit Pattern by taking 1's for Ontology Terms present in search string and 0's for not present in search string and length must be t;
3. return maskBitPattern;

Method 2.2: get Web-page list from IBAG based on the user given relevance range

getWebpagesFromIBAG(IBAG, relevanceRange, selectedOntology)

1. generate Web-page list by traversing IBAG for the user given relevance range and selected Ontology;
2. return Web-pageList;

Algorithm 2: Find Predicted Web-Page List

findPredictedWebpageList (IBAG, search string, relevance range, selected Ontology, Web-page bit pattern, number of search result)

1. $\beta := \text{getMaskBitPattern}$ (Search String, Ontology);
2. selectedWebpageList := **getWebpagesFromIBAG**(IBAG, relevanceRange, selectedOntology);
3. for each Web-page in selectedWebpageList perform 4-10
4. calculate $\mu = \alpha \wedge \beta$;
5. for each Ontology term in search string perform 6-9
6. if (Ontology term position in $\mu = 0$) then perform 7-9
7. add Web-page in predicted Web-page list;
8. predicted Web-page counter ++;
9. exit step-5 for loop;
10. if (predicted Web-page counter \geq number of search result) then exit step-3 for loop;
11. display predicted Web-page list;

Where,

μ = Resulted Bit Pattern

α = Traverse Page Bit Pattern for the considered Ontology

β = Mask Bit Pattern

Now, for a given search string we have to find whether Web-page ‘P’ (refer algorithm1 example) needs to be included in the predicted Web-page List or not. Based on our Method 2.1, we have to create a mask bit pattern. Suppose, 2nd position ontology term exists in search string then the mask bit pattern looks like (0100000... t times). Again we assumed that the Web-page ‘p’ belongs to user given IBAG relevance range and supports user selected Ontology. Then as per our Algorithm 2, we perform XOR (^) operation between bit pattern of Web-page ‘P’ and mask bit pattern, i.e., (0100100... t times) ^ (0100000... t times) and the resultant bit pattern becomes (0000100... t times). Now we check 2nd position of the resultant bit pattern, if it is ‘0’ then include the Web-page else discard Web-page ‘P’. In our example, 2nd

position of the resultant bit pattern showing zero (0), hence we include Web-page ‘P’ in predicted Web-page List.

5 Performance Analyses

Here we will explain our test setting and will also discuss some comparative study in our test result section.

5.1 Test Setting

In this section we have described weight table, Syntable and also explained our testing procedure. Weight table and syntable are used for calculating term relevance value. In Table 1 and Table 2 we have shown a sample weight table and Syntable for few Ontology terms.

Table 1. Weight Table

cricket	0.9
wicket keeper	0.8
umpire	0.4
bat	0.2
match	0.1

Table 2. Syntable

Match	competition,contest
Stamp	stick,wicket
Ball	conglobate,conglomerate
Umpire	judge,moderator,referee
Catch	capture

Testing Procedure. For experimental purpose, we have a set of search string, which we applied to both IBAG models; i.e., before bit masking and after bit masking, for our comparative study. First we have taken such an IBAG model which contains 1000 URLs. Now we applied all search strings to find search time taken and number of page retrieved by both models, which contains 1000 URLs in each model. Then we average search time for each model and plot the graph and also do the same for the number of pages. Same way we have taken 2000, 3000, 4000 and 5000 URLs to calculate search time and number of pages retrieved for plotting the graph. Finally from the graph we find the performance of our system.

5.2 Test Results

In this section we have generated some test results based on our test procedure and represented them by the graph plot. We have also verified accuracy of our search result after retrieval of predicted Web page list based on our given set of Search String. Accuracy measurement is determined based on some parameters like meaning of the Web page content, number of Ontology terms of that particular domain existing in the Web page content etc. Meaning of the Web page content is explained by seeing the content of the Web page and this is a manual process.

Average Number of Predicted Web Page List for a Set of Search String. In Fig. 3 we have shown average number of predicted Web pages retrieved from both IBAG

models; i.e., before bit masking and after bit masking, for a given set of search string and various relevance rage values. From the figure we found, number of Web pages retrieved from “after bit masking in IBAG Model” is lesser than number of Web page retrieved from “before bit masking in IBAG Model.”

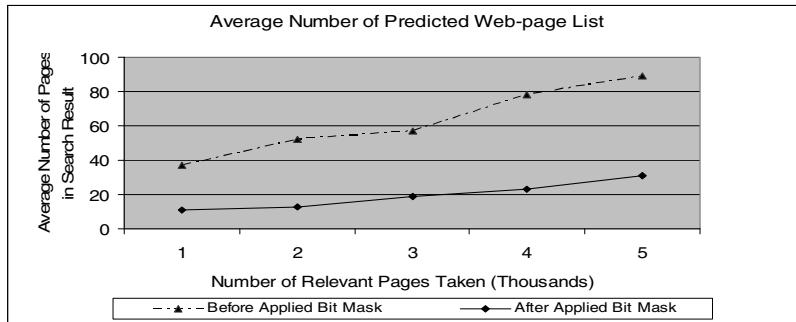


Fig. 3. Comparison between Average Number of Web-pages Retrieved from before and After Bit Masking in IBAG Model

Average Time Taken for a Set of Search String. In Fig. 4 we have shown average time taken by both IBAG models; i.e., “before bit masking” and “after bit masking,” for a given set of search string. From the figure we found that both IBAG models have taken near about same time but the accuracy of resultant predicted Web-pages list is better than before masking IBAG model.

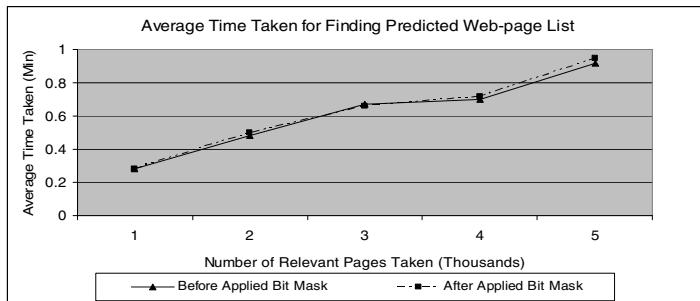


Fig. 4. Comparison between Average Time Taken for Searching Web-Pages from before and After Bit Masking in IBAG Model

Accuracy Measure. To measure accuracy we have used a metric called Harvest Rate (*HR*). We define *HR* such as given below:

$$HR: = T_{RelSR} / T_{RelSW}$$

Where, T_{RelSR} denotes average of search string term relevance value of all Web-pages exists in search result. T_{RelSW} denotes average of search string term relevance value of all Web-pages selected based on the user given relevance range. While measuring

accuracy we have chosen [Maximum Relevance Value, Minimum Relevance Value] as relevance range. Higher value HR denotes more accurate result. In Table 3, we have given an accuracy measure statistics for few search string and observed that accuracy varies on user given search string, but all the cases after bit mask we have achieved better accuracy.

Table 3. Accuracy Measure Statistics

Search String	Number of search result delivered from User Interface	Harvest Rate before bit mask	Harvest Rate after bit mask
ICC player rankings	20	0.296	1.467
	50	0.437	1.180
	100	0.296	1.063
best batsman in the world	20	0.590	2.128
	50	0.487	1.720
	100	0.744	1.462
ICC world cup 2011	20	0.358	1.490
	50	0.430	1.186
	100	0.358	1.100

Discussion of Average-Case Time Complexity for generating Search Results from both IBAG Model. To retrieve all the Web-pages in a particular level from IBAG model, we need to traverse $[(1+0) + (1+1) + (1+2) + \dots + (1+ (n/m - 1))]$ = $[1+2+3+ \dots + n/m]$ number of Web-pages. We assume that ‘n’ numbers of Web-pages are distributed in ‘m’ number of Mean Relevance Level. For finding all Web-pages from IBAG model, we need to traverse $[(1 + 2 + 3 + \dots + n/m) + (1 + 2 + 3 + \dots + n/m) + (1 + 2 + 3 + \dots + n/m) + \dots m \text{ times}]$ number of Web-pages. Now, finding a single Web-page from IBAG model in an average case scenario should be:

$$\begin{aligned}
 & (1/n) \sum_{\text{Level}=1}^m [1 + 2 + 3 + \dots + (n/m)] = (1/n) \sum_{\text{Level}=1}^m [(n/m)*(n/m+1)/2] \\
 & = \sum_{\text{Level}=1}^m [(n/m + 1)/(2*m)] = m * [(n/m + 1)/(2*m)] \\
 & = (n/m + 1)/2 < (n/m) \quad \forall n > 0, m > 0 \text{ and } n > m \approx O(n/m).
 \end{aligned}$$

Say, ‘k’ number of Web-pages selected from IBAG model for a user given search relevance range. Then the average case time complexity to retrieve ‘k’ number of Web-pages from IBAG model on which bit masking not applied is $k*O(n/m)$.

The average case time complexity for generating predicted Web-page list from IBAG model on which bit masking applied is given below:

$$p*c*k*O(n/m)$$

Where, $k*O(n/m)$ denotes average case time complexity of finding ‘k’ number of Web-pages from IBAG model based on the user given search relevance range. ‘p’ denotes number of Ontology term exist in search string. ‘c’ denotes time taken for bit operation. Generally, ‘p’ and ‘c’ are very very less than ‘k’ and $p*c*k \approx k$. That time the complexity of producing predicted Web-page list becomes $k*O(n/m)$.

6 Conclusions

In this paper, we have shown a prototype of multiple Ontology supported Web search engine, which filters search results again and again to present more accurate result to the end users. Basically it retrieves Web-pages from IBAG model. IBAG Web-pages are related to some domains, and our algorithm applied on this specified IBAG is also related to the same domains. Overall, the proposed algorithms have shown the mechanism to generate the bit pattern of all the Web-pages existing in IBAG and as a result prepare a predicted Web-page list using Boolean bit mask.

References

1. Berners-Lee, T.: *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. Harper, New York (1999)
2. Coffman, K.G., Odlyzko, A.M.: The size and growth rate of the Internet, AT&T Labs (1998), <http://www.dtc.umn.edu/~odlyzko/doc/internet.size.pdf> (retrieved May 21, 2007)
3. Khalil, F., Li, J., Wang, H.: Integrating Recommendation Models for Improved Web Page Prediction Accuracy. In: Thirty-First Australasian Computer Science Conference (ACSC), Wollongong, Australia (2008)
4. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.: Web usage mining: Discovery and applications of usage patterns from web data. *SIGDD Explorations* 1(2), 12–23 (2000)
5. Boole, G.: *An Investigation of the Laws of Thought*. Prometheus Books (2003), ISBN: 978-1-59102-089-9
6. Mukhopadhyay, D., Sinha, S.: A New Approach to Design Graph Based Search Engine for Multiple Domains Using Different Ontologies. In: 11th International Conference on Information Technology, ICIT 2008 Proceedings, Bhubaneswar, India, December 17-20, pp. 267–272. IEEE Computer Society Press, California (2008)
7. Gangemi, A., Navigli, R., Velardi, P.: The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet. In: Meersman, R., Schmidt, D.C. (eds.) CoopIS 2003, DOA 2003, and ODBASE 2003. LNCS, vol. 2888, pp. 820–838. Springer, Heidelberg (2003)
8. Mukhopadhyay, D., Kundu, A., Sinha, S.: Introducing Dynamic Ranking on Web Pages Based on Multiple Ontology Supported Domains. In: Janowski, T., Mohanty, H. (eds.) ICDCIT 2010. LNCS, vol. 5966, pp. 104–109. Springer, Heidelberg (2010)

Security Enhanced Digital Image Steganography Based on Successive Arnold Transformation

Minati Mishra¹, Sunit Kumar², and Subhadra Mishra³

¹ Department of I.& C.T., F.M. University, Balasore -19, Odisha
minatiminu@yahoo.com

² Department of Statistics, J. C. College, Kolhan University, Jharkhand
sunit.dba@gmail.com

³ Dept of CSA,CPGS,O. U. A. T. Bhubaneswar
mishra.subhadra@gmail.com

Abstract. Steganography is a process of secret communication where a piece of information or secret message is hidden in such a way that the very existence of the secret information remains concealed without raising any suspicion in the minds of the viewers and hence preventing its detection. This is generally achieved by embedding a piece of information inside another piece of innocent looking information and can be a spatial or time or transform domain method. All these methods hide information in different types of media such as text, image, audio, video etc. Amongst these varieties of available media, digital images are more commonly used for implementation of data hiding techniques because of their size and popularity. This paper uses a spatial domain LSB substitution method for information embedding and Arnold transformation is successively applied twice in two different phases in order to achieve higher security. The system is tested and validated against a series of standard gray scale images and the results thus obtained are found to be highly promising.

Keywords: Digital Image Processing, Steganography, Encryption, decryption, Arnold transformation, Cover image, Stegoimage, secret message.

1 Introduction

Covert or hidden communication is the process of hiding a piece of information in another information. There are a number of covert communication techniques such as: Cryptography, Steganography, Covert channel, Anonymity, Watermarking etc. Steganography is one of the effective means of data hiding that protects data from unauthorized or unwanted disclosure. It works by hiding secret messages into ordinary and innocent looking messages those are generally out of suspicion. Digital image Steganography procedures exploit the high capacity and widely used digital images for data hiding purposes.

A digital image is a two dimensional function $f(x, y)$ where, x and y are spatial coordinates, f is the amplitude at (x, y) , also called the intensity or gray level of the image at that point and x, y, f are finite- discrete quantities. Digital Image processing is the use of computer algorithms to perform image processing on digital images. It

allows a wide range of complex and sophisticated algorithms to be applied to digital images with ease and with a much effective way in comparison to analog signal processing [13].

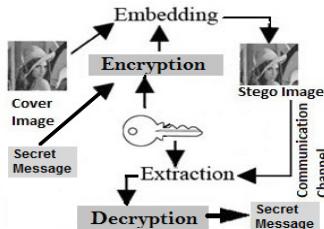


Fig. 1. General block diagrams of Steganography

The block diagram given in Fig. 1 above, describes the general procedure of image Steganography. In this process, at the transmission end, a secret message or a number of secret messages are first encrypted by the use of some suitable encryption method. Then the encrypted message(s) is/ are embedded to an innocent looking cover image to produce a stegoimage. The stegoimage, which is visually same as the original cover image is then transmitted over the communication channels without raising any suspicion in the minds of intermediate unintended receivers or viewers. At the receiving end the secret message is extracted by the authorized receiver using an extraction algorithm and a decryption process. The encryption and decryption methods used can be key-based or non-key. The key-based methods again can involve public or private keys depending upon the desired levels of privacy and security.

Steganography methods involve both spatial domain and transform domain procedures. The later procedures being more robust are commonly used for watermarking purposes while the former methods are generally used for Steganographic purposes due to their higher data hiding capacities. The least significant bit (LSB) substitution is a popular spatial domain method that replaces the lower order image bits which do not carry much useful image information by the secret message bits. In this paper we have used the spatial domain LSB substitution method for message embedding and Arnold Transformation is used successively in two phases to encrypt the message. The experimental findings show that the proposed method provides higher data hiding capacity and high security. The method also provides higher imperceptibility, the other important attribute of Steganography and preserves the quality of the cover.

2 Arnold Transformation

Arnold transformation or popularly known as Arnold's cat map is a chaotic map which when applied to a digital image randomizes the original organization of its pixels and makes the image chaotic or indistinguishable. It is a periodic mapping with a period p and if iterated p number of times, the original image reappears [5].

Definition: Arnold's transformation is a chaotic map from the torus into itself. Considering the torus T^2 as the quotient space R^2/Z^2 , the transformation $\Gamma : T^2 \rightarrow T^2$ can be given by the formula:

$$\Gamma : (x, y) \rightarrow (2x + y, x + y) \text{ mod} 1 \quad (1)$$

That is, if $P(x, y)$ is point in a unit square then applying Arnold's transformation to it, it can be transformed to another point $P' (x', y')$ such that:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} (\text{mod } 1) \quad (2)$$

Generalizing from unit square to a square of $N \times N$, equation (2) can be rewritten as:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} (\text{mod } N) \quad (3)$$

Where, N is the order of the digital image and $x, y \in \{0, 1, 2 \dots N - 1\}$. [8] Let the transform matrix in the equation (2) be A . If $(x, y)^T$ is the input and $(x', y')^T$ is the output then, the iterative procedure of the transformation can be given by:

$$\left. \begin{array}{l} P_{xy}^{n+1} = AP_{xy}^n (\text{mod } N) \\ P_{xy}^n = (x, y)^T \end{array} \right\} \quad (4)$$

Where, $n = 0, 1, 2 \dots$ is the number of iterations and with each iteration the image information move from one position to another within the image matrix. When, in this way, all the points of an image are manipulated once, a new image is produced [5].

Digital Image encryption can be achieved through Arnold's transformation following the steps as given below:

STEP1: Applying t times ($t \in [1, p]$) Arnold's transformation to all the pixels of a given $N \times N$ digital image (I). Where, p is the transform period of I.

STEP2: Obtain an encrypted scrambled image Γ .

STEP3: Obtain the original image back by applying $(p - t)$ times Arnold transformation to Γ

3 Proposed Method

The proposed method runs in two phases: the embedding phase and the extraction phase. During embedding, the secret message is first scrambled using Arnold transformation at two different levels to make it more secure against unauthorized extraction. In the first level the secret message/image is sliced into m messages/images of smaller sizes. Here m is a key that only will be known to the authorized participants. Then each smaller image is scrambled t' times using Arnold transform, where t is the period of each smaller image and $t' \in [1, t]$. Then these scrambled smaller blocks of the secret messages/ images are joined together in a certain order O (which again is only known to the authorized participants) to construct a bigger

scrambled image S . In the second level, again Arnold transform is performed on S for p' times and another further scrambled image S' is obtained, where p is the period of S and $p' < p$. Now this scrambled message is embedded into the cover image C to generate the stegoimage C' . C' is transmitted and at the receiving end the secret messages are extracted by following the extraction and decryption process. In this technique, the keys- n, p, p', t, t', O are kept secret between the authorized participants and extraction without valid keys results with noises only, making the procedure highly secure against unauthorized access. This proposed encryption process can be depicted through figure 2 as given below:

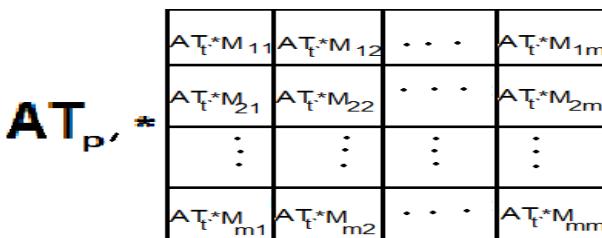


Fig. 2. Proposed encryption technique through Arnold transformation

Where $AT_t^*M_{ij}$ represents application of t times Arnold transformation to M_{ij} and M_{ij} are the partitioned matrices of the original matrix M .

3.1 Embedding Algorithm

INPUT: Cover image C of size $N \times N$. Secret messages/Images, let say, S_1, S_2, S_3 of $N \times N$ blocks, Keys: n, t, t', p, p' .

For each message/ image S_i , do step1 to 4

STEP1: Divide S_i into m blocks (M_1, M_2, \dots, M_m) of size $n \times n$ ($n < N$).

STEP2: Apply Arnold Transform to each M_i for t times ($0 < t < t'$, Where t is the period of M_i) to obtain scrambled blocks M_i' .

STEP3: Combine all the M_i' 's to get S_i' .

STEP4: Apply p' times Arnold transform to S_i' , where $0 < p' < p$ (the period of S_i') to obtain scrambled image/message S_i'' .

STEP5: Embed the scrambled messages/ images S_i'' to the LSB planes of the cover image C to get the stegoimage C' .

3.2 Extraction Algorithm

INPUT: Stegoimage C' of size $N \times N$.

Keys: p, p', t, t', n

STEP1: Retrieve S_i'' 's from C' .

For each S_i'' do

STEP2: Apply $(p - p')$ times Arnold transform to get back S_i' .

STEP3: Split S_i' into M blocks of size $n \times n$ to get back M_i' 's

STEP4: Apply $(t - t')$ times Arnold to each M_i to get back M_i s

STEP5: Join the M_i s, in order, to get back the secret message S_i

4 Results and Discussions

The proposed method is tested and validated taking a range of different standard gray scale images of size 128 x 128 including ‘woman’, ‘Lena’, ‘Baboon’ etc. as cover image. The secret messages/ images used for embedding include ‘Logo’, ‘Lion’, ‘Scorpio’ etc. - a number of binary images of size 128 x 128. Figure 3b and 3c show the stego images of original woman image after the secret messages are encrypted using the proposed method and embedded into three bit planes and four bit planes respectively, starting from the LSB plane of the cover image given in figure 3a. Figure 4 and figure 5 show the retrieved information without and with valid keys respectively. It is clear from the figures that the information retrieved without valid keys are completely random, undetectable and unsuspicious.



Fig. 3. Original woman Image and woman Image after embedding text into three and four LSBs respectively



Fig. 4. Three LSB planes of the Stego-woman Image retrieved without use of valid keys

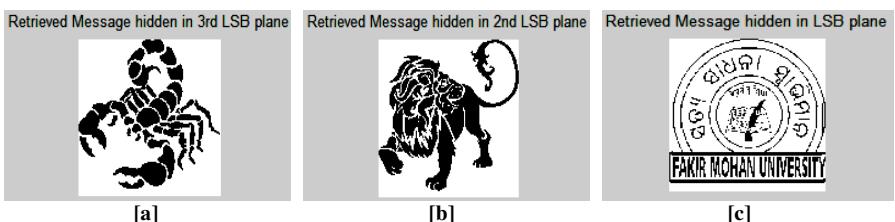


Fig. 5. Retrieval of information using valid keys [Keys: $n=64$, $p=64$, $p'=24$, $t=32$, $t'=20$]

In case of simple Arnold transformation based method, as Arnold transform is periodic in nature, the information can be retrieved by running the algorithm for a

certain number of iterations somewhere between 1 to $3p$ and observing the outputs even without knowing the period p of the image and p' - the number of times Arnold transform is initially applied to it. But in this proposed method the secret information remains highly secured and undetectable as the procedure involves a number of keys. It has been seen that it is not possible to reach at images of figure 5(a, b, c) by applying Arnold transformation to images given in figure 4 for a random number of iterations. Since the original messages in this case are nothing but another set of scrambled images therefore a random number of iterations at best will generate the initial noisy image from where the second phase has started but not at the actual message, retaining the secrecy of the messages against hit and trial extractions by unauthorized recipients. The message security against hit and trial retrieval can further be improved by introducing the following changes, as given in figure 6, to the proposed method:

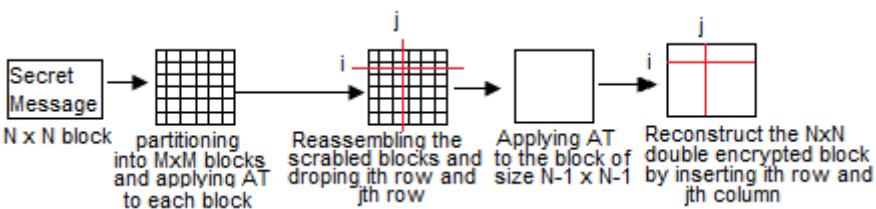


Fig. 6. Message encryption with removal of a selected row and a column

In step 3 of this modified approach, there are n^2 possible ways for i and j value selections which can add further security against hit and trial retrieval. This can further be made complex by using different Arnold Transforms, as given in equation (5) selecting different values for k , in step 2 and step 4 (of figure 6) instead of the simple AT of equation (3). [11]

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} k+1 & 1 \\ 1 & k \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} (\text{mod } N) \quad (5)$$

Where, $k \in \{1, 2, 3, \dots\}$

The data hiding capacity of the proposed method is also much higher in comparison the single LSB substitution method. Comparison of this method against simple 3 bit substitution method without encryption and encryption with simple Arnold Transform method can be summarized as follows:

Table 1. Comparison of Image Steganography Methods

Features	Simple LSB substitution method	Simple Arnold Transform method	Proposed method
Impeccability	High	High	High
Capacity	High	High	High
Security & hit and trial retrieval complexity	Low	Medium	High
Encryption	Low	Medium	High

It has also been observed that the bit preservation ratio of the proposed method is better (most of the times) in comparison to the methods involving unscrambled data insertion. The following table II gives the bit preservation values when same messages are embedded into different cover images using two different methods. It is clear from the values that the bit preservation ratio is almost 50% in each bit plane. This proves that the distortion to the original image is minimized against the unscrambled three bit substitution methods. The Pick Signal-to-noise ratio (PSNR) values after embedding data into 1, 2, 3, 4 bit planes of various test images are given in table-III, which show that the PSNR values are within acceptable range (acceptable range: 30-50 dB) even with 4-bits data insertion. Further, it has been discovered that the quality of the stego image is depended upon the cover and the message embedded and most of the time, in our example images, the PSNR value increased when the negative of a message is embedded instead of the image itself. In table III, the values given in the second rows against each image represent the PSNR when the negatives of the respective secret messages are embedded.

The formula used to calculate PSNR is as given in equation (6):

$$PSNR = 10\log_{10}\left(\frac{Max_i^2}{MSE}\right) \quad (6)$$

Where, Max_i is the maximum pixel intensity of the image. Many a time it is taken to be 255 for grayscale images as for those the maximum possible pixel intensity value is 255. But in our case we have taken this equal to the actual maximum pixel intensity of the image so as to get more accurate results. MSE is the mean squared error and the formula for MSE calculation is as given in equation (7):

$$MSE = \frac{1}{m.n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - J(i, j)]^2 \quad (7)$$

Where, I and J are the image pairs of size $m \times n$ for which PSNR is to be calculated. In our case m and n both are equal to 128.

Table 2. Bits Preservation Ratios

Image (128x128) size	Unscrambled message embedding method			Proposed method		
	LSB	2nd bit from LSB	3rd bit from LSB	LSB	2nd bit from LSB	3rd bit from LSB
Lena	8102	8165	8128	8234	8213	8182
Baboon	8828	8646	7502	8630	8698	7626
Miera	8148	8082	8139	8154	8106	8141
Siela	8059	8042	8348	8255	8280	8285
Munmun	8170	8110	8067	8224	8223	8156

Table 3. Text inserted into number of bit planes Vs PSNR

Image (128 x 128)	Embedding data into			
	One Bit	Two Bits	Three Bits	Four Bits
Tire	51.2396	44.4244	38.1998	30.4662
	-ve	51.3809	43.9752	37.4337
Baboon	51.3797	43.6852	37.0031	30.8176
	-ve	50.9058	43.3416	37.5430
Woman	50.5818	43.3098	36.6490	30.2325
	-ve	50.6475	43.3117	36.7118

5 Conclusions

Steganography systems are generally attributed by imperceptibility, capacity and robustness. These requirements being application dependent may vary from application to application and therefore, it is not that an easy task to develop a method that satisfies all these three requirements. In this paper we have implemented an algorithm that provides high capacity, high imperceptibility & high security. But, because this is a spatial domain LSB substitution method, it is not providing robustness against statistical attacks and image manipulations. However, we are trying to develop methods those will satisfy robustness and at the same time will provide even higher capacity and better security.

References

1. Reddy, A., Chatterji, B.N.: A new wavelet based logo-watermarking scheme. Pattern Recognition Letters 26, 1019–1027 (2005)
2. Piva, A., Barni, M., Bartolini, F., Cappellini, V.: DCT-based watermark recovering without resorting to the uncorrupted original image. In: Proc. IEEE Int. Conf. Image Processing (ICIP 1997), pp. 520–523 (1997)
3. Jain, A.K.: Fundamentals of Digital Image Processing. PHI (2005)
4. Chand, B., Majumdar, D.D.: Digital Image Processing and analysis. PHI, New Delhi (2002)
5. Zhang, C., et al.: Digital Image watermarking with Double encryption by Arnold Transform and Logistic. In: Fourth International Conference on Networked Computing & Advanced Information Management, pp. 329–334. IEEE Computer Society (2008)
6. Petitcolas, F.A., Anderson, R., Kuhn, M.: Information hiding: A survey. In: Proceedings of the IEEE 87,1062–1078 (July 1999)
7. Yu, G.J., Lu, C.S., Liao, H.Y.: Mean quantization-based fragile watermarking for image authentication. Optical Engineering 40, 1396–1408 (2001)
8. http://en.wikipedia.org/wiki/Arnold%27s_cat_map
9. Meena, M.K., et al.: Image Steganography tool using Adaptive Encoding Approach to maximize Image hiding capacity. IJSCE 1, 7–11 (2011)
10. Mishra, M., et al.: Steganography: the art of Secret Messaging through Digital Images. In: Proceedings of National Conference on Computational Intelligence and its Applications (NCCIA-2007), pp. 104–113 (July 2007)

11. Mishra, M., Routray, A.R., Kumar, S.: High Security Image Steganography with modified Arnold's cat map. IJCA 37(9), 16–20 (2012)
12. Johnson, N.F.: Steganography: Seeing the Unseen. George Mason University,
<http://www.iitc.com/stegdoc/sec202.html>
13. Gonzalez, R.C., Wood, R.E.: Digital Image Processing, 2nd edn. PHI, New Delhi (2006)
14. Zhao, X.F.: Digital image Scrambling based on baker's transformation. Journal of Northwest Normal University (Nature Science) 39, 26–29 (2003)

Impact of Bandwidth on Multiple Connections in AODV Routing Protocol for Mobile Ad-Hoc Network

K.G. Preetha¹, A. Unnikrishnan², and K. Paulose Jacob³

¹ Department of Information Technology,
Rajagiri School of Engineering & Technology, Rajagiri valley, Cochin, India
(Research Scholar, Cochin University of Science & Technology, Cochin, India)

preetha_kg@rajagiritech.ac.in

² DRDO, Cochin, India

unnikrishnan_a@live.com

³ Department of Computer Science, Cochin University of Science & Technology, Cochin, India
kpj@cusat.ac.in

Abstract. Mobile Ad-hoc Networks (MANETS) are infrastructure less self-organizing and self-configuring network. The nodes in MANETS are highly mobile and the routing mechanisms of these may vary depending on various applications. This poses serious challenges to routing and reliability. Reliability and reachability are the two important terms related to dissemination of data in MANETs. This paper studies the AODV protocol in MANET. And also try to understand the multiple connections in MANET under various situations. Several issues have been identified in the creation of multiple connections. The prominent issue recognized is the bandwidth of a node in creation of multiple connections. This poses major concern in connection establishment and connection maintenance. This paper also proposes a new idea of improving the simultaneous multiple connections with the consideration of bandwidth of each node.

Keywords: MANET, AODV, Multiple connections, Bandwidth.

1 Introduction

The increased usage of portable devices and advanced computing devices raised the importance of wireless and mobile computing. The term ‘mobile’ can turn the dream of networking at any place and at any time into reality. Mobile Ad-hoc network (MANET) is isolated, stand-alone networks with no connection to the internet. In MANETs all nodes are considered as source or router and the control of the network is distributed among nodes [1].

A major challenge lies in MANET communication is the unlimited mobility and more frequent failures. Because of frequent short lived disconnections the chances for collisions, transmission errors and the probability of missing data is more in MANETs. Researchers have developed many routing protocols in MANET [14] which can be classified as proactive and reactive. In the proactive or table driven protocol the up-to date routing information is maintained in each node and it is

independent of the requirement. In reactive or on demand protocol the route discovery process is done only when it is needed.

Ad-hoc On-demand Distance Vector (AODV) protocol [3] is one of the important on demand routing protocol in MANET. Researches towards the improvement of routing protocols have been going on for about a decade now. The objective of this paper is the study of AODV protocol and how multiple connections are implemented in AODV. This paper also exploits the importance of bandwidth of a node in multiple connections and also tries to propose an idea for improvement on the protocol. This ensures the maximum possibility of connection establishment and maintenance without any interference.

The rest of this paper is organized as follows. Section 2 gives how the data is disseminated using AODV. Section 3 describes the multiple connection scenarios in MANET. The current status and issues of the scenario and the impact of bandwidth is explained in section 4. The proposed improvement is discussed in section 5. Future enhancement is given in section 6 and a conclusion is given in section 7.

2 Dissemination of Data Using Ad-Hoc On-Demand Distance Vector Routing Protocol

Proactive or table driven routing protocol periodically propagates routing information to its neighbors. Each node updates its routing table accordingly. This creates an extra overhead in the network. Reactive or on demand protocols on the other hand find its root only when it is needed. This reduces the overhead but increase the time delay for determining the root. Reactive protocol is better than proactive in terms of routing overhead, packet delivery ratio and energy efficiency. Hybrid protocols combine the advantages of both table driven and on demand routing protocols. AODV, DSR, TORA are the examples of on demand routing protocols. This paper exploited the perception of on demand routing protocol AODV. Lots of researches have been done for the improvement of AODV with single connection only. The objective of this paper is the study of AODV, which supports multiple connections and the importance of bandwidth of a node in multiple connections.

AODV [2] is the extension of DSDV protocol. DSDV is a proactive protocol whereas AODV is reactive. In DSDV each node maintains complete routing table and update this information periodically. But in AODV each node calculates the route on an on demand basis. This is a type of next hope routing protocol. Each node keep a routing table contains next hope to be used to reach destination. On receiving data packet each node checks its routing table whether a valid route to the destination exists then only it forwards the packet to the next hope.

There are mainly two important phases in AODV protocol. These are Route discovery phase and route maintenance phase [1]. During the route discovery phase, source node floods a route request (RREQ) packet in the network. If a node receives RREQ packet and if it has the destination information then it send a route reply (RREP) message to the sender. Otherwise it sends the same RREQ packet to its neighbors. If there is no reply within certain time interval the source node assumes

that there is no route to the destination is available. In this case, after a certain interval of time source node again broadcast the RREQ message for route discovery. In order to avoid the duplication, each request contains a sequence number. When a node receives a RREQ message, it checks whether the sequence number has not seen before. Otherwise it discards the message. If the node have not seen the sequence number before then it sets up a reverse path towards the source node. The intermediate node updates its routing table with the address of the neighbor from which the first copy of the broadcast packet received. When this request reaches the destination it can send the reply through this reverse path. After that the data can disseminated through the reverse path is called the forward path. Each node is associated with a route timer. The forward path is deleted when the timer expires. Figure 1 represents the route discovery process in AODV.

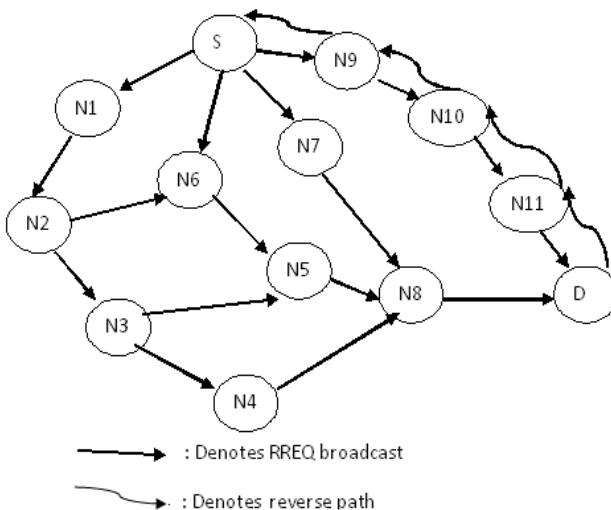


Fig. 1. The figure shows the route discovery process using AODV

The second phase in the AODV is route maintenance phase. In order to maintain the route each node periodically sends a HELLO packet to its neighbors. A node receives a HELLO packet it update its routing table as the originator of the packet is its neighbor. If HELLO packet reaches safely then the node send back the replay message. The failure of receiving HELLO packet indicates that the neighbor has moved away. So the linked to this neighbor is marked as broken in the routing table.

The main advantage of AODV over proactive protocols like DSDV is significant reduction of routing overhead. The ADOV achieve this by it's on- demand property for route discovery and route maintenance. The disadvantage of this protocol is the time delay of discovering routes. Whenever the source wants to send a message it first establishes the route to the destination. In DSDV all the updated routing information is readily available with each node. So there will not be any route calculation delay.

3 Multiple Connections in MANET-An Overview

Scientists from all over the world have been working on AODV algorithm from the beginning of this millennium to enhance its performance. Most of the researchers tried to improve the routing efficiency by considering only single connection. In all improvements they assume that only one connection between two nodes exists at a particular time. In real time scenario this is not true always. In multiprocessing environment one node can communicate with more than one destination.

Theoretically both single connection and multiple connection establishments are same [15]. Each connection is considered as a separate one. When any node moves from the network then the same connection establishment algorithm is implemented for each connection separately. Example of multiple connections is depicted in the figure 2.

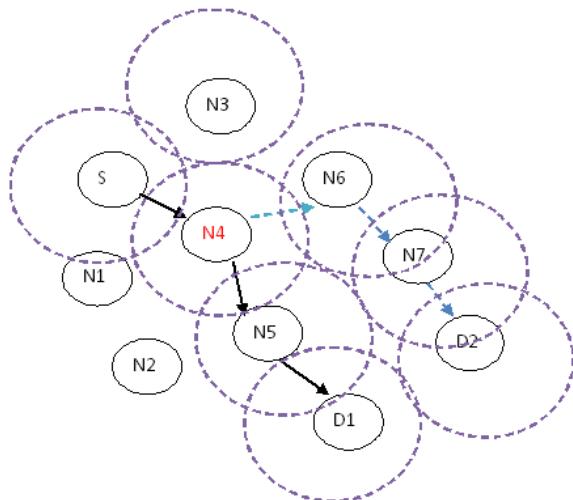


Fig. 2. The figure shows multiple connections from source S

The above figure shows the multiple connections established from source S to two destinations D1 and D2. N1 to N7 are the intermediate nodes in the network. Dotted circle represents the coverage area. S maintains a connection to D1 through the path $\langle S \text{ } N4 \text{ } N5 \text{ } D1 \rangle$ and to D2 through the path $\langle S \text{ } N4 \text{ } N6 \text{ } N7 \text{ } D2 \rangle$. So the source S simultaneously communicating to D1 and D2. If any intermediate node moves away from the network then reestablish the rout. Most of the proposed solutions for the route establishment problem are considering multiple connections in the network as single separate connections. There are so many issues arising due to multiple connections which will be discussed in the next session.

4 Issues and Challenges to Achieve Multiple Connections

Major challenge faced by the researchers in multiple connections is efficient implementation of simultaneous connection establishment. When a node moves from one

location to another the route should be reestablished. Most of the time simultaneous reestablishment creates so many problems. Reestablishment also depend on computational power, battery power and energy considerations. Simultaneous attempt to establish more than one connection could deteriorate the performance of the network adversely. If more than one connection exists from a node in the network then which connection should establish first is an important question. In order to avoid this crisis we can assign priorities to each process. Real time process has given higher priorities.

Other issue in the multiple connections is the bandwidth of a node. Each connection follows an entirely different path then it is similar to a single connection. In this battery power and computational power of the nodes can be utilized in its maximum. But if more than one route passing through the same node then the situation is different. Let us consider the figure 2. In that Node N4 is common node in the route of two connections S→D1 and S→D2. This is treated as a serious problem in the multiple connection environments. In this case bandwidth of the node is shared for different connections. In some cases, bandwidth of the node is not sufficient for all connections passing through it. In the route discovery process we have to consider the bandwidth of the node is a metric. If the bandwidth is ample enough to withstand all the connections then the connection is established. Otherwise search another path.

The above mentioned issues have been the inspiration of developing a new approach to establish the multiple connections in MANET.

5 Proposal for Improvement

Multiple connections in MANET aim at improving the performance of the network by introducing simultaneous route in the network. With this, one node can do simultaneous activities within the network at the same time. There are so many issues which have

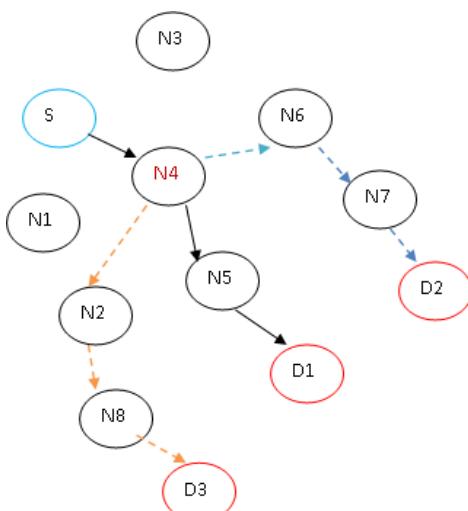


Fig. 3. The figure shows the impact of bandwidth in multiple connections

been discussed in the previous session regarding multiple connections. It has been observed that the bandwidth of a node is having a vital role during the creation of multiple connections in the network. This proposal gives a novel idea of selecting of node during the route discovery process by considering the bandwidth. In this new scheme, before establish a new route check whether the intermediate node is already participate in any other route. If so calculate its routing capability according to the number of routes it involved. In figure 3, node N4 is involved in three separate connections.

In the figure3, node S starts route discovery process for establishing route to different destinations. Node N4 is common for all routes from S. Consider the situation that there are two routes already existing to D1 and D2. Node S again wants to establish a route to D3. The optimal route to D3 is again passing through the node N4. But bandwidth of N4 is not sufficient to support all three connections. If this connection is established then this will disturb all other connections.

In order to avoid the problem mentioned above, when a node wants to establish a route to the destination then each node has to check the capability of intermediate nodes in the path before establishing the path. If the intermediate node is capable enough to withstand all the connections including the new one, then the route is established. Otherwise this node is discarded and route requesting process is continued to search another route even if this route is the optimal one. For achieving this idea each node includes the number of connections <Connection_No> of other nodes in the routing table. This represents the number of connection in which a particular node involves. Set a certain threshold <Root_limit> for the connections of each node. This indicates the maximum capability of the node that can support the number of connections. Before establishing the connection, node has to check whether all intermediate nodes are capable of supporting one more connection. If so, connection is established and the connection count is incremented by one and sends this new information to each of its neighbors. Then each node can update its routing table accordingly.

In the above example N4 is the intermediate node of two existing connections S→D1 and S→D2 then the routing table for S is given in Table 1.

Table 1. Routing table of node S while establishing a path

Node	Next_Hope	Connection_NO
N1	N2	0
N2	N8	1
N3	N6	0
N4	N2	2
N5	D1	1
N6	N7	1
N7	D2	1
N8	D3	1

If node S wants to establish a new connection to D3 by using AODV algorithm then it first flood RREQ message in the network. Then the RREP message came back through the reverse path <D3 N8 N2 N4 S>. In normal AODV the forward path is

established and data can send along this path. In this proposal, after the reverse path is established source node checks whether the connections of all the intermediate nodes in the reverse path is within the threshold. If any node in the path has exceed the limit then it marked as false node and again the route discovery process starts from the beginning. In the above example root_limit is 2. So the connection <S N4 N2 N8 D3> could not establish because N4 exceeds the maximum limit. In this case S again starts to find another root in the network.

This approach always considers the bandwidth of each intermediate node in the path in the route discovery phase. This is the main advantage of this proposal and this method ensures the maximum bandwidth of a node in the existing routes so as to maintain the existing route without any interference. In this if any new connection is established then the number of connections in each node in the path is incremented by one and floods this information to the network. All other nodes should update its routing table accordingly. This leads a small routing overhead than the normal AODV.

6 Future Enhancements

It has been demonstrated an idea to improve the multiple connections in AODV with the consideration of bandwidth of intermediate nodes in the route. In this approach assign the maximum number of connections that can be afford to each node. In addition to this limit, in future the bandwidth approach can consider not only the number of connections but also includes the probability of selecting the next node. If more than one node competes for a particular destination then select only one node depends on the packet delivery ratio and rate of packet discarding.

7 Conclusion

This paper pointed out the various issues in multiple connections in MANET. A lesser amount of bandwidth of an intermediate node in multiple connections is identified as a major problem in the connection establishment process in MANET. The improvement in AODV algorithm outlined above is always enhancing the performance in connection establishment process. The note to be kept is that the usage of this algorithm should be encouraged in all sorts of communications, so that it results in more beneficial wireless networks at all levels. The usage of this algorithm will lead to extend the connectivity in wireless networks thus creating an environment of communication in an easy way.

References

1. Aggelou, G.: Mobile Adhoc Networks. Tata McGraw-Hill (2009) ISBN:13:978-0-07-067748-7
2. Perkins, C.E., Royer, E.M.: Ad-hoc on-demand distance vector routing. In: The Proceedings of IEEE WMCSA, pp. 90–100 (1993)

3. Bai, R., Singhal, M.: DOA- DSR over AODV routing for Mobile Ad-hoc Networks. *IEEE Transactions on Mobile Computing* 5(10) (October 2006)
4. Macker, J., Corson, S.: Mobile Ad-hoc NETwork (MANET). IETF Working Group Charter (1997), <http://www.ietf.org/html.charters/manet-charter.html>
5. Broch, J., Maltz, D.A., Johnson, D.B., Hu, Y.-C., Jetcheva, J.: A Performance Comparison of Multi-Hop Wireless Ad Hoc Network Routing Protocols. In: Proc. MobiCom, pp. 85–97 (1998)
6. Jiang, H., Garcia-Luna-Aceves, J.J.: Performance comparison of three routing protocols for ad hoc networks. In: Proceedings of the Tenth International Conference on Computer Communications and Networks, pp. 547–554 (October 2001)
7. Perkins, C.E., Royer, E.M.: Ad-Hoc on Demand Distance Vector Routing. In: Proc. 2nd IEEE FVksp. Mobile Computing and Applications, pp. 90–100 (February 1999)
8. Bansal, M., Barua, G.: Performance Comparison of Two On Demand Routing Protocols for Mobile Ad hoc Networks. *IEEE Personal Communication* (2002)
9. Tomar, G.S.: Modified Routing Algorithm for AODV in Constrained Conditions. In: Second Asia International Conference on Modelling & Simulation. IEEE (2008)
10. Viswanath, K.: The Adaptive Routing For Group Communications in Multi-Hop Ad-Hoc Networks. The Dissertation, University of California, Santa Cruz (June 2005)
11. Tan, L., Yang, P., Chan, S.: An Error-Aware And Energy Efficient Routing Protocol In Manets. In: Computer Communications and Networks, ICCCN 2007 (2007)
12. Liu, T., Liu, K.: An Improved Routing Protocol in Mobile Ad Hoc Networks. In: IEEE 2007 International Symposium on Microwave, Antenna, Propagation, and EMC Technologies For Wireless Communications (2007)
13. Hubaux, J.-P., Gross, T., Le Boudec, J.-Y., Vetterli, M.: Toward self-organized mobile ad-hoc networks: the terminodes project. *IEEE Communications Magazine* (2001)
14. Royer, E.M., Toh, C.-K.: A review of current routing protocols for ad-hoc mobile wireless networks. *IEEE Personal Communications* (1999)
15. Yadav, M., Rishiwal, V., Arya, K.V.: Routing in Wireless Adhoc Networks: A New Horizon. *Journal of Computing* 1(1) (December 2009), <https://sites.google.com/site/journalofcomputing>, ISSN: 2151-9617
16. Perkins, C., et al.: Ad Hoc On Demand Distance Vector (AODV) Routing, draft-ietf-manet-aodv-10.txt (January 19, 2002)
17. Toh, C.K.: A Novel Distributed Routing Protocol to Support AdHoc Mobile Computing. In: Proceedings of the 1996 IEEE Fifteenth Annual International Phoenix Conference on Computers and Communications, pp. 480–486 (March 1996)
18. Espes, D., Teyssie, C.: Approach for Reducing Control Packets in AODV-Based MANETs. In: Fourth European Conference on Universal Multiservice Networks, ECUMN 2007, pp. 93–104 (February 2007)
19. Frikha, M., Ghandour, F.: Implementation and Performance Evaluation of an Energy Constraint Routing Protocol for Mobile Ad - Hoc Networks. In: The Third Advanced International Conference on Telecommunications, AICT 2007 (May 2007)
20. Rehman, H., Wolf, L.: Performance Enhancement in AODV with Accessibility Prediction. In: IEEE International Conference on Mobile Ad hoc and Sensor Systems, MASS 2007, October 8-11 (2007)
21. Safa, H., Artail, H., Karam, M., Ollaic, H., Abdallah, R.: HAODV: a New Routing Protocol to Support Interoperability in Heterogeneous MANET. In: IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2007, May 13-16 (2007)

22. Papadimitratos, P., Haas, J., Sirer, E.G.: Path set selection in mobile ad hoc networks. In: Proceedings of IEEE MobiHoc, pp. 1–11 (2002)
23. Kim, Y., Dehkanov, S., Park, H., Kim, J., Kim, C.: The Number of Necessary Nodes for Ad Hoc Network Areas. In: 2007 IEEE Asia-Pacific Services Computing Conference (2007)
24. Siva Ram Murthy, C., Manoj, B.S.: Ad hoc Wireless Networks, Pearson (2005) ISBN 81- 297-0945- 7
25. Krco, S., Dupecinov, M.: Improved Neighbor Detection Algorithm for AODV Routing Protocol. IEEE Communications Letters 7(12) (December 2003)

Conceptualizing an Adaptive Framework for Pervasive Computing Environment

Akhil Mohan and Nitin Upadhyay

Computer Science & Information Systems Group, BITS-Pilani Goa Campus
NH-17 B Zuari Nagar,

Goa, Goa-403726, India

{mohan.akhil, upadhyay.nitin}@gmail.com

Abstract. Pervasive computing is the next thrust area for the researchers, service providers and end users alike but the challenges associated for them are equally grilling. Ad-hoc availability of resources makes it extremely important for the stake-holders to keep their Software Architecture as flexible as possible in order to maintain the Quality of Service. This work provides a solution to the dynamic environment, posed by pervasive computing by suggesting a metadata driven adaptive framework, which can be modeled into existing Architecture Description Languages (ADLs) as well as directly implemented on existing platforms. This work reflects a conceptual framework, as an extension to existing work in this field of adaptive software architectures that can be further customized into more sophisticated system as per the requirements of a pervasive environment or any other similar application area.

Keywords: Self-Adaptive Frameworks, Dynamic Software Architecture, Pervasive Computing.

1 Introduction

The advancements, in Information and Communication Technology, Data-Networks and Hand-Held devices, have resulted in a huge impact on the society as a whole and the way we interact and maintain our social relations within the society [1]. One of the most important expectations from a pervasive environment is to handle lack of persistent infrastructure and pre-defined user expectations. Such environment not only handles the difficulty of variable availability of resources but also works invisibly with minimal user interference. The user defines only final goal as his expectations and the environment proactively manages the software and hardware services in order to deliver the result with acceptable quality [2]. As with time the miniature devices like smartphones and other embedded products are becoming cheaper, the demand of such pervasive environment is reaching more and more people with wide range of expectations. Such pervasive systems were initially deployed for Space Research and Extreme Defense Operation where human involvement was not always possible, but today pervasive computing is finding applications in almost every field from health-care to traffic management [3]. A Pervasive environment does not expect a static model with well-defined range of requirements and behavioural responses to such

requirements, instead it needs to be dynamic, in not only handling requirements of any nature but also must be capable of handling exceptional situations which may require re-composition of the environment itself.

Although, systems which support reconfiguration have existed in past but the depth of such repair in response to damage always used to be well defined within the scope of the system. Any exceptional or behavioural situations which were not pre-defined used to never deliver a solution to the user of the system beyond error messages. Another issue related gradual change in the system behaviour with time is that current software architectures point out the occurrence of exceptions and not the reason for the same. This reduces the visibility of the defective component in the architecture making its repair or replacement difficult [4]. Apart from this, the repair begins after occurrence of exceptions in current architectures and results in noticeable interruption of services for the user. Moreover, network is generally overlooked while designing the software architecture considering ideal conditions by adding assumptions and constraints pointing out the minimum network requirements to maintain the quality of service. It is important for the devices and applications to maintain an observational view of the network in order to take correct decisions and adapt to the changing environment.

This work presents a study of these missing links within the existing software architectures along with study of existing adaptive frameworks proposed in [3-7]. This work is closely based on [4] but adds to it, new functionalities and takes a completely new perspective on the approach of metadata driven dynamic component re-composition. The paper is further organized as follows: Section 2 of this paper describes the current work done in this field. Section 3 discusses the solution proposed in this paper and Section 4 presents a hypothetical case study showing scenarios and reactions of the framework in those scenarios. Section 5 is a discussion on the advantages achieved through the conceptualized framework proposed in the paper. The paper is finally concluded with future scope and conclusion in Sections 6 and 7 respectively.

2 Related Work

There has been an extensive study in the field of adaptive frameworks which can sustain in pervasive environments. While some of the works have proposed complete frameworks in [3-7] still a majority of papers have concentrated their focus on individual aspects of such adaptive environments and frameworks. The most important individual building block is the workflow management discussed in [8] which proposes a transactional model for web services providers to advertise transactional properties as non-functional attributes which can be mixed through a composition algorithm to meet transactional constraints keeping in mind the dynamics of those web services. A trust model has been discussed in [9] describing the factors for deciding the trustworthiness of an entity and formulae to calculate the trustworthiness. Further, it briefly discusses condition-action rules to take decisions on the calculations mentioned above. Apart from the trust model, [10] proposes a decentralized trust computation and maintenance to reduce communication and computational overhead without compromising security using direct and indirect neighbors' recommendations. Emphasis on privacy policies has been reflected in [11] by suggesting a new privacy

definition language PREFORM which can describe privacy setting for all architectural components. Importance of handling user preferences dynamically for every behavioural and functional change is described in [12]. Specific user preference about Quality of Service (QoS) majorly regarding wireless networks has been discussed in [2] assuring seamless transition between network options to maintain QoS. Use of TPSSMA for matching services (described in OWL-S) according to user preferences overcoming static keyword based discovery has been shown in [13]. Claimed as the first attempt [14] adds routing semantics into programming language in order to decide how service information should be propagated and routed across multiple and decentralized networks. Reflective approach has been shown to benefit dynamic software architectures in [15] by providing metadata about architectures back to them. The work in [16] deals with the structural as well as behavioural modeling of architecture using π -ADL. The work in [17] discusses a new context query which address heterogeneous representations of context information, supporting complex filtering through aggregation functions and enabling the use of knowledge represented through an ontology, and finally a completely new language PerLa has been introduced in [18-19] at middleware which provides SQL like interface to extract data from collaborating nodes within a pervasive system.

3 The Proposed Conceptualized Adaptive Framework

The conceptualized framework is based on cooperation amongst three layers namely, Processing Layer, Modeling Layer and Runtime Layer, as shown in Fig. 1 which can be implemented by any existing platforms as a library or inbuilt functionality depending upon the maturity and scalability of that platform. This work is an extension to [4] with major emphasis on metadata exchange as driver for the software architecture level adaption for dynamic component replacement.

This work introduces the flexibility in software architecture to adapt, repair or replace itself in order to answer the dynamics of a pervasive system. Although majority of the works mentioned above should be integral part of any adaptive framework which has to function in a pervasive environment, but this work is concentrated towards the critical issue about flexibility in component replacement on basis of user or environment based constraints.

Model Files introduced in the framework, describe the complete framework and its components which include process profiles, architecture profiles, constraints and dependencies along with non-functional requirements of the complete framework or a component within the framework. The model files have been discussed in detail in Section 3.1 through Fig. 3. The Model Files are specialized policy files in [6] but these are different in the sense that *instead of using XML for representation they use plain key, value pairs thereby reducing the average message size* during every communication within the framework as well as the interactions with the environment. The reason to avoid XML is that pervasive environment as well as this adaptive framework relies on, excessive metadata exchange and its flexibility to define various components, in order to function properly. This has been discussed in [20] but this framework uses these model files instead of Resource Description Framework (RDF) as recommended in the work.

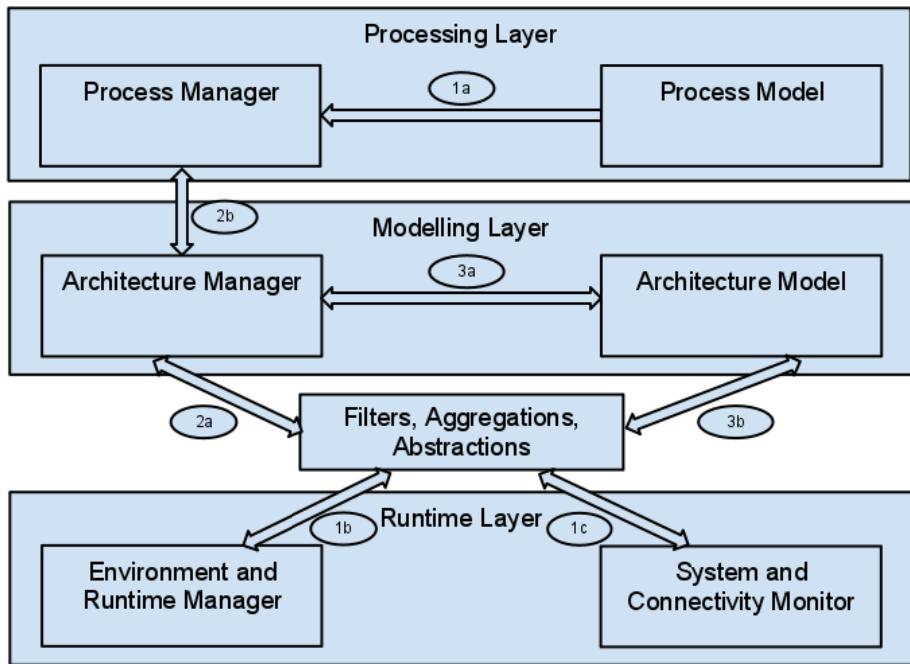


Fig. 1. Adaptive Framework

Achieving the goal, of dynamic component replacement at the level of software architecture, requires co-operation from three levels. The Processing Layer has knowledge of the kind of information a user requires and the QoS requirements for retrieving this information. This knowledge feeds into the Model Layer, so that relevant analysis can be performed to determine the appropriate configuration when a new process is created. The Modeling Layer then makes changes through the Runtime Layer, to the executing system to fulfill those requirements.

Existing concepts in mature programming languages like exception handling, reflection [15] and interfaces can be ported to the conceptualized framework with the added capability of specifying operating ranges instead of fixed values for operation. For e.g., if the network connectivity through one interface on a user device is about to be lost then the device needs to act in time in order to find a substitute connectivity else the user will have to face disconnection. This means that exceptions need to occur before the failure not after the failure. In order to work out and understand such cases a single value defined as threshold are not be sufficient and the exception handler should be able to predict failures by analyzing trends. This adds need of *monitoring systems* which can issue alerts based upon stats about the status and performance of hardware in recent past.

3.1 The Processing Layer

This layer interacts with the native Operating System of the device to fetch the information and metadata on currently executing process on the device. It also maps those

devices to various process models available locally within the framework or can fetch a new model over network which is applicable for a process with help of the metadata extracted from Operating System. Another aspect is the ability of the environment to feed models automatically directing and controlling the execution of the process. Refer Fig. 2 for internal components and their connections.

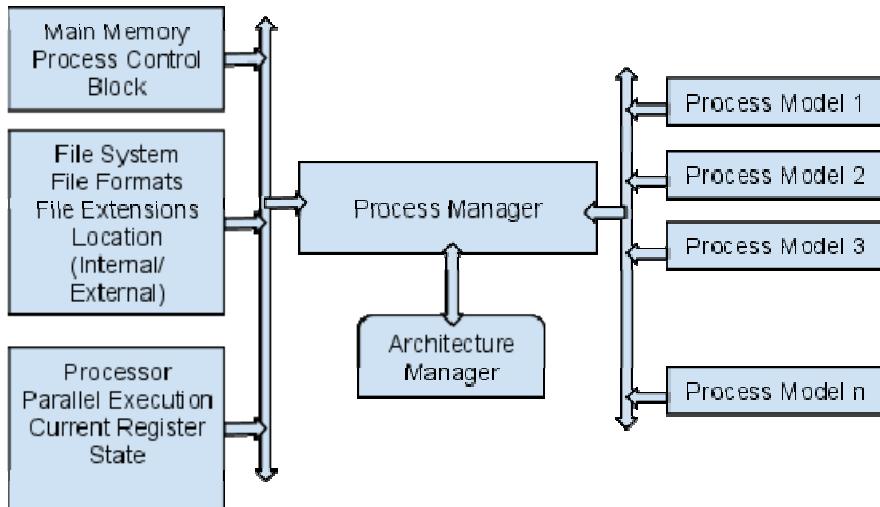


Fig. 2. Processing Layer

The Process Model is basically the model file, pointed out earlier, with (key, value) pairs describing various attributes of a process. The benefit here is that values can be simple and complex type. A simple value is an alpha-numeric string while complex value is actually reference to a key block defined within the parameter file. The key block is a set of {key1= value1, key2 = value2, ... , keyN = valueN} pairs scoped within curly brackets {} and assigned entirely as a block to the key. These parameter files can include other parameter files through use of special (include = /relative/path/to/file) key which takes relative path of other parameter file as its value. There should be minimal constraints on limiting value types and key types as every process may have unique expectations but specific application areas and environments may impose restrictions but such restrictions should not be generalized as required or not applicable within the framework by default. The Process Manager should be able to interface with native operating system to fetch the information on processes currently under execution and their supporting metadata. Once it has extracted that information the manager has to fetch the correct model which correctly describes the expectation of the next process which going to execute on the device. For e.g. if an X-Ray is to be displayed for a doctor then it may have minimum resolution expectations along with screen size.

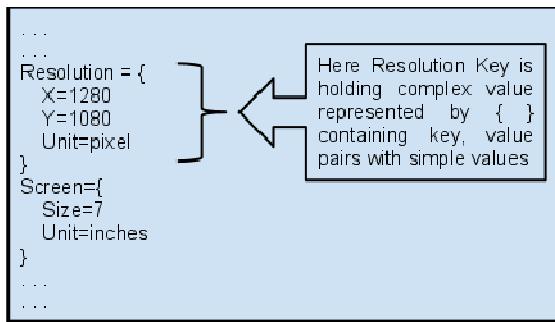


Fig. 3. Model File with Code Section for Complex Keys

If the device does not support any or both of them then the image needs to be directed to another output device within the network reach (well within the local environment). For this reason the manager needs to select the correct model to deliver this information. The process to model mapping could be done on basis of the file extensions, file format, and other such associated metadata that the manager has already extracted about the process. So, in case of this example the process manager will select something like x-ray.model giving details of the constraints like accepted resolution and screen size as shown in Fig. 3.

3.2 The Modeling Layer

The main function achieved on this layer is to match the Process expectations to hardware and runtime abilities. This layer basically tries to find the best match between the Processing and Runtime Layer to deliver the content on the device and keep it within permissible limits of quality. The internal structure is shown in Fig. 4.

The Architectural Model defines the architecture in terms of its quality, performance and other non-functional expectations. It does provide a robust functional description of the architecture but also provides equal emphasis on non-functional expectations to improve the approximation algorithm for choosing the correct architecture solution. The model can be defined in similar fashion as in case of process model with help of parameter file. In this case we need to define some (key, value) pairs as mandatory so that the Architecture Manager always gets sufficient data in order to evaluate and compare between possible solutions.

The Architecture Manager receives the information from the other two layers simultaneously and continuously. On one hand the Processing layer feeds the process models for execution, and on the other hand Runtime layer feeds the current hardware status on the device and about the environment around and approachable through the device. This manager then compares the process model with current hardware status to identify the best architecture models to be used. Amongst the possible solutions, it

also does a comparative analysis to select the optimal solution. Another important function of the manager is to check the dynamic changes in runtime layer and processing layer to identify either a mismatch between them or to identify need of replacement architecture to accommodate the changes. The manager also oversees seamless transition from one architectural model to another which means the user does not feel the shift. For e.g. if the user is moved from wire-line to wireless network the apps on device don't show network outage and at max result in a delay observed for a fraction of a second to accommodate the changes. Another important job of the manager is to make sure that moving from one architectural model to another model, should not interfere ongoing operations. For e.g., if the user executes some asynchronous SQL queries on some remote database which records IP addresses for security reasons then manager should direct the runtime layer to delay the connectivity changes (wire-line to wire-less or vice versa) unless same IP address assignment is assured over the new connection. This is must happen irrespective of the architecture expectations unless user/ environment go for an override or connectivity to remote object is permanently lost.

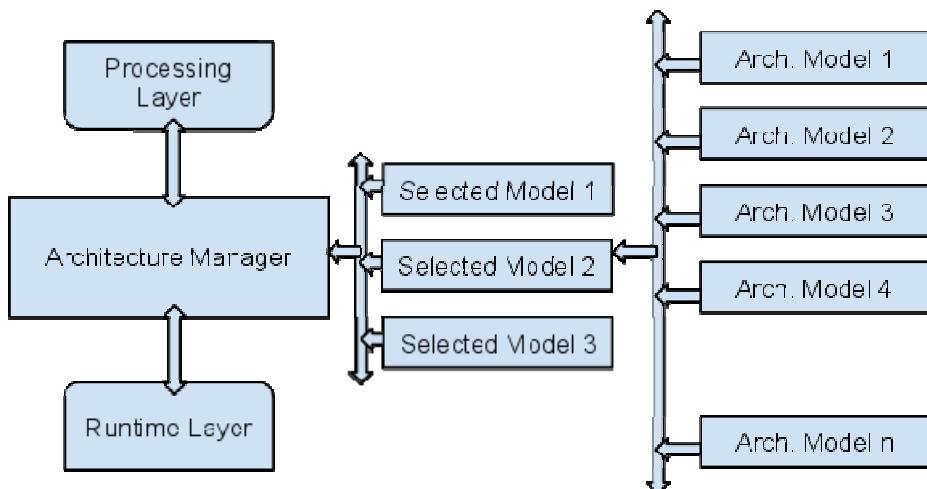


Fig. 4. Modeling Layer

3.3 The Runtime Layer

This layer maintains real-time status of all the hardware attached on the device as well as various possible options available in the environment [4]. This layer puts major emphasis on monitoring the network bandwidth using various connection interfaces.

The System and Connectivity Monitor is the actual monitoring system which has monitors installed as modules for various types of hardware attachments. As and when these attachments are detected by runtime manager these monitors are attached to them for further feedback. The only responsibility of these monitors is to report current status of attachments either on user device or in the environment.

The Environment and Runtime Manager is allocated the task for discovery of attachments available on device and the network. Once discovered the manager allocates a monitor to that attachment and configures it to report regular status, alerts and other such notifications back to manager. This manager also functions in tandem with architecture manager in order to support seamless transition during change of hardware attachments so that user does not feel any observable losses. The Manager works as an interface between the abstract components defined in architecture and a physical hardware showing similar properties. Fig. 1 shows the runtime layer.

The layered description of the framework has been covered in this section. In the next section a case study is considered to identify the feasibility of the conceptualized framework.

4 Case Study

This case study is based on a hypothetical Educational Institute looking forward to enable pervasiveness across the campus. The assumptions and the environment have been discussed in detail followed by specific scenarios for better understanding of the framework and to demonstrate how the adaption will take place in case of process or environmental changes.

Assumptions: The system expects a co-operation between the users for using communications systems like Bluetooth and Wi-Fi to be pooled and used by other users. Dead zones will not provide assured connectivity but will proactively inform availability of content on nearby device. The solution deals with assuring connectivity to content with best possible performance instead of other features like security, load balancing etc. as they are out of the scope of this work. Refer Fig 1 to relate the execution sequence.

Case: *An environment of a college where users are expecting a pervasive content sharing system for students, faculties and college administrators alike.*

Infrastructure: *The College has independent servers for faculty, students and administrators where the uploaded content can be shared with restricted visibility to All, Group or Selected Users who are enrolled on college records.*

Network: *The College is connected with wire-line as well as wireless end points across the campus but with only one of the two services allocated to a location. A vast majority of the campus landscape is dead zone with no network connectivity.*

Requirements: *Following requirements have been considered for the case study:*

1. *The system should support access to content across the campus including the dead zones, if possible.*
2. *User should not need to worry about the type of connectivity supported at their current location and connectivity should be maintained while changing zones without interrupting ongoing access. This applying to inter zone access.*
3. *The access should be provided to all users on all their devices.*

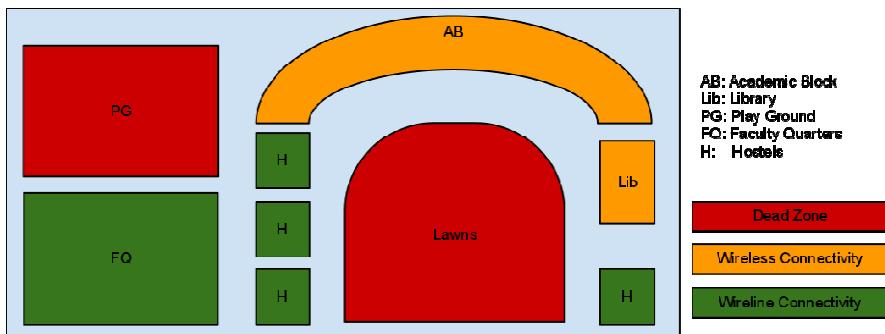


Fig. 5. Hypothetical Plot of the Institute depicting the connectivity options on campus

Scenario I

This scenario considers a stationary user who is requesting a presentation file through wire-line network on single device.

Solution: This is the simplest of all scenarios and the access is very close to most of the traditional systems. Refer flow control chart in Fig. 6. In this scenario following steps are required:

1. **Metadata Exchange** phase is reflected in Fig. 6.a as control flow chart.
 - (a) Process Model is selected for showing Presentation on user device is selected in Process Layer. (Fig. 1.1a)
 - (b) Runtime Manager searches for available hardware on device and network. (Fig. 1.1b)
 - (c) Monitors are attached to hardware. (Fig. 1.1c)
2. **Analysis** phase is reflected in Fig. 6.b as control flow chart.
 - (a) Metadata from Runtime Layer is filtered, aggregated and fed into Modeling Layer. (Fig. 1.2a)
 - (b) Process Model is fed from Processing Layer. (Fig. 1.2b)
 - (c) Architecture Model is selected. (Fig. 1.3a)
3. **Adaption** phase is reflected in Fig. 6.c as control flow chart.
 - (a) Architecture Model is read and required hardware information is sent to Runtime Layer. (Fig. 1.3b)
 - (b) Architecture Manager supervises hardware adaption as per the model. (Fig. 1.2a & Fig. 1.3a)
 - (c) Upon completion of hardware acquisition Architecture Manager intimates Processing Layer to allow process to execute. (Fig. 1.2b)

Scenario II

This scenario considers users moving across wireless network and expecting uninterrupted streaming of a video lecture till he remains close to at least one wireless link.

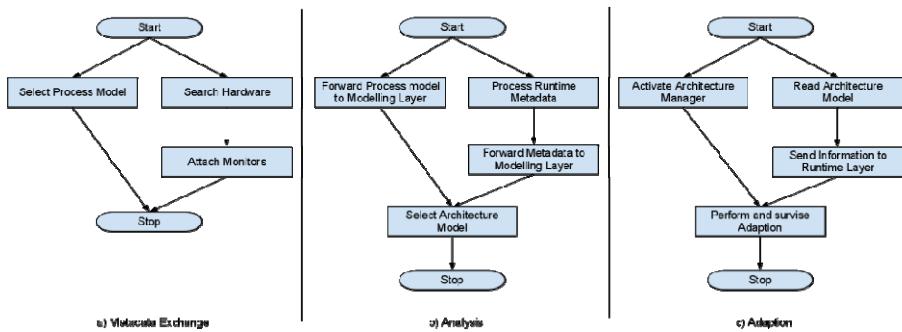


Fig. 6. Control Flow Chart for Scenario I showing the three phases a) Metadata Exchange, b) Analysis and c) Adaption

Solution: This solution starts exactly where last solution ends. We will directly reflect how system will adapt to change in wireless network zone. Refer flow control chart in Fig. 7. In this scenario following steps are required:

1. **Metadata Exchange** phase is reflected in Fig. 7.a as control flow chart.
 - (a) Monitors judge gradual decay in connectivity. (Fig. 1.1c)
 - (b) Runtime Manager searches for alternate connectivity options. (Fig. 1.1b)
2. **Analysis** phase is reflected in Fig. 7.b as control flow chart.
 - (a) Alert from Runtime Layer is filtered, aggregated and fed into Modeling Layer. (Fig. 1.2a)
 - (b) New connectivity options are selected subject to qualification for Process Model el. (Fig. 1.2b)
 - (c) Architecture Model is selected, if need for change is felt by Architecture Manager. In this scenario this step is neglected as alternate option is again similar wireless link. (Fig. 1.3a)
3. **Adaption** phase is reflected in Fig. 7.c as control flow chart.
 - (a) No Architecture Model is read and no hardware info is sent to Runtime Layer. (Fig. 1.3b)
 - (b) Architecture Manager directs Process Manager to send process into waiting state. (Fig. 1.2b)
 - (c) Architecture Manager supervises hardware adaption as per the existing model. (Fig. 1.2a & Fig. 1.3a)
 - (d) Upon completion of hardware acquisition Architecture Manager intimates Processing Layer to allow process to execute. (Fig. 1.2b)

Scenario III

This covers users who just moved into dead zone but the content they are requesting is available on one or more user device.

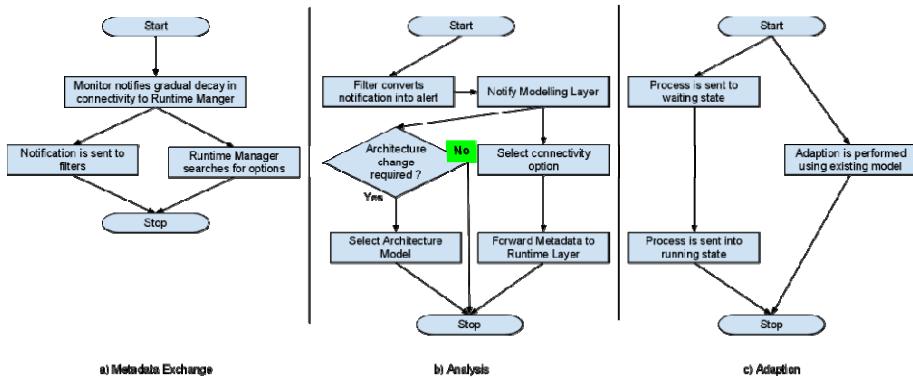


Fig. 7. Control Flow Chart for Scenario II showing the three phases a) Metadata Exchange, b) Analysis and c) Adaption. The difference from Scenario III is the selection of Architecture Model during b) Analysis where Decision gives negative result as **NO** highlighted in **GREEN**.

Solution: This is the most complex scenario where a lot of analysis needs to be done. Again, we start from the point of network adaption. Refer flow control chart in Fig. 8. In this scenario following steps are required:

1. **Metadata Exchange** phase is reflected in Fig. 8.a as control flow chart.
 - (a) Monitors judge gradual decay in connectivity. (Fig. 1.1c)
 - (b) Runtime Manager searches for alternate connectivity options. (Fig. 1.1b)
2. **Analysis** phase is reflected in Fig. 8.b as control flow chart.
 - (a) Alert from Runtime Layer is filtered, aggregated and fed into Modeling Layer. (Fig. 1.2a)
 - (b) New connectivity options are selected subject to qualification for Process Mod- el. (Fig. 1.2b)
 - (c) Architecture Model is selected, if need for change is felt by Architecture Man- ager. In this scenario this step is required as alternate option will be Bluetooth on user device which is considerably slower medium for data transfer. (Fig. 1.3a)
3. **Adaption** phase is reflected in Fig. 8.c as control flow chart.
 - (a) Architecture Model is read and required hardware info is sent to Runtime Layer. (Fig. 1.3b)
 - (b) Architecture Manager directs Process Manager to send process into waiting state. (Fig. 1.2b)
 - (c) Architecture Manager supervises hardware adaption as per the existing model. (Fig. 1.2a & Fig. 1.3a)
 - (d) Upon completion of hardware acquisition Architecture Manager intimates Processing Layer to allow process to execute. (Fig. 1.2b)

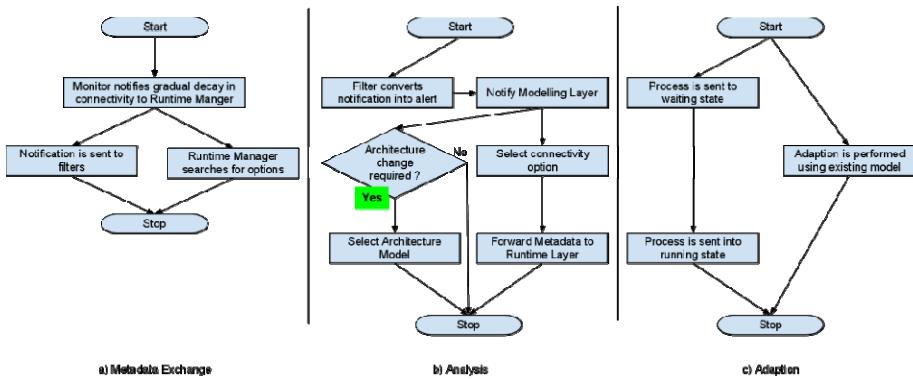


Fig. 8. Control Flow Chart for Scenario III showing the three phases a) Metadata Exchange, b) Analysis and c) Adaption. The difference from Scenario II is the selection of Architecture Model during b) Analysis where Decision gives positive result as YES highlighted in GREEN.

5 Advantages of the Proposed Conceptualized Adaptive Framework

This section gives a detailed description of the advantages of using this proposed conceptualized framework, in a pervasive environment, over other static software architecture models. On one hand it gives the system the ability work beyond the static model and on the other hand it adds following benefits:

1. It adds the scope to add the configuration through model files for specifying the dynamic behavior of the application, i.e. it allows the architectural elements to respond dynamically through modification/ selection of model files (discussed in Section 3) at runtime.
2. Evolution of the configuration which means the configuration is capable of performing new functionalities implying a modification or an evolution of the system's structure. This can be done through merging and selection of model files.
3. Scalability which empowers framework to adapt proposing re-configuration which evolves in size with time and environment changes. The architecture can cache new model files to be referred when required.
4. Definition for constraints related to configurations which describe the dependencies between components, as well as characteristics concerning dynamic assembling of components.
5. Definition for non-functional properties at the level of configuration, because these non-functional properties are not related to components. These properties depend on the environment of execution. They must be specified at the level of configuration.
6. The configuration of the framework allows profiling adaptable to mobility from one environment to another. These characteristics can be operating systems or hardware devices, which are heterogeneous.
7. The framework would represent the composition at different levels of details. That means a hierarchical composition which will permit us to specify the application

using a descending approach with different levels of refinement, starting with the most general level formed by main components, which are defined themselves by groups of components and connectors.

Along with these benefits, this conceptualized adaptive framework model, can also address the key characteristics of pervasive systems like client dependent adaptability, environment dependent adaptability and peer-to-peer cooperative computing using component based architecture definition. The components can be added and dropped at runtime as per the circumstances and expectations without involvement of the user as well as the service provider. Components here reflect the hardware attachments as well as software (processes) being provisioned as services by environment and native applications supported on the user devices.

6 Future Scope

Adaptive Frameworks are a new phase of software architecture which needs to mature and a lot of effort is needed for the same. In future, the work needs to be done to define a formal 3 layer system which should be implemented by all ADLs, Frameworks and Libraries to allow interoperability and COTS ideology [21] to integrate the components of framework to form an infrastructure of choice. Frameworks should not become the restriction in adoption of latest advancements in this field. Moreover, adding semantics, as loosely mentioned in [13] for web services, to the various components of the adaptive framework will not only improve the automation but also make the framework work across lingual barriers. Another important aspect of this framework is handling user preferences and privacy [12]. Such data is confidential to user but still needs to be shared with environment for better performance. A lot of emphasis needs to be given while handling user's sensitive data. The last point worth mentioning is about workflow or transaction management within pervasive environment [8]. As we know that due to pervasiveness the actual delivery might abstract the internal individual services which are required to produce the desired output. While this abstraction is mostly beneficial there is a need to make sure that all those internal services work together to deliver the final output.

7 Conclusion

It has been identified, in the current research work that Adaptive Frameworks for Pervasive Computing seem to be the formal building block of the infrastructure to make it truly scalable, flexible and extensible. This adaptively needs to accompany with self-awareness and extensive insight of the environment. Thus metadata exchange amongst collaborating systems and component plays a crucial role to describe the systems and their limitations. This work proposes the framework with central idea of metadata driven adaption where the functional as well as non-functional properties and constraints are described in model files as metadata. Any violation of constraints or better suitability of properties leads to an automated adaption governed by the architecture manager. This works has provided an extremely simple yet powerful and customizable configuration system for self-adaption through metadata exchange.

References

1. Nkumbwa, R.L.: Emerging Next Generation Communication Technology: Unveiling the Ubiquitous Society. In: International Conference on Education and Management Technology, ICEMT, pp. 1–5 (2010)
2. Yang, Y., Williams, M.H.: Handling Dynamic QoS Requirements in a Pervasive System. In: Proceedings of the International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies, ICNICONSMCL (2006)
3. Yared, R., Defago, X.: Software Architecture for Pervasive Systems. Japan Advanced Institute of Science and Technology
4. Cheng, S.W., Garlan, D., Schmerl, B., Sousa, J.P., Spitznagel, B., Steenkiste, P., Hu, N.: Software Architecture-Based Adaptation for Pervasive Systems. In: Schmeck, H., Ungerer, T., Wolf, L. (eds.) ARCS, pp. 67–82 (2002)
5. Indulska, J., Loke, S.W., Rakotonirainy, A., Witana, V., Zaslavsky, A.: An Open Architecture for Pervasive Systems. In: Third International Working Conference on New Developments in Distributed Applications and Interoperable Systems, pp. 175–187 (2001)
6. Ouyang, J., Shi, D., Ding, B., Feng, J., Wang, H.: A Framework for Self-Adaptive Scheme in Pervasive Computing. In: ICCS, pp. 750–755 (2008)
7. Qing, W., Weihua, H., Wen, D.: A Semantic and Adaptive Middleware Architecture for Pervasive Computing Systems. Journal of Software 4(10), 1061–1068 (2009)
8. Montagut, F., Molva, R., Golega, S.T.: The Pervasive Workflow: A Decentralized Workflow System Supporting Long-Running Transactions. IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews 38(3), 319–333 (2008)
9. Yin, S., Ray, I., Ray, I.: A Trust Model for Pervasive Computing Environments. In: International Conference on Collaborative Computing: Networking, Applications and Work-sharing, CollaborateCom, pp. 1–6 (2006)
10. Sun, T., Denko, M.K.: A Distributed Trust Management Scheme in the Pervasive Computing Environment, pp. 1219–1222 (2007)
11. Dehghantanha, A., Udzir, N.I., Mahmud, R.: Towards a Pervasive Formal Privacy Language. In: 24th International Conference on Advanced Information Networking and Applications Workshops, pp. 1085–1091 (2010)
12. Papadopoulou, E., McBurney, S., Taylor, N., Williams, M.H.: A Dynamic Approach to Dealing with User Preferences in a Pervasive System. In: International Symposium on Parallel and Distributed Processing with Applications, pp. 409–416 (2008)
13. Zhu, Y., Meng, X.: A Framework for Service Discovery In Pervasive Computing. In: 2nd International Conference on Information Engineering and Computer Science, ICIECS, pp. 1–4 (2010)
14. Suzuki, T., Pinte, K., Cutsem, T.V., Meuter, W.D., Yonezawa, A.: Programming Language Support for Routing in Pervasive Networks. In: Eighth IEEE International Workshop on Middleware and System Support for Pervasive Computing, pp. 226–232 (2011)
15. Peng, Y.E., Shi, Y., Jie, Y.W., Feng, Y.J., Bo, L.J., Lin, Z.L.: A Reflection-Based approach for Reusing Software Architecture (2008)
16. Oquendo, F.: Dynamic Software Architectures: Formally Modelling Structure and Behaviour with π -ADL. In: The Third International Conference on Software Engineering Advances, pp. 352–359 (2008)

17. Reichle, R., Wagner, M., Khan, M.U., Geihs, K., Valla, M., Fra, C., Paspallis, N., Papadopoulos, G.A.: A Context Query Language for Pervasive Computing Environments. In: Sixth Annual IEEE International Conference on Pervasive Computing and Communications, pp. 434–440 (2008)
18. Schreiber, F.A., Camplani, R., Fortunato, M., Marelli, M., Pacifici, F.: PERLA: a Data Language for Pervasive Systems. In: 6th Annual IEEE International Conference on Pervasive Computing and Communications, PERCOM, pp. 282–287. IEEE (2008)
19. Schreiber, F.A., Camplani, R., Fortunato, M., Marelli, M., Rota, G.: PerLa: a Language and Middleware Architecture for Data Management and Integration in Pervasive Information Systems. *IEEE Transactions on Software Engineering* (2011)
20. Decker, S., Melnik, S., Harmelen, F.V., Fensel, D., Klein, M., Broekstra, J., Erdmann, M., Horrocks, I.: The Semantic Web: the roles of XML and RDF. *IEEE Internet Computing*, 63–73 (2000)
21. Upadhyay, N., Despande, B.M., Agrawal, V.P.: Towards a Software Component Quality Model. In: Meghanathan, N., Kaushik, B.K., Nagamalai, D. (eds.) CCSIT 2011, Part I. CCIS, vol. 131, pp. 398–412. Springer, Heidelberg (2011)

Dynamic DCF Backoff Algorithm(DDBA) for Enhancing TCP Performance in Wireless Ad Hoc Networks

B. Nithya, C. Mala*, B. Vijay Kumar, and N.P. Gopalan

Department of Computer Science and Engineering,
National Institute of Technology, Tiruchirapalli 620 015, Tamil Nadu, India
 {nithya,mala,npogopalan}@nitt.edu, vijay.chaos@gmail.com

Abstract. Transmission Control Protocol (TCP) is the most commonly used transport protocol on the internet. The performance of TCP in wireless network is not satisfactory as it was originally designed for wired networks. This paper proposes Dynamic DCF Backoff Algorithm(DDBA) to improve the throughput of TCP in wireless mobile environment. The proposed DDBA minimizes TCP Instability problem initiated by wrong link failure information from the source node. The simulation in NS2 shows improved TCP performance in Grid, Cross and Random topology of wireless networks incorporating mobility of nodes.

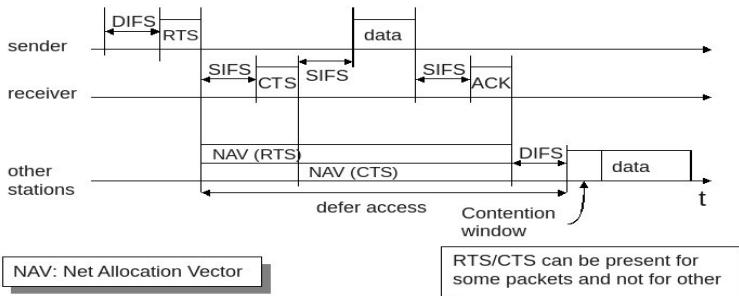
Keywords: TCP, IEEE 802.11, Back-off, TCP Instability, Mobility, Wireless Network, Hidden and Exposed Terminal Problem.

1 Introduction

The rapid technological advances and innovations in the past few decades have forced reliable end-to-end wireless communication from concept to reality. For the wireless communication, designing an efficient TCP protocol is a crucial problem as it faces heavy packet losses, high bit error rates, dynamism, path asymmetry, etc,. The bandwidth sharing among multiple TCP transfers can be very unfair. Link failures in wireless networks not only occur because of mobility but also due to exposed-terminal problem. As bit error rates are very low in wired networks, most of the TCP versions assume that packet losses are due to congestion [1]. In a wireless environment,when TCP responds to packet losses by invoking congestion control or an avoidance algorithm, a degraded end-to-end performance in wireless network results.

Several enhancements and optimizations (by modifying existing TCP flow control and error control mechanisms) have been proposed to improve TCP performance [1-6]. Surveys on improving TCP performance, issues and solutions are found in [7-15].Many proposals (by developing new techniques) have been suggested in the literature [16][25], which can be classified into three categories are as follows:

* Corresponding author.

**Fig. 1.** 802.11 DCF

1. End-to-End TCP protocols, where loss recovery, such as Explicit Loss Notification (ELN) option is performed by the sender.
2. Link layer protocols which provide local reliability, using techniques such as Forward Error Correction (FEC) and retransmission of lost packets in response to Automatic Repeat Request (ARQ) messages.
3. Split TCP connection protocol that breaks the end-to-end TCP connection into two parts at the base station, one between the sender and the base station and the other between the base station and the receiver.

The rest of the paper is organized as follows: Section 2 gives details about IEEE 802.11 and TCP Instability problem, Section 3 discusses about the proposed modified backoff algorithm and Section 4 depicts the simulation results and performance analyses. Section 5 concludes the paper.

2 IEEE 802.11 DCF and TCP Instability Problem

This section discusses about IEEE 802.11 Distributed Coordination Function (DCF) and TCP Instability Problem.

2.1 IEEE 802.11 DCF

IEEE 802.11 [18] uses a random access method with Carrier Sense and Collision Avoidance through random backoff (CSMA/CA). The station senses the medium with a Clear Channel Assessment(CCA) signal, if it is idle at least for a DCF Inter Frame Spacing (DIFS) duration, then that station can access the medium. If it is not, the random backoff time is used, in order to avoid collisions among stations that transmit after DIFS period. This backoff time (multiples of slots) is chosen from the Contention Window(CW).

A station willing to transmit a packet will first transmit a Request To Send(RTS) control packet which will include the source, destination and the duration of the following transaction(i.e., packet and respective ACK). The destination station will respond (if the medium is free) with a Clear To Send(CTS)

control packet which has the same duration as shown in Figure 1 [20]. All stations receiving either the RTS and/or the CTS, will set their Virtual Carrier Sense indicator ,called Network Allocation Vector(NAV) for the given duration, and will use this information together with the physical carrier sense, while sensing the medium. The complete description is given in [20].

Depending on the size of the CW, there are two possible cases [19]:

Case 1: The random values can be too close together. So some of the stations may choose the same random value, causing too many collisions.

Case 2: If the random values are too high, then the station has to unnecessarily defer, even though medium is idle.

Under light load condition , small CW ensures shorter access delay. Under heavy load condition, large CW provides greater resolution power of the randomized scheme. So always there will be a trade off between CW size and performance.

2.2 The TCP Instability Problem

TCP instability problem [24] is mainly due to failure of a node to reach its next hop due to the exposed terminal problem [19]. This triggers a route failure. If this node is an intermediate node, it drops all queued packets to its next hop and reports the route failure to the source. After receiving this message, the source again starts the route discovery process. Before a route is found, no data packets can be sent out. So the TCP session needs to wait before the routes become available. Since no data packet is sent out during this period, performance of TCP throughput will degrade considerably.

Our work is based on [17], in which a node considers two contention window sizes 128 & 256 for DSR routing protocol with string topology.The link failure problem due to mobility is not considered. This factor motivates us to further enhance TCP performance which incorporates the dynamism.

3 Proposed Work

This section discusses about proposed Dynamic DCF Backoff Algorithm (DDBA). The proposed algorithm chooses backoff times based on CW size ranges from 32 to 1024 as in DCF algorithm. By dynamically adjusting the random backoff times, transmission of RTS packet is appropriately spaced. Whenever the source station is not receiving CTS packet from the destination station, it assumes the packet loss in vicinity of destination. Instead of transmitting RTS again after a short period, source station has to choose the random backoff time such that it avoids packet loss due to collision around the destination. The proposed DDBA also ensures that the station is not choosing longer backoff time which induces more latency as in DCF algorithm. Thus it reduces the unnecessary retransmission and delay incurred thereby improving TCP performance.

3.1 Proposed Dynamic DCF Backoff Algorithm(DDBA)

Parameters used:

1. Backoff Time(BO), Random Number(R), DCF Inter Frame Spacing(DIFS) as in [22]
2. Contention Window(CW)

Step 1: $BO = (R \% CW) * slot_time$

Step 2:

2.1: if $CW \leq 64$

2.1.1: if $BO < 64\mu sec$, then $BO + = 64\mu sec$

2.1.2: while ($BO \geq 128\mu sec$), $BO - = 64\mu sec$

2.2: if $CW > 64 \ \&\& \ CW \leq 128$

2.2.1: if $BO < 128\mu sec$, then $BO + = 128\mu sec$

2.2.2: while ($BO \geq 256\mu sec$), $BO - = 128\mu sec$

2.3: If $CW \geq 256$

2.3.1: if $BO < 256\mu sec$, then $BO + = 256\mu sec$

2.3.2: while ($BO \geq 512\mu sec$), $BO - = 256\mu sec$

Step 3: Station waits upto $BO + DIFS$ period and proceed similar to DCF algorithm

3.2 Working of Proposed Dynamic DCF Backoff Algorithm (DDBA)

Contention Window(CW) sizes range from 32 ,64,128,....,1024 similar to DCF algorithm. To have CW values within this range, random number is chosen [21] and mod operation is performed then multiplied by slot time[22]. Random numbers are generated [20] by sequentially selecting numbers from a stream of pseudo-random number using Combined Multiple Recursive Generator[24].According to Step2, if CW size is less than or equal to 64, then BO time will be in the range from $64\mu sec$ to $127\mu sec$. If BO is less then $64\mu sec$ as in DCF, the station may repeat the transmission of RTS immediately which further increases collision at destination because of Exposed terminals. To avoid such repeated retransmission, DDBA forces the stations to wait for longer duration in order to allow the transmissions to be completed around the destination.

If CW size is increased which indicates high traffic, then BO time is also increased to resolve the contention.The BO time will be in the range of $128\mu sec$ and $255\mu sec$ for the CW size ranges from 65 to 128. Similarly for window sizes greater than 256, then backoff time is in the range of $256\mu sec$ and $511\mu sec$. This limited BO time ensures that the station will not wait unnecessarily for the medium, hence the latency is reduced. Increased BO time in DCF makes the

station to wait even though medium is idle. Once this BO time is elapsed, the station waits for the duration of DIFS to get the medium access.

Case 1: CW is small. For CW in the range 64- 128, the proposed DDBA gives longer waiting time compared to DCF algorithm thereby reduces collisions due to exposed terminals. Therefore the station will get CTS within RTS Threshold (maximum RTS transmission is 7), This avoids wrong link failure information to TCP. This behavior of DDBA is essential while broadcasting route discovery control packets as they collide with other packets at the destination and the source again starts the retransmission of route discovery control packets after retransmission timer expires. But DDBA avoids such false alarms from the MAC layer telling that the intended neighbour is not in its range even though neighbours do not move. Hence these unnecessary retransmissions are minimized by DDBA. As a result it reduces incurred delay and enhances the throughput. But in DCF, whenever collision is occurred, CW size gets exponentially doubled. This further increases unnecessary waiting time and hence minimizes throughput.

Case 2: CW is large. For CW in the range 512-1024, the proposed DDBA gives lesser BO time(from $256\mu\text{sec}$ to $511\mu\text{sec}$) compared with DCF algorithm. Thus DDBA makes longer waiting stations to defer for small period. Once the waiting time is elapsed, the station can access the medium immediately without wasting resources. But DCF forces stations to wait more time even though medium is free.

4 Simulation and Performance Analysis

Proposed DDBA in Section 3 is simulated using NS2.34 [22]. The simulation environment parameters and the analyses of the graphs obtained by Xgraph[22] are discussed in the following subsections.

4.1 Simulation Environment

The standard NS2 simulator is used with the parameters listed in Table 1. The simulation is performed for analyzing the TCP performance incorporating mobility for the following topologies:

- 1.Cross Topology 2.Grid Topology 3.Random Topology

4.2 Simulation Parameters

Parameters used for simulation and analyzing the TCP performance are as follows:

- 1.Throughput 2.End-to-End Delay 3.Packet Loss

Table 1. Simulation Parameters & their values

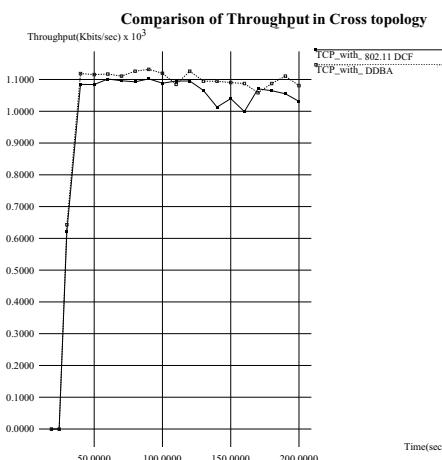
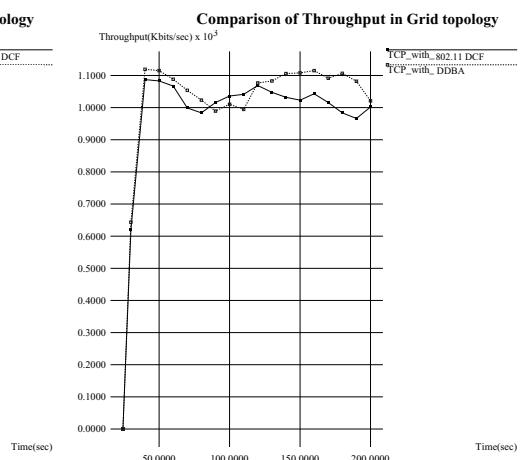
Propagation model	Two Ray Ground
Link Bandwidth	2 Mbps
Transmission range	250 m
IFQ length	32
Routing protocol	AODV
TCP window size	32
TCP Packet size	512 bytes
Traffic Pattern	FTP
Simulation Time	200 sec
Number of nodes	15
Slot time	20sec
SIFS	10sec

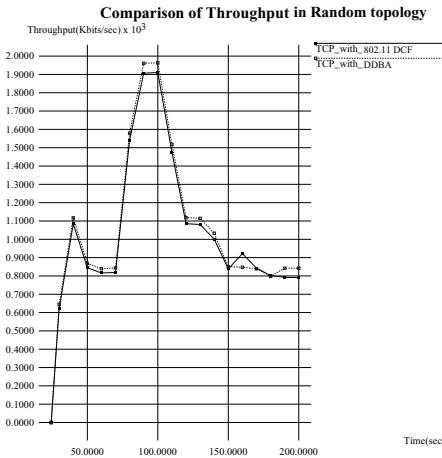
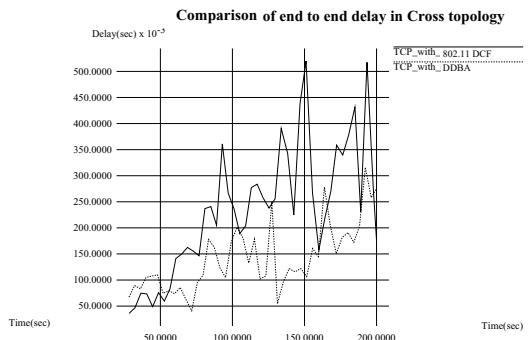
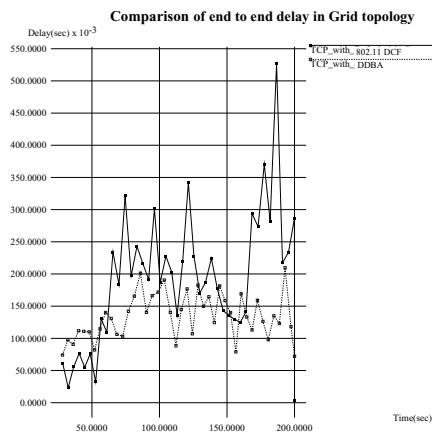
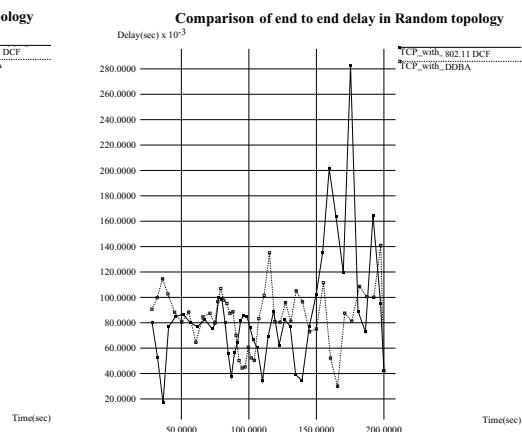
4.2.1 Throughput

Throughput is a measure of how fast the stations can send data through the network. By choosing appropriate BO time using DDBA, stations either transmit data successfully or refrain from the medium access to allow other transmissions to be completed. So collisions due to exposed terminal are reduced at the destination side. Since the packets are transmitted at correct time and received promptly, throughput is increased and delay is reduced as shown in Figures 2,3 and 4 for Cross, Grid and Random topologies..

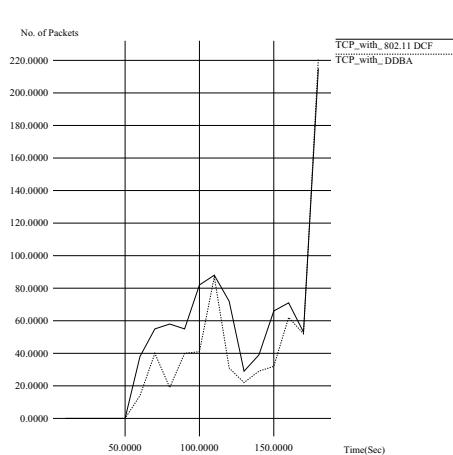
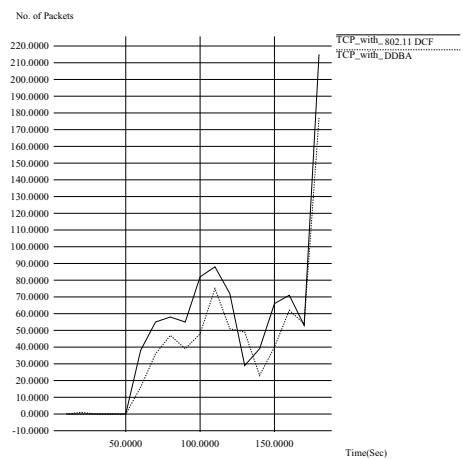
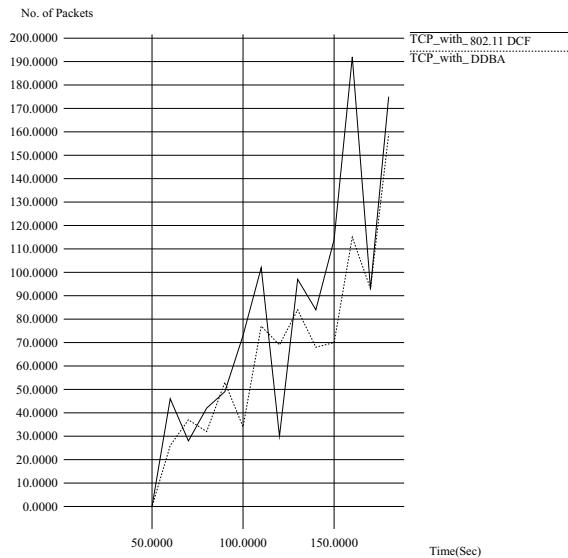
4.2.2 End-to-End delay

End-to-End delay is the sum of delays experienced at each node from source and destination. It includes transmission delay, propagation delay and processing time. Figures 5,6 and 7 exhibits end-to-end delay for different topologies. From

**Fig. 2.** Throughput in Cross Topology**Fig. 3.** Throughput in Grid Topology

**Fig. 4.** Throughput in Random Topology**Fig. 5.** End-to-End delay in Cross Topology**Fig. 6.** End-to-End delay in Grid Topology**Fig. 7.** End-to-End delay in Random Topology

these results, it is inferred that the proposed DDBA algorithm selects BO time from $64\mu\text{sec}$ to $128\mu\text{sec}$ at the beginning of the simulation. But DCF algorithm may select BO time less than this range and DDBA gives more delay during that period (from 0 to 50 sec). After 50 sec, BO time is chosen according to CW size, which avoids unnecessary retransmissions due to exposed terminals. From the simulation results, it is also inferred that the nodes can take modified BO time up to $511\mu\text{sec}$ and not beyond that, but DCF algorithm may choose BO time greater than $511\mu\text{sec}$ which makes the nodes to wait unnecessarily even though the medium is idle. So DDBA shows slightly more delay during the initial period and after that it is less compared with DCF algorithm.

Comparison of Packet Loss in Cross topology**Fig. 8.** Packet loss in Cross Topology**Comparison of Packet Loss in Grid topology****Fig. 9.** Packet loss in Grid Topology**Comparison of Packet Loss in Random topology****Fig. 10.** Packet loss in Random Topology

4.2.3 Packet Loss

Since the proposed DDBA permits the stations to transmit only when the destination is ready to accept the data, dropping of packets are reduced. As the

throughput is increased with minimum delay, it is obvious that packet loss is reduced. But if the network has more number of nodes with high traffic load, packet loss will be more due to congestion.

5 Conclusion

With the increasing importance of reliable end-to-end TCP in wireless networks, TCP should be optimized to deal with the problems caused due to mobility. The proposed DDBA algorithm is simulated using NS2 and its performance is tested in Cross, Grid and Random topologies of Wireless Mobile Network. With the appropriate values of CW, DDBA avoids sending wrong link failure information to TCP, thereby minimizing TCP instability problem. From the simulation results , it is concluded that DDBA outperforms DCF by improving the throughput and minimizing the latency,thereby enhancing TCP performance in Wireless mobile environment.

References

- [1] Chiang, M.: Balancing Transport and Physical Layers in Wireless Ad Hoc Networks: Jointly Optimal TCP Congestion Control and power control. *IEEE JSAC* 23(1), 104–116 (2005)
- [2] Leung, K.K., Klein, T.E., Mooney, C.F., Haner, M.: Methods to Improve TCP Throughput in Wireless Networks With High Delay Variability. In: *Vehicular Technology Conference, VTC 2004 (Fall 2004)*
- [3] Cordeiro, C., Das, S., Agrawal, D.: COPAS: Dynamic Contention-Balancing to Enhance the Performance of TCP over Multi-Hop Wireless Networks. In: *Proc. IC3N*, Miami, USA, pp. 382–387 (October 2003)
- [4] Altman, E., Jimenez, T.: Novel Delayed ACK Techniques for Improving TCP Performance in Multihop Wireless Networks. In: *Proc. Pers. Wireless Commun.*, Venice, Italy, pp. 237–253 (September 2003)
- [5] Fu, Z., et al.: Impact of Multihop Wireless Channel on TCP Throughput and Loss. In: *Proc. IEEE INFOCOM*, San Francisco, USA (2003)
- [6] Cali, F., Conti, M., Gregori, E.: Dynamic Tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit. *IEEE/ACM Transactions on Networking* (December 2000)
- [7] Mast, N., Owens, T.J.: A survey of performance enhancement of transmission control protocol (TCP) in wireless ad hoc networks. *Mast and Owens EURASIP Journal on Wireless Communications and Networking* (2011)
- [8] Tayade, M., Sharma, S.: Performance Comparison of TCP Variants In Mobile Ad-Hoc Networks. *(IJCSIS) International Journal of Computer Science and Information Security* 9(3) (March 2011)
- [9] Opara, F.K., Okorafor, G.N.: Survey of Transmission Control Protocol (TCP) Over wireless networks: issues, challenges and solutions. *Nigeria International Journal of Academic Research* 3(1) (January 2011)
- [10] Sachan, A., Rajput, A.: Comparison of TCP Performance on WLAN (April 13, 2010)

- [11] Shohidul Islam, M., Kashem, M.A., Sadid, W.H., Rahman, M.A., Islam, M.N., Anam, S.: TCP Variants and Network Parameters: A Comprehensive Performance Analysis. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, IMECS 2009, Hong Kong, March 18-20, vol. I (2009)
- [12] Chen, X., et al.: A survey on improving TCP performance in wireless Networks. *Networking Theory & Applications* 16 (July 2006)
- [13] Al Hanbali, A., Altman, E., Philippe: A survey of TCP over Ad Hoc Networks. *IEEE Communications Surveys & Tutorials* (Third Quarters, 2005)
- [14] Tian, Y., Xu, K., Ansari, N.: TCP in wireless Environments: Problems abd Solutions. *IEEE Radio Communications* (March 2005)
- [15] Holland, G., Vaidya, N.: Analysis of TCP Performance over Mobile Ad Hoc Networks. In: Proc. ACM, MobiCom 1999, Seattle, WA (August 1999)
- [16] Floyd, S., Mahdavi, J., Mathis, M., Podolsky, M.: An Extension to Selective Acknowledgement (SACK) Option for TCP. RFC 2883 (July 2000)
- [17] Krishna Kanth, T., Ansari, S., Mehkri, M.H.: Performance Enhancement of TCP on Multihop Ad Hoc Wireless Networks. In: TCP Performance over Multihop Wireless Networks, IEEE ICPWC 2002 (2002)
- [18] IEEE standard for Wireless LAN-Medium Access Control and Physical Layer Specification, p. 802.11 (November 1997)
- [19] Schiller, J.H.: Mobile Communication, 2nd edn. Pearson Education
- [20] Brenner, P.: A Technical Tutorial on the IEEE 802.11 Protocol, BreezeCOM (1997)
- [21] Issariyakul, T., Hossain, E.: Introduction to Network Simulator NS2. Springer (2009)
- [22] <http://isi.edu/nsnam>
- [23] Forouzan, B.A.: Tcp/Ip Protocol Suite, 3/E. Tata McGraw-Hill Education
- [24] Ecuyer, P.L.: Good parameters and implementations for combined multiple recursive random number generators. *Operation Research* 47 (1999)
- [25] Kliazovich, D., Bendazzoli, M., Granelli, F.: TCP-Aware Forward Error Correction for Wireless Networks. In: Chatzimisios, P., Verikoukis, C., Santamaría, I., Laddomada, M., Hoffmann, O. (eds.) MOBILIGHT 2010. LNCS, vol. 45, pp. 68–77. Springer, Heidelberg (2010)

Hybrid Cluster Validation Techniques

Satish Gajawada and Durga Toshniwal

Department of Electronics and Computer Engineering,

Indian Institute of Technology Roorkee, Roorkee, India

gajawadasatish@gmail.com, durgafec@iitr.ernet.in

Abstract. Clustering methods divide the dataset into groups of similar objects called as clusters. Two objects in different clusters are dissimilar and objects in the same cluster are similar. Evaluation of clustering results is known as cluster validation. Cluster validation can be of different types. Internal cluster validation indices measure the quality of the clusters based on the intrinsic properties of the data. External cluster validation is based on external information about the data. The advantage of internal validation is that external information is not required. But using small amount of external information can make unsupervised clustering technique using internal cluster validation for finding optimal clustering solution achieve better results. The advantage with supervised clustering technique using external validation is that clusters confirming to class distribution are obtained. But using intrinsic information present in the data can prevent over fitting of data by supervised learning technique using external validation. In this paper we propose various hybrid cluster validation indices using internal and external cluster validation indices. The advantage with hybrid indices is that validation is done using both intrinsic information of data and available external information. In this work we focus on hybrid cluster validation indices for semi-supervised clustering.

Keywords: Semi-supervised clustering, hybrid cluster validation indices, cluster validation.

1 Introduction

Clustering refers to the division of the dataset into groups called clusters. The objects in the same cluster are more similar and objects in different clusters are dissimilar. Using validity indices is a common approach for evaluation of clustering results. Internal validation is based on the information present in the data. External validation is based on external information about the data [1]. Most clustering methods need the number of clusters to be provided as an input parameter but it is difficult to predict the correct number of clusters [2]. To know optimal clustering solution, a clustering algorithm can be executed several times, with different number of clusters as input each time. The clustering solution with optimum validity index is selected as the best partition [3]. Identification of partition of clusters for which a quality measure is optimal is the main goal of cluster validation indices [2]. There are various cluster validation indices defined in literature [4, 5, 6, 7]. One single cluster validation index may not always give better result compared to other indices. Fusing various cluster validation indices can yield better results compared to using single cluster validation

index for predicting correct number of clusters. Hence different internal validation indices like Davies-Bouldin index (DB index) and Dunn index can be fused for comparing various clustering solutions to get optimal clustering solution [14]. Although using multiple cluster validation indices which are based on intrinsic information present in the data can give better results but available external information is not used in validating clustering solution. Hence using hybrid indices which are based on internal and external information have the advantage of using the internal information present in the data together with available external information. Clustering algorithms which optimize hybrid cluster validation index for finding optimal clustering solution come under semi-supervised clustering algorithms [16].

In this paper, we propose various new cluster validation indices. The proposed hybrid indices use both internal information present in the data together with available external information. Results obtained on synthetic datasets for an internal validation index and a proposed hybrid index are compared.

The rest of the paper is organized as follows: Related work is presented in Section 2. Section 3 contains proposed work. Section 4 contains results and discussion along with data sets used. Conclusion and future work are given in Section 5.

2 Related Work

Bolshakova et al. [8] applied cluster validity methods to estimate number of clusters in cancer tumor datasets. A weighted voting technique was used to improve the prediction of number of clusters. Dimitriadiou et al. [9] examined 14 cluster validation indices for determining number of clusters in artificial datasets. Dudoit et al. [10] developed a new method to estimate number of clusters in the dataset. Halkidi et al. [11] presented a review of cluster validity measures available in literature. Erendira Rendon et al. [1] compared various internal and external validation indices. Satish et al. [12] optimized an internal validation index using Genetic algorithm to find optimal level of cutting the dendrogram obtained by hierarchical clustering on input dataset. Zheng-Yu Niu et al. [13] presented a document clustering method based on cluster validation. Krzysztof Kryszczuk et al. [14] estimated number of clusters using multiple internal cluster validation indices. It was shown that fusion of multiple cluster validation indices can lead to significant gains in accuracy in estimating the number of clusters. Pihur et al. [15] used a Monte Carlo approach for weighted rank aggregation of various cluster validation measures. Demiriz et al. [16] used a hybrid index based on Davies-Bouldin (DB) index and gini index.

Recently, Patil et al [17] proposed effective framework for prediction of disease outcome using clustering and classification. Labels have been assigned through clustering (unsupervised) and assigned labels were matched with given labels to pre-process the data. Classification has been done on pre-processed data. The proposed framework obtained promising classification accuracy as compared to other methods found in literature. Hence from work [17] it is clear that using internal information present in the data can improve classification accuracy of classifier. So, using internal validation which is based on intrinsic properties of data can improve results of supervised clustering technique using external validation. Hence there is need to create hybrid cluster validation indices for semi-supervised clustering techniques for achieving better results.

3 Proposed Hybrid Cluster Validation Indices

In this section our proposed hybrid cluster validation indices are explained. Clustering solution of a dataset can be considered as a partition at a particular node in decision tree and impurity measures like gini index can be used to determine impurity of such partition [16]. Impurity measures like gini index, entropy index, classification error index and information gain ratio index have been used in literature to determine impurity of certain split in decision trees [18].

Demiriz et al. [16] used hybrid DB-Gini index shown in Equation (1) for semi-supervised clustering using Genetic algorithms. W1 represents weight given to internal validation component and W2 represents weight given to external validation component.

$$\text{DB-Gini index} = W1 * \text{DB index} + W2 * \text{Gini index} \quad (1)$$

The gain obtained in impurity measure by splitting parent node into child nodes is defined in [18] and hence Gini gain, Information gain, Classification error gain corresponds to using impurity measures Gini index, Entropy index, Classification error index respectively in the gain criterion defined in [18] that can be used to determine goodness of the split. Impurity measures such as entropy index, gini index tend to favour splits that have more number of nodes. To overcome this problem Gain ratio index [18] is defined which is ratio of information gain and split information. This index is referred as Information Gain Ratio index (IGR index) in this paper.

Similarly, we can obtain Gini Gain Ratio index (GGR index), Classification Error Gain Ratio index (CEGR index) by dividing Gini gain, Classification error gain respectively with split information.

In this paper, we define various hybrid indices using internal validation indices {Dunn index, DB index, Silhouette index, C index} and external validation indices {IGR index, CEGR index, GGR index}.

Equation (2), Equation (3) and Equation (4) shows hybrid indices based on Dunn index and {IGR index, CEGR index, GGR index} which are external validation indices. W1 represents weight given to internal validation component and W2 represents weight given to external validation component. Both internal validation component and external validation component are to be converted such that either both components are to be maximized or both components are to be minimized. The index values of both components are to be scaled such that minimum and maximum values of both components are equal.

$$\text{Dunn-IGR index} = W1 * \text{Dunn index} + W2 * \text{IGR index} \quad (2)$$

$$\text{Dunn-CEGR index} = W1 * \text{Dunn index} + W2 * \text{CEGR index} \quad (3)$$

$$\text{Dunn-GGR index} = W1 * \text{Dunn index} + W2 * \text{GGR index} \quad (4)$$

Similarly we can obtain other hybrid indices based on external validation indices {IGR index, CEGR index, GGR index} and other internal validation indices like silhouette index, DB index and C index.

4 Experimental Results

We obtained optimal number of clusters in the dataset by using a proposed hybrid cluster validation index and an internal validation index for cluster validation. k-means clustering has been used to obtain clustering solutions. The clustering solutions obtained for various values of number of clusters parameter are validated with a proposed hybrid cluster validation index and an internal validation index. Results have been obtained on synthetic datasets. First dataset is organized into 15 clusters in 3 dimensional space. Each cluster is given a separate class label. Hence there are 15 classes in the first dataset. There are 5 groups of 3 close clusters in the dataset. Second dataset is created by giving a single class label to each group of 3 close clusters. So, second dataset has 5 labels because there are 5 groups each of which is given a separate class label. Class labels have been assigned to only a part of the datasets because complete external information may not always be available. Both datasets contain labelled data part and unlabelled data part. Internal index has been calculated using all points in the dataset. External index has been calculated using labelled data part of the dataset. Results have been obtained using Dunn index and hybrid Dunn-CEGR index with equal weights on two datasets. Results obtained are discussed below.

Figure 1 shows the index plot of Dunn index for both the datasets. As both datasets differ only in class labels Dunn index plot will be same for both the datasets. Figure 2 shows the index plot of dataset 1 with hybrid Dunn-CEGR index. Figure 3 shows the index plot of dataset 2 with hybrid Dunn-CEGR index. In Figure 1-Figure 3 x-axis represents number of clusters and y-axis represents index value.

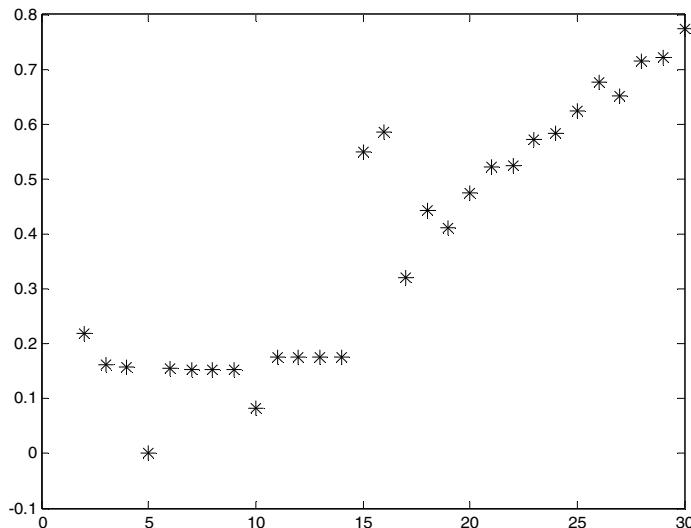


Fig. 1. Index plot of dataset 1 and dataset 2 with Dunn index

Table 1 shows optimal number of clusters obtained with Dunn index and hybrid index when applied on two datasets. Table 1 is created by taking minimum (optimal values) of cluster validation index values. These optimal index values can be obtained from figures Figure 1 to Figure 3.

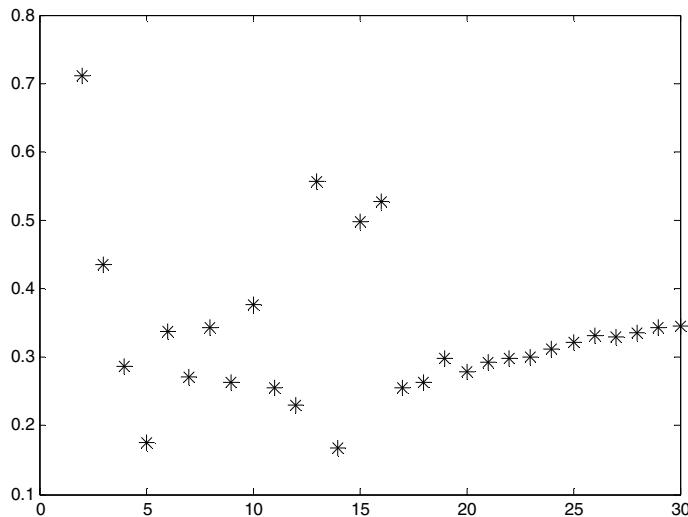


Fig. 2. Index plot of dataset 1 with hybrid Dunn-CEGR index

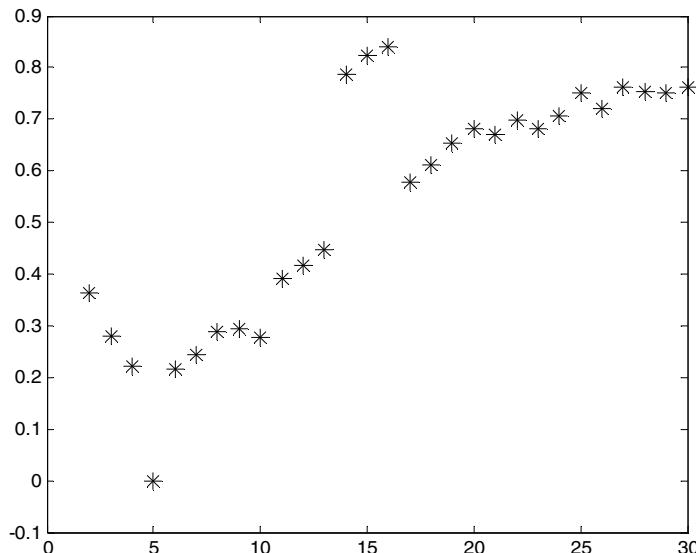


Fig. 3. Index plot of dataset 2 with hybrid Dunn-CEGR index

Table 1. Optimal number of clusters obtained

Dataset / index	Dunn index	Hybrid Dunn-CEGR index
Dataset 1	5	14
Dataset 2	5	5

From Table 1 we can observe that Dunn index gave 5 clusters as optimal for both datasets. This is because there are 5 groups of 3 close clusters and we get 5 clusters by clubbing 3 close clusters to a single cluster and Dunn index value is optimal at 5 clusters. Result obtained by Dunn index is correct for dataset 2. For dataset 1 difference between expected result (15 clusters) and obtained result (5 clusters) is 10 clusters. Hybrid Dunn-CEGR index gave 14 clusters for dataset 1 and hence error in number of clusters is just 1 cluster. For dataset 2 hybrid index gave correct result of 5 clusters. This is because we are using some class labels available for cluster validation in addition to Dunn index.

Both datasets used differ only in class labels. When there are group of close clusters in the dataset then all these close clusters may be clubbed and considered as single large cluster or each cluster in the group of close clusters can be considered as separate cluster. But using only internal information present in the dataset will give same result for both cases. Hence using little external information available for cluster validation in addition to internal information present in data can give different result for both cases according to available external information.

There is scope for creating various new cluster validation indices by combining other internal validation and external validation indices.

5 Conclusion and Future Work

In this paper, we proposed various hybrid cluster validation indices for semi-supervised clustering. We obtained results on some synthetic datasets by using an internal validation index and a hybrid index for cluster validation. Results obtained on datasets used show that when only internal validation is used error in number of clusters expected and number of clusters obtained is higher compared to using proposed hybrid index. In hybrid cluster validation benefits of internal and external validation are combined to get better results compared to using internal validation and external validation separately. Our future work includes creating hybrid cluster validation techniques for subspace and projected clustering methods.

References

- [1] Rendon, E., Abundez, I., Arizmendi, A., Quiroz, E.M.: Internal versus External cluster validation indexes. International Journal of Computers and Communications 5(1), 27–34 (2011)
- [2] Bolshakova, N., Azuaje, F.: Cluster validation techniques for genome expression data. Signal Processing 83(4), 825–833 (2003)
- [3] Bolshakova, N., Azuaje, F.: Machaon CVE: cluster validation for gene expression data. Bioinformatics 19(18), 2494–2495 (2003)

- [4] Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comp. App. Math.* 20, 53–65 (1987)
- [5] Dunn, J.: Well separated clusters and optimal fuzzy partitions. *J. Cybernetics* 4, 95–104 (1974)
- [6] Davies, D.L., Bouldin, D.W.: A cluster separation measure. *IEEE Transactions on Pattern Recognition and Machine Intelligence* 1(2), 224–227 (1979)
- [7] Hubert, L., Schultz, J.: Quadratic assignment as a general data-analysis strategy. *British Journal of Mathematical and Statistical Psychologie* 29, 190–241 (1976)
- [8] Bolshakova, N., Azuaje, F.: Estimating the number of clusters in DNA microarray data. *Methods of Information in Medicine* (2006)
- [9] Dimitriadou, E., Dolnicar, S., Weingessel, A.: An examination of indexes for determining the Number of Cluster in binary data sets. *Psychometrika* 67(1), 137–160 (2002)
- [10] Dudoit, S., Fridlyand, J.: A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology* 3(7) (2002)
- [11] Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On Clustering Validation Techniques. *Intelligent Information Systems Journal* 17(2), 107–145 (2001)
- [12] Gajawada, S., Toshniwal, D., Patil, N., Garg, K.: Optimal clustering method based on genetic algorithm. In: Deep, K., Nagar, A., Pant, M., Bansal, J.C. (eds.) *Proceedings of the International Conf. on SocProS 2011*. AISC, vol. 131, pp. 295–304. Springer, Heidelberg (2012)
- [13] Niu, Z.-Y., Ji, D.-H., Tan, C.-L.: Document Clustering Based on Cluster Validation. In: *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*, CIKM 2004 (2004)
- [14] Kryszczuk, K., Hurley, P.: Estimation of the Number of Clusters Using Multiple Clustering Validity Indices. In: El Gayar, N., Kittler, J., Roli, F. (eds.) *MCS 2010. LNCS*, vol. 5997, pp. 114–123. Springer, Heidelberg (2010)
- [15] Pihur, V., Datta, S., Datta, S.: Weighted rank aggregation of cluster validation measures: A Monte Carlo cross-entropy approach. *Bioinformatics* 23(13), 1607–1615 (2007)
- [16] Demiriz, A., Bennett, K.P., Embrechts, M.J.: Semi-supervised clustering using genetic algorithms. *Artificial Neural Networks in Engineering*, 1–20 (1999)
- [17] Patil, B.M., Joshi, R.C., Durga, T.: Effective framework for prediction of disease outcome using medical datasets: clustering and classification. *Int. J. Computational Intelligence Studies* 1(3) (2010)
- [18] Tan, P.N., Steinbach, M., Kumar, V.: *Introduction to Data Mining*. Pearson Education (2009)

Energy Efficient and Minimal Path Selection of Nodes to Cluster Head in Homogeneous Wireless Sensor Networks

S. Taruna, Sheena Kohli, and G.N. Purohit

Computer Science Department,
Banasthali University, Rajasthan, India
staruna71@yahoo.com, sheena7kohli@gmail.com,
gn_purohitjaipur@yahoo.co.in

Abstract. Wireless sensor networks (WSN) provide the availability of small and low-cost sensor nodes with capability of sensing physical and environmental conditions, data processing, and communication. These sensor nodes have limited transmission range, processing and storage capabilities as well as their energy resources. Routing protocols for wireless sensor networks are responsible for maintaining the energy efficient routes in the network and have to ensure extended network lifetime. In this paper, we propose and analyze a new approach of routing by clustering. The new cluster head selection approach by a homogeneous sensor node (having same initial energy) in wireless sensor network has been proposed, which involves choosing the cluster head which lies closest to the midpoint of the base station and the sensor node. Our proposed routing algorithm is related with energy and distance factors of each nodes. This scheme is then compared with the traditional LEACH protocol which involves selecting the cluster head which is nearest to the particular node. We conclude that the proposed protocol effectively extends the network lifetime with less consumption of energy in the network.

Keywords: Wireless Sensor network, Cluster head, Routing protocol, Network lifetime, Alive nodes.

1 Introduction

A wireless sensor network (WSN) consists of a large number of small autonomous devices called sensors or nodes, capable of sensing the environment, processing information locally and sending it to the point of collection through wireless links in a particular geographical area. WSNs are scalable and smart. The sensors can communicate directly among themselves or to some base station deployed externally in the area. But being autonomous nodes , they have limited battery , processing power and bandwidth. Of all the resources constraints, limited energy is most concerning one . One of the main design goals of WSNs is to carry out energy efficient data communication while trying to prolong the lifetime of the network.[6]

Routing in wireless sensor networks is very challenging due to the essential characteristics that distinguish wireless sensor networks from other wireless networks. It is highly desirable to find the method for energy efficient route discovery and relaying of data from sensor node to base station so that lifetime of network is maximized.

Much research has been done in recent years and still there are many design options open for improvement. Thus, there is a need of a new protocol scheme, which enables more efficient use of energy at individual sensor nodes to enhance the network survivability.[5]

In this paper, we analyze energy efficient homogeneous clustering head selection algorithm by a sensor node for WSN. We first describe the new distance based scheme and its different scenarios , and then the simulation results in MATLAB[2].

Further, the performance analysis of the proposed scheme is compared with benchmark clustering algorithm LEACH[4].

2 Related Work

Routing is a process of selecting a path in the network from source to destination along which the data can be transmitted. Various protocols [3] like LEACH, HEED, PEGASIS, TEEN, APTEEN are available to route the data from node to base station in WSN.

Sensors organize themselves into clusters and each cluster has a leader called as cluster head(CH), i.e. sensor nodes form clusters where the low energy nodes called cluster members (CM) are used to perform the sensing in the proximity of the phenomenon. For the cluster based wireless sensor network, the cluster information and cluster head selection are the basic issues. The cluster head coordinates the communication among the cluster members and manages their data.[1]. The process of clustering in routing provides an efficient method for maximizing the lifetime of a wireless sensor network by rotating the role of cluster head.

Low-energy adaptive clustering hierarchy (LEACH)[4] is a popular energy-efficient clustering algorithm for sensor networks. It involves distributed cluster formation. LEACH randomly selects a few sensor nodes as CHs and rotate this role to evenly load among the sensors in the network in each round. In LEACH, the cluster head (CH) nodes compress data arriving from nodes that belong to the respective cluster, and send an aggregated packet to the base station. A predetermined fraction of nodes, p, elect themselves as CHs in the following manner. A sensor node chooses a random number, r, between 0 and 1. If this random number is less than a threshold value, $T(n)$, the node becomes a cluster-head for the current round. The threshold value is calculated based on an equation that incorporates the desired percentage to become a cluster-head, the current round, and the set of nodes that have not been selected as a cluster-head in the last $(1/p)$ rounds, denoted by G. It is given by:

$$T(n) = \frac{p}{1 - p(r \bmod (1/p))} \quad \text{if } n \in G$$

Here, G denotes the set of nodes involved in the selection of CH. Each elected CH broadcasts a message to the rest of the nodes in the network to inform that it is the new cluster-head. A sensor node or non- CH selects the CHs which is nearest to it .

LEACH clustering terminates in a finite number of iteration, but does not guarantee good cluster head distribution. Some nodes may choose a cluster so that the distance between its CH and sink (base station) is even further than the distance between the node itself and the sink. According to the energy model of LEACH

protocol, the energy cost will increase as the distance increases. Battery power being limited in the sensor nodes, let the nodes to expire on full consumption of energy.

3 The Proposed Algorithm

In order to save the total energy cost of the sensor networks and prolong its lifetime, we propose a distance-based clustering protocol, LEACH-MP (LEACH-minimal path). The basic idea of the protocol is as follows:

Firstly some assumptions are addressed in this paper:

- All nodes can send data to Base station (BS).
- The BS has the information about the location of each node. It's assumed that the cluster heads and nodes have the knowledge of its location.
- Data compression is done by the Cluster Head.
- In the first round, each node has a probability p of becoming the cluster head.
- All nodes are of same specification.
- All nodes in the network are having the same energy at starting point and having maximum energy.
- Energy of transmission depends on the distance (source to destination) and data size.
- Nodes are uniformly distributed in network in a random manner.

Like LEACH, the operation of LEACH-MP is also divided into rounds. Each round begins with a set-up phase and steady phase. We do not change the way LEACH elects its cluster heads but changed the cluster formation algorithm. After the cluster heads are selected, cluster-heads broadcast an advertisement message that includes their node ID as the cluster-head ID and location information to inform non-cluster head nodes. Non-cluster head nodes first record all the information from cluster heads within their communication range. Then the node finds the cluster head which is closest to the middle-point between the node itself and the sink and joins that cluster. In other words, we changed the way how nodes join the cluster in order to decrease the total energy cost of the network and prolong the network lifetime.

Next the mathematical analysis will be given about the new scheme.

4 Network and Energy Consumption Model

The transmission energy of transmitting a k-bit message over a distance t is given by:

$$\begin{aligned} E_{TX(k,t)} &= E_{TX-elec(k)} + E_{TX-amp(k,t)} \\ &= kE_{elec} + kE_{fs}t^2 \end{aligned} \quad (1)$$

E_{elec} is the transmitter circuitry dissipation per bit , E_{amp} is the transmit amplifier dissipation per bit and E_{fs} is the dissipation energy per bit.

The receiving energy cost is:

$$\begin{aligned} E_{RX(k)} &= E_{RX-elec(k)} \\ &= kE_{elec} \end{aligned} \quad (2)$$

The total energy cost of a network is given by:

$$E_{\text{total}} = E_{\text{TX}} + E_{\text{RX}} + E_{\text{I}} + E_{\text{S}} \quad (3)$$

which needs to be minimized i.e. Min (E_{total})

Here, E_I is the energy cost during idle state. E_S is the energy cost while sensing.

Generally the three cost except the transmission cost are constant for a node. Only E_{TX} needs to be considered. So, we have to find Min (E_{TX}).

The energy cost mainly depends on the distance t, taking all other variable in the equation constant. Thus, we derive that we have to optimize Min (t²). The distance between a sensor node and a cluster head is denoted as dNtoCH and that between a cluster head and a sink as dCHtosink.

According to the energy model, we further simplify the optimization goal to minimize t² as Min (dNtoCH² + dCHtosink²).

As shown in the Fig. 1, let the triangle SNC depicts the position of node, CH and sink or Base Station respectively, where, S is the BS, C the CH and N the sensor node. The distance between sensor node N and sink S is dNtoSink = z, dNtoCH = y and dCHtosink = x.

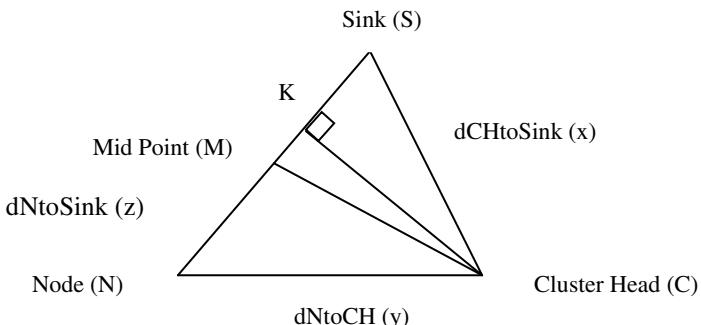


Fig. 1. The basic concept

Further a perpendicular is drawn from C on line SN at point K. The length of this perpendicular is s. M is the mid point between node and sink. The distance between mid point of node and sink and that of CH is given by t.

Thus. KM=p, CM=t, CK=s

From the rule of trigonometry, applying pythagoras theorem,

In Δ SKC,

$$\begin{aligned} dCHtoSink^2 &= CK^2 + (dNtoSink / 2 - p)^2 \\ x^2 &= s^2 + (z/2 - p)^2 \end{aligned} \quad (4)$$

In Δ NKC,

$$\begin{aligned} dNtoCH^2 &= CK^2 + (dNtoSink / 2 + p)^2 \\ y^2 &= s^2 + (z/2 + p)^2 \end{aligned} \quad (5)$$

Combining both equations (4) and (5),

$$\begin{aligned} x^2 + y^2 &= s^2 + (z/2 - p)^2 + s^2 + (z/2 + p)^2 \\ &= 2s^2 + z^2/2 + 2p^2 \end{aligned} \quad (6)$$

From ΔMKC ,

$$\begin{aligned} t^2 &= p^2 + s^2, \text{ so substituting } p^2 = t^2 - s^2 \\ \text{we get, } x^2 + y^2 &= z^2/2 + 2t^2 \end{aligned} \quad (7)$$

We can see that when the value of dNtoSink is fixed, $dNtoCH^2 + dCHtosink^2$ is only related to t i.e. $\text{Min}(dNtoCH^2 + dCHtosink^2)$ is equivalent to $\text{Min}(t^2)$. As a result, if a node chooses its CH which is closest to the mid point of this node and the sink, the squared distance of their communication is smallest.

$\text{Min}(dNtoCH^2 + dCHtosink^2)$ is to actually minimize the distance between the CH and the midpoint of a node and the BS when the distance between the node and the BS is fixed. Thus, in LEACH-MP, non-cluster nodes select the CH which is nearest to the midpoint between itself and the BS as its communication CH for minimizing the communication cost.

5 Simulation and Performance Evaluations

We choose MATLAB for simulation. The protocol is compared to the LEACH algorithm giving results on the comparison of energy consumption and network lifetime under different scenarios.

5.1 Simulation Parameters

Table 1. Simulation Parameters

Parameter	Values
Simulation Round	2000
Number of nodes	100
CH probability	0.1
Fusion rate (cc)	0.6
Initial node power	0.5 Joule
Nodes Distribution	Nodes are uniformly distributed
Packet size (k bits)	4000
Energy dissipation (Efs)	10*0.000000000001 Joule
Energy for Transmission (E_{TX})	50*0.000000000001 Joule
Energy for Reception (E_{RX})	50*0.000000000001 Joule
Energy for Data Aggregation (EDA)	5*0.000000000001 Joule

5.2 Simulation Result

5.2.1 Energy Consumption with Different Sink Location

The energy consumption of LEACH-MP protocol has been compared with that of LEACH by changing the location of Base station. The simulation was done on a network of area 200 x 200 and the energy consumed by the network was calculated by letting BS to be at different locations i.e. (100,100), (100,200),(100,250) and (100,300).

The graph shown in Fig.2. depicts that in each case, the energy consumption of LEACH-MP is always less than that of LEACH, even on changing different BS locations in the network.

Table 2. Energy consumption on changing BS position

Sim.Run	Energy consumption in μJ							
	BS(100,100)		BS(100,200)		BS(100,250)		BS(100,300)	
	leach	MP	leach	MP	leach	MP	leach	MP
1	33.62	42.21	23.1	125.662	125.79	162.24	149.61	273.61
2	16.8	75.89	10.57	71.73	71.96	338.99	115.58	177.6
3	42	91.78	43.1	212.7	24.62	216.03	112.08	156.21
4	37.5	104.3	33.21	235.65	16.61	101.52	193.39	410.42
5	25.9	56.11	26.61	167.71	38.34	153.9	92.61	458.33
6	15	26.1	16.71	38.09	192.43	310.09	87.82	377.91
7	62.7	208.09	18.01	38.8	74.65	122.02	32.81	200.62
8	45.25	98.12	50.98	197.06	54.8	232.97	84.44	371.97
9	75.75	185.1	64.34	128.34	16.1	82.58	108.82	391.55
10	46.23	66.42	70.12	169.23	28.97	64.26	136.93	182.06

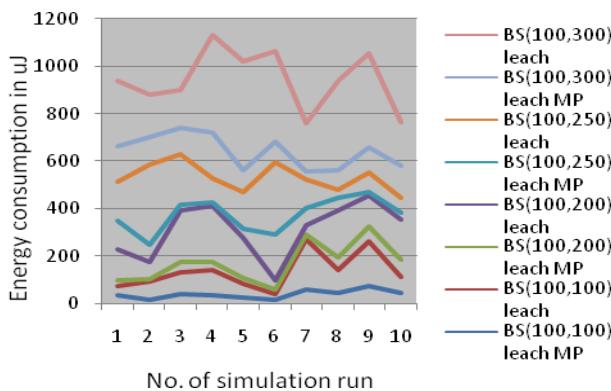


Fig. 2. Energy consumption vs simulation run on changing BS position

5.2.2 Network Lifetime with Different Sink Location

For analyzing the network lifetime, the number of nodes which remain alive after each simulation run was compared for both the protocols.

We can conclude that LEACH-MP extends the network lifetime as compared to LEACH, as the number of nodes which remain alive in the end of each simulation is more in LEACH-MP than that in LEACH, no matter where the sink is located, keeping network size as 200x200. The result is shown in Fig. 3.

Table 3. Network lifetime on changing BS position

Sim. Run	No. of alive nodes			
	BS(100,100)		BS(100,200)	
	leach	leach MP	leach	leach MP
1	42	96	19	54
2	36	90	17	51
3	46	99	11	23
4	40	100	10	26
5	43	95	21	58
6	37	96	20	52
7	59	94	25	56
8	40	95	19	50
9	59	98	27	44
10	50	97	18	49

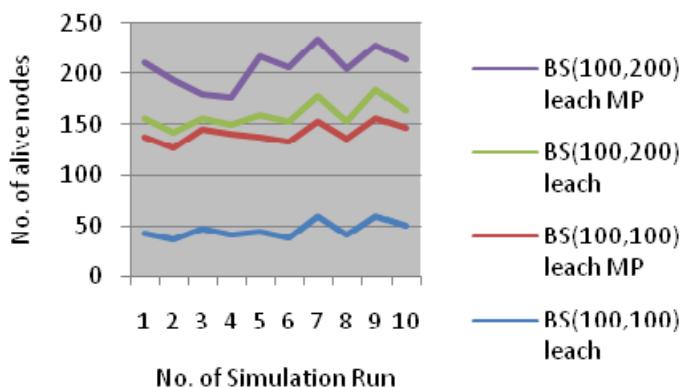


Fig. 3. No. of alive nodes vs simulation run on changing BS position

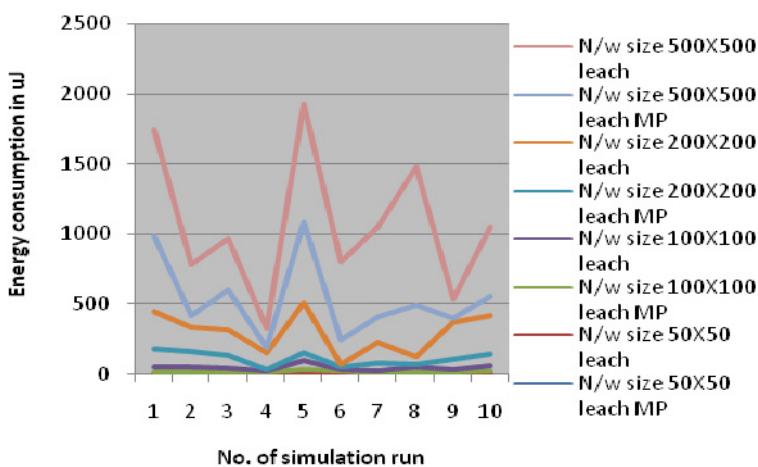
5.2.3 Energy Consumption with Different Network Size

Further the energy consumption of the network was calculated by changing the size of the network area. i.e. (50x50), (100x100), (200x200) and (500x500). The BS is located at the centre in each case.

The simulation results are shown in Fig.4. The consumption of energy is more in LEACH as compared to LEACH-MP protocol in each case, even on changing the network size.

Table 4. Energy consumption on changing network size

Sim.Run	Energy consumption in μJ							
	N/w size 50X50		N/w size 100X100		N/w size 200X200		N/w size 500X500	
	leach	MP	leach	MP	Leach	MP	Leach	MP
1	2.48	6.07	11.2	39.7	122.21	259.93	539.64	763.5
2	3.92	6.26	10.59	31.25	107.43	172.88	88.99	367.39
3	2.28	5.68	13.28	21.8	93.02	178.58	289.07	363.14
4	2.58	8.1	7.04	12.32	7.53	112.22	42.22	138.91
5	4.2	14.56	19.9	61.34	55.51	352.21	577.64	840.96
6	8.48	11.17	6.25	14.18	11.37	20.08	175.02	554.44
7	1.92	6.01	2.64	17.7	53.19	140.38	185.17	639.78
8	3.38	4.52	10.96	38.71	19.04	50.78	364.85	991.69
9	1.87	5.2	13.31	20.71	66.2	261.34	30.78	144.05
10	1.35	8.31	18.38	31.67	82.84	275.05	141.52	493.33

**Fig. 4.** Energy consumption vs simulation run on changing network size

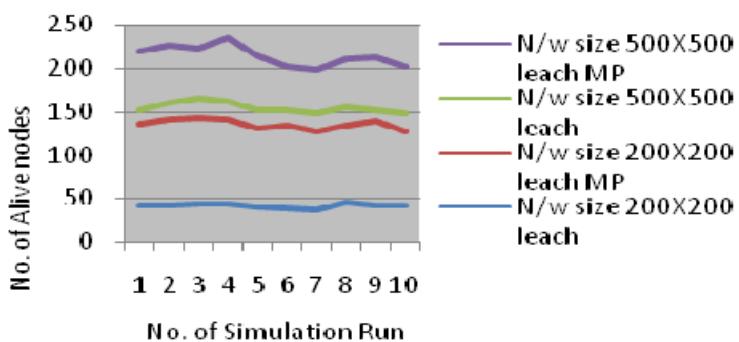
5.2.4 Network Lifetime with Different Network Size

Similarly, we changed the network area size again and calculated the lifetime of the network with both the protocols. The results are shown in Fig.5.

It was concluded that the number of nodes remaining alive at the end of each simulation run was more in LEACH-MP as compared to LEACH in each case, whatever the network size be.

Table 5. Network lifetime on changing network size

Sim. Run	No. of alive nodes			
	N/w size 200X200		N/w size 500X500	
	leach	leach MP	leach	leach MP
1	43	93	15	68
2	42	100	18	65
3	44	99	23	56
4	45	97	21	71
5	40	92	19	63
6	39	95	17	52
7	37	91	20	51
8	46	88	22	54
9	43	97	12	60
10	42	86	20	55

**Fig. 5.** No. of alive nodes vs simulation run on changing network size

6 Conclusion and Future Work

Energy consumption is the main design issue in routing of Wireless Sensor Networks. We concluded that energy consumed for the cluster head selection is less in the proposed algorithm, where we choose the cluster head which lies closest to the midpoint of the base station and the sensor node , which directly shows the increased network survivability. Further the network lifetime of the proposed algorithm has greater span than the LEACH protocol, even on changing the network size and sink position. The proposed algorithm is for the homogeneous network and we propose to extend this work for the heterogeneous network.

References

1. Al-Karaki, J., Kamal, A.: Routing Techniques in Wireless Sensor Networks: A Survey. *IEEE Communications Magazine* 11(6), 6–28 (2004)
2. <http://www.mathworks.in>
3. Joshi, A., Lakshmi Priya, M.: A Survey of Hierarchical Routing Protocols in Wireless Sensor Network. In: International Conference on Information Systems
4. Heinzelman, W.R., et al.: Energy-Efficient Communication Protocol for Wireless Micro-sensor Networks. In: Proceeding of the 33rd Hawaii International Conference on System Sciences, pp. 1–10 (January 2000)
5. Taruna, S., Lata, J.K., Purohit G.N.: Zone Based Routing Protocol for Homogeneous Wireless Sensor Network. *International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC)* 2(3) (September 2011), doi: 10.5121/ijasuc.2011.2307 99
6. Chandramathi, S., Anand, U., Ganesh, T., Sriraman, S., Velmurugan, D.: Energy Aware Optimal Routing for Wireless Sensor Networks. *Journal of Computer Science* 3(11), 836–840 (2007) ISSN 1549-3636 © 2007 Science Publications

Texel Identification Using K-Means Clustering Method

S. Padmavathi, C. Rajalaxmi, and K.P. Soman

Amrita School of Engineering, Coimbatore, Tamil Nadu, India

s_padmavathi@cb.amrita.edu, crajiz@yahoo.co.in, kp_soman@amrita.edu

Abstract. Identifying the smallest portion of the image that represents the entire image is a basic need for its efficient storage. Texture can be defined as a pattern that is repeated in a specific manner. The basic pattern that is repeated is called as Texel(Texture Element). This paper describes a method of extracting a Texel from the given textured image using K means clustering algorithm and validating it with the entire image. The number of gray levels in an image is reduced using a linear transformation function. The image is then divided in to sub windows of certain size. These sub windows are clustered together using K-means algorithm. Finally a heuristic algorithm is applied on the cluster labels to identify the Texel, which results in more than one candidate for Texel. The best among them is then chosen based on its similarity with the overall image. The similarity between the Texel and the image is calculated based on then Normalized Gray level co-occurrence matrix in the maximum gradient direction. Experiments are conducted on various texture images for various block sizes and the results are summarized.

1 Introduction

The storage of images in a large database requires enormous amount of memory. Retrieval and transmission of these images consumes more time for larger image size. Efficient representation of the images is thus a greater demand. The content based image retrieval algorithms extracts features to represent the image. These features have limited capacity and accuracy in representing the images. Instead of extracting features from an image if the image is represented with a smallest version of itself, the storage and querying on an image will be improved greatly. In this regard this paper aims in extracting such a representation for a textured image. A textured image consists of repeating patterns. The smallest pattern identified to be repeating, which is capable of generating a larger version through translation and/or rotation is called a Texel. This implies that a Texel is capable of synthesizing larger textures. These Texels could be used as primitives. The position and placement of these Texels could be described by productions, leading to Structural representation of textures. Given a Texel, textures can be generated using various tiling methods as discussed by Paul Bourke in [12]. Identifying the Texel in an image is an area which is not much explored. In this paper images with regular textures are considered for experimentation. The image is subdivided into smaller blocks and grouped based on their similarity by K-means clustering algorithm. The block size is kept smaller in order to avoid

missing out a Texel smaller than the block. The combination of these blocks will give rise to Texel. Since large numbers of such combinations are available, a selection procedure is applied to assess its capability to represent the entire texture. The one that closely represents the textured image is then given as output.

2 State of Art

2.1 Texture Analysis

Image texture is defined as a function of the spatial variation in pixel intensities as shown in Fig 1. Texture is the most important visual cue in identifying homogeneous regions, such as cotton canvas, straw matting, raffia, pressed calf leather etc. It is also useful in analysis of aerial images, biomedical images and seismic images as well as the automation of industrial applications.

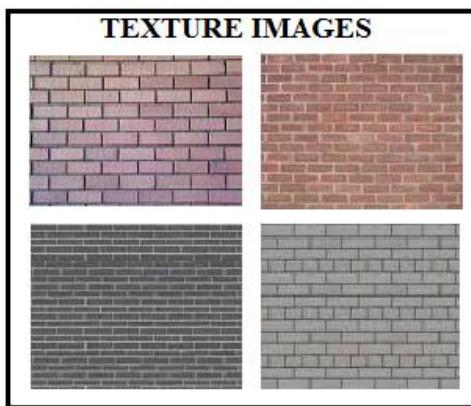


Fig. 1. Texture images

Texture can be represented using statistical, structural or spectral methods. In statistical approaches statistics extracted from the gray-level histogram or Gray-Level Co-occurrence Matrix (GLCM) are used for describing textures. The GLCM methods describe texture better than histograms [3]. GLCM is specified for a specific orientation and distance parameter. The drawback of this method is that the size of the matrix depends on the number of gray levels. The number of such matrices for representing the texture depends on the different orientations and the distance parameter. These parameters again depend on the input texture. For proper representation of texture more matrices need to be considered and hence more computations are required. The application of GLCM could be found in [7] and [10]. In spectral approaches frequency based representations are used. Gabor filters methods dominate the texture representation in the frequency domain. Determining the values for the parameters like scale, orientation and frequency plays a key role in Gabor representation. In [9] Gabor filters are used for texture recognition. Identifying the repeating

spatial pattern is difficult in such representation. The statistical and spectral methods could be considered as bottom up approaches since they are extracted from the image. In structural approaches, the basic element of texture, the Texel, and a set of grammar rules are used for representing the texture. The grammar rules specify the placement rule of the Texel to generate the entire texture pattern. Since it is used to generate the image this could be considered as a top down approach. For example a chessboard image can be considered as the texture image and the texels are the white square and the black square. The placement of these texels becomes the grammar to generate the chessboard. This is a highly powerful method of representing texture. But it is seldom used in practical image application since identifying the Texel is a difficult task. Other variations of texture representation could be found in [2],[8] and [11].

2.2 Clustering

Clustering refers to the process of grouping samples so that the samples are similar within the group. These groups are called as clusters. Data clustering is a common technique for statistical data analysis, which is used in many fields including machine learning, data mining, pattern recognition, image analysis and bio-informatics. In image analysis, the pixels are clustered based on their similar gray level which is used for segmenting the image based on color.

The computational task of classifying the data set into ‘k’ groups is often referred to as k- clustering. K-means is one of the simplest unsupervised learning algorithms used for it. Unsupervised learning is a type of machine learning where cluster labels of the input are not known priorly. K means algorithm is an iterative, non-hierarchical approach used to form desired number of clusters ‘k’ from ‘n’ samples of data, where $k < n$. More details on clustering algorithm could be found in [1], [5] and [13].

3 Proposed Technique

A texture image is given as input. Since the Texels are larger than a pixel, the image is divided into sub windows of a specific size. If all the sub-windows are similar that should be the Texel of the image. Hence the problem reduces to identifying the smallest sub-windows that is repeated in the image. K-means clustering algorithm is used to group the similar sub windows which are given a common cluster label. To cluster the sub-windows the positional gray levels are used for finding the similarity. The gray levels are hence reduced by using a linear transformation given in equation 1. This process is called as binning.

$$I_b = (\max_b - \min_b) * ((I - \min_i) / (\max_i - \min_i)) + \min_b \quad (1)$$

Where I_b represents the binned image and I represents the input image, \min_i and \max_i represent the maximum and minimum gray level in the input image and \min_b and \max_b represents the minimum and maximum gray level in the binned image. \min_b is kept as zero and \max_b is kept to be the number of gray levels required in the binned image.

3.1 K Means Clustering

Initially k clusters are taken arbitrarily and the remaining samples are assigned to one of the ‘k’ clusters so as to minimize the measure of dispersion within the clusters. The objective function J, given in equation 2 is minimized.

$$J_{sse} = \sum_{i=1}^c \sum_{x \in D_i} \|x - \mu_i\|^2 \quad (2)$$

Where

$$\mu_i = \frac{1}{n_i} \sum_{x \in D_i} x \quad (3)$$

represents the mean of i^{th} cluster, x represents the input sample, c represents the number of clusters, D_i represents i^{th} cluster, n_i represents the number of samples in i^{th} cluster. The algorithm is given as follows:

1. Begin with ‘k’ clusters, each consisting of the first ‘k’ samples. For each of the remaining ($n-k$) samples, find the mean closer to it. Assign the samples to the cluster identified with this nearest mean. After each sample is assigned, recompose the mean of the altered cluster
2. Go through the data for the second time. For each sample, find the mean nearest to it. Alter the samples in the cluster accordingly.
3. If no samples changed clusters, stop the process.
4. Re-compute the means of altered clusters and go to step 2.

An explanatory diagram is shown in Fig. 2 where X represents the cluster means and the dotted lines represent the cluster.

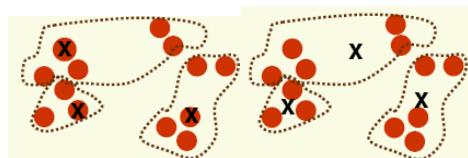


Fig. 2. Shows the initial assignment and change of means after re-computing

The Silhouette plot is used to plot cluster data and the cluster indices. The Silhouette Plot (Fig 3) displays a measure of how close each block in one cluster is to blocks in the neighboring clusters.

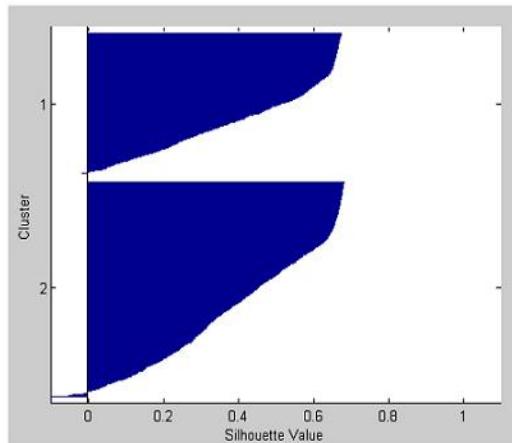


Fig. 3. Silhouette plots of the clusters

A value of +1 indicates blocks that are very distant from neighboring clusters. A value of 0 indicates blocks that are not distinctly in one cluster or another. A value of -1 indicates blocks that are probably assigned to the wrong classifier. If the cluster is identified as negative then the value of k is incremented and the process is repeated again. In another variation, when the value occurs as -1 the size of the sub window is increased and the experiment is repeated.

3.2 Texel Identification

The cluster indices of each block of pixels in the input image are taken as a matrix. For each block the cluster indices of the 8 neighbors are obtained. A count of the indices that are different from the central one is stored. If this count is greater than a Threshold, the central block together with the 8 neighbors form a candidate of Texel. This results in more than one candidate for texel. Among these set of texels, duplicated texels based on the cluster index are identified and eliminated. Since more than one Texel is available the one which is very close to the original image is chosen as the Texel.

To find the similarity with the original image the Gray level co occurrence matrix (GLCM) is calculated for the input texture image I. A GLCM is a second order statistic method. It is a $G \times G$ matrix P , in which G represent the set of possible grey level values in the image. The GLCM is defined by:

$$P_d[i, j] = n_{ij} \quad (4)$$

where n_{ij} is the number of occurrences of the gray levels (i, j) with a specific displacement vector d . The displacement vector depends on the distance and orientation between the pixels. GLCM is normalized for image size invariance as $N[i, j]$, defined by:

$$N[i,j] = \frac{P[i,j]}{\sum_i \sum_j P[i,j]} \quad (5)$$

The distance parameter is fixed based on the size of the Texel. For fixing the orientation parameter gradient operator is applied on the input image. If G_x and G_y represents the gradient along the x and y direction respectively the gradient angle is calculated as in Equation 6.

$$\Theta = \tan^{-1}(G_y/G_x) \quad (6)$$

The angle are quantized to the nearest of 0, 45, 90 and -45 degrees. The frequently occurring gradient angle in the image is calculated as in Equation. 7

$$M_\theta = \max\{ n_\theta \} \quad (7)$$

Where n_θ represents number of occurrence of the angle θ . M_θ is taken as orientation parameter. GLCM of the input image is calculated for this orientation and distance. Similarly the GLCM for each of the unique Texel image is calculated. Euclidean distance S_e given in equation 8 and correlation metrics S_c given in equation 9 are calculated between each Texel and the image. The Texel that has minimum S_e or maximum S_c is chosen as the best Texel.

$$S_e = \sqrt{\sum_{i=1}^G \sum_{j=1}^G \| P_{Dij} - p_{Dij} \|^2} \quad (8)$$

$$S_c = \frac{\sum_{i=1}^G \sum_{j=1}^G (P_{Dij} - \bar{P}_D)(p_{Dij} - \bar{p}_D)}{\sqrt{\left(\sum_{i=1}^G \sum_{j=1}^G (P_{Dij} - \bar{P}_D)^2 \right) \left(\sum_{i=1}^G \sum_{j=1}^G (p_{Dij} - \bar{p}_D)^2 \right)}} \quad (9)$$

Where P_D and p_D represents the GLCM of the image and texel candidate respectively in the direction of M_θ . \bar{P}_D and \bar{p}_D represents their corresponding mean values.

4 Experimental Results

Brodatz texture images[6] and few other images are considered for experimentation. The graylevels of the images were binned using equation 1 for various sizes. Binning

with 8 gray levels gives better result for the data set. The input image \mathbf{l} of size $(M \times N)$ is subdivided into blocks, each of size $(r \times c)$. The input image is resized to a size $(m \times n)$, which is multiple of the block's size. This converts the given image into a $(r * c) \times (m/r * n/c)$ matrix (\mathbf{X}) , where each element represents a sub-window. This matrix is given to the k-means algorithm. The algorithm partitions the blocks given into ' k ' mutually exclusive clusters. It returns a vector (\mathbf{L}) of dimension $(m/r * n/c) \times 1$ consisting of cluster indices of each block. The number of clusters ' k ' has to be passed as parameter to this function. ' k ' is initialized to 2 because in the beginning there should be at least two clusters. Euclidean distance is used to measure the closeness of the sample to the cluster. The sub window size is varied as specified in section 3.1. The vector (\mathbf{L}) is converted to a matrix of dimension $(m/r) \times (n/c)$ called as (\mathbf{T}) . So (\mathbf{T}) now refers to the cluster indices of each block of pixels in the input image. The cluster indices of sub-window together with its 8 neighboring sub-windows are checked and a candidate Texel is chosen as specified in section 3.2. Fig. 4 shows a sample texture image with all candidates of texels displayed on the right. The texels with similar cluster labels are removed as shown in Fig.5. From the remaining Texels S_e and S_c are calculated and best texel is chosen. A snapshot is shown in Fig. 6. Fig 7a and Fig. 7b shows the images, the texels chosen based on the two factors. The binned gray levels and sub-window sizes are also shown Fig 8 shows few images with their texels. The sub window size is specified near the Texel.

The computational complexity of calculating the GLCMs for each Texel image is considerably reduced by binning the gray levels and fixing the parameters based on the gradient angle.

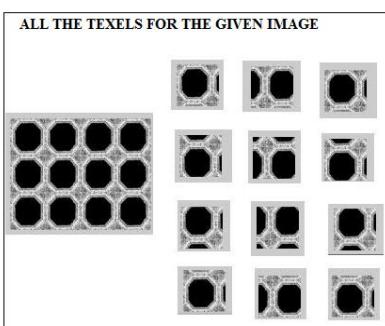


Fig. 4. Texels given by K-means algorithm

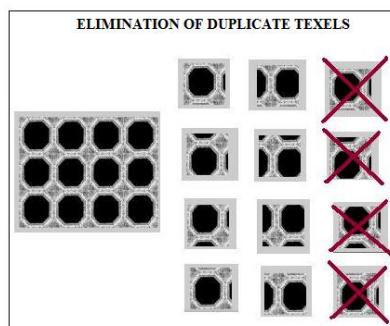


Fig. 5. Duplicate Texels removed

```

MATLAB
File Edit Debug Desktop Window Help
Current Directory: C:\Users\CRAJIZ\Desktop\Image_Processing_Papers\texel_ider
Shortcuts How to Add What's New

EUCLEDIAN DISTANCE BETWEEN TEXELS AND ORIGINAL BINNED IMAGE
0.0820    0.0609    0.0965    0.0923    0.0971    0.0921    0.0835    0.0607

ans =
0.0607

BEST TEXEL ACCORDING TO EUCLEDIAN DISTANCE IS
FIGURE
8

CROSS CORRELATION BETWEEN TEXELS AND ORIGINAL BINNED IMAGE
0.9758    0.9818    0.9316    0.9271    0.9303    0.9275    0.9778    0.9811

ans =
0.9818

BEST TEXEL ACCORDING TO CROSS CORRELATION IS
FIGURE
2


```

Fig. 6. Snapshot showing selection of best texel using Euclidean and cross correlation metric

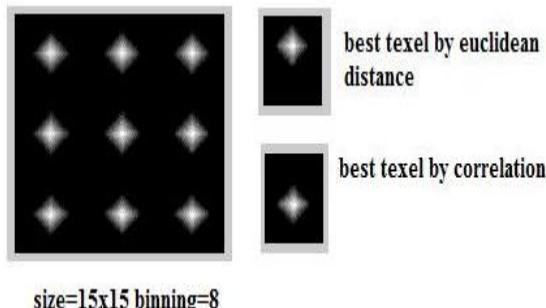


Fig. 7a. Image with best Texel

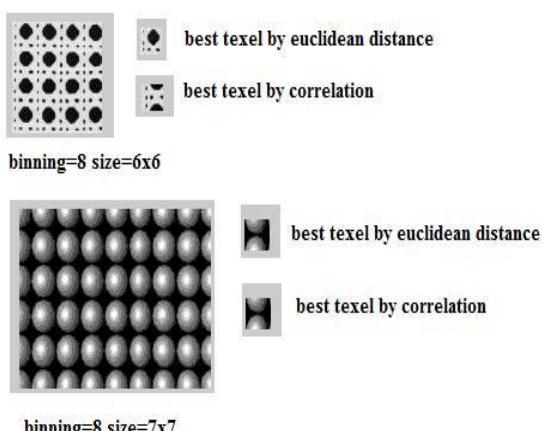


Fig. 7b. Image with best Texel

		size=6 x 6			size=6 x 6			size=10x10
		size=6 x 6			size=6 x 6			size=10x10
		size=6 x 6			size=6 x 6			size=10x10
		size=6 x 6			size=10x10			

Fig. 8. Table showing images and their Texels generated by the algorithm

This method has been implemented only on intensity component of the image; the color component is not taken into consideration. This method is applicable only for single textured images. K-means is an unsupervised method so the number of clusters needs to be specified. The initial window size has to be specified and the accuracy depends on it.

5 Conclusion and Future Work

This paper describes a method of identifying texel using k-means clustering. Instead of performing clustering on individual pixels, our method performs clustering on a group of pixels. The Texel is identified after clustering which has similar characteristic as the original texture. This ensures that the texel obtained is the repeating pattern that best represents the texture in the image.

This method provides a basic foundation for texture representation of practical images using structural methods. The texel extracted could be stored efficiently in large databases. By using Neural Networks a suitable window size could be generated instead of providing the window size. Further this method can be extended to classify multiple textures in a single image.

References

- [1] Gose, E., Johnsonbaugh, R., Jost, S.: Pattern Recognition and Image Analysis, pp. 199–215. Prentice Hall
- [2] Lee, B.: A New Method for Classification of Structural Textures. IEEE Transactions on International Journal of Control, Automation, and Systems 2(1) (March 2004)
- [3] Tuceryan, M., Jain, A.K.: Texture Analysis. IEEE Transactions on Systems, Man, and Cybernetics 2 (March 1998)
- [4] Veksler, O.: Lecture notes on Pattern Recognition,
<http://www.csd.uwo.ca/~olga/Courses//Winter2006//CS434.../index.html>
- [5] Hartigan, J.A.: Clustering Algorithm. Wiley (1975);
- [6] Brodatz Textures, <http://www.ux.uis.no/~tranden/brodatz.html>
- [7] Soh, L.-K., Tsatsoulis, C.: Texture Representation of SAR Sea Ice Imagery Using Multi-Displacement Co-Occurrence Matrices,
http://www.ittc.ku.edu/publications/documents/Soh1996_igarss96-6.pdf
- [8] Kruizinga, P., Petkov, N.: Grating cell operator features of oriented texture segmentation. Appeared in Proc. of the 14th Int. Conf. on Pattern Recognition, Brisbane, Australia, August 16-20, pp. 1010–1014 (1998)
- [9] Clausi, D.A., Deng, H.: Fusion of Gabor Filter and Co-occurrence Probability Features for Texture Recognition,
<http://www.eng.uwaterloo.ca/~dclausi/Papers/Clausi>
- [10] Hosseini Aria, E., Saradjian, M.R., Amini, J., Lucas, C.: Generalized Co-occurrence Matrix to classify IRS-1D Images using Neural Network,
<http://www.isprs.org/istanbul2004/comm7/papers/23.pdf>
- [11] Kuan, J.: Complex Texture Classification,
<http://www.mmrg.ecs.soton.ac.uk/publications/archive/kuan1997b/html/node1.html>
- [12] Bourke, P.: Tiling textures on the plane,
http://www.paulbourke.net/texture_colour/tiling
- [13] Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A review,
<http://www.nd.edu/~flynn/papers/Jain-CSUR99.pdf>

Single and Multi Trusted Third Party: Comparison, Identification and Reduction of Malicious Conduct by Trusted Third Party in Secure Multiparty Computing Protocol

Zulfa Shaikh¹ and Poonam Garg²

¹ Faculty of Computer Applications, Acropolis Institute of Technology & Research, Indore (M.P.)

shaikh.zulfa@gmail.com

² Deptt. of Information Technology, Institute of Management Technology, Ghaziabad pgarg@imt.edu

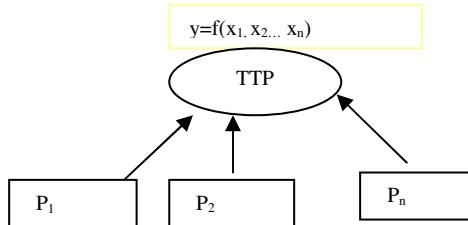
Abstract. SMC is a problem of n parties with inputs $(x_1, x_2 \dots x_n)$, hand over their inputs to third party for computation $f(x_1, x_2 \dots x_n)$ and third party announces the result in the form of y. During joint computation of inputs, all the organizations involved in computation wish to preserve privacy of their inputs. So need is to define a protocol which maintains privacy, security and correctness parameters of SMC. In this paper, single third party and multi third party model are defined and compared. The probabilistic evidences for single and multi third party SMC model have been analyzed with security analysis graphs. In this paper, we have also worked on identification and reduction of malicious conduct of TTPs in multi TTP environment.

Keywords: Secure Multiparty Computation (SMC), Trusted Third Party (TTP), Single TTP, multi TTP, privacy, security, and correctness.

1 Introduction

With the current trend in data growth and the corresponding needs for cooperative computation, it has become a challenge for organizations to maintain privacy, security and correctness during joint computation. SMC helps to jointly compute the private inputs of respective parties and announce the result of computation. SMC is a problem in which parties involved during computation hand their inputs $(x_1, x_2 \dots x_n)$ to third party for computation. Third party performs computation $f(x_1, x_2 \dots x_n)$ and announces the result in the form of y. In real world scenario, trust on third party performing the computation is doubtful, so need is to design and develop a protocol where parties privacy can be maintained and malicious conduct of third party can be identified and reduced. In figure 1, the general SMC framework has been defined.

In this paper, SMC model with single third party and multi third party have been defined. In single third party SMC framework, all the parties involved in computation hand their inputs to the single third party for computation whereas in multi third party

**Fig. 1.** General SMC Model

SMC framework, same computation is performed by number of trusted third parties selected at runtime. The paper also helps in identifying malicious conduct by TTPs during computation in multi TTP model and reduces the malicious conduct.

2 Background

SMC problem is the problem of n parties to compute a private function of their inputs in a secure method, where security means the correct result computed by the TTPs for maintaining the privacy of the parties as some of the parties may want to misuse the other party's data. We assume that the inputs are x_1, x_2, \dots, x_n where x_i is the data of party P_i and the TTP will compute a function $f(x_1, x_2, \dots, x_n) = y$ and announce the result y [1]. Security is meant to achieve correctness of the result of computation and keeping the party's input private even if some of the parties are corrupted. In figure 1, trusted third party is used for doing the computation on the inputs provided by the parties. According to [2], the major problem with this approach is that it is difficult to find the third party which is trusted by all the parties providing the inputs and to control the function of adversaries.

Yao's introduced the SMC problem in [3].The first solution uses a centralized TTP which is selected by majority of honest party, which shows synchronous system with cryptography [4].After handing up the inputs to a trusted third party, security increases but there is a chance that the trusted third party behaves like a malicious adversary. It was demonstrated analytically as well as experimentally, the performance characteristics and security and proved that for the range of numbers; Yao's protocol is secure [5]. The idea was extended to multiparty computation by many researchers [6]. They used circuit evaluation protocols for secure computation. Earlier research focused on theoretical studies. Later, some real life applications emerged like Private Information Retrieval (PIR) [7, 8], Privacy-preserving data mining [9, 10], Privacy-preserving geometric computation [11], Privacy-preserving scientific computation [12], Privacy preserving statistical analysis [13] etc. A detailed review of SMC research is provided by Du et al. in [14] where they developed a framework for problem discovery and converting normal problem to SMC problem. A review of SMC with special focus on telecommunication systems is given by Oleshchuk et al. in [15].

Aiming at privacy preserving computing of statistical distribution, which is frequently encountered in statistics, and based on the intractability of computing discrete logarithm and using rigorous logic, they proposed the solution.[16] presented the protocols allowing the players to securely solve standard computational problems in linear algebra

such as determinant of matrices product, rank of a matrix, and determine similarity between matrices. [17] Presented TASTY, a novel tool for automating, i.e., describing, generating, executing, benchmarking, and comparing, efficient secure two-party computation protocols. They used TASTY to compare protocols for secure multiplication based on homomorphic encryption with those based on garbled circuits and highly efficient multiplication. [18] Presented a hybrid-secure MPC protocol that provides an optimal trade-off between IT robustness and computational privacy.

3 Proposed Work

The basic need during SMC is to obtain correct results of computation maintaining privacy and security in the protocol. As trust on third parties performing joint computation is doubtful in real scenario. So the need is to define more secure protocol announcing the right result of computation. The objectives of present study are:-

- To define and compare single and multi third party computing model.
- To analyze and store the behavior of third parties in several rounds of computation.
- To identify and reduce malicious conduct of trusted third parties in order to obtain correct results of computation.

3.1 Architectural Framework

In figure 2, the architectural framework of SMC model is designed. The model works in two different environments: one is the single TTP, selected for performing the computation from a pool of TTPs and second, multiple TTPs performing computation on single function.

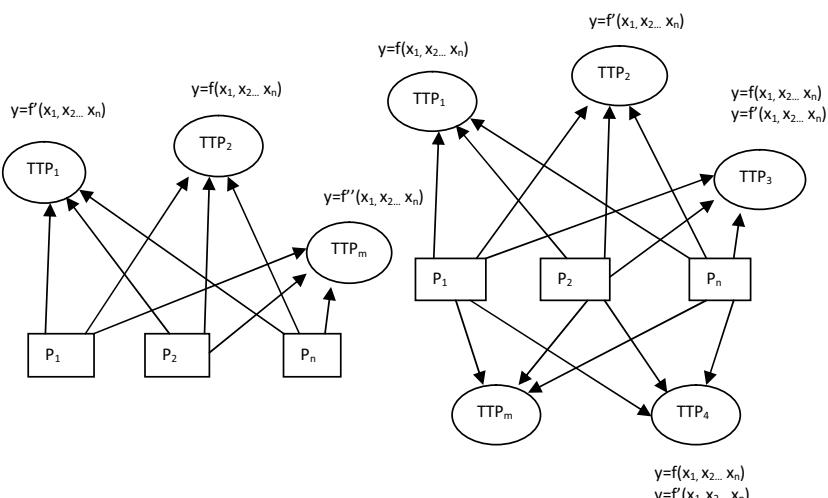


Fig. 2. SMC_ Single and Multi TTP Computation Model

The assumptions in both the models are:

- Due to critical mission data no party will share its information with other parties involved in computation.
- Parties provide their inputs to TTP or TTPs for computation.
- TTPs are selected at runtime from a pool of TTPs.

The major concern with this computation is that, what if TTPs involved in computation are malicious? The results will be:

- Correctness in the output cannot be ensured.
- The malicious TTPs may affect security and privacy issues of SMC as well.

Comparing single TTP with multi TTP model the advantage of multi TTP model is that the protocol does not rely on one TTP but on majority providing the identical results. This will give more clear identification in correctness of results. On the other hand, in single TTP model the protocol selects a single TTP for computation at runtime from a pool of TTPs. This makes the protocol inefficient as if, a TTP computing the function is malicious then it may announce the incorrect result of computation and there is no alternative, other than relying on the malicious TTP.

3.2 Probability Analysis

Case 1: Probability of malicious conduct in single and multi TTP model before selection of TTPs

In single TTP model if the TTPs are selected at runtime using randomization function (R_f) then the probability of malicious conduct is:

$$P(TTP_{single}) = 1/m \quad (1)$$

Here m is the total number of TTPs involved in the model.

Multi TTP model will have the probabilistic analysis of malicious conduct by TTPs, during selection, is:

$$P(TTP_{multi}) = r/m \quad (2)$$

where m is the total number of TTPs in the model and r is the number of TTPs that will perform computation.

Case 2: Probability of malicious conduct in single and multi TTP model after selection of TTPs

Probability of malicious conduct in single TTP model is:

$$P(TTP_{single}) = 1 \quad (3)$$

Probability of malicious conduct in multi TTP model is:

$$P(TTP_1) = P(TTP_2) = P(TTP_3) = 1/3$$

Here TTP_1 , TTP_2 and TTP_3 are selected for computation on a particular function f.

In generalized form, suppose m is the total number of TTPs performing the computation on a particular function f and r is the number of TTPs that can perform malicious conduct out of m TTPs selected at runtime is:

$$P(TTP_{multi}) = r/m \quad (4)$$

If all the TTPs performing the computation are malicious then $r=m$, hence

$$P(TTP_{multi}) = m/m = 1 \quad (5)$$

If this is the case single TTP and multi TTP behavior will be identical and more often multi TTP model will have high computation cost with no effectiveness.

In single TTP model the probability of malicious conduct is either 0 or 1. Contrary in multi TTP model it increases gradually as malicious TTPs increases.

3.3 Graph Analysis

After selection of TTPs

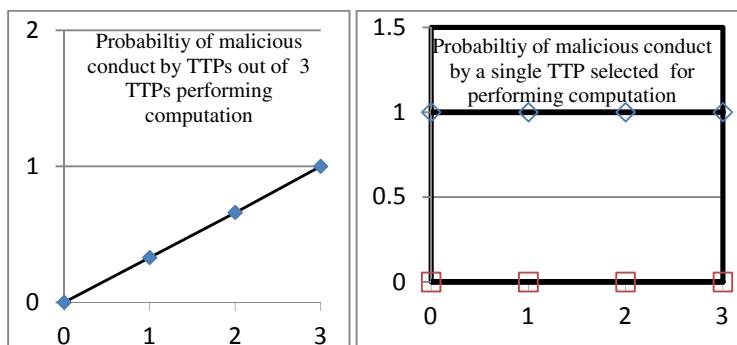


Fig. 3. Graph analysis for multi and single third party computation

In single TTP model the probability of malicious conduct is either 0 or 1. Contrary in multi TTP model it increases gradually as malicious TTPs increases. In graph, 3 TTPs are performing computation on particular function. So the probability of being malicious during computation will be $1/3$, $2/3$ and $3/3$. In this paper, we have also worked on identification and reduction of malicious conduct of TTPs in multi TTP environment. There are following cases considering different behaviors of Third party.

Case 1: (number of TTP= m)> 1/2 is providing identical and correct results

If this is the result of computation where number of TTPs $>50\%$ is giving identical and correct results and remaining TTPs are providing some other results then in real model, we have to consider the majority providing the same and correct results.

Case 2: m>1/2 (Majority is providing identical but wrong results of computation)

If this is the case of computation then it is difficult to identify the correctness in result as majority of TTPs are providing same wrong outputs. This case leads to protocol failure, and this kind of scenario is rarely possible in real world where majority of TTPs are giving same wrong results.

Case 3: $m=1/2$ (half of the TTPs providing identical and correct results and remaining half giving identical and wrong outputs)

This case also leads to system unacceptability and failure as equal number of TTPs providing same results in both the half, and so almost difficult to identify the correct results.

Case 4: Majority of TTPs giving identical and correct results and remaining TTPs giving different outputs in groups.

In this case it is almost reliable to go with majority.

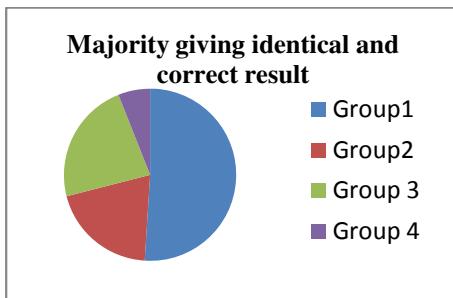


Fig. 4. Majority giving identical and correct results

3.4 Identification and Reduction of Malicious Conduct by TTPs

Case 1: $m>1/2$ (providing identical and correct results of computation)

Here our aim is to identify the malicious TTP resulting in protocol disruption in multi TTP computing model. The TTP performing malicious conduct is identified in several rounds of computation.

Consider a scenario where total number of TTPs used in the model=5. The number of TTPs performing the computation out of 5 TTPs is selected at runtime.

Table 1. TTPs performing computation in different rounds and their analysis

Round	Name of TTPs selected	Identical results in majority	Correct result	Malicious conduct	No of TTPs giving same result	Groups of TTPs
(No of TTPs=3) I	TTP ₁	✓	✓	x	2	A
	TTP ₂	X	x	✓		B
	TTP ₃	✓	✓	x		A
(No of TTPs=4) II	TTP ₂	X	x	✓	3	A
	TTP ₃	✓	✓	x		B
	TTP ₄	✓	✓	x		B
	TTP ₅	✓	✓	x		B

Note: Correct result parameter depends on majority of TTPs giving identical results.

The identification of TTPs is made on the basis of following steps:

1. Find the name of TTPs not in majority of correct results in several rounds of computation.
2. Now the TTPs not in majority of identical results are considered to be marked.
3. Find the number of times that TTP is out of majority in several rounds of computation.
4. The TTP is then stored in SMC_multi_trouble zone with the value (number of times) parameter.

In the above scenario if TTP_2 is not in the majority of identical results in several rounds of computation, the TTP_2 is considered to be malicious trying to deviate from the protocol. From table 1, following observations are drawn:

Round I- TTP_2 is not in the majority of identical results and is marked.

Round II- TTP_2 is not again in the majority of identical results.

In this way TTPs deviating from the protocol and behaving as an “Odd Man Out” can be identified by saving their all records of computation performed in several rounds.

Now on the basis of parameters used in table 1 we divided the TTPs to lie in either of two zones: SMC_Multi_Safezone and SMC_Multi_Troublezone.

Table 2. Behavior identification of TTPs

Name of TTP	SMC_Multi_Troublezone	SMC_Multi_Safezone
TTP_1	-	1
TTP_2	2	-
TTP_3	-	2
TTP_4	-	1
TTP_5	-	1

This gives identification of TTP behaving maliciously in the protocol and therefore its involvement in computation has to be reduced if the value parameter of TTP in trouble zone is highest and computation should be performed with the TTPs giving highest value in safe zone.

Case 2: Majority providing incorrect and identical output of computation

From table 2, Round III- TTP_2 is in the majority of identical results but it is the case when computation done is incorrect. In this case, where majority is giving identical but wrong output of computation then protocol will go with the majority and this scenario leads to system failure as correctness parameter of SMC does not work.

The only solution is to see the behavior of TTPs after each round of computation in “Table: Behavior identification of TTPs” if there are entries of TTPs performing the computation.

Table 3. Majority giving identical but wrong output of computation

Round	Name of TTPs selected	Identical results in majority	Correct result	Malicious conduct	No of TTPs giving same result	Groups of TTPs
(No of TTPs=3) III	TTP ₁	✓	x	✓	2	A
	TTP ₂	✓	x	✓		A
	TTP ₃	x	✓	x		B

- The reference to “Behavior identification of TTPs” is made after every computation.
- Find the number of times a TTP involved in computation is in the SMC_Multi_Troublezone (TTP₂= highest trouble zone value=2).
- Find the number of times a TTP involved in computation is in the SMC_Multi_Safezone (TTP₃= highest safe zone value=2).
- The trust on SMC_Multi_Safezone column with high values will be more than SMC_Multi_Troublezone.

The behavior of TTPs performing computation is identified through table “Behavior--“. From the table, the conclusions are, TTP₂ has the highest troublezone value as 2 whereas TTP₃ has the highest safe zone value as 2. So the trust on TTP₃ will be high.

- Now, the same computation has to be re-performed with other remaining TTPs in the pool with highest safezone value.
- If the remaining output matches with TTP₃ and is in the majority then correctness can be ensured.
- The entries have to be updated in Table 3 for the computing TTPs.

Case 3: Equal number of TTPs giving identical results.

Table 4. Equal number of TTPS giving identical results

Round	Name of TTPs selected	Identical results in majority	Correct result	Malicious conduct	No of TTPs giving same result	Groups of TTPs
(No of TTPs=4) IV	TTP ₂		x	✓	2,2	A
	TTP ₃	identical	x	✓		A
	TTP ₅		✓	x		B
	TTP ₁	identical	✓	x		B

In this case it is almost difficult to identify the correct result but if there is an entry in trouble and safe zone column of the table 3, then reference to table is the only solution in identification of wrong conduct and steps of case 2 has to be followed.

4 Results and Conclusion

In today's fast growing Internet environment, when most of the operations are jointly performed, there is a need of more secured protocols which can maintain privacy and assure correctness. In this paper single and multi third party SMC environment is defined, compared and analyzed. The need of using multi TTPs computing model is that of privacy concern as parties providing inputs for computation may not be able to know the third party performing computation as the TTPs are selected at runtime from the pool of TTPs. While using multi third party environment for computation, different cases were studied for identification of malicious conduct by TTPs .The behavior of TTPs is analyzed, in several rounds. Analyzing the behavior of TTPs, by looking at the highest count in trouble_zone column of *Behavior Identification of TTPs*, the involvement of that TTP in computation is reduced and the highest safe_ zone count TTP is given more rights at computation. This reduces the malicious TTPs and increases the system acceptability.

References

- [1] Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., Michael, Y.: Tools for privacy preserving distributed data mining. *SIGKDD Explorations* 4(2), 1–8 (2002)
- [2] Vaidya, J., Clifton, C.: Leveraging the Multi in Secure Multi-Party Computation. In: Proceeding of the 2003 ACM Workshop on Privacy in Electronic Society. ACM Press (2003)
- [3] Yao, A.C.: Protocol for secure computations. In: Proc. 23rd IEEE Symposium on the Foundation of Computer Science (FOCS), pp. 160–164. IEEE (1982)
- [4] Goldreich, O., Micali, S., Wigderson, A.: How to play any mental game- a complete theorem for protocol with honest majority. In: Proceeding of 19th ACM Symposium on the Theory of Computing (STOC), pp. 218–229 (1987)
- [5] Ioannidis, I., Grama, A.: An efficient protocol for Yao's Millionaires Problem. In: Proceeding of 36th Hawaii International Conference on System Sciences, HICSS 2003, pp. 6–11. IEEE Press (2003)
- [6] Goldreich, O., Micali, S., Wigderson, A.: How to play any mental game. In: Proceedings of the Nineteenth Annual ACM Conference on Theory of Computing, STOC 1987, pp. 218–229. ACM, New York (1987)
- [7] Chor, B., Gilboa, N.: Computationally Private Information Retrieval (Extended Abstract). In: Proceedings of 29th Annual ACM Symposium on Theory of Computing, El Paso, TX, USA (1997)
- [8] Chor, B., Kushilevitz, E., Goldreich, O., Sudan, M.: Private Information Retrieval. In: Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science, Milwaukee, WI, pp. 41–50 (1995)
- [9] Lindell, Y., Pinkas, B.: Privacy Preserving Data Mining. In: Bellare, M. (ed.) CRYPTO 2000. LNCS, vol. 1880, pp. 36–54. Springer, Heidelberg (2000)
- [10] Agrawal, R., Srikant, R.: Privacy-Preserving Data Mining. In: Proceedings of the 2000 ACM SIGMOD on Management of Data, Dallas, TX, USA, pp. 439–450 (2000)
- [11] Atallah, M.J., Du, W.: Secure Multiparty Computational Geometry. In: Proceedings of Seventh International Workshop on Algorithms and Data Structures (WADS 2001), Providence, Rhode Island, USA, pp. 165–179 (2001)

- [12] Du, W., Atallah, M.J.: Privacy-Preserving Cooperative Scientific Computations. In: 14th IEEE Computer Security Foundations Workshop, Nova Scotia, Canada, June 11-13, pp. 273–282 (2001)
- [13] Du, W., Atallah, M.J.: Privacy-Preserving Statistical Analysis. In: Proceedings of the 17th Annual Computer Security Applications Conference, New Orleans, Louisiana, USA, pp. 102–110 (2001)
- [14] Du, W., Atallah, M.J.: Secure Multiparty Computation Problems and Their Applications: A Review and Open Problems. In: Proceedings of New Security Paradigm Workshop, Cloudfcroft, New Mexico, USA, pp. 11–20 (2001)
- [15] Oleshchuk, V., Zadorozhny, V.: Secure Multi-Party Computations and Privacy Preservation: Results and Open Problems. Teletronikk: Telenor’s Journal of Technology 103(2) (2007)
- [16] Zheng, Q., Shan Luo, S., Xin, Y.: Research on the Secure Multi-Party Computation of some Linear Algebra Problems. Applied Mechanics and Materials. Trans. Tech. Publication 20-23, 265–270 (2010)
- [17] Henecka, W., Ogl, S.K.: TASTY: tool for automating secure two-party computations. In: The Proceedings of the 17th ACM Conference on Computer and Communications Security (2010)
- [18] Lucas, C., Raub, D., Maurer, U.: Hybrid-secure MPC: trading information-theoretic robustness for computational privacy. In: Proceeding of the 29th ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing, PODC 2010 (2010)

Ubiquitous Medical Learning Using Augmented Reality Based on Cognitive Information Theory

Zahra Mohana Gebril, Imam Musa Abiodunde Tele, Mohammed A.Tahir,
Behrang Parhizkar, Anand Ramachandran, and Arash Habibi Lashkari

Faculty of Information & Communication
Technology, LIMKOKWING University
Cyberjaya, Selangor, Malaysia

{zahra.gebril,hani.pk,anand}@limkokwing.edu.my,
musa.imam@aol.com, mohammed_tahir@rocketmail.com,
a_habibi_L@hotmail.com

Abstract. Attention has been drawn to mobile devices, which show that the success will be the portability of applications from one platform to the other. With integration of 3D virtual objects into real environment in real time thus allowing student to relate with their physical environment and also making the subject more interesting. This study carry out based on the understanding of different types of learning theories, concept of mobile learning and mobile augmented reality and discusses how applications using these advanced technologies can improve learning process.

Keywords: Screen Finger Interaction, Augmented Reality, Human Anatomy, Tablet.

1 Introduction

M-Learning (mobile learning), is a subset of e-Learning (electronic learning) that uses a wireless network, portable and handheld technologies including laptops, table computers, Smartphone's and other electronic computing devices that can be used to provide learning experience in more dynamic environments [1]. The advantages of learning anytime and anywhere have long been near the top of the benefits listed by proponents of online education, but until the advent of m-Learning technologies it was not really an anytime, anyplace environment. [2]

The increasing use of wireless technology and mobile phones suggests that training and education cannot ignore the use of handheld devices in the training process. Furthermore, in the mobile environment, frequent data updates with handheld devices through broadcast or multi-cast will reduce the performance of the application system and mobile applications have to scale dramatically to keep up with the growth in the number of users [3].

Mobile learning, instructional paradigms Using appropriate design methodologies, educational application have demonstrated that it is possible to offer a flow experience that immerses learners in active learning and are able to boost the intrinsic motivation level of a learner by means of highly engaging challenges [4].

2 Mobile Learning in Medical

The rapid proliferation of mobile devices nationally coupled with high levels of individual ownership both nationally and internationally provides many opportunities for exploring ways in which students own mobile devices might be integrated into teaching and learning activities in higher education. Explored the implementation of mobile devices and rich media in the Health Sciences and concludes that mobile learning can make valuable contributions to linking different learning environments. The ability to listen "on the go" using a portable media player, PDA (personal digital assistant), cell phone, or personal computer has made podcasting an attractive tool for learning. A preclinical practice survey identified that the majority of students were willing to use their mobile devices for teaching and learning activities. In this project, a number of activities were designed to encourage students to use their mobile technologies to reflect on their clinical practice activities and relate these to other aspects of the course. Importantly, students also felt that the ability to practice techniques on their mobile devices improved their competency in relation to the use of the voice therapy techniques particularly in demonstrating the use of the voice therapy techniques to patients. The flexibility and mobility offered by using the mobile devices greatly increased their 'anytime, anywhere' learning. These case studies clearly demonstrate that the use of mobile technologies and rich media can enhance the learning process in clinical practice settings for both on-campus and off-campus students. [5]

3 Mobile Learning Using Augmented Reality in Medical

The fact that augmented reality has introduced new ways of presenting information. The medical world can be re-innovated to be represented in a mobile augmented reality way. Augmented reality can well represent these medical data or information in its utmost visual form possible unlike anything before. According to [6], augmented reality is the extension of virtual reality which is the synthetic reality makeup of our real world, adding more details to what is already there. The smartphone is come equipped with sensors and camera which simply bring AR to life. Such application of sensor allows accurate context information to be provided to the environment aware situation allowing doctors to collect information, exchange, interpret and recognize procedures [7]. Doctors will have control over patients whom require continuous monitoring, temperature, heartbeats, etc. This information can be presented immersive in AR.

3.1 Architecture

Being a part of the medical industry, keeping the patient informed all the time is important and educating them is being part of it [8]. The increasing use of mobile phones has enable developers to create multi-functional application that pushes information to patient's mobile. Patient has become inter-connected, no matter when and where they are. The mobile has become a new learning tool in creating awareness and adding value to existing knowledge of the patient. AR will have further visual implication on it as it will visualize this hard-to-consume information for the patient. For example, the patient who needs information about asthma can just retrieve it on their mobile and push into AR.

3.2 Appl

One of the applications of this system is to turn it into a problem-based learning approach to medical education [9] The medical field could use a problem-based learning system to such as identify the most common and rare medical conditions. Mobile phones with equipped AR technology can be used as an electronic clinical-log, which is useful for both students and medical faculty. It is beneficial in terms of self-assessment, monitoring students' clinical experiences, future curriculum evaluation and development, and future research in medical education.

4 Problem Statement

Learning takes place all the time and now with mobile devices people are constantly learning on their own and connecting to each other to improve their knowledge. In this process our focus was useful because it got us to think more about how we could use the future and the most common mobile operating system in the market such as Android, to be able to provide an application that can be run on the most smartphones in the market. As we can see that technology is changing every day, this as shown us how uniquely mobile technologies impact our lives and learning. More also we have seen companies, schools (both higher institution and Tertiaries school) are making wave in using the Mobile learning to acquired educate. Technology has given us a great deal of new knowledge from difference perspectives and some fantastic virtual learning environments and teaching tools; however, there are still a lot of challenges ahead mainly fuelled by the fact that some still see technology as a threat in the classroom. Technology has given us more freedom from the four walls of the classroom but it does mean we have to approach teaching in a different ways to get its full benefit. These are some of the Threats been appeared in Mobile learning, using mobile learning for the right reasons. Measuring the effectiveness of mobile learning (this been measured according to the finding of the answers from the Question 9(What do you think about the idea of applying M-Learning for medical course in this university) and Question 10(What do you think about the idea of applying M-Learning in this university to enhance the way of teaching effectively and improve the student performance by providing interactive multimedia contents as mobile applications that can be implemented on your Cell phone or PDAs) from the questionnaire we distributed at Master Skills University College of science and health.

5 Problem Solving with ARM

According to the researched carry out at Master Skill University College of health sciences in Malaysia ,the findings show that implementing Mobile learning for the medical student is going to be a great improvement in their learning skills, Questionnaires and interviews are conducted to both lecturers and the student. In this researched we found out that most of this students never heard about Mobile learning. According to the name Mobile learning this as shown that M-Learning is characterized by the ability to learn through portable devices known as PDA, Smart phones,

and other device that are mobile which will enhance learning skills for both the students and their lecturers. In the demographic which was stated above show the rate in which the correspondent respond to the implementation of the application, this show that only 20% of lecturers and student know what is called mobile learning, and 15% have read about it but they have not witness it, while 65% have not heard about it. As it was shown in the pie chart this researched was conducted in medical school. In developing this App is to teaching and training the end user though mobile technologies and devices. At Demographic the age, 96% are between 18-20years of age and 4% are there lecturers. This show that the respondent range from 18-20 years are still new to know about Mobile learning, so this will required a lot of train for the student to basically to understand and know how to use the device properly, as it was show from the Demographic, only 79% of the students don't use their school website while 29% only access studies materials from the school website and for this show that students are not be encourage by the school management, show that student read only textbook base on what they are taught in class, they limit their knowledge only on what it been taught in class. Conducting interviews with some of their lecturers (name withheld), said this will be a great improvement for the next generation and make life easier for teaching and learning. In conclusion they talk about the advantages and this advances Mobile Learning, the advantages is few than disadvantages, advantages it provide a group discussion for both the lecture and their students, reading with easy, read ahead of the class and It just a pocket fit device. The disadvantages are it makes student lazy to read broad (less students will be seen in the Library), Student will capitalized on the negative part.

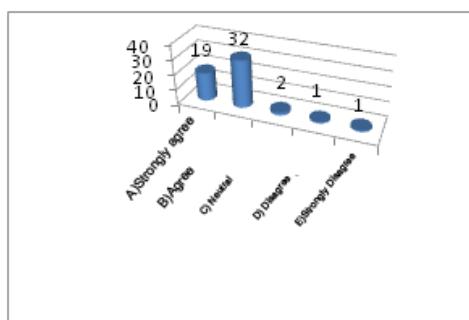


Fig. 1. Students of Master Skill University College opinion about the idea of applying M-Learning for medical course

6 Iterative Visual Cognitive Based Learning Theory in ARM Application

(i) Learning Theory

George A. Miller has provided two theoretical ideas that are fundamental to the information processing framework and cognitive psychology more generally. The first

concept is 'chunking' and the capacity of short term (working) memory [10]. Presented the idea that short-term memory could only hold 5-9 chunks of information (seven plus or minus two) where a chunk is any meaningful unit. A chunk could refer to digits, words, chess positions, or people's faces. The concept of chunking and the limited capacity of short term memory became a basic element of all subsequent theories of memory. The second concept, that of information processing uses the computer as a model for human learning. Like the computer, the human mind takes in information, performs operations on it to change its form and content, stores and locates it and generates responses to it. Thus, processing involves gathering and representing information, or encoding; holding information or retention; and getting at the information when needed, or retrieval. Information processing theorists approach learning primarily through a study of memory.

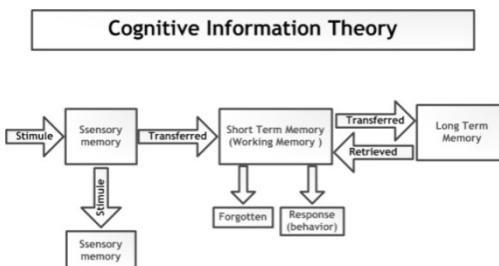


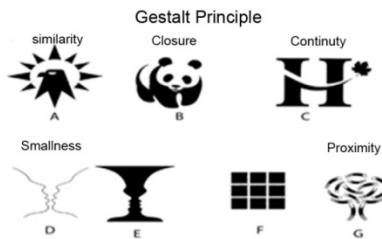
Fig. 2. Cognitive Information Theory

The principle known as the multimedia principle states that people learn more deeply from words and pictures than from words alone [11]. However, simply adding words to pictures is not an effective way to achieve multimedia learning. The goal is to instructional media in the light of how human mind works. This is the basis for Mayer's cognitive theory of multimedia learning. This theory proposes three main assumptions when it comes to learning with multimedia:

1. There are two separate channels (auditory and visual) for processing information (sometimes referred to as Dual-Coding theory);
2. Each channel has a limited (finite) capacity (similar to Sweller's notion of Cognitive Load);
3. Learning is an active process of filtering, selecting, organizing, and integrating information based upon prior knowledge

(ii) Gestalt Principle

Gestalt theory is a broadly interdisciplinary general theory which provides a framework for a wide variety of psychological phenomena, processes, and applications. Human beings are viewed as open systems in active interaction with their environment.

**Fig. 3.** Gestalt Principle

It is especially suited for the understanding of order and structure in psychological events, and has its origins in some orientations of Johann Wolfgang von Goethe, Ernst Mach, and particularly of Christian von Ehrenfels and the research work of Max Wertheimer, Wolfgang Köhler, Kurt Koffka, and Kurt Lewin, who opposed the elementistic approach to psychological events, associationism, behaviorism, and to psychoanalysis.

(i) Platform

We followed the visual cognitive, so that the application hold video images which able to interact with novice. However, student can perform the visualization by command on touch screen as many as they want.

Table 1. Software Specification for ARM application

Software Specification		
No	Item	Specification
1	OS	Windows 7
2	Developer	Eclipse
3	Language	Java, OpenGL ES
4	Phone OS	Android
5	3D Modeler	3Ds Max10
6	2D Modeler	Photoshop

(ii) Hardware requirement for its design and Architecture

The development of the method contains the ARM. The system been constructed on tablet. The main target devices are tablet. Since hardware devices are always upgraded so this application will perform on the smart devices operated with Android OS-3.0 version. The phase develops the six design phases which is 3D objects, 3D animation, 2D Images ,text ,audio ,phone touch screen and finger interaction on display.

(iii) Design of model prototype for ARM

The process of designing the prototypes of ARM comprised of four main components such as: resources, objectives, strategies & approach and interactivity. It can be observed in Figure 5



Fig. 4. Example of Augmented Reality medical School practical books from student view on phone display.

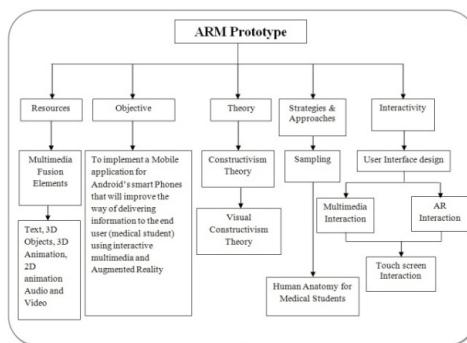


Fig. 5. ARM prototype

7 Methodology

The development methodology for this research was based on the Interactive –Visual cognitive software development life cycle for Augmented Reality for medical (I-CSDLC-ARM). This application was develop on an augmented reality (AR) system for I-CSDLC-ARM; there was a need for planning in terms of design and development approach to make researchers understand the augmented reality technology as well as the use of this technology for intermediate medical students, Interactive – Visual cognitive software for Augmented Reality for medical (I-CSDLC-ARM), radically enhance both the quality and the bandwidth of educational processes. This allows Students to read, flip pages by touching the screen of the tablet, which also give the students to walk through a human body and observing the functioning of biological system. Thus, the I-CSDLC-ARM, is design for ARM.

8 Objective

Our objective is to explore the cognitive, social and cultural aspects of learning and to design innovative technologies and environments for medical student using Android and Augmented Reality so they can make uses of these devices everywhere and any time. Perhaps as the definition goes mobile learning as learning that follow us. Whenever we are, whatever we are doing, there it is.

- i. To study the suitable method of designing and developing medical contents in the mobile phone based on Constructivism Method of learning.
- ii. To develop the Mobile Augmented Reality application for Medical (ARM) students to enhance the learning in medical students based on Iterative - Visual Cognitive Software Development Life Cycle (ARM-I-VCSDLC).
- iii. To evaluate and test the strength and weakness of MAR-MED using Simple Usability (S-Usability) Testing.

9 Conclusion

The main aim of this project concern on developing a mobile based augmented reality for medical students based on cognitive information theory. This application is an integrated research within ubiquitous technology and augmented reality. The ARM project has been applied for all android platforms smart phones as well as all tablet devices that using android operating system.

Acknowledgement. The special thank goes to our helpful supervisor, Mr. Behrang Parhizkar and our advisors Mrs. Zahra Mohana Gebril and Dr. Arash Habibi Lashkari for their supervision and guidance in the progression of our dissertation.

References

- [1] Devices and Rich Media: line learning. In: Second International Conference on Mobile, Hybrid, and On-Line Learning (2010)
- [2] Jason, G.C.: The Growth of m-Learning and the Growth of Mobile Computing: Parallel developments. The International Review in Open and Distance Learning 8(2) (2007), <http://www.irrodl.org/index.php/irrodl/article/view/348/873>
- [3] Zhang, J., Levy, D., Chen, S.: A Mobile Learning System For Syndromic Surveillance and Diagnosis. In: 10th IEEE International Conference on Advanced Learning Technologies (2010)
- [4] Ghazvini, F.: Designing Augmented Reality Games for Mobile Learning using an Instructional-Motivational Paradigm. In: International Conference on CyberWorlds (2009)
- [5] Andrews, T., Smyth, R., Caladine, R.: Utilizing Student' own Mobile (2010)
- [6] Arusoae, A.: Augmented Reality. In: 12th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (2010)
- [7] Kim, S.: Design of Mobile u-Healthcare service System. In: International Symposium on Computer Science and its Applications (2008)
- [8] Finkelstein, J.: Mobile eLearning Platform for Interactive Patient Education. In: International Conference on Mobile, Hybrid, and On-line Learning (2009)
- [9] Luanrattana, R.: Data security and Information Privacy for PDA Accessible Clinical-Log for Medical Education in Problem-Based Learning (PBL) Approach. In: IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (2010)
- [10] Mayer, R.E., Heiser, J., Lonn, S.: Cognitive constraints on multimedia learning: When presenting more material results in less understanding. Journal of Educational Psychology 93, 187–198 (2001)
- [11] Mayer, R.E.: The promise of educational psychology. Teaching for meaningful learning, vol. 2. Prentice Hall, Upper Saddle River (2002)

A Secured Transport System by Authenticating Vehicles and Drivers Using RFID

C.K. Marigowda¹, J. Thriveni², and Javid K. Karangi¹

¹ Department of Information Science & Engineering,
Acharya Institute of Technology, Bangalore

² Department of Computer Science & Engineering,
University Visvesvaraya College of Engineering, Bangalore University, Bangalore
{marigowda, javidkarangi.ise.07}@acharya.ac.in,
thrivenijgowda@yahoo.co.in

Abstract. The Secured Transport System by authenticating Vehicles and Drivers using RFID is aimed at providing a secure transportation mechanism. As the vehicular traffic is growing exponentially, monitoring the authenticity of drivers and vehicles is a hurricane task for RTO (Regional Transport Office). Though monitoring is achieved to some extent by Global Positioning Systems (GPS), currently it is restricted to taxi cabs. Our proposed work utilizes server-client methods used in other areas. Each vehicle will be embedded with a client module circuitry and RTO offices will host servers. Vehicles' and Drivers' information gathered by client is validated by Servers and thus RTO can keep track of various credentials like vehicle insurance, driving license of every vehicle plying on streets.

Keywords: Radio Frequency Identification(RFID), Regional Transport Office (RTO), Global Positioning System(GPS), ARM 920T S3C2440, Linux Cross Platform Tool – Qt.

1 Introduction

Vehicular traffic in city roads is intensive as a result tracking the authenticity of vehicles is a complex task for city police. Also tracking vehicle theft cases in spite of registration numbers and vehicle insurance is a cumbersome process for RTO and traffic police. An automated method of tracking the vehicle documents and validating the credentials of drivers/owners is proposed in this paper. The proposed system works on client server concept with RTO regional offices hosting the server machines and vehicles embedded with a circuitry to act as clients. Every vehicle is tracked and validated for license, insurance, emission test certificates and so on. This system doesn't rely upon GPS system and uses RFID, Linux platform and Wi-Fi connectivity as a cost effective alternative.

1.1 System Working Overview

Vehicle user(s) have to place driving license in the card holder to use the vehicle. User's license and vehicle will be validated by sending all the details to nearest server by

wireless mode. Once the vehicle registration number, insurance and driving license are found valid, user is allowed to drive the vehicle. While the vehicle is in movement and has met with an accident or stopped by a traffic police the same will be detected by catch module of the system.

1.2 Main Components of Vehicle Security System

- ARM9 Board
- RFID Reader
- Linux Operating System
- Wi-Fi Connection
- Qt Framework

1.3 ARM9 Overview

ARM is one of the most licensed and thus widespread processor cores in the world. Used especially in portable devices due to low power consumption and reasonable performance (MIPS / watt) Several interesting extensions available are in development like Thumb instruction set and Jazelle Java machine ARM9 is an ARM architecture 32-bit RISC (Reduced Instruction Set Computers) CPU family. With this design generation, ARM moved from a von Neumann architecture (Princeton architecture) to a Harvard architecture with separate instruction and data buses (and caches), significantly increasing its potential speed. Most silicon chips integrating these cores will package them as modified Harvard architecture chips, combining the two address buses on the other side of separated CPU caches and tightly coupled memories [1]. The specification of ARM9 board which we are using are: **ARM 920T S3C2440** (Single board Computer), **CPU:** 400 MHz Samsung S3C2440A ARM920T (max freq. 533 MHz), **RAM:** 64 MB SDRAM, 32 bit Bus, **Flash:** 64 MB NAND Flash and 2 MB NOR Flash with BIOS, **EEPROM:** 1024 Byte (I2C) [2].

1.4 Radio Frequency Identification (RFID)

Radio-frequency identification (RFID) is a technology that uses communication through radio waves to exchange data between a reader and an electronic tag attached to an object, for the purpose of identification and tracking. Furthermore, passive RFID tags (those without a battery) can be read if passed within close proximity to an RFID reader. It is not necessary to "direct" them to it, as with a bar code. In other words it does not require line of sight to "see" an RFID tag, the tag can be read inside a case, carton, box or other container, and unlike barcodes RFID tags can be read hundreds at a time. Bar codes can only read one at a time [3].

Some RFID tags can be read from several meters distance and beyond the line of sight of the reader. The application of bulk reading enables an almost-parallel reading of tags. Radio-frequency identification involves the hardware known as *interrogators* (also known as *readers*), and tags (also known as *labels*), as well as RFID software or RFID middleware. Most RFID tags contain at least two parts: one is an integrated circuit for storing and processing information, modulating and demodulating a

radio-frequency (RF) signal, and other specialized functions; the other is an antenna for receiving and transmitting the signal.

1.5 Linux Operating System

The goal is to present *abstract* architecture of Linux kernel. This is described by Sony as being the conceptual architecture [4]. By concentrating on high-level design, this architecture is useful for entry-level developers who have to verify the high level architecture before understanding where their changes fit in. In addition, the conceptual architecture is a good way to create a formal system vocabulary that is shared by experienced developers and system designers. This architectural description may not perfectly reflect the actual implementation architecture, but can provide a useful mental model for all developers to share. Ideally, the conceptual architecture should be created before the system is implemented, and should be updated to be an ongoing system conscience in the sense of Monroe 1977 showing clearly the load-bearing walls as described in Perry 1992[5].

1.6 Wi-Fi Connections

Wi-Fi' is not a technical term. However, the Alliance has generally enforced its use to describe only a narrow range of connectivity technologies including wireless local area network (WLAN) based on the IEEE 802.11 standards, device to device connectivity [such as Wi-Fi Peer to Peer AKA Wi-Fi Direct], and a range of technologies that support PAN, LAN and even WAN connections. Derivative terms, such as Super Wi-Fi, coined by the U.S. Federal Communications Commission (FCC) to describe proposed networking in the UHF TV band in the US, may or may not be sanctioned by the alliance.

1.7 Qt Framework

Qt is a cross-platform application framework that is widely used for developing application software with a graphical user interface (GUI) (in which cases Qt is referred to as a *widget toolkit*), and also used for developing non-GUI programs such as command-line tools and consoles for servers [6].

Qt can also be used in several other programming languages via language bindings. It runs on all major platforms and has extensive internationalization support. Non-GUI features include SQL database access, XML parsing, thread management, network support, and a unified cross-platform API for file handling.

Features: Intuitive C++ class library, Portability across desktop and embedded operating system, Integrated development tools, with cross platform IDE, High runtime performance and small footprint on embedded [7].

The rest of this paper is organized as follows. In **Section 2**, presents some important difficulty of monitoring the authenticity of drivers and vehicles. In **Section 3**, describes hardware and software requirement for the proposed model implementation. In **Section 4**, discusses in details the design of proposed model. In **Section 5**, presents the execution procedure and results. Finally in **Section 6** concludes the paper and also highlights some future scope of work.

2 Motivation

As the vehicular traffic is growing exponentially, monitoring the authenticity of drivers and vehicles is a hurricane task for RTO (Regional Transport Office). Hence there is a need for an RTO automation system which helps in monitoring various credentials of the driver and vehicle.

3 Requirement Specification

Hardware and software requirements for the proposed system is dealt in this section. The requirement is to implement an effective client console at the vehicle end and server access points at RTOs. The system has to generate appropriate messages indicating access permissions and denials at both client side and server side so that traffic monitoring personnel and RTO officers can take appropriate decisions. Here is a brief description of hardware, software and network features of the proposed system:

Hardware Requirements for the proposed work is as follows: Radio Frequency Identification (RFID), ARM 920T S3C2440 (Single board Computer), Processor: 400-533 MHz Samsung S3C2440A ARM920T, primary Memory: 64 MB SDRAM, flash Memory: 64 MB NAND Flash and 2 MB

Software Requirements for the proposed work is as follows: Programming Language- ANSI C, Operating System - Linux, Cross Platform Tool - Qt(framework) NOR Flash with BIOS. Connection between the server and the vehicle will be provided by the Wi-Fi, in our proposed system. Further, the connectivity may be changed to other means of communication i.e. GPS, infrared, Bluetooth etc.,

3.1 Feasibility Study

The assessment is based on an outline design of system requirements in terms of Input, Processes, Output system by the automation of RTO and vehicles of a modern and technological advanced Transport platform. Diagnosis of large and complex software systems is a challenging task that can highly benefit from monitoring of the high-level functional requirement. This research identifies the potential of applying requirements monitoring for a software system of high complexity: the Vehicle Security Simulations show that detailed diagnosis of a complex software system as a VSS is feasible. It also demonstrate that the combination of requirements monitoring and rule-based reasoning provide a solid foundation for various levels of autonomy in an existing Vehicle Security System.

This concentrates on implementing network centric vehicle security system with the use of client-server architecture to make the automation to the present scenario. Using a unique methodology, the project provides the following:

- Characterization of the Transport System
- Description of the design principles applied
- Conceptual design.

4 Design

The Context Diagram Fig 1 depicts the proposed system viz “Vehicle Security System”, it has external input entities in form of vehicles, which provide inputs at given rate. The output of the proposed system acts as input to RTO server. Following content explains the functionality of each sub-system.

Vehicular Module: Comprises of RFID tag and RFID reader. A driver has to validate authenticity by scanning RFID tag in RFID reader on validation by RTO server. User will be allowed to switch on the ignition key

Proposed System: It will obtain vehicle registration number, licence number, insurance details from vehicle module and send it to nearest RTO server for validation.

RTO Server: The server collects the message logged and verifies all the credentials like vehicle insurance, rules broken and complaints on vehicle. If all the credentials are legitimate vehicle will be given access.

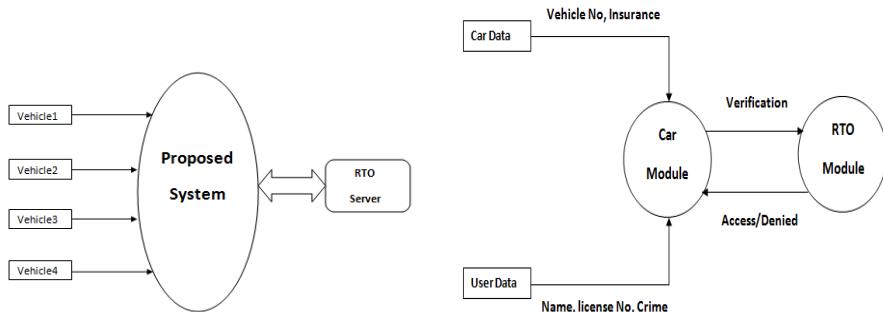


Fig. 1. Level 0 Context Diagram

Fig. 2. Level 1 Data Flow Diagram

The functionality in each module in Data flow diagram in Fig.2 is as follows

Car data: It includes the vehicle registration number and the insurance details

User data: It includes the name, license number and crime information if any by the user of the vehicle.

Car module arm9: This module represents the vehicular module initially, the user enters into the vehicle by scanning the RFID tag into the reader and the car will use the access provided by the RTO server and allows ignition of the vehicle.

RTO module: The details of the user is verified against the information located in the RTO server based on which access granted or denied message will be sent back.

Verification data: It is user details, gets verified against the information located in the RTO server.

The Data flow diagram Fig 3 deals with information about the data flow across various modules

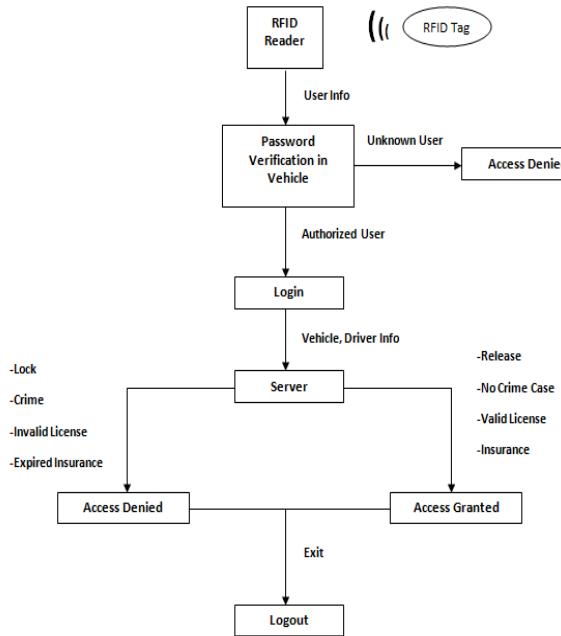


Fig. 3. Level 2 Data Flow Diagram.

5 Implementation and Results

The results of testing were recorded. The execution of our project has following steps.

Phase 1: We have to Load libraries into ARM9 board.

Phase 2: On the client side, the RFID tag of the driver is detected by the RFID reader which is placed in the vehicle. On the ARM9 interface (user interface) the user name will be displayed and it will password for Primary Authentication.

Phase 3: Now the driver will click **LI** (Login) Button after entering the password shown in Fig 4. If the password authentication is successful then login successful message will be displayed shown in Fig 5. Driver and Vehicle Details will be sent to the RTO server.

Phase 4: On the server side (RTO server) the details sent by the client (vehicle) will be verified against existing database shown in Fig 6. If the three conditions including valid license, no crime and valid insurance evaluate to true then access granted message will be sent to the vehicle by clicking **RELEASE** button in the server shown in Fig 7.

Phase 5: During the time of login if the password is invalid or password length exceeds then the following message will be displayed as shown in the Fig 8.

Phase 6: On the Server side if the user license found to be invalid or if any crime cases are booked or if the insurance is expired or if any combination of above conditions

occur then access denied message will be sent to the vehicle by clicking **LOCK** button in server. As in Fig 9.

Phase 7: If the license alone is found to be invalid with remaining valid conditions then access can be granted by clicking **RELEASE** button, in case of which security check personnel can catch the vehicle which will be implemented as a future enhancement of our project.

Phase 8: When vehicle is switched off, the driver should logout by clicking **LO** (logout) button on the ARM9 interface as in Fig 10.



Fig. 4. Primary Authentication after RFID Tag Detection



Fig. 5. Successful Login in Client Side

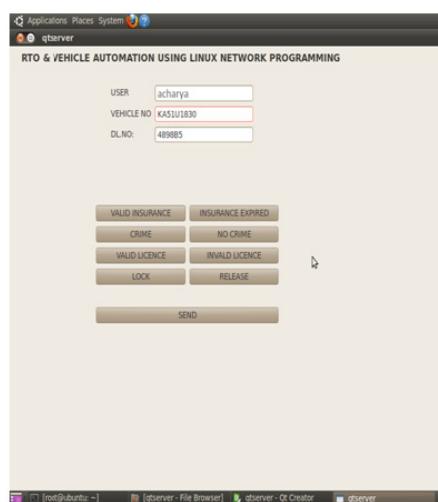


Fig. 6. Server Side Verification

**Fig. 7.** Access Grant Message on Client Side**Fig. 8.** Invalid Password Entry on Client Side**Fig. 9.** Access Denied Message on Client Side**Fig. 10.** Logout Message on Client Side

6 Conclusion and Future Enhancement

The Secure Transport System regulations can be used to induct strict traffic rules among drivers and vehicle users. It also helps in preventing vehicle thefts and also can be used to monitor the vehicle density in various parts of the city remotely. Also any crimes carried out using vehicles can also be detected.

The proposed system can be enhanced in future by incorporating the following

Biometric Scans: At present we are using the RFID concepts to provide the security in our project. But, in order to provide greater efficiency in security, we can use biometric scans which provide security to vehicle and the driver with greater efficiency.

Satellite Communication: Now, in our project we use the Wi-Fi connections to establish the communication between all the modules. When we think of the actual scenario the vehicles will be spreads in the large geographical areas at that time the connectivity establishment between the vehicles, RTO server and the police module by Wi-Fi or the Bluetooth is not at all possible.

So, we need to think about the connections that can afford the future needs, the answer to these questions is Satellite Communication. This will cover the greater extends, the communication will be stronger, good coverage, better security and the availability will be more.

References

- [1] Ragavan, P., Lad, A., Neelakandan, S.: Embedded Linux system design and development, 5th edn. McGraw Hill
- [2] Ball, B., Smoogen, S.: Sams Teach Yourself LINUX in 24 hours, 2nd edn. Sams Teach
- [3] Welbourne, G., Khoussainova, N., Suciu, D.: Specification, Detection, and Notification of RFID Events with Cascadia, 2nd edn. Pearson Publishers
- [4] Performance of the ARM9TDMI and ARM9E-S cores, ARM Ltd.
- [5] Petazzoni, T.: Porting the Linux kernel to an ARM Board, 6th edn. Tata McGraw Hill
- [6] Molkentin, D.: The Book of Qt 4: The Art of Building Qt Applications, 3rd edn. Harper Collins
- [7] Summerfield, M.: Advanced Qt Programming: Creating Great Software with C++ and Qt 4, 4th edn. Pearson

Virtualization of Large-Scale Data Storage System to Achieve Dynamicity and Scalability in Grid Computing

Ajay Kumar¹ and Seema Bawa²

¹ Department of Computer Engineering, Thapar University, Patiala, India
ajaycpp@gmail.com

² Department of Computer Engineering, Thapar University, Patiala, India
seema@thapar.edu

Abstract. Data storage management is one of the most challenging issues for Grid resource management since large amount of data intensive applications frequently involve a high degree of data access locality. Grid applications typically deal with large amounts of data. In traditional approaches high-performance dedicated servers are used for data storage and data replication. This allows opportunistic grids to share not only the computational cycles, but also the storage space. This paper explains new mechanism for Dynamic and Scalable Storage Management (DSSM) in grid environments is proposed. The storage can be transparently accessed from any grid machine, allowing easy data sharing among grid users and applications. The concept of virtual ids that, allows the creation of virtual spaces has been introduced and used. The DSSM divides all Grid Oriented Storage devices (nodes) into multiple geographically distributed domains and to facilitate the locality and simplify the intra-domain storage management. Grid service based storage resources are adopted to stack simple modular service piece by piece as demand grows. To this end, we propose four axes that define: DSSM architecture and algorithms description, Storage resources and resource discovery into Grid service, Evaluate purpose prototype system, dynamically, scalability, and bandwidth, and Discuss results. Algorithms at bottom and upper level for standardization dynamic and scalable storage management, along with higher bandwidths have been designed.

Keywords: Data, Data Locality, DSSM, GOS, GRID, Virtualization, Web Services, Virtual Organization.

1 Introduction

A Grid is a collection of machines, sometimes referred to as “nodes”, “resources”, “members”, “clients”, “hosts”, “engines”, and many other terms[1]. Grid environments are increasingly being used by applications such as particle physics, climate modelling, whether forecasting or astrophysics. They all contribute any combination of resources to the Grid as a whole. These applications make use of unique, high-end supercomputers and large-scale data storage-systems and produce large multi-dimensional data sets. This type of work is increasingly being performed in collaborative efforts between geographically distributed scientists and organizations, utilizing shared resources that

are also widely distributed. Network based storage systems such as Network Attached Storage (NAS) [9] and Storage Area Network (SAN) [3] offer a robust and easy method to control and access large amounts of storage. However, with the steady growth of client access demands and the required data sizes, it is a challenge to design an autonomous, dynamic, large-scale and scalable storage system which can consolidate distributed storage resources to satisfy both the bandwidth and storage capacity requirements [2]. Some resources may be used by all users of the Grid while others may have specific restrictions. In this research paper we are focusing to achieve dynamicity and scalability for large scale data storage system in Grid environments. The Dynamic and Scalable Storage Management (DSSM) architecture to organize Grid Oriented Service (GOS) devices into a large –scale and geographically distributed data storage system to meet the requirements imposed by all kinds of Gris applications [2].

2 Data Management in Grid Computing

In this section we explain the evolution of the data management within the Grid and investigate the various implementations of the data Grids that are in place today as well as those currently emerging.

2.1 General Traditional Data Management

Electronic data management has a long and rich history dating back to the 1950s many data management systems have tried to make their way into the mainstream of information technology, some more successfully than others; hierarchical, network, object, and in-memory are only a few examples. The most successful data management system has been the relational data management technology [10]. We will start at this point to look at its development and what has made it succeed and also how far forward it has moved in comparison to any other data management system.

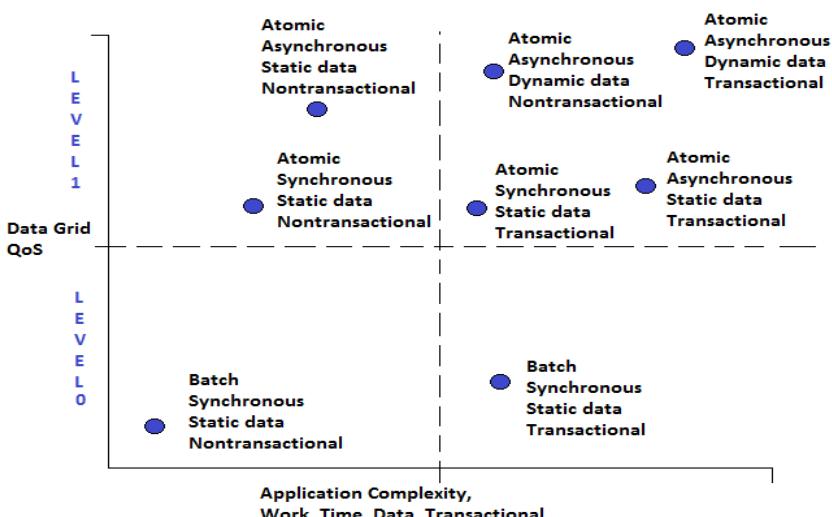


Fig. 1. Date Grid Quality of Service vs. Application Demand and Requirement [1]

3 Related Work

3.1 Proposed Architecture

The proposed architecture named “Dynamic and Scalable Storage Management (DSSM)” to organize Grid storage devices into a large-scale and geographically distributed storage system to meet the requirements imposed by all kinds of Grid applications. The DSSM divides Grid storage devices into multiple geographically distributed domains to facilitate the data access locality [2][5]. The architecture consists of two levels. The bottom level adopts multicast to achieve dynamic, scalable, and self-organized physical domains. The method significantly simplifies the intra-domain storage resource management. The upper level is a virtual domain that consists of geographically distributed and dynamic agents selected from each physical domain.

3.2 Data Locality

Data locality is a measure of how well data can be selected, retrieved, compactly stored, and reused for subsequent accesses [2]. In general, there are two basic types of data locality: temporal and spatial. Temporal locality denotes that the data accessed at one point in time will be accessed in the near future. Temporal locality relies on the access pattern of different applications and can therefore change dynamically [6][7]. Spatial locality defines that the probability of accessing data is higher if the data near it was just accessed (e.g. pre-fetch). Unlike the temporal locality, spatial locality is inherent in the data managed by a storage system, and is relatively more stable and does not depend on applications, but rather on data organizations which is closely related to the system architecture. Reshaping access patterns can be employed to improve temporal locality [2]. Data reorganization is normally adopted to improve spatial locality. Many research efforts have been invested in exploiting the impact of access pattern and data organization of applications on the data locality to achieve performance gains. Fig.-3.1 illustrates the DSSM architecture.

3.3 Domain Division

Choosing a suitable criterion to form GOS nodes into domains is an important factor of the DSSM architecture. DSSM employs the distance criterion in a geographical way to divide GOS nodes into multiple domains to facilitate the data access locality and limit the amount of communication traffic seen by each node with performance guarantee. GOS devices belonging to the same geographical area (e.g. the same enterprise or the same LAN) are formed into a domain, because the GOS devices in the area are normally placed geographically close [11][12]. GOS nodes divide into multiple domains based on:

- Distance
- Quantitative (bandwidth)
- Qualitative (performance analysis)

With a domain based GOS Network, the scalability is achieved from a two-tiered architecture, namely, intra-domain and inter-domain. Within a domain, each GOS device is equal in functions and capabilities, and can leave or join the domain dynamically. Any node can communicate with anyone else directly. Each domain selects a domain agent from the domain members to cooperate with other domains according to domain formation algorithm [2].

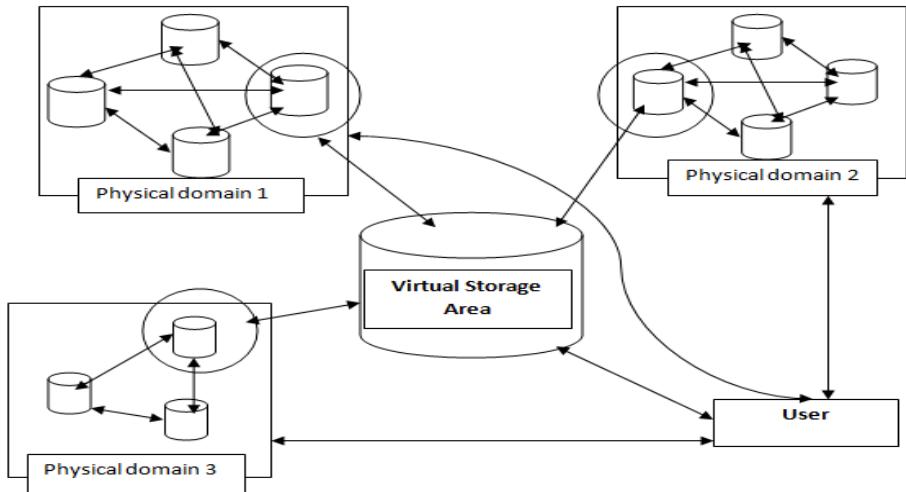


Fig. 2. The DSSM Architecture

4 Data Management in Grid Computing

The DSSM architecture consists of two levels. The bottom level adopts multicast to achieve dynamic, scalable, and self-organized physical domains. The upper level is a virtual domain that consists of geographically distributed and dynamic GOS agents selected from all domains. The algorithm consists of knowledge of the neighbours of each node in a domain to organize the domain members and select a domain agent from all candidate members. Several distributed domain formation algorithms have been devised over the years. One is the lowest-ID algorithm. The other is the highest-connectivity (degree) algorithm.

4.1 Algorithm Formulation at Bottom Level

DSSM architecture organized as a scalable storage system. Assign with a multicast IP address to provide a single entry point the storage space. Each GOS device maintains an Adjacent Information Table (AIT) which keeps a list of all active GOS devices with their related resource information. Information consists of IP address, reminder storage capacity, processing power. A data structure is adopted to describe the entry of AIT. The data structure should be maintained in memory so that they can be accessed with little overhead. The data structure takes 32 bytes per entry.

For a physical domain which has 1000 GOS devices, it takes only

$$1000 \times 32 = 32000 \text{ bytes}$$

to track the whole physical domain.

Compared with the size of main memory, the storage capacity of AIT is negligible.

Table 1. Adjacent Information Table (AIT)

Field	Size	Key Value
AIT ID	4 bytes	Alphanumeric
IP address	32-bits	Alphanumeric
Storage capacity	8 bytes (in MB)	Floating points
Processing power	8 bytes (in MHz)	Floating points

Algorithm#1: When a new GOS device wants to join the domain:

Step1: Send “JOIN” (the device sends a one hop probing multicast “JOIN” request of its coming and the corresponding local resource information such as the processing power, storage capacity, and waits for the response of other online GOS devices).

Step2: Receive “JOIN” request (the online GOS devices which receive the “JOIN” request add the oncoming device’s information to their AIT).

Step3: Send ack “ACCEPT” (online GOS devices send an “ACCEPT” ack back to the probing GOS device using unicast)

Step4: Create AIT (the oncoming device constructs its own AIT in terms of the ack messages.)

Algorithm#2: Leaving of a GOS device

DSSM is based on stable and trusted GOS devices; the leasing of a GOS device is normally caused by maintenance, upgrade and other reasons.

Step1: Send “LEAVE” (multicast leave message in the domain)

Step2: Receive “LEAVE” (once the online GOS devices receive the message, the devices delete the leaving device’s information)

Step3: Update AIT

Step4: Repeat Step2 (if message is not received then the device is assumed to have failed and other reminder GOS devices delete the device from their AIT).

4.2 Algorithm Formulation at Upper Level

Algorithm#3: Select agent from particular physical domain

Step1: Select MAX[PP] (select highest processing power agent within a domain)

Step2: Repeat Step1 (if fail to select highest processing power agent)

Step3: Compare agent to another GOS

Step4: Repeat Step1 (highest processing power will always reselect as an agent)

Step5: In case of same processing power unchanged.

5 Implementation and Experimental Evaluation

We constructed a prototype with three different Thapar University's networks, one High End Computer Lab (HECL) Network emulator, Software Engineering Lab (SEL) Network and several client machines of PG Hostel, J Hostel's networks. All components were connected through a 1000/100M adaptive switch. Table-1 shows the system configurations of the prototype. The WAN emulator forwards packet at line rate and has user-settable delay and drop probability.

5.1 Dynamicity and Reliability Evaluation

The dynamicity and reliability of the DSSM architecture were measured at two stages. At the first stage, we set the IP address of three GOS devices within one network segment to denote a single physical domain. To illustrate the dynamic scalability and reliability of the architecture, we first configured one GOS device in HECL Network, and then the other two GOS devices joined the domain SEL Network, PG Hostel Network respectively, finally, the three GOS devices continuously left and joined the domain.

Table 2. System Configurations of the Prototypes

H/W	GOS Device	WAN Emulator and Clients
CPU	Intel Xeon 2.8GHZ	Intel Pentium IV 2.66 GHZ
Memory	3GB	512MB
NIC	Two Broadcom 100/1000M	Intel(R) PRO/100M
Disks	Six IBM FRU 32P0730	Seagate ST340014A
OS	Red Hat(Kernel 2.4.21)	Red Hat (Kernel 2.4.21)/MS Windows 7

We repeated the above process for more than 50 times, it did not cause any problems in our experiment. Because all GOS devices in the domain have the same processing power, the first GOS device of the domain is selected as an agent by default even it has no other GOS agents to cooperate with. At the second stage, to simulate two physical domains (HECL Network, SEL Network), we configured two network segments by setting the IP address of the three GOS devices. The first domain consisted of two GOS devices, and the second domain had only one GOS device.

5.2 Bandwidth Evaluation

This paper focuses on a dynamic and scalable storage management architecture which may involve hundreds or even thousands of distributed GOS devices. The goal of this work is to support long-distance and bulk data access of large-scale and complex Grid applications. Grid service combines the Web service and WSRF to provide a service based Grid environment that enables heterogeneous environments to be integrated and reconciled to accomplish complex tasks [8]. A set of files appropriate range were transferred over the emulated WAN to measure the bandwidth.

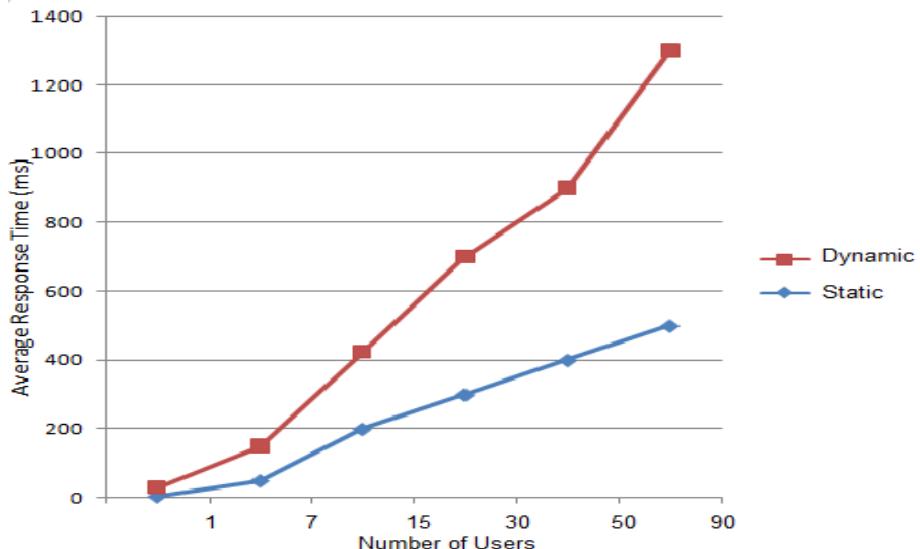


Fig. 3. Average Response Time of the Static and Dynamic Storage System.

6 Conclusion

In this research paper, based on an existing Grid environment, we propose and design a DSSM architecture which organizes GOS devices into different domains to enhance the data access locality in terms of geographical and distributed areas. Storage expansion is achieved by simply adding services. The DSSM architecture avoids the hierarchical or centralized approaches of traditional Grid architecture, eliminates the flat flooding of unstructured P2P system, and provides a dynamic storage pool in Grid environment. The proof-of-concept prototype gives useful insights into the architecture behaviour of DSSM architecture.

7 Future Scopes

The DSSM can be regarded as a virtual two-tiered hierarchy, but it replaces a few root nodes of the traditional hierarchy with many GOS agents which can be expanded to a

large-scale, thus avoiding the single point of failure and the potential performance bottleneck. The DSSM architecture does incur some additional overhead which results in performance penalty on the GOS agents in three scenarios. The first, if hundreds of thousands query requests go to a particular domain agent simultaneously, the agent could become a potential system bottleneck. The reason is that all query traffic going to a domain has to pass through the domain agent. The second, an intelligent domain agent is able to pool the storage resources and keep the load balance among the GOS devices in the domain. It takes some additional computing overhead to do this job. A dynamic and transparent data replication mechanism which automatically places the data replicas across domains where they are needed is able to alleviate the performance impact on the GOS agents, it is also crucial to the overall performance.

Acknowledgements. I sincerely acknowledge the guidance rendered to me by Dr. Seema Bawa, Professor, at Computer Science and Engineering Department, Thapar University, Patiala. Words are not enough to describe her invaluable support; inspiration, guidance, encouragement and making me understand easily anything, anytime during the thesis work.

I would like to thank to Dr. Maninder Singh, Professor, Thapar University, Patiala who really put me in a real technical frame and bent of mind – that allowed me to conceptualize and materialize this concept.

I am deeply indebted to our Sr. Grid Members (Ms Shashi, Ms Rajani, Ms Seemu, Ms Pankajdeep Kaur, Ms Ratinder Kaur) who constantly encouraged me tender my utmost gratitude and appreciation for her invaluable guidance and suggestions.

References

- [1] Di Stefano, M.: *Distributed Data Management for Grid Computing* (2005) ISBN 0-471-68719-7
- [2] Deng, Y., Wang, F., Helian, N., Wu, S., Liao, C.: Dynamic and scalable storage management architecture for Grid Oriented Storage device. *Parallel Computing* 34, 17–31 (2008)
- [3] Shoshani, A., Sim, A., Junmin: Storage Resource Managers: Middleware Components for Grid Storage. Gu Lawrence Berkeley National Laboratory Berkeley, California 94720
- [4] Yang, B., Garcia-Molina, H.: Improving search in peer-to-peer networks. In: *Proceedings of the 22nd International Conference on Distributed Computing Systems (ICDCS 2002)*, Vienna, Austria, pp. 5–14 (July 2002)
- [5] Deng, Y., Wang, F.: A heterogeneous storage Grid enabled by Grid service. *ACM SIGOPS Operating Systems Review*, Special Issue: File and Storage Systems 41(1) (2007)
- [6] The Open Grid Services Architecture, <http://www.globus.org/ogsa/>
- [7] Jansen, F.W., Reinhard, E.: Data locality in parallel rendering. In: *Proceedings of the 2nd Eurographics Workshop on Parallel Graphics and Visualisation*, pp. 1–15 (1998)
- [8] Web Services Resource Framework, <http://www.globus.org/wsrf/>
- [9] Deng, Y.: Deconstructing Network Attached Storage systems. aEMC Research China, Beijing 100084, PR China (2008)

- [10] McHugh, J., Quass, D., Widom, J.: Lore A Database Management System for Semistructured Data,
[http://citeseervx.ist.psu.edu/viewdoc/download?
doi=10.1.1.81..pdf](http://citeseervx.ist.psu.edu/viewdoc/download?doi=10.1.1.81..pdf)
- [11] Kumar, A., Bawa, S., Sharma, V.: Dynamic and Scalable Data Storage Management in Grid environments. In: National Conference on Emerging Trend in Engineering and Sciences. Samrat Ashok Technological Institute, India (December 2010)
- [12] Kumar, A., Bawa, S.: Performance Modeling and Run Time Estimation for Large-Scale Data and Computational Grid. In: International Conference on Advance in Modeling, Optimization and Computing, IIT Roorkee (U.K.), India (2011) ISBN 81-86224-71-2

Behavioral Profile Generation for 9/11 Terrorist Network Using Efficient Selection Strategies

S. Karthika, A. Kiruthiga, and S. Bose

Department of Computer Science and Engineering
College of Engineering Guindy, Anna University, Chennai-600025
sk_mailid@yahoo.com, kiruthiga312@gmail.com, sbs@cs.annauniv.edu

Abstract. In recent days terrorism poses a threat to homeland security. It's highly motivated by the “net-war” where the extremist are organized in a network structure. The major problem faced is to automatically identify the key player who can maximally influence other nodes in a large relational covert network. The nodes and links are represented in the form of a directional semantic graph where each node is related with more than one relationship with the other node. The behaviors of nodes are analyzed based on the semantic profile generated. This analysis helps the crime analyst to judge the key player for a criminal activity. The semantic profile is obtained by choosing carefully the path types that suits best for a specific node. The selection strategies can be generalized as path equivalence and constraint based. The strategy further supports the variable relaxation approach by grouping all the paths with the same sequence of relations as a single path type. This can also be made as user friendly by letting the user to represent their own preferences on the nodes and links.

Keywords: Social Network Analysis (SNA), Terrorism, Unsupervised learning, Selection Strategy, Semantic profile.

1 Introduction

An event that brought a worldwide attention towards terrorism is the unforgettable 9/11 disaster [14] [15]. This provoked a need for awareness on anti-terrorism based national security. From then on a lot of research has been done on the terrorist networks. This covert network is seen as a social network with a lot of secrecy and influence. It is supposed to be covert or hidden but still has to manage the communication between them periodically. There should be some structure maintained within this network at the higher levels and at the lower hierarchy, they manage a cell structure.

In a covert social network one of the critical problems is to identify a set of key players who are highly influential. The problem of determining the importance of nodes is resolved based on the researches done in node centrality, group centrality and structural measures. The covert networks emphasizes on the importance of the links

which presents the relations among the nodes [12]. When a set of nodes are given, automatically identifying the key players helps significantly in homeland security issues. Generally the characteristics of a person are understood based on his/her behavior. These behaviors can be represented in the form of a profile.

The profile generation uses the semantic graph as its input. A semantic graph is a graph that is used to represent semantic structure in terms of nodes and relations between them. A semantic structure has two nodes that are linked at least by one relation. A node can be any type of object like a person involved in an activity or a location or an event etc. and the link is the relation between the nodes. In a semantic graph the links can exist between different types of nodes and there can be multiple relationships i.e. heterogeneous relations between heterogeneous nodes. For example the nodes A and B can be colleagues as well as neighbors. Hence such a semantic graph can be called as Multi-Relational Networks (MRN).

The sample 9/11 network such as the one shown in Fig.1 is an MRN that represents hijackers, other associates and the locations involved in this attack as nodes along with multiple relations between them as links.

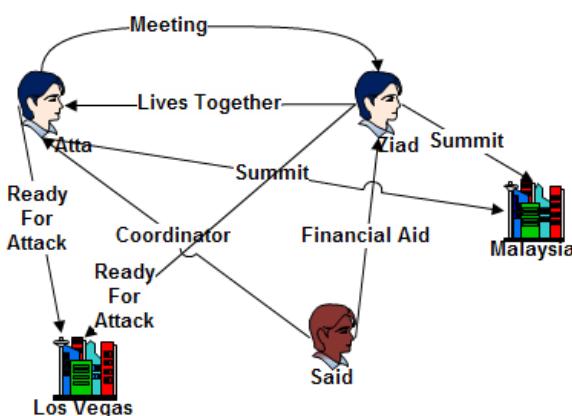


Fig. 1. Sample MRN of 9/11 attack

In this paper an efficient technique for analyzing the behavior of the nodes is presented by generating the profile of each node in the MRN. The profile illustrates the importance and the contribution of that specific node to the event. The links in the MRN generates various paths among the nodes and these paths are condensed to different formats using the variable relaxation approach. Many efficient selection strategies form path types from the condensed paths. The profile is a collection of such path types that describes specific characteristic about the node.

This overcomes the issues of the earlier methods like the ensemble problem in which the structural equivalence of the nodes is not efficient enough to determine the key player. Secondly the goal problem where not all the nodes with high centrality

values are key players and the distantly connected nodes have less effect on the influential nodes[1][10].

The remainder of this paper is structured as follows. Section 2 provides the various supporting works for the developed system. Section 3 illustrates the problem statement along with the assumptions made and Section 4 explains about the various efficient selection strategies used to generate the profile. In section 5, we present our behavioral profile generation algorithm SoNMine based on selection strategies and in section 6 the analysis and evaluation of the selection strategy is done. Section 7 concludes the paper and discusses the future work.

2 Related Work

The semantic graph has been used by Shou-de Lin et.al [2] in an unsupervised framework UNICORN, to generate a profile from which suspicious nodes could be detected and uses a novel explanation system to verify the profiles using natural language processing. Shou-de Lin et.al [3] has also used the interestingness measure to determine the rarity using the node path and node loop discovery strategies. The same author has implemented the interestingness measure using rarity analysis for the bibliography dataset in [4].

Stephen P.Borgatti[1] discusses about the key player problem which determines the highly important node set in the given network and a cut set of nodes that could fragment the entire network into individual entities. This method uses the distance based metrics to determine the centrality of the nodes and maintains the minimum size for the subset to defragment the network.

Valdis E.Krebs [5] has analyzed the network of 9/11 attack and has used the basic centrality measures to determine the importance of nodes. The author has also focused on identifying the task and trust ties between the conspirators by using the shortcuts and without using it.

Mohammad Al Hassan et.al [6] has focused on link prediction from the supervised learning perspective. The author handles the link prediction problems using classification techniques and uses the information gain, gain ratio and average rank as performance metrics. The author faces the problem in finding trained samples due to the incompleteness and fuzzy boundaries.

Bin Zhao et.al [7] proposes the use of relational Markov networks to describe the entities and the relations among them in an affiliation network. The author has used Profile In Terror (PIT) data base to study the entity and relationship labeling. The author faces the problem in using a supervised database as the random sampling created falls into different subsets.

Lei Zou et.al [8] proposes a sub- graph matching query system in which the similarity of given query is compared to that of the target query using score variable. This system has extensive scalability and quick response time.

3 Problem Definition

In this paper we focus on analyzing the behavior of the nodes in a covert network by generating a profile for each using the selection strategies. In other words we use unsupervised methodology to determine the node's dependency through the contribution value generated with respect to each activity.

We assume that the dataset collected is complete and the time order of the occurrence of the event is not important. We also assume that the social network under consideration is a single community and not heterogeneous. Given these assumptions we define the problem as to determine efficient selection strategies for choosing the path types to be included in the profile of a node. These strategies judge the involvement of a node in a specific activity.

4 Selection Strategies

The selection strategies have been considered here because it helps us to choose a path type which is collection of paths based on which the profile is generated [2]. These strategies are used basically in two circumstances namely path oriented selection strategy and constraint oriented selection strategy. The path selection strategy is more related to the syntax of a path and the constraint strategy is based on the criteria with which a path belongs to a specific path type.

Definition of selection strategies: A selection strategy $S(V, R, E^{-1})$ is a path where V is a finite set of nodes, R is a finite set of relations and E is a finite set of edges $E \subseteq V \times R \times V$. E^{-1} is a finite set of inverse edges such that $(v_2 R v_1) \in E$.

4.1 Path Selection Strategy

The path selection strategy is classified into three types namely

1. Same sequence of relations
2. Single relation
3. Loop based strategies

The same sequence of relations has all the paths with the same type of consecutive relations as elements in the path and it's defined as

Definition: A selection strategy $S(V, R, E^{-1})$ is a path where V is a finite set of nodes, R is a finite set of relations such that $R \subseteq r_1, r_2, \dots, r_n$ in the same sequence among the V and E^{-1} is a finite set of inverse edges such that $(V_2 R V_1)$ where $E \subseteq V \times R \times V$.

The single relation strategy has paths of only single type of relation among the nodes. It's defined as

Definition: A selection strategy $S(V, r, E^{-1})$ is a path where V is a finite set of nodes, r is a single relation such that $R \in r$ and E^{-1} is a finite set of inverse edges such that $(V_2 R V_1)$ where $E \subseteq V \times R \times V$.

The loop based selection strategy has the paths which starts with a node and ends with the same node. It's defined as

Definition: A selection strategy $S(v, R, E^{-1})$ is a path where v is the starting and ending node of the path where $v \in V$, R is a finite set of relations and E^{-1} is a finite set of inverse edges such that $(v_2 R v_1)$ where $E \subseteq V \times R \times V$.

4.2 Constraint Based Selection Strategy

The constraint based strategies are classified into six types namely

1. Exclusive node
2. Selection node
3. Exclusive relation
4. Selection relation
5. Path length based
6. Score based

The exclusive node based strategies chooses the paths such that the path doesn't have a mentioned node as its element.

Definition: A selection strategy $S(!v, R, E^{-1})$ is a path where V can be any node other than v , R is a finite set of relations and E^{-1} is a finite set of inverse edges such that $(v_2 R v_1)$ where $E \subseteq V \times R \times V$.

The selection node strategy forms the path type with the paths having a specific node whose occurrence is at least once in the path.

Definition: A selection strategy $S(v^+, R, E^{-1})$ is a path where $v \in V$ and v occurs at least once, R is a finite set of relations and E^{-1} is a finite set of inverse edges such that $(v_2 R v_1)$ where $E \subseteq V \times R \times V$.

The exclusive relation based strategies chooses the paths such that the path doesn't have a mentioned relation as its element.

Definition: A selection strategy $S(V, !r, E^{-1})$ is a path where V is a finite set of nodes, R can be any relation other than r and E^{-1} is a finite set of inverse edges such that $(v_2 R v_1)$ where $E \subseteq V \times R \times V$.

The selection relation strategy forms the path type with the paths having a specific relation whose occurrence is at least once in the path and its defined as

Definition: A selection strategy $S(V, r^+, E^{-1})$ is a path where V is a finite set of nodes, $r \in R$ and r occurs at least once and E^{-1} is a finite set of inverse edges such that $(v_2 R v_1)$ where $E \subseteq V \times R \times V$.

The path length based selection strategy is a very needful aspect for generating the profile because it determines the length of the paths that should be considered in the path types [13]. We assume that all the paths in the path type are of uniform length.

Hence only, the inverse relation is to be used in path generation. The limits of the path length is curtailed because farther the node it has lesser impact on the starting node. So if a node is distantly connected then it is less influential on the source node. The path length for 9/11 dataset is limited to four.

The score based strategy assigns a score for each path based on the similarity of the sequence of relations/labels in the path with that of the target path [8]. All the paths with the second highest score are also considered when the path type is determined. The Table 1 shows the various selection strategies and their path type formats. The path type is obtained by reducing the path using variable relaxation approach [2].

Table 1. Different types of Selection Strategies

SELECTION STRATEGIES	PATH TYPE FORMAT
1. PATH BASED	
Same Sequence of Relation	$x(?,?)^y(?,?)^z(?,?)$
Single Relation	$x(?,?)^x(?,?)^x(?,?)$
Loop based	$?x(?,?)^?x(?,?)^?x(?,?)$
2. CONSTRAINT BASED	
Selection Node	$?^?x(?,?)^?$ Or $?^?x(?,?)^?$
Exclusion Node	$?^?(!x,?)^?$ Or $?^?(?,!x)^?$
Selection Relation	$?^?x(?,?)^?$
Exclusion Relation	$?^!x(?,?)^?$
Score based	Highest score: $x(?,?)^y(?,?)^z(?,?)$ Second highest score: $x(?,?)^y(?,?)^z(?,?)$
Path length based	Maxi relations: 4 $x(?,?)^y(?,?)^z(?,?)^a(?,?)$

5 Behavioral Profile Generations with SoNMine (Social Network Mining)

In the earlier section the various selection strategies for profile generation have been discussed. Based on these an unsupervised system SoNMine has been presented below which features on the various attributes of node types and relation types. The following pseudo code describes the algorithm formally:

<pre> function SoNMine (M, ss, k){ // M is an MRN <V,E,R > // ss is a selection strategy for selecting // path types // k is the maximum path length for path // types; k=4 var path_set[V , R , PT] PT := path_types(M, ss, k) for n ∈ V && for pt ∈ PT profile[n, pt] := dependency_val(M, n, pt) return profile } </pre> <pre> function path_types(M,ss, k) PT := {R₁, R₂, . . . , R_R} for path_length := 1 to k – 1 { // path_length ss { for pt ∈ PT where length(pt)=path_length for r ∈ R { // same_seq ss{ for path_length := 1 to k – 1 if for all R:= {r₁,r₂,r₃} //single_rel ss{ if for all R:= r} //loop ss{ if start_node:=end_node for v ∈ V} //selection_node ss{ V:=v at least for any one V} </pre>	<pre> //exclusion _node ss{ V !=v for all paths } //selection_rel ss{ R := r at least for any one R} //exclusion_rel ss{ R !=r for all paths } //score_based ss{ if target_path:=test_path score:= high; else if target_path is partially same to test_path score:= min;} if path_exists(M, pt') and satisfies(pt', ss) PT := PT U pt' return PT </pre> <pre> function dependency_val(M, s, pt) Calculate Contribution; Normalize the Contribution; return Contribution </pre>
---	--

6 Evaluation of SoNMine

SoNMine generates various path types for the 9/11 attack through a synthetically generated dataset and by following the above mentioned selection strategies which is illustrated in Table 2.

SoNMine finds the influential node based on their semantic profile which has contribution value of the nodes towards a specific behavioral path type. UNICORN is an already existing framework for the above mentioned purpose [2]. It has done its evaluation based on the relation only selection strategy and used the path length as 5 [11]. The following Table 3 compares the performance of UNICORN with SoNMine based on the score based selection strategy and has shown the results for a sample of

Table 2. Path types with and without variable relaxation approach for all Selection Strategies

SELECTION STRATEGY	PATH GENERATED	PATH TYPE WITH VARIABLE RELAXATION APPROACH
Same sequence of Relation	BrotherOf(saleem,nawaf)^ReadyForAttack(nawaf,hani)^CarVisitWith(hani,atta)	BrotherOf(?,?)^ReadyForAttack(?,?)^CarVisitWith(?,?)
Single relation	Meeting(atta,ziad)^Meeting(ziad,said)^Meeting(said,atta)	Meeting(?,?)^Meeting(?,?)^Meeting(?,?)
Loop based	Meeting(hamza,marwan)^Pilot_Training(marwan,ziad)^Friends(ziad,atta)^ReadyForAttack(atta,hamza)	?(hamza,?)^?(?,?)^?(?,?)^?(?,hamza)
Selection node	Meeting(saleem,nawaf)^Attack(nawaf,hani)^LivesIn(hani,salem)	?(saleem,?)^?(?,?)^?(?,?)
Exclusion node	Atta	-
Selection relation	BrotherOf(ahmed,hamza)^Attack(hamza,hani)^Meeting(ziad,hani)	BrotherOf(?,?)^?(?,?)^?(?,?)
Exclusion relation	Meeting	BrotherOf(?,?)^ReadyForAttack(?,?)^CarVisitWith(?,?)^FigthsWith(?,?)^LivesWith(?,?)
Score based	Meeting(saleem,nawaf)^Attack(nawaf,hani)^LivesIn(hani,salem) Meeting(ahmed,hamza)^Attack(hamza,hani)^Meeting(ziad,hani)	HighestMeeting(?,?)^Attack(?,?)^LivesIn(?,?) Meeting(?,?)^Attack(?,?)^Meeting(?,?)
Path length based	K=4 Before: BrotherOf(saleem,nawaf)^ReadyForAttack(nawaf,hani)^CarVisitWith(hani,atta)^Meeting(atta,ziad)^LiveWith(Ziad,Marwan) After: BrotherOf(saleem,nawaf)^ReadyForAttack(nawaf,hani)^CarVisitWith(hani,atta)^Meeting(atta,ziad)	Before: BrotherOf(?,?)^ReadyForAttack(?,?)^CarVisitWith(?,?)^Meeting(?,?)^LivesWith(?,?) After: BrotherOf(?,?)^ReadyForAttack(?,?)^CarVisitWith(?,?)^Meeting(?,?)

nodes and path types. By using the score parameter it not only chooses the path that has the exact match to target path type but also considers the second highest matching paths. This increases the number of paths considered for the contribution value computation and increases the contribution value of the nodes towards the path type. Hence this system provides more accurate contribution values and detailed profile for the nodes.

Table 3. Performance evaluation of SoNMine and UNICORN based on score Selection Strategy

PATH TYPE		NAWAF	KAM	ATTA	HANI
With Score(SoNMine) / Without Score (UNICORN)					
[READY FOR ATTACK,READY FOR ATTACK,CAR VISIT WITH,MEETING,]	SoNMine	0.148147 30232934 75	0.33503 8529803 38594		
	UNICORN	0.113635 8	0.30136 9549113 1174		
[CAR VISIT WITH,CAR VISIT WITH,MEETING,M EETING,]	SoNMine	0.004287 09248134 0961		0.277233 22121202 29	0.626770 0931716 268
	UNICORN	- 0.036052 60862126 335		0.250867 11968343 68	0.602869 4636957 58
![MEETING,MEETING,MEETING,MEE TING,]	SoNMine	- 0.330364 04565141 11	0.04383 4825377 27326	.0252689 52568649 517	0.351159 0496541 419
	UNICORN	- 0.384261 59	0.02562 8246327 478024	0.062670 15319226 307	0.400448 5268824 518
[SALEEM,]	SoNMine		- 0.16119 2123479 79814	0.190513 18118439 525	- 0.135376 9159010 9716
	UNICORN		- 0.20823 5947889 5923	0.160983 58425586 06	- 0.176794 7894343 2814

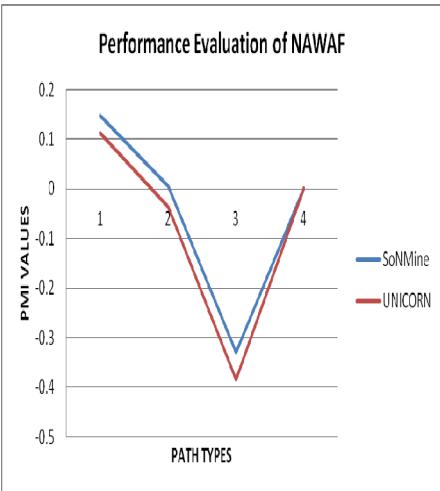
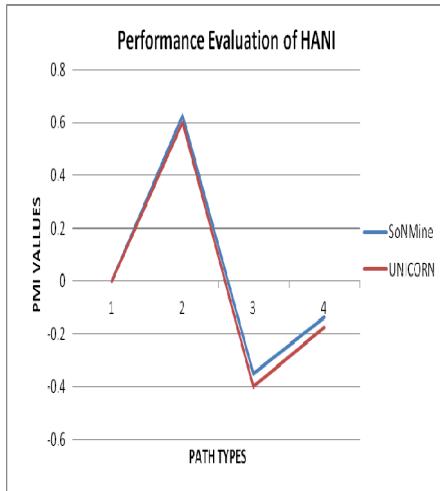
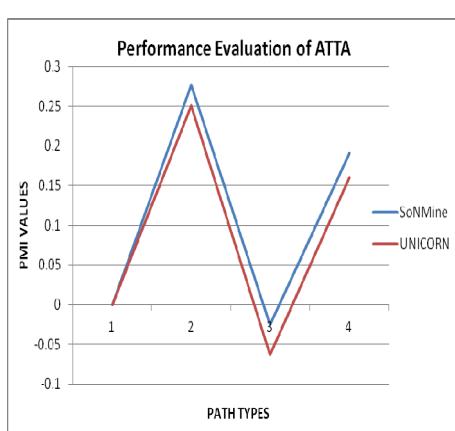
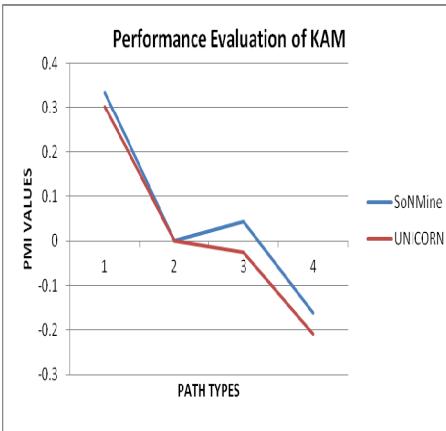
**Fig. 2a.****Fig. 2c.****Fig. 2b.****Fig. 2d.**

Figure 2.a, 2.b, 2.c, 2.d shows the performance of the four nodes Nawaf, Atta, Hani and KAM respectively.

The profile of the node [9] is generated based on the above discussed selection strategies. The Table 4 below discusses about the profile of the node Nawaf who has been involved in various activities like car visit, meeting, ready for attack with higher PMI based contributonal values.

Table 4. Profile of Nawaf

PROFILE OF : NAWAF	
[READY FOR ATTACK,READY FOR ATTACK,CAR VISIT WITH,MEETING,] = 0.9662304535905736	
[NAWAF,NAWAF,] = 2.4703078503668476	
[READY FOR ATTACK,CAR VISIT WITH,MEETING,MEETING,] = 1.6718001541490761	
![MEETING,MEETING,MEETING,MEETING,] = -2.1546649629174235	
[CAR VISIT WITH,CAR VISIT WITH,MEETING,MEETING,] = 0.02796081499764329	

7 Conclusion

In this paper we have discussed about an unsupervised system called as SoNMine which uses many selection strategies to analyze the profile of the nodes involved in the 9/11 attack. The selection strategies are based on the relation types and the constraints. We have determined the importance of the node in a profile based on the contribution value found. Since it's an unsupervised system there is no need for any training data set but the relations that are to be considered should be already selected. To further improve the performance of the SoNMine system the time order of the event occurrence could be included in the data set.

References

1. Borgatti, S.: Identifying sets of key players in a social network. *Comput. Math. Organiz. Theor.* 12, 21–34 (2006)
2. Lin, S.-D., Chalupsky, H.: Discovering and explaining abnormal nodes in semantic graphs. *IEEE Transactions on Knowledge and Data Engineering* 20(8), 1039–1052 (2008)
3. Lin, S.-D., Chalupsky, H.: Unsupervised Link Discovery in Multi-relational Data via Rarity Analysis. In: Third IEEE International Conference on Data Mining (2003)
4. Lin, S.-D., Chalupsky, H.: Using Unsupervised Link Discovery Methods to Find Interesting Facts and Connections in a Bibliography Dataset. *ACM SIGKDD Explorations Newsletter* (2003)
5. Krebs, V.E.: Mapping Networks of Terrorist Cells. *Connections* 24(3), 43–52 (2002)
6. Hasan, M., Chaoji, V., Salem, S., Zaki, M.: Link prediction using supervised learning. In: Proceedings of the Workshop on Link Discovery: Issues, Approaches and Applications (2005)
7. Zhao, B., Sen, P., Getoor, L.: Entity and relationship labeling in affiliation networks. In: ICML Workshop on Statistical Network Analysis (2006)
8. Zou, L., Chen, L., Lu, Y.: Top-k subgraph matching query in a large graph. In: PIKM 2007: Proceedings of the ACM First Ph.D. Workshop in CIKM, pp. 139–146. ACM, New York (2007)
9. Kocsis, R.N.: Criminal Profiling: International Theory, Research, and Practice, pp. 169–188. Humana Press Inc., Totowa (2007)
10. Memon, N., Harkiolakis, N., Hicks, D.L.: Detecting High-Value Individuals in Covert Networks: 7/7 London Bombing Case Study. *Computer System and Application* (2008)

11. Memon, N., Larsen, H.L., Hicks, D.L., Harkiolakis, N.: Detecting Hidden Hierarchy in Terrorist Networks: Some Case Studies. In: ISI 2008 Workshops, pp. 477–489 (2008)
12. Memon, N., Wiil, U.K., Alhajj, R., Atzenbeck, C., Harkiolakis, N.: Harvesting covert networks: a case study of the iMiner database. *Int. J. Networking and Virtual Organisations*
13. Boyer, R.S., Strother Moore, J.: A Fast String Searching Algorithm. *Communications of the ACM* 20(10) (1977)
14. Kean, T.H.: The 9/11 Commission Report. National Commission on Terrorist Attacks upon the United States, USA (2004)
15. Responsibility for September 11 attacks (November 2010),
http://en.wikipedia.org/wiki/Responsibility_for_the_September_11_attacks

A Quantitative Model of Operating System Security Evaluation

Hammad Afzali¹ and Hassan Mokhtari²

¹ Operating System Security Lab (OSSL), Alzahra University, Tehran, Iran

² ICT Department, Malek-ashtar University of Technology, Tehran, Iran

{afzali,mokhtari}@ce.sharif.edu

Abstract. Operating System (OS) as the root of trust for all applications running on the computer systems and plays an important role in information security. If the network or the software in application level that are executing on the operating system be unsecure, it is expected that OS as a defensive layer or in another words as the last defensive layer protects the security of information. In this paper, first, we extract and classify a vast spectrum of security features which has been used in multi-purposed OS, and then attribute them to three levels of low, medium and high. Our case study indicates how it is possible to evaluate OS security and specify the security level of an OS.

Keywords: OS security, Security mechanisms, Security levels, Evaluation Methodology.

1 Introduction

Recent years have witnessed a steady increase in the widespread and variety of OS vulnerabilities, resulting in data manipulating, tampering kernel, and so on [14, 17].

Security of OS is one of the most complicated security issues due to the reason of being execution bed of all other programs and having administration responsibility of the resources like memory, I/O and etc. Also it is expected that if the implemented security mechanism in the higher levels of an organization, network and programs go out of commission, at least OS can avoid from more damages. In other side, any OS vulnerabilities, it will influence on the other security factors of system or even tamper the whole information system security [20, 30].

Till now many activities has been done in the OS security field thorough which we can point to using virtualization in security control [23, 26], various models of secure file system [24], administration of memory in order to avoid common attacks [22], secure oriented coding guidelines [21], implementing strong access control model of MAC in SELinux [8] and security architecture like FLASK [2].

Evaluation of OS Security has been considered from long time ago and vast researches have been done in this regard. The most comprehensive treatment of this topic is assessment of OS security based on the standards such as ITSEC, TCSEC and above all CC [13]. Based upon the latter standard some Protection Profiles [29] for different environment and OS have been represented and also some of OS were succeeded to obtain assurance levels of this standard. The considered general

approach [5, 7, 13, 20] in this standard is specifying environmental threat and using security functions to mitigate threats.

Another performed activity in evaluation of Os security is scattered actions that have been done for a specific OS [4, 24]. Our considerations indicate that still there exist no reliable and specific method for assessing OS Security and also most of OS producers do not access to any particular secure model such as those that exist in the software engineering like CLASP [1]. In this paper we tried to represent a specific methodology for evaluation of OS security so that somehow solves the above mentioned problems, although, this work is the beginning step in this regards.

2 Evaluation Methodology

In this section, we generally describe our OS security evaluation methodology by expressing a definition of security levels and the manner in which an OS achieve these levels. At first, we accumulate the complete list of mechanisms and capabilities implemented for increasing security in common, well-known OS. In the next, we classified the mechanisms and tried to extract the security parameters independent of details and methods of their implementation in specific OS. Afterwards, we categorized this information in a hierarchy structure in form of concepts of *class*, *sub class*, *criteria and mechanisms*.

As it's mentioned before, in this hierarchy it has been attempted to at most use abstraction. Therefore, the applied literature in this regard even in the lowest levels of the hierarchy is independent from a particular OS or specific implementation of a mechanism.

The granularity level in the leaves of proposed hierarchy is fine and detailed enough, so that one can easily separate mechanism in the leaves based on the power and the utilization rate of this mechanism in establishment of security. This classification helps to choose more powerful security mechanism for higher security level. In other words, separation and security categorization has been done in a way to be able to easily select relevant mechanism after raising the security level.

After parameters classification, based on some of experts' view points, the existing security mechanism have been attributed to 3 levels of *low*, *medium*, and *high*. In other words per each mechanism, we asked some experts: "with which security level, will you use this mechanism?" In fact, it has been distinguished which mechanisms must exist in these three levels of security. Along with this description, an OS has a security level if all the required mechanisms of that level have been applied.

Evaluation phases for specifying the security level of a typical OS are being introduced briefly in following stages:

1. Receiving OS along with the required documents.
2. Testing criteria and mechanisms according to the selected level for evaluating OS.
3. Determining maximum verified security level.
4. Announcing criteria for promoting to selected security level.

3 Evaluation Parameters

In this section, we are going to describe the concepts by which we formed the hierarchy. These concepts include: ***parameter, class, sub class, criteria and mechanism.*** These concepts are defined as following:

Class: According to the OS security literature we categorized all of security issues in to 10 classes. The classes are the highest level in our hierarchy and we described them in this section.

Subclass: In the second level of our hierarchy we have some subclasses. It means that each class is distributed to 69 sub classes in a lower level of detail.

Criteria: In next lower level each subclass distributed to some criteria. The derived hierarchy tree from the above distribution is not a balanced tree. In this hierarchy structure, we will find 108 criteria totally.

Mechanism: Mechanism is the leaves of our hierarchy and each criterion are described by some mechanism. All mechanism of same criteria tries to satisfy the purpose of criteria by different strength. So the mechanism has two main attribute: atomicity and security level. By atomicity we mean that mechanism is not detailed more and is so simple that one can readily verify if an OS has it or not. By security level we mean that having mechanism is required for each security level (i.e. OS in which level of security must implement the mechanism).

Parameter: Any node of the hierarchy is called a parameter.

As an example, about criteria of “access control to the password file” three mechanisms were being defined:

1. Creating restriction in accessing to passwords authentication data files.
2. Keeping passwords file in a special place.
3. Encryption of passwords before storage.

The third mechanism must be considered for an OS in low level security, the first and third ones have to consider in medium level and in high level all the mechanisms must be considered. Therefore, if an OS implements the above 3 mechanisms, it can be said, this OS meets the high level of security.

The number of mechanisms for every defined criterion varies. The total number of mechanisms in all criteria is over 200 ones. The reason behind this detailed dividing of the criteria and mechanisms is nothing but the obvious and repeatable evaluation. Because if the criteria are being generally stated similar to other analogical tasks such as CC, the exact and distinct assessment of criteria will face with problems and evaluator dependent.

Similar structure can be seen in definition of security requirements in second part of CC standard in the form of Class, family and component. Here we followed from this mentioned pattern and finally we achieved this hierarchy for OS security. In order to make sure of considering all the various faces of security and parameters, we did the following two tasks:

1. In a downward glance, the existing criteria in confirmed protection profiles of CC and scientific articles of OS Security field have been noticed and then used as classification of mechanisms thorough applying changes in order to have better relation with the available mechanism in real OS.
2. In an upward glance, all the existing mechanisms in Linux [4, 12, 16, 25-27], MAC OS [28, 29], and Windows [9, 15] have been gathered and then classified.

The ten extracted classes for OS security are as following:

- ***Authentication:*** the security requirements related to operation and security policies relevant to Authentication includes authentication methods and authorization in different levels. Each user must be authenticated based on the defined and allowed operation.
- ***Access control:*** by controlling access to system operation and user and system data, confidentiality and other security services in different level are achieved.
- ***User data protection:*** one of the main factors of a secure system is protection of user information against disparate damages.
- ***Protection of Meta data and OS security operations:*** this class contains operation requirement relevant to integrity and management of OS security mechanisms and its data. Protection of system's Meta data neutralizes most of system's attacks.
- ***Logging and auditing:*** in the case of separation of duty, recording accesses and systems' events can provide high security. Thorough logging and audit mechanisms at the time of attack incidence or after that, it can be distinguished that which security mechanism have been attacked and which user is responsible for the attack settlement. It needs to mention that recording users' activities can avoid from most of malicious attempts psychologically.
- ***Security management:*** diverse management roles are being defined in this section which covers managerial aspects of reminding domains. This area is used for management of data attributes and OS security operations.
- ***Trusted path:*** it is a mechanism for making sure that user is interacting with real agent and entity not faked one. Making trusted communication path between user and OS security function and also a reliant connection among OS security function and other IT output are the necessity of a security system.
- ***Resources management:*** vital resources such as processing and storing resources must be available. A system must endure against failure and be capable of prioritizing of services and also perform resource allocation rightly.
- ***Cryptography:*** one of the main purposes in every system is confidentiality i.e. preventing unauthorized user access to data. Usually this need meets via methods like cryptography. OS protective functions can use cryptography techniques for achieving high level security. Other relevant purposes are authentication, non-repudiation and trusted path. In this regards key management and security of key is the most important one.

- **Secure backup and restore:** in a secure system, some plans must exist for supporting data and also restoring those data which have been lost unwantedly or thorough sabotage tasks.

Table 1. Subclasses and criteria for authentication class.

Sub Class	Criteria
Password security	Security in generating password Access control to password file. Determining at minimum and maximum time for remaining the passwords constant. Impossibility of changing password of a user via other users.
Possibility of applying security requirement based on the user account level	Supervisor account security restriction Guest account security restriction
Re-authentication for critical operation	Installing program Changing user role Changing password idle user session Login after standby or hibernate state. Executing system programs. Export data to external medium. Loading a new operating system
Applying time and position entering the user into system	-
Controlling unsuccessful authentication	-
Non confidence to one authentication method	One factor authentication. Double factor authentication. Triple factor authentication.
Un revealing extra information at the authentication time of users	Verification of username and password in an atomic operation. Not showing the previous login information in login page. Not showing the reason for unsuccessful authentication
Authentication from trusted path	-
Authentication module	Authentication module transparency. Not supporting the user level authentication module.

As we mentioned before, these 10 classes have divided into 69 subclasses in a lower level and at the lower level of hierarchy to 108 criteria. At the lowest level the relevant mechanisms to these criteria are over than 200 mechanisms. The NIST uses such a measurement methodology to measure security in information system [8, 11] but our method is specific for OS security measurement. In contrast with Common Criteria approach [13], we can say that CC is not a methodology for measurement but can be used for that. Measurement and evaluation based on CC leads us to a heavy process and hard to repeat process. We used the Security Function of CC literature and dependability model [18] as input to make our hierarchy. Due to high volume of the hierarchy in the table 1, we represent the related detail for authentication class.

Although in the table 2, there are the mechanism and their assignment to security level for subclass 'Possibility of applying security requirement based on the user account level'.

Table 2. Mechanism and their assignment to security level for subclass: "possibility of applying security requirement based on the user account".

Criteria	Mechanism	Security Level
Supervisor account security restriction	Not using the phrase like administrator or root for privileged user	Medium or higher
	Restricting the number of privileged user	Medium or higher
	Mandatory change of username and password for privilege user at a short time period.	Medium or higher
	Multi-factor authentication	High
Guest account security restriction	Restricting the Guest user privileges	Low or higher
	Not using the phrase like Guest or Anonymous for guest user	Medium or higher

4 Case Study

In this section, we apply our evaluation methodology to the Fedora OS [10] because to show it is operational. For each criterion we determined that the OS have implemented mechanism up to which security level. Also we have determined if the OS activates this mechanism by default or it needs some effort to configure securely. In both last situations we suppose that the OS has implemented the mechanism, nevertheless if the OS doesn't implement the OS (neither be default nor by some configuration and effort) we suppose that the OS doesn't implement that mechanism.

Unfortunately because evaluation of some criteria needed some information about the OS development process and/or about the specific cryptographic standard specifically for each country, we could evaluate 91 criteria out of 108 criteria in Fedora OS. The result of our evaluation for these 91 criteria can be impressed as follow:

- In 43 criteria, out of 91 criteria, the Fedora is in HIGH security level. So it must implement some mechanism in 48 criteria ($91-43=48$) to completely promote to HIGH security level.
- Fedora in 65 criteria, out of 91 criteria, has MEDIUM security level. So if it add some security mechanism to 26 residual criteria it promote to MEDIUM security level.
- Fedora has all mechanism that an OS must have for a LOW level security.

5 Conclusion and Future Work

In this quest, we developed an OS security evaluation method that is specific for OS security measurement, opposite most of prior researches. Our approach showed how

one can measure the OS security, by putting and categorizing the OS security mechanism in a hierarchy and assigning security level to this mechanism. Briefly we can say that our methodology and measurement have following benefit:

- This method of measurement can be customized and adapted by changing the assignment of mechanism to security level according to organization and customer threats, concern and priorities. Changing this assignment doesn't need re-evaluating the OS which is a heavy with huge cost.
- Precise and detailed distinction of mechanism like what we have done in this work results a real and precise measurement.
- By this detailed hierarchy of mechanism and parameters, re-evaluation or repeating of some part of evaluation can be readily done, even after some change in the OS.

Measurement and evaluation based on CC leads us to a heavy process and hard to repeat process. We used the Security Function of CC literature and dependability model [18] as input to make our hierarchy.

We can defend our scoring approach with following justifications:

- Covering all mechanism by investigating we'll-known Oss
- Using expert view in assignment of security level to mechanism.
- Selecting logical mechanism instead of concrete and OS-dependent mechanism.
- Our scoring approach is practical and operational because it uses the best practices in OS security literature.
- Fine granularity and the elaborated hierarchy of security mechanism, criteria and parameters, guarantees the sound and complete scoring approach. Because the evaluator can precisely and unambiguously determine if the OS has passes the mechanism or not.

Nevertheless our method has some problem, but it is a first step toward a method for OS security measurement. In the future work we are going to promote our method by:

- Formally considering the customer security requirement.
- Using the threat modeling method, e.g. STRIDE [6] model or attack tree in assigning the mechanism to security level, instead of using only the expert knowledge and opinion. In this regard we can use dependability model [3] for deriving security requirement from threat [1].
- Making the parameter hierarchy more formal according to overlap of mechanisms and their impact in promoting different security properties of system like confidentiality, integrity and availability.

References

1. CLASP, <http://www.list.org/~chandra/clasp/OWASP-CLASP.zip>
2. Spencer, R., Smalley, S., Loscocco, P., Hibler, M., Andersen, D., Lepreau, J.: The Flask Security Architecture: System Support for Diverse Security Policies. In: Proc. of the 8th Conference on USENIX Security Symposium, SSYM 1999 (2009)
3. Nicol, D.M., Sanders, W.H., Trivedi, K.S.: Model-Based Evaluation: From Dependability to Security. The IEEE Transactions on Dependable and Secure Computing 1(1), 48–65 (2004)
4. Jeffery, H.: Security Evaluation of the OpenBS Operating system, (2002), <http://www.eduunix.ccut.edu.cn/index/pdf/Security%20Evaluation%20of%20the%20OpenBSD%20Operating%20System.pdf> (last accessed: October 2011)
5. Information technology – Security techniques – Evaluation criteria for IT security – Part 1: Introduction and general model ISO/IEC 15408-1 (2009)
6. Hernan, S., Lambert, S., Ostwald, T., Shostack, A.: Threat Modeling: Uncover Security Design Flaws Using the STRIDE Approach (2006), [http://msdn2.microsoft.com/hin/magazine/cc163519\(en-us\).aspx](http://msdn2.microsoft.com/hin/magazine/cc163519(en-us).aspx) (last Accessed: October 2011)
7. Chew, E., Swanson, M., Stine, K., Bartol, N., Brown, A., Robinson, W.: Performance measurement guide for information security. NIST Special Publication 800-55, Revision 1, Information Security (2008)
8. Security-enhanced Linux (SELinux), <http://www.nsa.gov/selinux>
9. Shostack, A.: Experiences Threat Modeling at Microsoft. In: Modeling Security Workshop. University, UK (2008), <http://blogs.msdn.com/b/sdl/archive/2008/10/08/experiences-threat-modeling-at-microsoft.aspx> (last accessed: October 2011)
10. Fuller, J., Ha, J., O'Brien, D., Radvan, S., Christensen, E.: Fedora 11 Security Guide: A Guide to Securing Fedora Linux. Red Hat Inc. (2008)
11. US National Institute of Standards, “Recommended Security Controls for Federal Information Systems and Organization”, NIST Special Publication 800-53 Revision 3, Information Security (2009)
12. Common Criteria EAL4+ Evaluated Configuration Guide for Red Hat Enterprise Linux 5 on HP Hardware, v 2.3 (2007)
13. US National Institute of Standards, Common Criteria for IT Security Evaluation, ISO Standard 15408 (1999), <http://csrc.nist.gov/cc/>
14. Mourani, G.: Securing and Optimizing Linux: Red Hat Edition, Open Network Architecture and Open Docs Publishing, v 1.3 (2000)
15. Scambray, J., McClure, S.: Hacking Exposed: Windows Security Secrets and Solutions. McGraw-Hill Prof. Med./Tech. (2007)
16. Mokhov, S.A., Laverdière, M., Benredjem, D.: Taxonomy of linux kernel vulnerability solutions. In: Innovative Techniques in Instruction Technology, E-learning, Eassessment, and Education, Proceedings of CISSE/SCSS 2007, pp. 485–493 (2007)
17. Kong, J.: Designing BSD Rootkits: An Introduction to Kernel Hacking. No Starch Press Inc., San Francisco (2007)
18. Trivedi, K.S., Kim, D.S., Roy, A., Medhi, D.: Dependability and Security Models. In: Proc. DRCN 2009 Improving Dependability by Revisiting Operating System Design (2009)
19. Gligor, V.: Architectures for practical security. In: Proc. of the 15th ACM Symposium on Access Control Models and Technologies, SACMAT (2010)

20. Rushby, J.M.: Design and verification of secure systems. In: Proceedings of the Eighth ACM Symposium on Operating Systems Principles, December 14-16, pp. 12–21 (1981)
21. Klein, G., Elphinstone, K., Heiser, G., Andronick, J., Cock, D., Derrin, P., Elkaduwe, D., Engelhardt, K., Kolanski, R., Norrish, M., Sewell, T., Tuch, H., Winwood, S.: sel4: formal verification of an OS kernel. In: Proceedings of the ACM SIGOPS 22nd Symposium on Operating Systems Principles, October 11-14 (2009)
22. Rhee, J., Riley, R., Xu, D., Jiang, X.: Kernel Malware Analysis with Un-tampered and Temporal Views of Dynamic Kernel Memory. In: Jha, S., Sommer, R., Kreibich, C. (eds.) RAID 2010. LNCS, vol. 6307, pp. 178–197. Springer, Heidelberg (2010)
23. Seshadri, A., Luk, M., Qu, N., Perrig, A.: SecVisor: A tiny hypervisor to provide lifetime kernel code integrity for commodity OSes. In: Proc. of the 21st ACM Symposium on Operating Systems Principles, SOSP (October 2007)
24. Hughes, J.P., Feist, C.J.: Architecture of the Secure File System. Storage Technology Corporation (2001)
25. Song, J., Hu, G., Xu, Q.S.: Operating System Security and Host Vulnerability evaluation. In: Management and Service Science, MASS 2009 (2009)
26. Nguyen, A.M., Schear, N., Jung, H.D., Godiyal, A., King, S.T., Nguyen, H.D.: MAVMM: Lightweight and Purpose Built VMM for Malware Analysis. In: 2009 Annual Computer Security Applications Conference, pp. 441–450. IEEE (2009)
27. Weidner, K.: Common Criteria EAL4+ Evaluated Configuration Guide RedHat Enterprise Linux 5 on HP Hardware (2007)
28. National Security Agency, “Apple Mac OS X v10.3.x “Panther” Security Configuration Guide”, ver 1.1, SNAC (2004)
29. Mac OS X: System Hardening Guidelines for Faculty and Staff Desktops, <http://www.info.apple.com/kbnum>
30. COTS Compartmentalized Operations Protection Profile – Operating Systems (CCOPP-OS), ver. 2.0 (2008), <http://www.commoncriteriaportal.org>

Energy Management in Zone Routing Protocol (ZRP)

Dilli Ravilla¹ and Chandra Shekar Reddy Putta²

¹ Senior Assistant Professor, Department of Electronics and Communication Engineering,
Manipal Institute of Technology, Manipal

dilli.ravilla@gmail.com

² Professor Coordinator, Dept. of Electronics and Communication Engineering,
JNT University, Hyderabad, A.P, India
drpcreddy@gmail.com

Abstract. Ad hoc networks are wireless networks without a fixed infrastructure, and are usually established on a temporary basis for a specific application like emergency rescue or battle field communication. Energy management in wireless networks is the process of managing the sources and consumers of energy in a node or in the network as a whole for enhancing the lifetime of the network. Since, most of the mobile nodes in the network are equipped with low power batteries, it could be difficult for a mobile device to sustain for a long time if it send and receive data more often. To solve this problem here we describe the power management issues in mobile nodes using modified Zone Routing Protocol (ZRP) and it was simulated using NS2 simulator.

Index terms: Ad hoc Networks, Energy Management, Zone Routing Protocol (ZRP), NS2 simulator.

1 Introduction

A mobile ad hoc network (MANET) is comprised of mobile hosts that can communicate with each other using wireless links. In this environment a route between two hosts may consist of hops through one or more nodes in the MANET. An important problem in a mobile ad hoc network is finding and maintaining routes since host mobility can cause topology changes. [1] MANETs have been employed in scenarios where an infrastructure is unavailable, the cost to deploy a wired networking is not worth it, or there is no time to set up a fixed infrastructure. Some scenarios where an ad hoc network can be used are conferencing, emergency services, home networking, sensor dust, embedded computing Algorithms for a MANET must self-configure to adjust to environment and traffic where they run, and goal changes must be posed from the user and application. Ideally, a routing algorithm for an Ad hoc network should not only have the general characteristics of any routing protocol but also consider the specific characteristics of a mobile environment—in particular, bandwidth and energy limitations and mobility Routing algorithms and protocols need to save both bandwidth and energy and must take into account the low capacity and limited processing power of wireless devices.[2] [3].

Based on the routing information update mechanism, Ad hoc wireless network routing protocols are basically divided into pro-active routing and re-active protocols.

The Proactive routing algorithms aim to keep consistent and up-to-date routing information between every pair of nodes in the network by proactively propagating route updates at fixed time intervals. The pro-active routing protocol learns the network topology before a request comes in for forwarding. Since the proactive routing algorithms maintain routing tables for all nodes in the network, a route is found as soon as it is requested. The advantage of these protocols is low latency in discovering new routes and minimizes the end-to-end delay. Examples of proactive protocols are Destination-Sequenced Distance Vector (DSDV) [9], Optimized Link-State Routing (OLSR) [7], Cluster-Head Gateway Switch Routing Protocol (CGSR) [11], Wireless Routing Protocol(WRP)[11] and Topology-Based Reverse Path Forwarding (TBRPF) [8] Protocols.

Reactive or also called on-demand routing algorithms establish a route to a given destination only when a node requests it by initiating a route discovery process. Once a route has been established, the node keeps it until the destination is no longer accessible, or the route expires. The re-active routing protocol becomes active only when a node is willing to forward a request. Reactive protocols tend to be more efficient than proactive protocols in terms of control overhead and power consumption because routes are only created when required. Some of the re-active routing protocols are Dynamic Source Routing Protocol (DSR) [6], Ad Hoc On-Demand Distance-Vector Routing Protocol (AODV) [4] [5], Temporally Ordered Routing Algorithm (TORA) [10], Associativity-Based Routing (ABR) and Preferred Link-Based Routing Protocol (PLBR) [9] [10].

In spite of a reactive protocol gives the low overhead of control messages, it has higher latency in discovering routes as it determine the route using flooding route request packet in the network and builds the route on demand from the responses it receives. On the other hand, proactive protocols need periodic route updates to keep information updated and valid, also many available routes might never be needed all these increases the routing overhead and consume large amounts of bandwidth [3].

2 Zone Routing Protocol (ZRP)

Zone Routing Protocol (ZRP) [12] is a well-known hybrid routing protocol that is most suitable for large-scale networks. The ZRP framework is designed to provide a balance between the contrasting proactive and reactive routing approaches. Its name is derived from the use of “zones” that define the transmission radius for every participating node. ZRP uses a proactive mechanism of node discovery within a node’s immediate neighborhood, while interzone communication is carried out by using reactive approaches. ZRP utilizes the fact that node communication in ad hoc networks is mostly localized, thus the changes in the node topology within the vicinity of a node are of primary importance. ZRP makes use of this characteristic to define a framework for node communication with other existing protocols. Local neighborhoods, called *zones*, are defined for nodes. The routing zone of a given node is a subset of the network, within which all nodes are reachable within less than or equal to *zone radius* hops. The size of a zone is based on ρ factor, which is defined as the number of hops to the perimeter of the zone. There may be various overlapping zones, which helps in route optimization. [13]

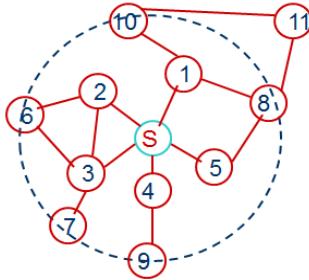


Fig. 1. A Routing Zone with Radius $\rho = 2$ hops

An example of a routing zone for node S of radius 2 is shown in figure 1[14]. The nodes from 1 to 10 belong to the routing zone of S, but not node 11. The nodes 6 to 10 are called peripheral nodes because hop distance from S is equal to radius of the routing zone. The information about neighbors is required to construct a routing zone of a given node. A neighbor is defined as a node with whom *direct* communication can be established. Neighbor discovery is accomplished by simple “Hello” packets (periodic transmission of beacon packets (active discovery) or with promiscuous snooping on the channel to detect the communication activity (passive discovery)) [15].IARP [16] is proactive approach and always maintains up-to-date routing tables. Route queries outside the zone are propagated by the route requests based on the perimeter of the zone (i.e., those with hop counts equal to ρ), instead of flooding the network. The Interzone Routing Protocol (IERP) [17] uses a reactive approach for communicating with nodes in different zones. Route queries are sent to peripheral nodes using the Bordercast Resolution Protocol (BRP) [18]. Since a node does not resend the query to the node in which it received the query originally, the control overhead is significantly reduced and redundant queries are also minimized.

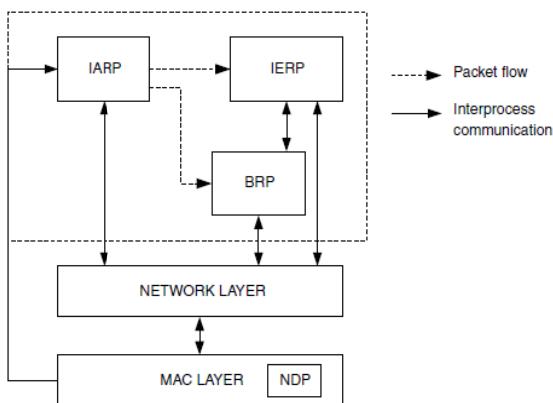


Fig. 2. Design Diagram of ZRP

ZRP provides a hybrid framework of protocols, which enables the use of any routing strategy according to various situations. It can be optimized to take full advantage of the strengths of any current protocols [12]. Neighbor discovery information is used as a basis for proactive monitoring of routing zones through the IntrAzone Routing Protocol (IARP) [16]. Since ZRP assumes that local neighbor discovery is implemented on the link-layer and is provided by the Neighbor Discovery Protocol (NDP) [15] [33], the first protocol to be part of ZRP is the IntrAzone Routing Protocol, or IARP [16]. Hence the larger the routing zone, the higher the update control traffic. The paths to the nodes which are outside the routing zone can be achieved by IERP [17].

If the destination belongs to its own zone, then it delivers the packet directly. Otherwise, source node bordercasts the *Route Request* to its peripheral nodes. If any peripheral node finds the destination node in its *routing zone*, then it sends a *Route Reply* back to source node indicating the path; otherwise, the node rebroadcasts the *Route Request* packet to the peripheral nodes and this procedure continues till the destination is identified [12].

3 Query Control Mechanisms

In ZRP, due to the large overlapping of node's routing zones there is higher control overhead. The main aim of Query control mechanisms is to avoid redundant or duplicate route request that are forwarded. ZRP has three schemes for query control. These note that redundant querying occurs when a route request packet arrives in a previously queried zone. In this section, we introduce a collection of query control mechanisms so called Query Detection (QD), Early Termination (ET) and Selective Bordercasting (SB) which meet the basic design objectives [19] [20].

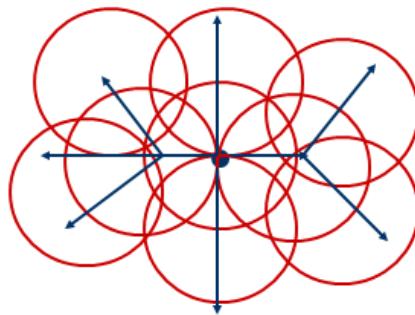


Fig. 3. Guiding the Search in InterZone Routing

When a node receives a route request message, it records the message in its list of route request messages that it has received. If this node receives the same route request message once again, then it does not forward that route request packet [20].

3.1 Query Detection (QD1/QD2)

Redundant querying occurs when a query message reappears in the routing zone of a node that has already broadcasted the query. Clearly, a broadcasting node is aware that

its own zone has been queried. If the query message were relayed from a bordercasting node to its peripheral nodes via IP, the query would travel through the routing zone, undetected by ZRP. Here, Bordercast Routing Protocol (BRP) [18] is performing query detection in two levels. The first level of query detection would allow nodes to detect queries as they relay them to the edge of the routing zone. (QD1). Thus, these nodes will maintain some info with regards to the query and discard duplicate queries if seen. The second level of query detection is called extended query detection (QD2) which detect overheard queries as they are propagated (e.g. node 5 in figure 4). Node make note of overheard queries and thus, discard duplicate queries if they are received. QD2 can be implemented using IP and MAC layer broadcasts. Other query control mechanisms may require QD to record additional information contained in the route query packet. Of particular importance is the ID of the node that most recently bordercast the query. As we will see in the next section, this information provides valuable insight into the local coverage of the query, which can be used to terminate or prevent redundant queries [20].

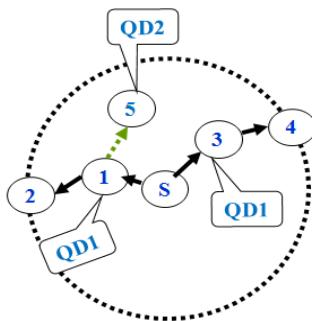


Fig. 4. Advanced Query Detection (QD1 and QD2)

3.2 Early Termination (ET)

As per IEEE 802.11 standards, a node can overhear passing traffic when it is operating in promiscuous mode. If a given node is already covered by the query packet, the protocol drops the query packets which come again using Early Termination.

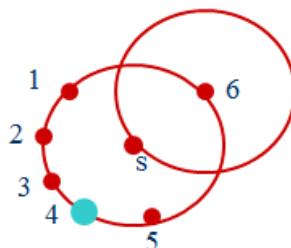


Fig. 5. Early Termination of Unnecessary RREQs

In the Fig. 5, a node ‘s’ has a list of nodes 1, 2, 3, 4, 5 such that the RREQ message has already arrived in the routing zones of the nodes 1, 2, 3, 4, 5. Now ‘s’ receives a request to forward a RREQ message from another node 6. This may happen when ‘s’ is a peripheral node for the routing zone of node 6. ‘s’ receives a RREQ from node 6 since ‘s’ is a peripheral node for the routing zone of node 6. ‘s’ does not broadcast the RREQ to 1, 2, 3, 4, 5 but only to 4 which is not in its list. Through advanced query detection and knowledge of the local topology, each node is able to identify surrounding regions that have already been covered by the query. Nodes can steer queries away from those areas by early termination of stray messages, encouraging the search to proceed outward. In some cases, delaying the early termination processing for a random period of time provides a valuable opportunity to detect recent additions in query coverage.

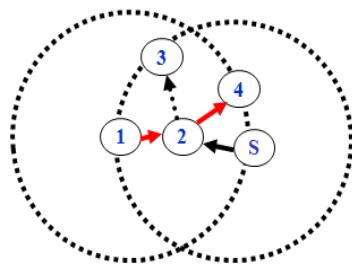


Fig. 6. Early Termination (ET)

In the Fig. 6, Node 2 has seen a broadcast packet from 1(Sent to 4). Now, later on, it gets a packet from S to be broadcast to node 3. Node 2 would note that node 3 belongs to the previously queried zone (of node 1) and will withhold transmission. It would need to know that node 3 was in node 1’s broadcast tree. The absence of hierarchies eliminates definitive points of congestion. A node will not relay a query packet to a broadcast recipient either if that recipient lies inside the routing zone of a previously broadcast node or if it has already relayed the query to a recipient. This scheme is called Early Termination. To identify a node that lies inside the routing zone of a previously broadcast recipient, an extended routing zone has to be maintained.

When a node broadcasts a query, all nodes within its routing zone are effectively covered by the query. Any further query messages directed into this region are redundant and represent a potential inefficiency of broadcasting. In general, it is not possible to guide the query perfectly outward into uncovered regions of the network. Fortunately, information obtained through advanced query detection (QD1/QD2), combined with knowledge of the local topology, can support Early Termination (ET) of many query messages that otherwise would stray inward. When a node relays a query along a broadcast tree, it can safely prune any downstream branches leading to peripheral nodes *inside* covered regions of the network. The relaying node can use the known topology of its extended routing zone (or standard routing zone plus cached broadcast trees, in the case of root directed broadcast) *interior* routing zone members of each previously broadcast node in the Detected Queries Table. Relaying the same query message to a peripheral node for a second time would not add to the overall query coverage.

3.3 Random Query Processing Delay (RQPD)

When a node initiates a bordercast to its peripheral nodes, the node's routing zone is instantly covered by the query. However, it takes some finite amount of time for the query to make its way along the bordercast tree, and be detected through the QD mechanisms. The routing zone may vulnerable to query overlap from the nearby bordercasts during the bordercast propagation. Although this bordercast propagation of vulnerability is not very large, it can be a real problem when nearby nodes initiate bordercasts at roughly the same time. In single-channel networks the above problem is common when neighboring peripheral nodes receive a query message and simultaneously re-bordercast the message farther out into the network [20]. This problem of "simultaneous" bordercasts can be addressed by spreading out the bordercasts with a Random Query Processing Delay (RQPD). Specifically, each bordercasting node schedules a random delay prior to bordercast tree construction and ET. During this time, the waiting node benefits from the opportunity to detect the added query coverage from earlier bordercasting nodes. This, in turn, promotes a more thorough pruning of the bordercast tree (through ET) when it is time for the waiting node to bordercast. Increasing the average RQPD can significantly improve performance, up to a point. Once the bordercast times are sufficiently spread out, further increases in delay have a negligible impact on query efficiency [20].

4 Issues in ZRP

Here we address two major issues that need to be considered and they are outlined below

4.1 Power Management

In ZRP, the packets are forwarded with full power without considering the node's position inside the zone. According to Inverse Square Law, the power received by the receiving node is inversely proportional to square of the distance between the nodes (i.e)

$$\gamma = P_t / 4\pi r^2$$

The node could waste power if the distance between the sender and the receiver node is less.

4.2 Bandwidth Utilization

As the distance between the sender and border nodes increases, the zone area will also increase, which means the radio coverage of the sender node will not be able to reach the border nodes in the zone. Due to that reason, the sender node will increase the number of broadcasts to find the border nodes in the zone, which will obviously increase the bandwidth utilization.

5 Modified ZRP

5.1 Power Factor in ZRP

Whenever the node forwards a packet to the intermediate or border node in the zone it uses the maximum power to reach the destination [5]. By following this approach the node will lose its full power in a very short period of time. To avoid this problem, the ZRP protocol is modified to create zones with respect to two power levels, for example 20mW and 50mW. The reason for creating a zone with two power level is that, if a node is elected as a border node or as a intermediate node and if the node is moving at a particular speed of 2 m/sec. Then the corresponding nodes should be in that respective state (intermediate or border node) for a particular period of time to avoid the rapid fluctuation from border node to intermediate node or vice-versa. So by doing this, the node can avoid generating unnecessary routing updates or change its state more frequently.

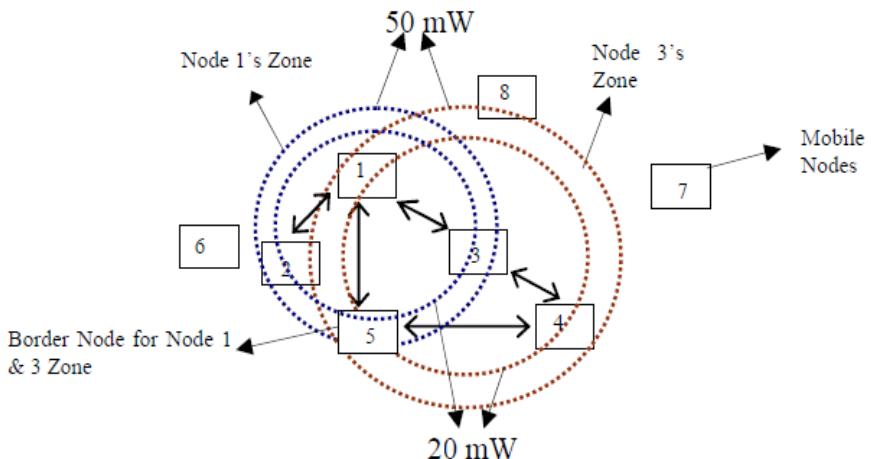


Fig. 7. Routing Zones of Node 1 & 3, Threshold powers are 20 & 50mW

From the above figure it can be seen that every node creates their own routing zones and initially when the node switches ON, it creates the zone with 20mW and 50mW, since that is the threshold power level set initially by the protocol. But if a node is unable to find a border node since the node's threshold power level is high (20 & 50mW), then the corresponding node will start reducing its threshold power level until it's able to find the border node. The reason for creating a dynamically changing zone is that, if a node has no border nodes elected but full of intermediate nodes elected then the intermediate nodes inside the zone will not be able to talk with its neighboring zone nodes. Because according to this protocol one zone can communicate to another zone through the border nodes only [1]. If we consider the above diagram, if node 1 wants to talk with node 4 then node 1 should pass through one of its border nodes to reach the neighboring zone, they are nodes 2, 5 or 3.

To calculate the power consumption, consider node 1 wants to forward a packet to destination node 8. The source node sends a broadcast with 50mW to all its border nodes (i.e) nodes 2, 5 and 3. Then the corresponding nodes check their own routing table and in that node 3 can reach node 8 since it is the border node of node 3's zone. After seeing that, node 3 sends a unicast packet to destination node 8 with 50mW. Therefore, the source node found the destination node by shedding only 50mW in the modified ZRP protocol. But in the actual ZRP protocol the node would have spent 100mW to reach the destination since all the nodes form zone with respect to hop count and it always forwards the packet with full power level (100 mW) [1]. So as the number of broadcasts increase, the power usage will also increase according to the formula

$$P=C*N,$$

Where C=Transmit power and N=Number of Broadcasts. [2]

6 Simulation Results and Analysis

The ZRP was modified to test the power utilization of the node and it was simulated in NS2 simulator. The simulation parameters are shown in the table below.

Table 1. PE ZRP Simulation Parameters

Channel Bandwidth	2 Mbps
Power Levels Used	20,30, 50 and 100 Mw
Transmission Range	Obstruction: 300 to 600 ft Without Obstruction: around 32ft
Packet Size and Rate	150 Bytes, 5 packets/sec

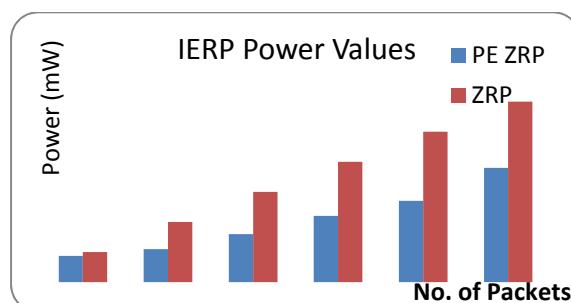


Fig. 8. IERP Power Consumption

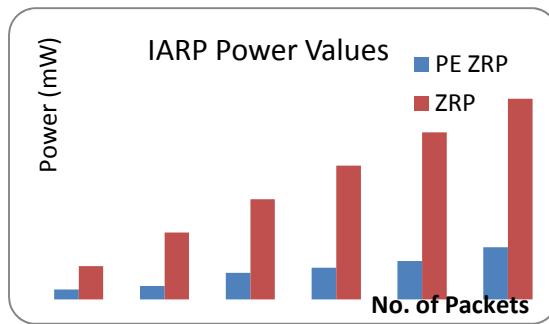


Fig. 9. IARP Power Consumption

From the above simulation results, we can observe that ZRP protocol consumes more power compared to the Power Efficient (PE) ZRP protocol, which is the modified version of ZRP. Since ZRP protocol forwards packet with 100mW constantly, the node wastes more power as the number of packets increases. But in the case of PA ZRP, the power consumption was less; because packets were send with 20 or 50 mW power levels. Since the simulation is in initial stages, the protocol was tested only for power consumption and with less number of packets.

7 Conclusions

In this paper a new method was proposed to reduce the power consumption using ZRP protocol and since the design and simulation is in initial stages, the protocol was tested with less number of packets and only for packet level power consumption. The future scope of this protocol is to successfully simulate it for a very large network and implement it with a voice application to study the performance and efficiency of the PE ZRP protocol.

References

- [1] Boukerche, A.: Performance Evaluation of Routing Protocols for Ad Hoc Wireless Networks. *Mobile Networks and Applications* 9, 333–342 (2004); ONT K1N- 6N5, Canada
- [2] Geetha, V., Aithal, S., Chandrasekaran, K.: Effect of Mobility Over Performance of the Ad Hoc Networks. IEEE Conference, December 20-23, pp. 138–141 (2006)
- [3] Toh: *Ad Hoc Mobile Wireless Networks: Protocols And Systems*. Pearson Education, India (2007)
- [4] Perkins, C., Belding-Royer, E., Das, S.: Ad hoc On-Demand Distance Vector Routing Protocol (February 2007), <http://www.ietf.org/rfc/rfc3561.txt>
- [5] Luo, Y.-H., Wang, J.-X., Chen, S.-Q.: An Energy-efficient AODV routing protocol based on link stability. *Journal of Circuits and Systems* 13(6), 141–147 (2008)
- [6] Johnson, D., Hu, Y., Maltz, D.: The Dynamic Source Routing Protocol (DSR) for Mobile Ad Hoc Networks for IPv4 (February 2007), <http://www.ietf.org/rfc/rfc4728.txt>

- [7] Clausen, T., Jacquet, P.: Optimized Link State Routing Protocol, IETF Internet Draft (January 2003),
<http://menetou.inria.fr/draft-ietf-manet-olsr-11.txt>
- [8] Ogier, R., Templin, F., Lewis, M.: Topology Dissemination Based on Reverse-Path Forwarding (TBRPF) (February 2004),
<http://www.ietf.org/rfc/rfc3684.txt>
- [9] Mahdipour, E., Rahmani, A.M., Aminian, E.: Performance Evaluation of Destination-Sequenced Distance-Vector (DSDV) Routing Protocol. In: IEEE International Conference on Future Networks, pp. 186–190 (March 2009)
- [10] Kuppusamy, P., Thirunavukkarasu, K., Kalavathi, B.: A study and comparison of OLSR, AODV and TORA routing protocols in ad hoc networks. In: 2011 3rd IEEE International Conference on Electronics Computer Technology (ICECT), pp. 143–147 (2011) (issue Date: April 8-10, 2011)
- [11] Arora, V., Rama Krishna, C.: Performance evaluation of routing protocols for MANETs under different traffic conditions. In: 2010 2nd IEEE International Conference on Computer Engineering and Technology (ICCET), vol. 6, pp. V6-79–V6-84 (April 2010)
- [12] Haas, Z.J., Pearlman, M.R., Samar, P.: IETF, The Zone Routing Protocol (ZRP) for Ad Hoc Internet Draft, draft-ietf-manet-zone-zrp-04.txt (July 2002)
- [13] Thipchaksurat, S., Kirdpipat, P.: Position-based Routing Protocol by Reducing Routing Overhead with Adaptive Request Zone for Mobile Ad Hoc Networks. In: Communication Systems Wireless Mobile Communications & Technologies Paper 10 1420 Thailand 10520
- [14] Prasun, S., Srikanth, K., Son, D.: Scalable Unidirectional Routing with Zone Routing Protocol (ZRP) Extensions for Mobile Ad Hoc Networks. IEEE Transactions on Networking 4 (2006)
- [15] Lee, J.-C., Han, Y.-H., Shin, M.-K., Jang, H.-J., Kim, H.-J.: Considerations of Neighbor Discovery Protocol (NDP) over IEEE 802.16 Networks. In: The 8th IEEE International Conference on Advanced Communication Technology (ICACT), vol. 2, pp. 951–955 (2006)
- [16] Haas, Z.J., Pearlman, M.R., Samar, P.: Intrazone Routing Protocol (IARP), IETF Internet Draft, draft-ietf-manet-iarp-01.txt (June 2001)
- [17] Haas, Z.J., Pearlman, M.R., Samar, P.: Interzone Routing Protocol (IERP), IETF Internet Draft, draft-ietf-manet-ierp-01.txt (June 2001)
- [18] Haas, Z.J., Pearlman, M.R., Samar, P.: The Bordercast Resolution Protocol (BRP) for AdHoc Networks, IETF Internet Draft, draft-ietf-manet-brp-1.txt (June 2001)
- [19] Buhari, A., Othman, M.: Efficient Query Propagation by Adaptive Bordercast Operation in Dense Ad Hoc Network. IJCSNS International Journal of Computer Science and Network Security 7(8), 101–108 (2007)
- [20] Haas, Z.J., Pearlman, M.R.: The Performance of Query Control Schemes for the Zone Routing Protocol. IEEE/ACM Transactions on Networking 9 (August 2001)
- [21] Xiao, B.-L., Guo, W., Liu, J.: Pseudo gossip routing algorithm based link stability in mobile ad hoc networks. Journal on Communication 29(6), 26–33 (2008)
- [22] Tang, C.-G., Chen, S.-Q., Gong, X.-X.: Steady Routing Protocol with Prediction in Mobile Ad Hoc Networks. Journal of Chinese Computer Systems 28(1), 9–14 (2007)
- [23] Zhang, H., Dong, Y.-N.: Link stability metric based on mobility prediction model in mobile ad hoc networks. Journal of Communication 28(11), 30–37 (2007)
- [24] Xie, Z.-P., Zhang, Q.: Truthful Mechanisms for Maximum Lifetime Routing in Wireless Ad Hoc Networks. Journal of Software 20(9), 2542–2557 (2009)

- [25] Samar, P., Pearlman, M.R., Haas, Z.J.: Independent Zone Routing: An Adaptive Hybrid Routing Framework for Ad Hoc Wireless Networks. *IEEE/ACM Transactions on Networking* 12(4) (August 2004)
- [26] Zhou, J., Cheng, Y., Lu, J.: Velocity based Adaptive Zone Routing Protocol. In: *Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems*, Xiamen, China, November 28-December 1 (2007)
- [27] Giannoulis, S., Katsanos, C., Koubias, S., Papadopoulos, G.: A hybrid adaptive routing protocol for ad hoc wireless networks. In: *Proceedings of 2004 IEEE International Workshop on Factory Communication Systems*, pp. 287–290 (2004)
- [28] Yang, C., Tseng, L.: Fisheye zone routing protocol for mobile ad hoc networks. In: *Second IEEE Consumer Communications and Networking Conference, CCNC 2005*, pp. 1–6 (January 2005)
- [29] Jaiswal, A.K., Singh, P.: New Scheme of Adaptive Zone Routing Protocol. *International Journal of Computer Science & Communication* 1(2), 207–210 (2010)
- [30] Jaiswal, A.K., Singh, P.: Optimizing Velocity Based Adaptive Zone Routing Protocol. In: *IEEE International “Conference on Computer and Communication Technology (ICCCT)”, September 17-19* (2010)
- [31] Shih, T.-F., Yen, H.-C.: Location-Aware Routing Protocol with dynamic adaptation of request zone for mobile ad hoc networks. *Wireless Networks* 14, 321–333 (2008) (published online: October 9, 2006)

8 Biographies



Dilli Ravilla received the B.Tech. degree in Electronics and Communication Engineering from Jawaharlal Nehru Technological University(JNTU), Hyderabad, India, in 2003 and the M.E degree in Electronics and Communication Engineering from Satyabama University, Chennai, India, in 2006. He is working toward the Ph.D. degree in the Electronics and Communication Engineering at Jawaharlal Nehru Technological University, Hyderabad, India. His research interests include ad hoc network routing. His research has focused on the design of hybrid routing protocols and its effects on performance optimization in ad hoc networks.



Dr Chandra Shekar Reddy Putta received the B.Tech. degree in Electronics and Communications Engineering from JNTUH, Hyderabad, India and M.E from Bharatiyar Deemed University. He received M.Tech and Ph.D from JNT University. Hyderabad, India. He joined as faculty in JNTU, Currently he is working as Professor Coordinator in JNTUH, Hyderabad, India .He is an author of numerous technical papers in the fields of high-speed networking and wireless networks. His research interests include mobile and wireless communication and networks, personal communication service, and high-speed communication and protocols.

A New Approach for Vertical Handoff in Wireless 4G Network

Vijay Malviya, Praneet Saurabh, and Bhupendra Verma

Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India
vijaymalviya@gmail.com, praneetsaurabh@gmail.com,
bk_verma3@rediffmail.com

Abstract. Future generation 4G wireless network is designed for flawless and continues connections between devices of several independent wireless networks like WLAN, GPRS and so on. Heterogeneous wireless networks will be dominant in the next-generation wireless networks with the integration of various wireless access networks. One of the most challenging problems for coordination is vertical handoff, which is the decision for a mobile node to handoff between different types of networks in heterogeneous network. Handoff is based on received signal strength comparisons; in this paper we develop a Vertical Handoff strategy for 4G network based on bandwidth and Power (Signal Strength) between different networks .we compare the results with handoff mechanism depending upon only signal strength. Results show a significant performance improvement of the proposed system over the signal strength driven handoff or connectivity driven handoff.

Keywords: 4G, QOS, Vertical Handoff, Bandwidth in 4G network, Signal Strength in 4G Network, Client driven Handoff.

1 Introduction

Connectivity among multiple Wireless networks is becoming an increasingly important and popular way for providing seamless connectivity to the mobile users [1]. Current technologies vary widely in their bandwidths, latencies, frequencies, and media access methods [2][3]. In simple words there are several wireless networks in place and in use with different architecture, MAC protocol, services, Bandwidth, Cost and accessibility. Unfortunately, no technology in and of itself makes possible the best available network at all times. For example a WLAN may be better at internet connectivity over GSM cellular network [4], but when it comes to voice calls, the second clearly outperforms the first. No single network technology simultaneously provides a low-latency, high-bandwidth, wide-area connection for all the services over the entire access time to a large number of users [5]. Therefore 4G networks are built around the idea of making the best network available to an end user. In conventional 2G and 3G horizontal handoff system, a handoff or passing the services from one peer to another peer was designed merely to support the mobility [6] and the handoff is generally a designed phenomenon between two base stations such that when a mobile device leaves a cell, the current base station handoffs the connection to the base station of the cell that the mobile device is currently on [6].

But in a Heterogeneous Network like 4G, Several connections may be available to a node at the same time for a particular service. Therefore protocols are designed to support the Vertical Handover which can be initiated by the end user or the network node [5][8].

Few studies focus on selecting or opting for a handover depending upon the access cost. For example for internet connections, a Cellular Gateway and a WLAN gateway are available to the end node. When the end node needs high speed connections, it can use WLAN which will cost more and when he merely requires a low speed low cost connection, he can switch back to cellular gateway [21].

With increased competition among the wireless service providers and the infrastructure cost coming down, cost becomes a secondary metric to the quality in connections [7].

Therefore in this work we focus on designing a system for measuring and quantifying the Quality of Service [8] at the end node and selecting an appropriate network based on the best QOS available from a network[11].

We estimate the quality of service by calculating the bandwidth between links from available gateways to the end node and by measuring the received signal power.

Even though Handoff is easy for theoretical claims, it passes on several challenges which include detection of handoff need and initiating the handoff process

2 Related Work

Pedro et al. proposes[1] two different procedures in the handover preparation phase in IEEE 802.11, candidate access technologies discover and resources availability check, can be performed only in one procedure, optimizing the handover process. Inwhee Joe et al. [2] propose a mobility-based prediction algorithm with dynamic LGD (Link Going Down) triggering for vertical handover by applying the IS (information server) of IEEE 802.21 MIH (media independent handover). The proposed algorithm predicts a possible moving area (PMA) of the mobile terminal based on mobility information (the velocity, coordinate values, position, movement detection, etc) in IS. Since the PMA indicates a next target cell for handover, it can advance the LGD trigger point dynamically to prepare for handover beforehand. Khan [3] proposes an approach build on IEEE 802.21 standard for service negotiation. SIP and IPv6 based flow management approaches are discussed, the later approach is implemented using OPNET modeller simulator. The performance of our approach is compared with Long-term contractual approach in terms of user's throughput, users' cost, operators' revenue and call blocking probability. Buiati et. Al. [7] proposed a new Media Independent Handover (MIH) standard using the IEEE 802.21; this letter proposes a new neighbor network discovery mechanism, considering a hierarchical view of the network information. Atanasovski [8] discusses novel methods for resource management in heterogeneous network. They proposed a technique for media handover and management especially in wireless management. Guang Lu [9] presents a solution using IEEE 802.21 to enable seamless mobility for data and video streaming sessions. The solution was implemented and evaluated using commercial wireless networks and mobile devices. Lab and field trial results show minimal handover delay and improved

user experience. V. Kumar [10] presents a novel vertical handover scheme applicable for IEEE 802.11 (WLAN) and Universal Mobile Telecommunication System (UMTS) based on IEEE 802.21. UMTS is a 3GPP (Third Generation Partnership Project) technology, which is being used in cellular networks. WLAN is a LAN technology which supports smaller coverage area than the cellular networks. Vertical handover between these two is needed because they have different RAT. Jiann-Liang Chen [11] developed a novel IEEE 802.21 MIH (Media Independent Handover) mechanism for next generation vehicular multimedia network. which discussed adaptive QoS management mechanism. By obtaining received signal strength parameters, the proposed MIH framework can determine the best available network. The adaptive QoS mechanism substantially improves the performance of real-time multimedia applications. Simulation results show that average handover time is slower than both UMTS and WiMAX when the MIH mechanism is used in vehicular network Jun Yuan et al. [12] proposed a novel scheme using IEEE 802.21 MIH services to improve packet loss performance by utilizing the active links to maintain the data flow. The MIH services, Link_Action and MIH_Link_Action, are extended and an MIH event named Link_PDU_Receive_Status is added. A complete message exchange in handover procedure is provided. Numerical analysis shows that the proposed scheme performs better in terms of packet loss comparing with the traditional independent FMIPv6 scheme. Obreja [13] presents a solution for mobility management in heterogeneous networks which is based on the MIH framework. It is presented the architecture and the simulation testbed used for system validation. QualNet simulator was chose to implement the proposed solution. Vaca Ramírez , Ramos[14] present extensive numerical simulations of a novel vertical handover mechanism, which is compared with well known mechanisms: one based on the RSS (received signal strength) and the other one based on the traditional AHP (Analytic Hierarchical Process) method. The proposed scheme considers that a mobile node carries more than one traffic class; it also integrates an MIH (Media Independent Handover) QoS Model, the AHP method, and cost functions to a fuzzy MADM (Multiple Attribute Decision Making) handover decision algorithm. We show that our proposed scheme overperforms the other two schemes.

3 Problem Formulation

Handoff can be defined as “transferring an ongoing session from one network to another network in geographically heterogeneous wireless network as the user is in motion”. Handoff process can be seen as having two stages: (a) Handoff detection, and (b) Handoff execution. Handoff detection includes network discovery and handoff decision [12]. Which kind of handoff metrics should be used and how to apply them to make the handoff decision are the main problems in handoff detection [13]. In handoff execution, the mobility management plays an important role. To achieve seamless and fast handoff, these two stages should be paid attention. The time when the handoff decision is made can affect the overall performance of the handoff process such as packet loss [14]. A heavy signaling overhead in handoff management leads to large handoff latency [15].

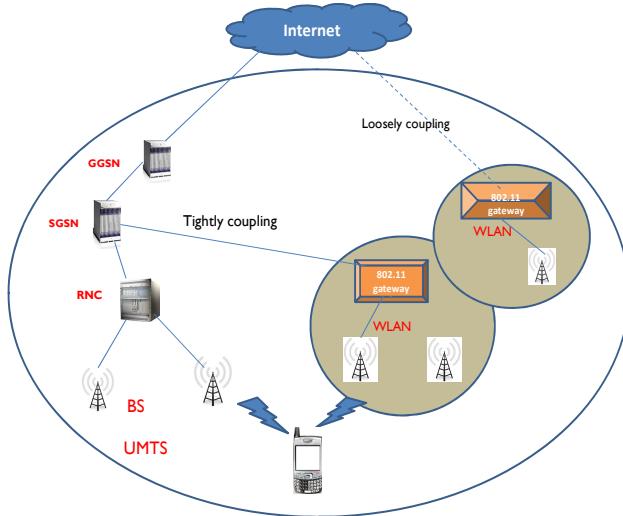


Fig. 1. Integrated architecture of UMTS and WLAN

UMTS and WLAN can be integrated in two different ways such as tight coupling and loosely coupling depending on architecture loose coupling is the type of coupling the UMTS and WLAN network[9] uses the same database for authorization, accounting, and authentication (Refer to fig. 1). Rapid introduction of the nodes from one network to other is possible .no impact will be on the internal nodes. Disadvantage of scheme is poor handoff. In Tight coupling UMTS and WLAN network [17] are connected to each other by special interface .Rapid introduction of nodes from one network to other is possible. No impact will be on internal node .advantage is this scheme is good handoff and this type of interconnection is possible only when the operator has control over the two network. So we are considering tightly coupled architecture [18].

The Wireless LAN Access Point is an integral part of the UMTS network, directly connected to the *Serving GPRS Support Node* (SGSN) and thus represents an alternative radio access network to the existing cellular one[12]. The *Mobile Equipment* (ME) itself is equipped with two interfaces, a Wireless LAN interface and a UMTS interface, which are connected to each other. Whenever the ME moves out of the coverage area of a Wireless LAN cell, it indicates measurement reports to the SGSN, the vertical handover is initiated, The Radio Network Controller (RNC) is the governing element in the UMTS radio access network (UTRAN) [20][21] responsible for control of the Node Base Stations (BS), that is to say, the base stations which are connected to the controller. The RNC carries out radio resource management. The *Serving GPRS Support Node* (SGSN) is a main component of the GPRS network [22], which handles all packet switched data within the network, e.g. the mobility management and authentication of the users. The *Gateway GPRS* [23] *Support Node* (GGSN) is a main component of the GPRS network. The GGSN is responsible for the interworking between the GPRS network and external packet switched networks like the Internet and X.25 networks [19].

4 Proposed System

In proposed architecture, we will use a dynamic database at the network which will contain the prevailing conditions of the network such as available bandwidth, network load etc.

4.1 Architecture

In UMTS network database is connected to 3GPP server and When MN attached to AP at WLAN needs to switch to UMTS, it sends the request to WAG, it forwards the request to SGSN of UMTS were mobile node need to switched, SGSN checks all the RNC connected to it gets the network conditions of all RNC connected from 3GPP server were database is connected and sends it back to WAG and finally to MN through AP. MN gets the all condition and finally execute handoff. When mobile node attached RNC of UMTS network need to switch to WLAN it send request to SGSN and then to WAG and collect network condition from server were database is connected and sends back the information to MN through RNC and finally handoff initiated at MN.

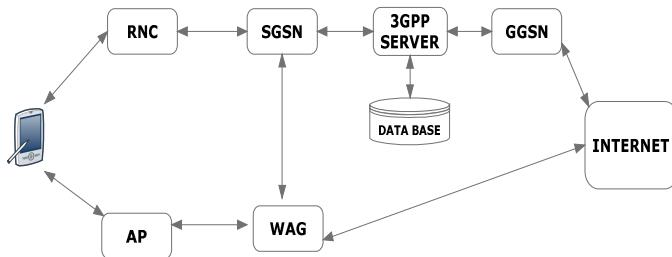


Fig. 2. Proposed architecture

4.2 Link Quality Assessment

SNR is a good measure of the link quality. High SNR ensures higher packet delivery ratio. In a signal level measurement, calculation of SNR [35] is typically complex and needs model like kalman filter and so on. It is also calculated from the power spectrum of the received signal. But in this work we propose a packet level estimation of SNR which can be directly derived from the bit error rate.

SNR or signal to noise ratio is typically given by

$$\left[\frac{\mathcal{E}_b}{\mathcal{N}_{0_{\text{eff}}} } \right]_{ij} = \frac{P_{Rj}/D}{\mathcal{N}_0 + P_{Ij}/W}, \quad (1)$$

where D is the data rate in bits per second, W is the system bandwidth in hertz, \mathcal{N}_0 is the power spectral density of the thermal noise, P_{Ij} is the power of the interference at

node j due to all nodes excluding node i , and PR_{ij} is the received power at node j due to node i .

Table 1. Relationship between Signal to Noise Ratio and BER

SNR(dB)	BPSK&QPSK	CCK5.5(5.5Mbps)	CCK11(11Mbps)
4	9e-2	1.7e-2	4e-2
5	9e-2	3.5e-3	8e-3
6	3e-2	6.2e-4	1.2e-3
7	4.1e-3	7e-5	1.2e-4
8	1.01e-3	5e-6	1.01e-5
9	2e-4	2e-7	3.9e-7
10	2.02e-5	4e-9	7e-9
11	1.8e-6	7e-11	2e-10
12	5.9e-8	7e-11	2e-10
13	1.4e-9	7e-11	2e-10

Where simulation results discussed in [16], that the results obtained are very closer to the theoretical values. Erroneous bits are calculated [34] using parity check of the received packet. Now a node stores the number of packets it has received and the current measure of bit error rate. As the simulation is using 20 packets to 50 packets per second as the data transmission rate, maximum error rate that can be observed per second is e^{-7} . Thus any value over it may be considered as infinitely low BER and infinitely high SNR. Further an argument can be easily put forward about selecting SNR as the performance measurement criteria instead of BER where it is derived directly from the table. Error rate are more instantaneous values.

Though the table represents the Signal to Noise ratio for 802.11 channels[34], we assume that the trend of the ratio is universal. Exact value of SNR may vary from network to network but Link Quality measurement is similar in terms of BER and SNR.

It is also understandable from the table that Bit error rate can not be measured for small data or burst of small data. In a 4G network if an end user is using certain connection, then it can measure the BER and hence SNR in that connection. But the proposed technique needs the SNR of the other available connections to also be known.

Therefore we propose a proactive technique for SNR calculation which is even used for measuring the bandwidth. We assume that the neighboring node broadcast a periodic advertisement signal called HELLO packets with its ID. When a gateway receives Hello packet from a node, it calculates the SNR and Bandwidth available between it and the end node (as explained in the next section) and notifies to SGSN which passes the value to 3GPP server which in turn save the value in the database. every single estimation at wireless gateway updates the value in the database.

4.2.2 Bandwidth Estimation

Bandwidth is calculated as inverse function of latency between two nodes and is essentially the link bandwidth here. Every Hello packet is time stamped at the sender

side. When a node receives the hello packet, it calculates the total delay from the sender by subtracting the sending time from current time [33]. Let the delay be T.

Then bandwidth between two nodes is given by

$$BW = K * \text{Packet_Size} / T \quad (2)$$

Where K is the proportionality constant and are consistent for a specific network. Adaptation of this bandwidth function in a heterogeneous network can be well argued as different links belongs to different network. Not let us assume that in any arbitrary link an ideal bandwidth is 200Kbps. then if 100 bits are received in 10 milliseconds, effective transmission rate or bandwidth is .1kbps. As the Hello packets are short, actual link speed cannot be measured through this technique. Now assume that there is congestion in the network. Then the same packet may take 50 milliseconds to arrive which results in calculation of effective bandwidth being .02kbps. Instead of Hello packet, if a data frame were transmitted then also delays would have been the same as in a given slot a node may transmit minimum one frame of payload. As the measurement occurs at the gateway, the gateway replaces K with the frame size of the network it represents and measures the bandwidth.

Thus equation 2 can be simplified as (3) which is also the mean link bandwidth measured over a period N.

$$\text{Link Bandwidth} = 1/N \sum_{n=1}^N (\text{FrameSize} / T(n)) \quad (3)$$

4.3 Cost Estimation

Whenever multiple metrics are used in any network, it is essential to device a relationship between them such a single cost can represent the entire link quality [29]. In any multi-metric decision, cost can be calculated as

$$\text{Cost} = \alpha * (\sum (1/(1+x)) + (1-\alpha)(\sum(1-1/(1+y)))) \quad (4)$$

Where x and ye are the metrics. α is network dependent and here it is assumed as .5. Hence (4) can be simplified as (5) by considering our metrics.

$$\text{Cost} = .5/(1+\text{SNR}) + .5(1 - 1/(1+\text{BW})) \quad (5)$$

5 Methodology

1. Measure Cost at NIC and at Wireless Gateway, when Cost falls below threshold then generate a signal to request SGSN for all available connections.
2. SGSN checks if any other connection or network has better QOS than the current one. If available, it notifies the mobile node with available QOS and networks
3. Node selects the best Cost Network and Request for a handoff to its connected gateway.

4. The gateway passes on the request to the requested Network's Gateway.
5. Upon Receiving the Handoff acknowledgement, the connection is passed to the other gateway.

6 Simulation and Results

Proposed system is simulated with Omnet3.3, an even based simulator. Call drop probability and Call block probabilities are calculated as given in [17]. The results are verified for proposed system and conventional system that performs handoff based on mobility (i.e. handoff is performed only when a mobile moves away from its current gateway).

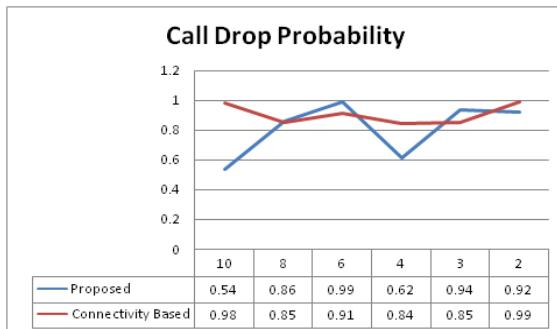


Fig. 3. Number of calls per second v/s Call Drop Probability

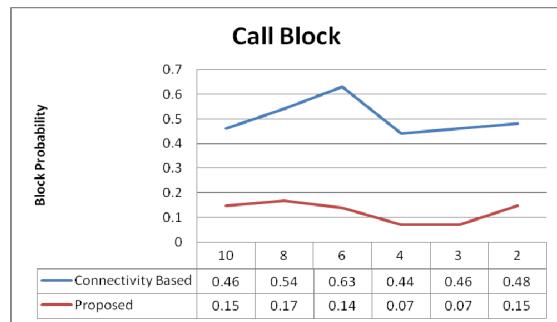


Fig. 4. Number of Call Block Probability

7 Conclusion and Future work

Assuring quality of service to the links and especially to the end nodes in heterogeneous 4G network is a major challenge due to different signaling, inherent quality and services provided by different interconnected network. Besides being

difficult to quantify and provide an optimized framework for handoff, it is quite difficult to propose and formulate a suitable structure for the same. Conventional mobility and connectivity based approach for handoff may not be suitable as seen from results (figure 3, 4). It is clear from the result that with increase in network load (i.e. number of calls per second) the drop and block probability increases in the conventional technique. In case of proposed solution, the need of the handoff is determined way before the quality of a link enforces a handoff. It is also proved through the result which clearly shows the improved performance, especially in block probability. This technique can be further modified by incorporating other network resources as QOS parameters like Jitter, Number of Flows in the nodes and so on.

References

1. Neves, P., Soares, J., Sargent, S.: Dynamic media independent information server. In: IEEE Symposium on Computers and Communications (ISCC), pp. 865–872 (2010)
2. Joe, I., Shin, M.: A Mobility-Based Prediction Algorithm with Dynamic LGD Triggering for Vertical Handover. In: IEEE 7th Consumer Communications and Networking Conference, CCNC (2010)
3. Khan, M.A., Toseef, U., Marx, S., Goerg, C.: Auction based interface selection with Media Independent Handover services and flow management. In: 2010 European Wireless Conference (EW), pp. 429–436. IEEE (2010)
4. Ferrus, R., Sallent, O., Agusti, R.: Interworking in heterogeneous wireless networks: Comprehensive framework and future trends. IEEE Wireless Communications 17(2), 22–31 (2010)
5. Lampropoulos, G., Skianis, C., Neves, P.: Optimized fusion of heterogeneous wireless networks based on media-independent handover operations. In: IEEE Wireless Communications 17(4) (2010) (Accepted from Open Call)
6. Yang, S.-J., Chen, S.-U.: QoS-based fast handover scheme for improving service continuity in MIPv6. In: 2010 IEEE International Conference on Wireless Communications, Networking and Information Security (WCNIS), pp. 403–408. IEEE (2010)
7. Buiati, F., Villalba, L.J.G., Corujo, D., Soares, J., Sargent, S., Aguiar, R.L.: Hierarchical Neighbor Discovery Scheme for Handover Optimization. In: IEEE Communications Letters 14(11), 1020–1022 (2010)
8. Atanasovski, V., Rakovic, V., Gavrilovska, L.: Efficient resource management in future heterogeneous wireless networks: The RIWCoS approach. In: Military Communications Conference, MILCOM 2010. IEEE (2010)
9. Lu, G.: Enable multimedia mobility with IEEE 802.21. In: IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks, WoWMoM (2010)
10. Kumar, V., Tyagi, N.: Media independent handover for seamless mobility in IEEE 802.11 and UMTS based on IEEE 802.21. In: 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), pp. 474–479 (2010)
11. Chen, J.-L., Ma, Y.-W., Huang, Y.-M., Yang, Q.-T.: An Adaptive QoS mechanism for multimedia applications over next generation vehicular network. In: 2010 5th International ICST Conference on Communications and Networking in China (CHINACOM) (2010)
12. Yuan, J., Wang, Y., Liu, F., Zheng, L.: Optimized Handover Scheme Using IEEE 802.21 MIH Service in Multi-Service Environment. In: IEEE 71st Vehicular Technology Conference (VTC 2010- Springer) (2010)

13. Obreja, S.G., Fratu, O., Vulpe, A.: A simulation testbed for a MIH enabled system. In: 2010 8th International Conference on Communications (COMM). IEEE (2010)
14. Vaca Ramírez, R.A., Ramos, R.V.M.: Handing Multiple Communications Sessions for the Next Generation of Wireless Networks. In: 2010 Fifth International Conference on Systems and Networks Communications (ICSNC), pp. 249–254. IEEE (2010)
15. Andrei, V., Popovici, E.C., Fratu, O., Halunga, S.V.: Development of an IEEE 802.21 Media Independent Information Service. In: IEEE International Conference on Automation Quality and Testing Robotics (AQTR), vol. 2. IEEE (2010)
16. Wu, X.: SOC, NUS, Simulate 802.11b Channel within NS2
17. Ramanath, S., Kavitha, V., Altman, E.: Impact of mobility on call block, call drops and optimal cell size in small cell networks, pp. 157–162
18. Alzubi, M., Amr, A., Anan, M.: An Efficient Handover Technique for 4G Networks, pp. 79–83. IEEE (2010)
19. Hwang, I.-S.: An integrated ISV Call Management Strategy in Heterogeneous Wireless Networks. In: 22nd International Conference on Advanced Information Networking and Applications, pp. 989–994. IEEE (2008)
20. Shiao, C.-M.: Performance Analysis of Algorithms with Multiple Attributes for Adaptive Call Admission Control in Heterogeneous Wireless Networks, pp. 303–308. IEEE (2009)
21. Gao, Z.: Performance Analysis of Cooperative Handover in Heterogeneous Wireless Networks. IEEE (2010)
22. Venes, J.: Performance of a Heterogeneous Network with UMTS, Wi-Fi and WiMAX
23. Kaci, N.: Performance of wireless heterogeneous networks with always-best-connected users. IEEE (2009)
24. Yang, X.: Research on the Mobility Management Scheme in heterogeneous network, pp. 4759–4763. IEEE (2010)
25. Márquez-Barja, J.: Evaluation of a technology-aware vertical handover Algorithm based on the IEEE 802.21 standard, pp. 617–622
26. Lee, S., Sriram, K.: Vertical Handoff Decision Algorithms for Providing Optimized Performance in Heterogeneous Wireless Networks. IEEE (2009)
27. Saboji, S.V., Akki, C.B.: A Client-Based Vertical Handoff in 4G Wireless Systems (November 2010)
28. Song, W., Zhuang, W.: Interworking of 3G cellular networks and Wireless LANs. Int. J. Wireless and Mobile Computing, 237–247 (2007)
29. Krendzel, A.: Cost and reliability estimation of radio access network structures for 4G systems
30. NandaKumar, S.: Performance of UMTS Interworking with WLAN to Provide Consistent Services (May 2011)
31. Patil, M.B.: Vertical Handoff in Future Heterogenous 4G Network. International Journal of Computer Science and Network Security (October 2011)
32. Kassar, M.: An overview of vertical handover decision strategies in heterogeneous wireless networks. Computer Communications 31, 2607–2620 (2008)
33. Koutsonikolas, D.: On the feasibility of bandwidth estimation in wireless access networks
34. Zhang, J., Marsic, I.: Link Quality and Signal-to-Noise Ratio in 802.11 WLAN with Fading: A Time-Series Analysis
35. Liyanage, M.: Steady-state Kalman filtering for channel estimation in OFDM systems utilizing SNR

An Analysis on Critical Information Security Systems

A Technical Review Tour and Study of the Sensitive Information Security Methods and Techniques

Sona Kaushik and Shalini Puri

Birla Institute of Technology, Mesra
Ranchi, India

sonakaushik22@gmail.com, eng.shalinipuri30@gmail.com

Abstract. Information's security and delicacy make the electronic systems in a challenging and chasing phase when it is passed on from an end to other via Internet. In today's world, as the necessity and importance of sensitive information are growing with new advancements and technologies, they become more prone to attacks and oriented towards the insecure environment. Attackers always want just an attack on such information to either to use it or intercept it. In this direction, many researchers have put their good efforts not only just to provide information security, instead they also consider the related primary security concerns, like confidentiality, access control, integrity along with the quality issues of reliability, robustness, usability etc. This effort included and presented a technological advancement based comparative study of various Sensitive Information Security (SIS) models and their proposed methods, algorithms, and experimental results. This study provides a great forum and better understanding of these models with their respective advantages and disadvantages.

Keywords: sensitive information models, data security models, concerns of security systems, confidentiality, integrity, reliability, denial of service, usability, robustness.

I Introduction

Sensitive Information Security [1] is sensitive and critical but unclassified information obtained or developed in the conduct of security activities, the public disclosure of which would constitute an unwarranted invasion of privacy, reveal trade secrets or privileged or confidential information, or be detrimental to the *security of nation*.

Information sensitivity is the control of access to information or knowledge that might result in loss in level of security if disclosed to others who might have low or unknown trustability or undesirable intentions. Loss, misuse, modification or unauthorized access to sensitive information can adversely affect the privacy or welfare of an individual, trade secrets of a business or even the security, internal and foreign affairs of a nation depending on the level of sensitivity and nature of the information.

Various examples of sensitive information include personal data such as Social Security Number (SSN); trade secrets; system vulnerability information; pre-solicitation procurement documents, and law enforcement investigative methods; similarly, detailed reports related to computer security deficiencies in internal controls are also sensitive information because of the potential and impending damage that could be caused by the misuse of this information.

With the exception of certain types of information protected by statute, there are no specific federal criteria and no standard terminology for designing the security systems. Such designations are left to the discretion of each individual federal agency.

As, with the advancement of technology and fast access speed with high use of unsecured network, like Internet, the sensitive information and data essentially need to be provided high security either at the user end, at the receiver end or on the communication channel [2] – [10]. As the demand of the information transfer is increasing day-by-day, its security and protection mechanisms and methods are also increased. A lot of research work has been done and still going on Sensitive Information Security (SIS) models, related techniques and methodologies. In this direction, many different methods have been introduced and successfully used by different organizations according to their system needs.

Many of the security models include the security at the hardware level; i.e. the Operating System level and where the trust OS is implemented [1]. The most promising and common way of security provision is at the user site when critical and delicate data and information are provided hard level or soft level security. Many researchers work on the security at the user site, so they design the models, techniques and algorithms to provide the best data security at this level only.

Section 2 discusses the background of sensitive information security including the survey information on cyber-crimes threatening the industrial security. Section 3 discusses the comparative study showing the different research methodologies. Section 4 discusses and shows the analysis consequences of the related primary security issues and concerns for different underlying studied models. Finally, section 5 gives the conclusion of the study of the work.

2 Background

With the research and development, there are some challenges and primary issues related to the security of critical and sensitive data which requires a great level of attention and focus. These research questions to be answered in this context are:

- How can the systems be secured completely while transferring the sensitive information and data?
- Can the large security systems be described and categorized in a systematic and meaningful way?
- How can the system achieve best confidentiality and integrity at the first priority?
- In what way, other concerns of security, access control and availability are provided?

- Can these systems provide high quality and its measures as well with the primary concerns?
- Using the defined criteria and factors for security, usability and external influence factors, can a formal and innovative security framework be designed?
- Which external factors, directly or indirectly influence the security of the underlying system?

The quest presented includes most of the features and desired characteristics for the secure systems. Including all these concerns, the primary concerns, data confidentiality, integrity, access control, and availability need a lot of attention in different models and algorithms without sacrificing them in any case. Attackers always keep an eye to break such system security. In this view, data confidentiality is broken up when the sensitive information and data cannot be confidential and secured from the attacker's eye and is visible to him. The loophole to integrity is through when the receiver side does not receive the same data as it was transmitted to it. The data has been modified or lost during the data transmission process, which is considered as one of the big loop holes of the security system. Thirdly, in the security systems, the access control mechanisms provide access and usage of the data and information to the user. If the user is not authenticated to access some particular and important data, then the provided security mechanism controls him/her to happen to do so. Access control techniques make it possible to use the information and provide the privileges to the genuine or authenticated user; thereby reducing and avoiding the cases of accessing the information by anyone extra. With this, the authenticated user is provided the information for which he is asking for; i.e. the data must be made available to him. Many researchers are working against the Denial of Service (DoS) attack so that the original user can access its data.

The Ernst and Young (E & Y) Global Information Security is one of the longest running surveys, respectable and recognized of its own kind [20]. E & Y surveyed and analyzed by the people how the risk environment in which they operate have changed in last month, as shown in figure 1. As organizations “digitize,” move into the cloud and become “borderless,” the risk landscape changes as well. 72% of respondents see an increasing level of risk due to increased external threats. At the same time, however, only about a third of respondents have updated their information security strategy in the past 14 months to respond to these enhanced threats. In addition, 46% of organizations have also identified increased threats within their own organizations.

Cyber Crime Statistics

The 2001 Computer Security Institute (CSI) Computer Crime and Security Survey found that financial losses caused due to cyber-crimes amounted to more than \$37 million for the nearly 200 companies that participated in the survey. Now, in 2011, FBI reports have indicated that more than 350,000 complaints of cyber-crimes were received this year alone. This identified that most the cyber-crimes remain unreported.

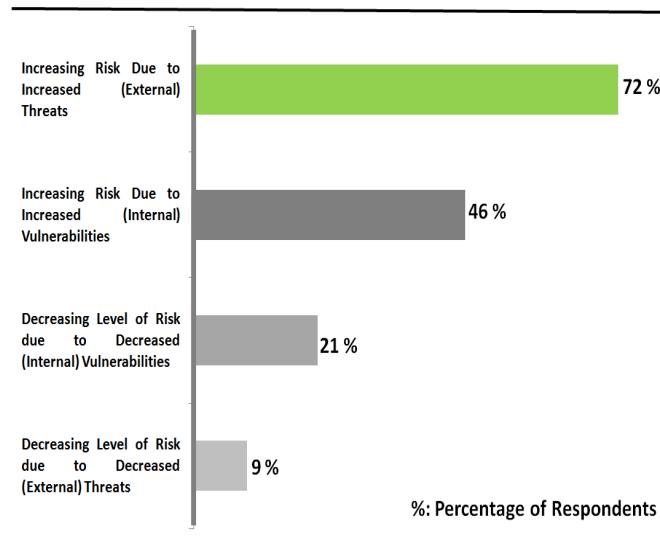


Fig. 1. Graph Depicting Respondents on Cyber Crimes

Figure 2 shows that just over a third of organizations do not have the information security strategy, and for the rest, plans continue to evolve despite a sense they may not be effective.

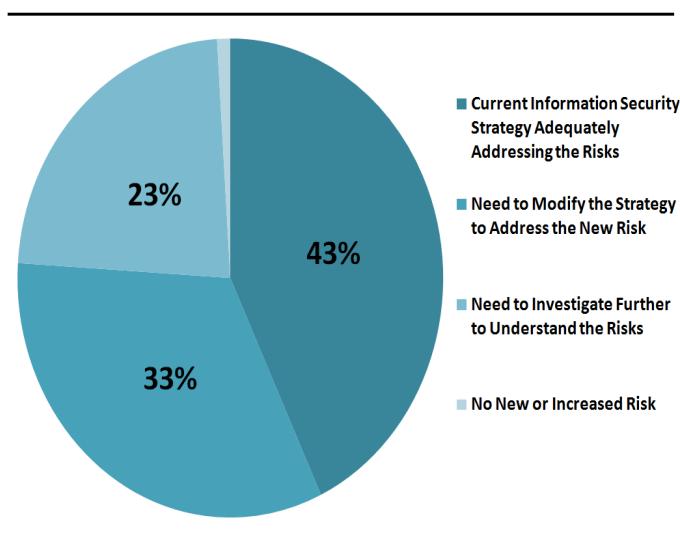


Fig. 2. Graph Depicting Organization's Security Strategy

Thus, this can be concluded that more and more key concerns must be addressed while architecture development of the security systems addressing higher number of threats and vulnerabilities.

3 Comparative Study

Based on the brief introduction given on the topic and substantiated by the background of the next section, a comparative study is provided on the basis of technological advancements, methodologies and techniques proposed.

A. Comparative Study Based on Technology Advancements

The comparative study shown in table 1 includes different techniques, and algorithms proposed for security of sensitive information when it is transmitted over on the communication channel. These models and their inherited concepts show the requirement of these models with their concerns. Table 1 also includes the experimental results and finally the conclusion and limitations of the respective models. This study provides us the brief overview and summary to show their usage in the current insecure data transmission environment.

Table 1. A Comparative Analysis of Various Proposed Security Methods and Techniques

REF NO.	AU-THOR(S)	PROB-LEM AREA	SOLUTION PROPOSED	PRIMA-RY CON-CERNs	CON-CEPTS USED	EXPERI-MENTAL RESULTS	CONCLU-SION AND LIMITA-TIONS
[1]	Wu, X.; Le, P. D.; Srinivasan, B.	Need to reduce key distribution problem and enhancing key security.	Security Architecture for Sensitive Information Systems.	• Reliability	-	Development of the body of knowledge surrounding sensitive information protection.	Involves the design of new cryptographic algorithms to enhance the security of sensitive information systems.
[2]	Páez, R.; Satizábal, C.; Forné, J.	Problem in critical systems to protect distributed networks.	A security scheme to verify the entities' integrity inside the IDS architecture named Cooperative Itinerant Agent (CIA).	• Scalability • Throughput • Delay • Reliability. • Integrity	• Watermarking • Fingerprinting • Intrusion Detection	Time spent by a CIA agent allows the system to issue several CIA agents without overloading the network and it is limited only by the network delay and throughput.	• Necessary to use other protection mechanisms such as software fingerprinting to identify each entity inside the IDS. • Reduction of traffic network.

Table 1. (continued)

[3]	Zhang, W.; Yan, X.	Need of high level security in network applications.	Achieved by multiple verification and frozen sensitive data and code ensuring network safety of network applications even if intercepted during its transport.	<ul style="list-style-type: none"> • Authentication • Integrity • Confidentiality 	<ul style="list-style-type: none"> • Mobile Agent (MA) • Frozen Agent • Agent Transport Security 	MA in the process of transport is intercepted or lead towards the illegal destination host, they can't produce safety threat to the system because its key data and codes are in frozen state.	<ul style="list-style-type: none"> • Improves the safety of the Mobile Agent (MA) in the process of transport. • Solves Safety problem between MA and destination host.
[4]	Chen, E. Y.; Ito, M.	Problem of Evil Twin attacks.	Proposed “end-to-middle security,” to be adopted by mobile users.	<ul style="list-style-type: none"> • Authentication • Reliability 	-	Generalized an end-to-middle security model, which mobile users can adopt to protect themselves against Evil Twin attacks.	Can implement true end-to-middle security systems by addressing issues with VPN, web proxies and voice gateways.
[5]	Chen, H.	Problem of existing defects of traditional VPN in constructing enterprise network.	Design of secure enterprise network to put forward solution of DMVPN (Dynamic Multipoint VPN) technique to solve the unsolved problems of traditional VPN.	<ul style="list-style-type: none"> • Authentication • Reliability 	<ul style="list-style-type: none"> • VPN • DMVPN • UDP • NHRP • Enterprise network 	Puts forward the solution and implementation mechanism for enterprise constructing the safe network.	Enhanced Security provided.
[6]	Xu, S.; Li, X.; Parker, T. P.; Wang, X.	Need of exploiting Trust-Based Social Networks for Distributed Protection of Sensitive Data.	Proposed taking advantage of real-life social trust between average users (“trust-based social networks”) as well as threshold cryptography, leading a complex system.	<ul style="list-style-type: none"> • Reliability • Psychological soundness • Availability • Confidentiality 	<ul style="list-style-type: none"> • Anonymity • Availability • Complex system • Social network • Social trust • Threshold cryptography 	The idea of exploiting trust-based social networks for distributed protection of sensitive data, especially cryptographic keys.	<ul style="list-style-type: none"> • Impact of the distribution of Sybil nodes while also considering the node degree distribution. • Re-examining countermeasures against Sybil attacks through the lens of the system.
[7]	Lijun, G.; Lul, Z.	Major barrier in RFID system is the issue of security and cost.	Reviewed the existing RFID security protocols and analyzed the flaw of these protocols.	<ul style="list-style-type: none"> • Authenticity 	<ul style="list-style-type: none"> • RFID • Security-Protocol 	Proposed a suitable protocol model which can adapt to the RFID system environment.	Identified the shortcomings in the existing RFID security protocol and the problems of international standard incompatibility.

Table 1. (continued)

[8]	Bhutta, A.; Fo-roosh, H.	Need of security provision to combine encryption for multiple data streams.	To combine the encryption of multiple data streams by generating a single encrypted stream, from which any of the original data streams can subsequently be decrypted.	• Authenticity • Reliability	-	Demonstrated the reuse of key matrices without compromising the security of encryption technique	• Presented an attack study to evaluate the security of the algorithm. • Proposed an algorithm to combine the encryption of multiple data streams into a single encrypted stream. • Computationally efficient.
[9]	Proan˜o , A.; Lazos, L.	Problem of selective jamming attacks in wireless networks.	Providing Packet-Hiding Methods.	• Availability	• Denial - Of-Service • Wireless Networks • Packet Classification	Addressed the problem of selective jamming attacks in wireless networks.	Improved performance with very low effort.
[10]	Ukil, A.	Inadequacy of the traditional authorization mechanisms to secure distributed systems.	A trust and reputation model for Secure Trust Management in Distributed Computing Systems to compute the trustworthiness of individual nodes.	• Reliability	• Trust • Reputation • Distributed systems • Security • Wireless sensor networks	Trust management modeling for distributed systems like WSNs.	Finds communication overhead and modeling other malicious behavior patterns to make the distributed system more reliable.
[11]	Liu, Y.; Corbett, C.; Chiang, K.; Archibald, R.; Mukherjee, B.; Ghosal, D.	SIDD: A Framework for Detecting Sensitive Data Exfiltration by an Insider Attack.	System of a high-speed transparent network bridge located at the edge of the protected network.	• Confidentiality	-	• Developed a systematic approach to address the key problems of detecting the exfiltration of sensitive content. • A multilevel framework that composed of application detection, content signature generation and detection, and covert channel detection was proposed.	• Able to significantly extend the state-of-the-art to address a critical problem in network security. • Broad range of exfiltration is not investigated.

Table 1. (*continued*)

[12]	Yong-sheng, L.; Guan-gyu, L.; Jing, L.; Cheng-cheng, L.	Problem of illegal accessing, misusing and tempering in service oriented computing paradigm.	Study on Access Control Model of Service-Oriented Computing to protect the services.	• Authenticity	<ul style="list-style-type: none"> • Service computing • Access control • Role hierarchy • Binding Context 	<ul style="list-style-type: none"> • Proposed a service oriented computing access control model based on RBAC. • Introduced the concept of Binding Context. • Provided support for fine granularity implementation of the strategy. <p>Can be further refined to enhance the relationship between the domains of members.</p>
[13]	Wang, S.; Li, X.	Problem to protect sensitive information flows based on trusted computing technologies to be applied in sensitive organizations.	Secure Information Flows Control Model (SIFCM) based on trusted computing technologies to be applied in sensitive organizations.	• Confidentiality	<ul style="list-style-type: none"> • Trusted computing • Information Security • Cryptography • Key Management • Information Flow Control 	<p>Applied in sensitive organizations to ease the difficulty to protect and manage the keys before trusted computing technologies.</p> <p>Incorporates Access control mechanism.</p>
[14]	Yong, Z.; Ji-Qiang, L.; Zhen, H.; Chang-Xiang, S.	Necessity of an Operating System Trusted Security Model For Important Sensitive Information System.	An Operating System Trusted Security Model for Important Sensitive Information System to control and adjust the information flow using trusted measurement, and to control bi-direction information flow without reducing information system security.	• Confidentiality • Integrity • Reliability	<ul style="list-style-type: none"> • OS trusted security • Trusted entities for information flow. 	<ul style="list-style-type: none"> • Reliability of integrity to adjust the subject's integrity level. • Better complexion that integrity level equals confidentiality level for bi-direction information flow. <p>Used to protect the system confidentiality and integrity.</p>

Table 1. (continued)

[15]	Mejia-Nogales, J. L.; Vidal-Beltrán, S.; López-Bonilla, Y. J. L.	Need to design and Implementation of a Secure Access System to Information Resources for IEEE802.11 Wireless Networks.	A secure system to get access to privileged data for medium access of wireless network based on IEEE 802.11 standard.	• Authentication • Confidentiality	• Network secure system • Captive portal • Authentication server • Access point	• Independence between subsystems • Transparent procedures • Flexibility and Compatibility • Use freeware • Easy implementation • Robust • Private Addressing	• Currently designed and tested in an academic environment. • Can be installed in libraries, enterprises or any scenario to protect critical information • Requires user authentication to get access to the networks resources.
[16]	Shi, W.; Fryman, J. B.; Gu, G.; Lee, H.S.; Zhang, Y.; Yang, J.	Problem of theft of sensitive information such as passwords, encryption keys, and other private data.	Proposed a unified and lightweight solution Info-Shield: A Security Architecture for Protecting Information Usage in Memory to strengthen application with a minimal performance impact.	• Authentication	• Cyber theft • Non-volatile memory information	Architectural and programming support to protect usage of sensitive information against many documented attacks and exploiting on data privacy including memory scan, pointer/array index manipulation, integer, format string attacks, and password-stealing Trojans.	Provides good security.
[17]	Hussain, K.; Rajan, S.; Ad-dulla, N.; Moussa, G.	Problem to prevent capture of sensitive data by users that have administration privileges.	Proposed a no capture hardware feature for securing sensitive information.	• Confidentiality	Sensitive Information	Implementation of no-capture hardware security feature to prevent capturing of sensitive information.	Prevents the capturing of sensitive data by users that have administration privileges.

Table 1. (continued)

[18]	Benjamin, C.; Fung, M.; Xiong, L.	Requirement of Service-Oriented Architecture for High-Dimensional Private Data Mashup.	Mashup, a web technology to allow different service providers to flexibly integrate their expertise and to deliver highly customizable services to their customers.	• Reliability • Confidentiality	• Anonymity • Data mashup • Data integration • Service-oriented architecture • High dimensionality	Implemented a data mashup application for the online advertising industry in social networks, and generalize their privacy and information requirements to the problem of privacy-preserving data mashup for the purpose of joint data analysis on the high-dimensional data.	Industry primary concern is whether or not the anonymous data is still effective for data analysis; solutions that solely satisfy some privacy requirement are insufficient for various data analysis tasks.
[19]	Wan, K.; Su, R.; Li, Z.; Cai, Z.; Zhou, L.	Study of Secure Complicated Information System Architecture Model.	Model to conduct the construction or reconstruction of CIS using layered method, it divides CIS into modules and reduces system complexity.	• Authentication • Confidentiality • Reliability • Scalability	• Security • Interoperability • Extensibility	Presented a secure CIS architecture model to guide the construction or reconstruction of CIS.	Based on the model, China has approved the national e-government application support platform criterion and e-government security support platform criterion.

4 Analysis Consequence

The study provided in the last section includes the variety of approaches and concepts for security related concerns. The key concerns while building a security architecture or system are the attributes they have to be focused upon. The attributes to be covered and with what priority are dependent on the core objective of building the security structure.

The analyzed architectures include various security key concerns for their specific requirements which are categorized to eight only by the authors. The graph shown in figure 3 gives the graphical analysis of the key concepts used in the architectures.

The study provided in the last section includes the variety of approaches and concepts for security related concerns. The key concerns while building a security architecture or system are the attributes they have to be focused upon. The attributes to be covered and with what priority are dependent on the core objective of building the security structure.

The analyzed architectures include various security key concerns for their specific requirements which are categorized to eight only by the authors. The graph shown in figure 3 gives the graphical analysis of the key concepts used in the architectures.

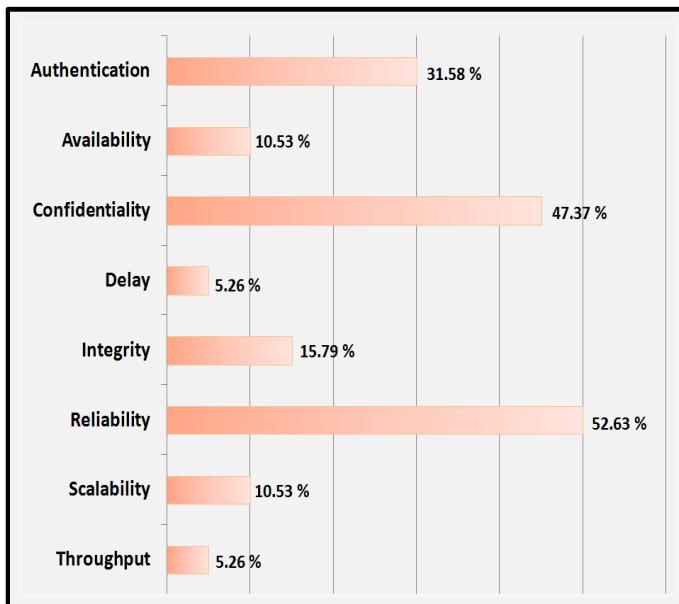


Fig. 3. An Analysis Graph of Security Key Concerns

5 Conclusion and Future Scope

Offline or online information, if sensitive, has to be secured in the best efficient manner. Primary concerns are identified and taken care to build secure system. Many systems have been developed by now securing at their best. The comparative study puts an effort to check their key areas and solutions for the given problems. This research has presented a security overview that shows the existing security approaches in protecting sensitive information.

This current research has enabled the formal description of critical information security and regulated the cryptographic properties. An analysis is presented in the work realizing the current primary objectives of the security systems and identifying integrity as the key concern to be focused more in the future systems to make the system robust against attacks and thus, more reliable.

Future work might involve the testing of these definitions to further demonstrate their appropriateness.

Acknowledgment. We would like to acknowledge and give our special thanks to Asst. Prof. Pankaj Gupta, Dept. of Computer Science, Birla Institute of Technology, Noida, India and Dr. Vikas Saxena, Dept. of Computer Science, Jaypee Institute of Information Technology, Noida, India for their continuous encouragement, help and guidance.

References

- [1] Wu, X., Le, P.D., Srinivasan, B.: Security Architecture for Sensitive Information Systems. In: Convergence and Hybrid Information Technologies, pp. 239–266 (March 2010)
- [2] Paez, R., Satizabal, C., Forne, J.: A Performance Model to Cooperative Itinerant Agents (CIA): A Security Scheme to IDS. In: The Second International Conference on Availability, Reliability and Security(ARES), pp. 791–798. IEEE (2007)
- [3] Wei, Z., Xiaofei, Y.: Agent Transport Security Based On Freezing Mode. In: International Conference on Communications and Intelligence Information Security (ICCIIS), pp. 60–63. IEEE (2010)
- [4] Chen, E.Y., Ito, M.: Using End-to-Middle Security to Protect against Evil Twin Access Points. In: IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks & Workshops (WoWMoM), pp. 1–6 (2009)
- [5] Chen, H.: Design and Implementation of Secure Enterprise Network Based on DMVPN. In: International Conference on Business Management and Electronic Information (BMEI), pp. 506–511. IEEE (2011)
- [6] Shouhuai, X., Xiaohu, L., Parker, T.P., Xueping, W.: Exploiting Trust-Based Social Networks for Distributed Protection of Sensitive Data. IEEE Transactions on Information Forensics and Security 6(1), 39–52 (2010)
- [7] Lijun, G., Zhang, L.: Low-Cost RFID Security Protocols Survey. In: Cross Strait Quad-Regional Radio Science and Wireless Technology Conference (CSQRWC), vol. 2, pp. 1068–1070. IEEE (2011)
- [8] Bhutta, A., Foroosh, H.: On Combining Encryption For Multiple Data Streams. In: IEEE 9th International Multitopic Conference (INMIC), pp. 1–6 (2005)
- [9] Proano, A., Lazos, L.: Packet-Hiding Methods for Preventing Selective Jamming Attacks. IEEE Transactions on Dependable and Secure Computing 9(1), 101–114 (2012)
- [10] Ukil, A.: Secure Trust Management in Distributed Computing Systems. In: Sixth IEEE Symposium on Electronic, Design, Test and Application (DELTA), pp. 116–121 (2011)
- [11] Yali, L., Corbett, C., Ken, C., Archibald, R., Mukherjee, B., Ghosal, D.: SIDD: A Framework for Detecting Sensitive Data Exfiltration by an Insider Attack. In: 42nd Hawaii International Conference on System Sciences (HICSS), pp. 1–10. IEEE (2009)
- [12] Zhang, Y., Liu, G., Li, J., Li, C.: Study on Access Control Model of Service-Oriented Computing. In: International Forum on Computer Science-Technology and Applications (IFCSTA 2009), vol. 1, pp. 239–242. IEEE (2009)
- [13] Shi-Hua, W., Xiao-Yong, L.: A Security Model to Protect Sensitive Information Flows Based on Trusted Computing Technologies. In: International Conference on Machine Learning and Cybernetics, vol. 7, pp. 3646–3650. IEEE (2008)
- [14] Zhao, Y., Liu, J.Q., Han, Z., Shen, C.X.: An Operating System Trusted Security Model for Important Sensitive Information System. In: The First International Symposium on Data, Privacy, and E-Commerce (ISDPE), pp. 465–468. IEEE (2007)
- [15] Mejia-Nogales, J.L., Vidal-Beltran, S., Lopez-Bonilla, J.L.: Design and Implementation of a Secure Access System to Information Resources for IEEE 802.11 Wireless Networks. In: Electronics, Robotics and Automotive Mechanics Conference, vol. 1, pp. 58–63. IEEE (2006)
- [16] Shi, W., Fryman, J.B., Gu, G., Lee, H.-H.S., Zhang, Y., Yang, J.: InfoShield: A Security Architecture for Protecting Information Usage in Memory. In: The Twelfth International Symposium on High-Performance Computer Architecture, pp. 222–231. IEEE (2006)

- [17] Hussain, K., Rajan, S., Addulla, N., Moussa, G.: No-capture Hardware Feature for Securing Sensitive Information. In: International Conference on Information Technology: Coding and Computing (ITCC), vol. 1, pp. 697–702. IEEE (2005)
- [18] Fung, B., Trojer, T., Hung, P., Xiong, L., Al-Hussaeni, K., Dssouli, R.: Service-Oriented Architecture for High-Dimensional Private Data Mashup. IEEE Transactions on Services Computing (2011) (unpublished)
- [19] Wang, K., Su, R., Li, Z., Cai, Z., Zhou, L.: Study of Secure Complicated Information System Architecture Model. In: First International Conference on Semantics, Knowledge and Grid (SKG 2005). IEEE (2005)
- [20] Ernst, Young: Into the cloud, out of the fog. Ernst & Young Global Information Security Survey (December 2011)

Emergency Based Remote Collateral Tracking System Using Google's Android Mobile Platform

Prabhu Dorairaj, Saranya Ramamoorthy, and Ashok Kumar Ramalingam

Department of Electrical Engineering
Blekinge Institute of Technology, Sweden, 2011

Abstract. Introduction of Smart phones redefined the usage of mobile phones in the communication world. Smart phones are equipped with various sophisticated features such as Wi-Fi, GPS navigation, high resolution camera, touch screen with broadband access which helps the mobile phone users to keep in touch with the modern world. Many of these features are primarily integrated with the mobile operating system which is out of reach to public, by which the users can't manipulate those features. Google came up with an innovative operation system termed as ANDROID, which is open system architecture with customizable third party development and debugging environment which helps the user's to manipulate the features and to create their own customizable applications.

In this paper, 'Emergency Based Remote Collateral Tracking System' application using Google's Android Mobile Platform is addressed. Emergency is divided into three categories: heart beat based emergency, security threats like personal safety and road accidents. This application is targeted to a person who is driving a vehicle. Heart rate monitoring device is integrated with our application to sense the heart beat of a person driving the vehicle and if there is any abnormalities in the heart beat, then our application performs a dual role. One in which, application uses a GPS to track the location information of the user and send those location information as a message via SMS, email and post it on Facebook wall Simultaneously, an emergency signal is sent to Arduino Microcontroller.

Road accidents are quite common, this application is also designed to detect the accident using the sensors in the Android Mobile. Security threat can occur anywhere, our application also answers for personal safety, when the user interacts with the application by pressing the button, then automatically the application generates the geographical information and sends that location information via SMS and email to a pre-stored emergency contact and the same information will be posted on user's Facebook wall. This application is written in JAVA programming language which runs on Eclipse Integrated Development Kit.

Keywords: Android, Arduino microcontroller, Emergency, GPS, Heart rate device.

1 Introduction

Instauration of mobile devices gave birth to lot of innovative technology, and exchanging information globally has become more prominent. Smart phones gave a

new dimension to the usage of mobile phones for the users. Apart from basic functionality such as messaging, calling and cameras, smart phones laid a way to portray a personal computer. Not only the mobile phone looks newer, it's the operations system and the applications which are built to meet the various features of the hardware made difference.

The mobile phone has now become a major source of information device which can be seen almost in everyone's hand in the world. Mobile devices with computing process ability have been widely used to access network via mobile communication network. Different categories of application such as games, social networks, and health care are being developed to meet the user's requirements. Each mobile user is of unique kind, one wants to use the basic functionality of the smart phones, the other want to use the built in application, the most advanced user who wants to play with the hardware and to develop his own customizable application. To answer each kind of user, Google mustered up a groundbreaking product called as "ANDROID", which includes an open source operating system, middleware and a user-interface [1, 3].

2 What Is Android?

In 2005, Google acquired Android from Android Inc. which was found in year 2003 by Andy Rubin and they dealt with developing software for mobile devices. Later, OHA which comprises of 79 companies along with Google developed their new mobile platform for mobile devices. This alliance was formed so as to develop open technologies for mobile devices and make those applications easily available in the market. This new open source technology was named as **Android** [3, 4].

Android is an open source architecture which is used for developing applications for mobile devices. Android works on Linux Kernel. It has an operating system, middleware and key applications. Android announced its code under the license of free software/open source in the year 2008. Android comes up with an API for mobile devices. This Linux Kernel supports Java Virtual Machine which favours Java to be most suitable programming Language for development of the applications. Google provides a SDK to all developers which include libraries, debugger and a handset emulator in Eclipse IDE [5, 6]. The application which is developed in Android can be tested using this emulator which works similar to a mobile phone.

2.1 Arduino Microcontroller

Arduino Microcontroller is an open source prototyping platform which can sense the environment by the sensors which is given as the input to it. The programming of the controller is done using Arduino Programming language. The language used for programming is C/C++. Various Arduino microcontrollers are available in market such as Arduino Extreme, Arduino Mini, Arduino Nano, Arduino Bluetooth, Arduino Diecimila, Arduino Duemilanove, Arduino Mega and so on. Each of these microcontrollers have their own significance. Arduino Bluetooth is found as best choice for our project. As the name suggests, this microcontroller has in built Bluetooth module which lacks in other controllers [9].

Arduino Bluetooth (Arduino BT) microcontroller as in Fig.1. works on principle of Atmega168 and the Bluegiga WT11 Bluetooth module. It has 14 input/output pins and it supports serial communication over Bluetooth. The operating voltage is 5V which makes the controller very fragile and hence the voltage should not be exceeded else it would result in the damage of the microcontroller. It has 16kb flash memory for the storage of the code. The reset option is at pin number 7 which is connected to the reset of bluegiga WT11 module. The Bluetooth communication is provided by Bluegiga WT11 module on Arduino BT which can connect to any devices which has Bluetooth connectivity. It should be configured and should be detected by the device to which it is connected. It works on the baud rate of 112500. The controller is connected to another device by pairing and the name of the device suggested by Arduino is ARDUINOBT and the passcode is 12345. This is the default setting of the device [9]. Arduino Bluetooth microcontroller is connected to the Android mobile device via Bluetooth.

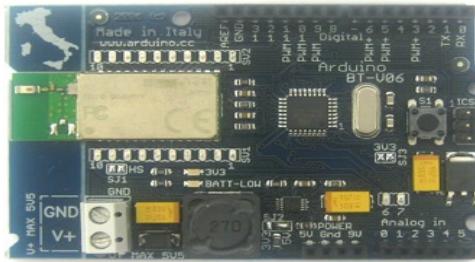


Fig. 1. Arduino Microcontroller

The challenge is the compatibility of the microcontroller to the mobile phone enabled with android. The programming of the android mobile device should work well with arduino BT microcontroller which is proper integration of them. This amalgamation of the components is done using Amario.

2.2 Heart Rate Monitoring Device

Heart is the main organ in a human's body. One can't live without it. It's because of this anything in the world is compared to heart, that's the importance of such a vital organ. Heart rate is an important factor to be considered in a human body. Heart rate tells us how many times heart beats in a minute. It is usually measured by feeling the pulse on any area near the artery. This measure signifies the blood pressure of a person. The blood pressure either low or high is dangerous to health. Hence it has to be kept under control and also by constant monitoring [10].

Heart rate monitoring is an important aspect of a human being. This monitoring is usually done by a regular health checkup at any hospital. This is a normal scenario, but there are situations where the heart rate is not monitored while driving any vehicle or while exercising and so on. Hence a heart rate monitoring device is very essential. Heart rate monitoring is done at any hospitals using devices like ECG. Even though it

is accurate, this device is costly and also regular visit to hospital should also be carried out. Also a person with heart disease complaint should be able to monitor his condition continuously. To solve all these criticalities, a heart rate monitoring device has to be purchased and maintained for personal care. These days heart rate monitor is been used commonly by normal person rather than in a hospital [11, 12].

Heart rate monitor helps to detect the abnormalities in the heart and would display it to the person who is using it. This feature has inspired us to use such a device for our emergency conditions especially when a person is driving a car and is suddenly met with heart attack. To get situation under control, this monitor device would send an alert to the android enabled mobile phone which will in turn halt the car to avoid further causalities. There are various heart rate monitoring device are available in the market such as Zephyr HR Bluetooth heart rate monitor, polar Bluetooth heart rate monitor, Wahoo Fitness ANT plus Dongle and so on [13].

Zephyr HR Bluetooth enabled heart rate monitor as shown in Fig.2. is best suited for our project for various reasons such as the Zephyr programming is easier and it is open source. It is a device with Bluetooth connectivity which avoids wired connection and reduces the hardware cost for it. It also has a fabric sensor which detects the data irrespective of any fabric. Speed, distance is also displayed using this device which helps to see a pictorial representation of a person's heart rate. The best thing about this device is that it can tolerate any extreme motion of the body like running, jumping, jogging and so on. Also the transfer of data is via Bluetooth [13].



Fig. 2. Zephyr Bluetooth Heart Rate Device

2.3 Integration of Arduino with Android-Amarino

Every request sent has its own response. The same is the case with a mobile phone. For instance, a phone call is alerted to the user by a ringtone, a text message received is displayed on the screen, a photo clicked with the help of the camera is saved in memory of the phone and so on. These events are generated on the phone itself. The same event can also be viewed somewhere else like in our room, through a sensor like accelerometer or on a microcontroller. To such a situation to occur, Amarino is used.

Amarino is a tool kit which helps in integration of android with arduino. It consists of the Android application and libraries required for arduino. Amarino helps to connect a mobile device enabled with android and an arduino microcontroller via Bluetooth [14].

2.4 Research Questions

1. How can the location of a person be tracked and notified using Google's Android platform in case of Emergency?
2. How to monitor the heart rate of a person and to manipulate those data to detect a critical situation using Android mobile platform?
3. How to integrate Arduino microcontroller with android mobile device?

2.5 Expected Outcomes

The expected outcomes of this paper are

- The exact location of a person will be tracked using GPS and location information will be sent to a pre-stored number via SMS, email and message will be posted on Facebook wall in case of emergency.
- Heart rate of a person will be monitored and will be notified under critical situation to android mobile device and using that information the location of a person will be tracked and will be sent to an emergency contact number/email/Facebook wall.
- Under critical situation of heart rates, the android mobile send a signal to a microcontroller and LED in that microcontroller blinks to make an alert of risk signal.

2.6 Main Contribution

- To design an application for android enabled mobile device and to track the location of a person using GPS and to send the location information to a pre-stored emergency contact number.
- Design an application and user interface using Java program to integrate SMS functionality to send geographical information to another remote emergency email address and also message which has to be posted on Facebook wall.
- Integrating Arduinio microcontroller with android mobile phone, to make an alert of risk signal under critical situation.

3 Heart Rate Based

3.1 Problem Statement

Nowadays, mobile devices started to integrate with various third party hardware's to provide more functionality to the users, which also leads to the integration of a heart rate device which will monitor the heart beat of an user. But how the heart rate device can be integrated with Android mobile, so that the android enable mobile can monitor the heart beat of a person, and also how to use that heart rate to manipulate a person under emergency? The Zephyr heart rate monitoring device is used to fetch the heart rate of a person and that device is integrated with the android mobile with help of programmable application, developed using android SDK and, this application will decide about the critical situation with respect to heart rate and sends a message to a

pre-stored emergency contact number which also contains the geographical location of the user.

3.2 Scenario

The design of this project deals with a person/user driving in a vehicle. A typical scenario is, when a person driving in isolated roads, wearing the Zephyr heart rate device around the chest. This heart rate device will send the heart rate every second to the android mobile via Bluetooth by which it is monitoring the heart beat of the driving person. The heart rate is normal between 60-100, if it is less than 60 it's called has *Bradycardia* and if it's more than 100 it is called as *tachycardia*. In most condition the heart beat becomes less when there is dehydration, decreased protein intake and it becomes more in uncontrolled hypertension.

3.3 Modeling

Our application is designed to sense this heart rate, and if there are any abnormalities in the heart rate like, if the heart rate goes below 60 or above 100, automatically the android mobile will send a signal to an Arudino microcontroller which is connected to android mobile via Bluetooth. This Arudino microcontroller will make an alert signal, in our case the alert signal is indicated by blink of a led. Simultaneously our application will track the location information of the user who is under emergency and send that location information to a remote pre-stored emergency contact number. This scenario is shown in Fig.3.

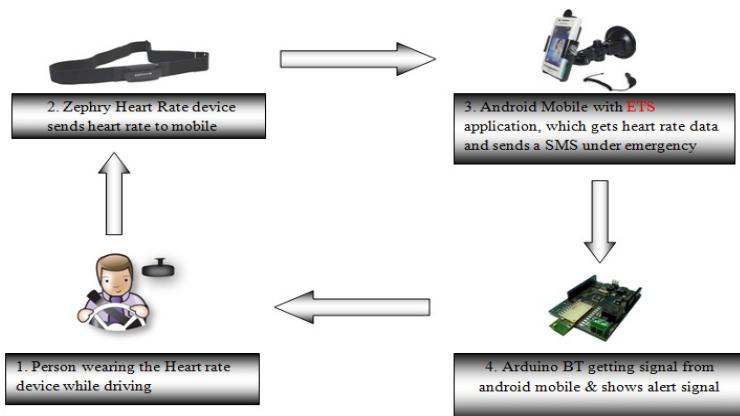


Fig. 3. Flow diagram of Heart Rate Based Scenario

3.4 Implementation and Validation

3.4.1 Implementation of API with Heart Rate Device

Implementation of communication between a zephyr heart rate (HR) device and android mobile starts with designing an API. Fig.3below shows the Zephyr Bluetooth

communicates with a mobile device over the Bluetooth link. The Zephyr Bluetooth HR device uses a Bluetooth SPP (Serial Port Profile) to communicate with the low level protocol such as

- a) 115,200 baud rate b) 8 data bits c) 1 stop parity bits d) No parity

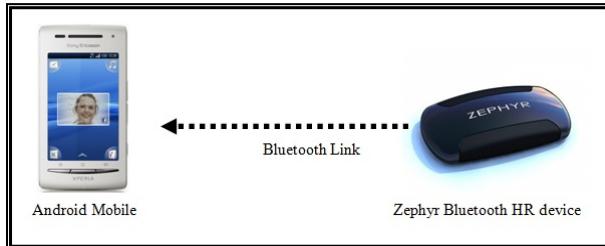


Fig. 4. Zephyr Bluetooth device communication with Android phone

Our application employs the API and enables the HR device to transmit the different packet types such as heart rate, speed and distance packets. The following steps below are a description of the most important aspects of the source code in our application used to enable the General Packet and display the data on the Android phone.

1. On clicking the Connect button, a Bluetooth adaptor type object is created and passed to an Object of the *BT Client* class type. The *BT Client* object is essentially a thread that manages the overall Bluetooth connectivity of the phone with the HR device.

2. Next, an object of the *NewConnectedListener* class will need to be created which essentially implements the *ConnectedListener* interface, and one that extends the *ConnectedListener* class. This object is responsible for reacting differently to different kinds of messages. In this object we override the parent class's *connected* method and define our own method. In this method we create a *ZephyrProtocol* object and call its *addZephyrPacketEventListner* method. This method takes a *ZephyrPacketListener* argument, in whose *ReceivedPacket* method we define what message we are interested in, and how we want the data to be displayed on the phone screen.

3. This *ConnectedListenerImpl* object needs to then be connected to the *BTClient* object type via *addConnectedEventListner* function call to tie this object to respond to a received packet from the HR device.

4. Calling the *start* function of the *BTClient* thread kicks off the communication of the Application with the HR device.

Validation

The HR device powers on automatically when worn. If there is insufficient skin conductivity (excessively dry skin and/or strap sensor pads), the wear-detect circuitry may not trip. Moisten skin and sensor pads with water. The heart rate device has to be worn around the chest region just near the sternum as shown in the Fig.5.

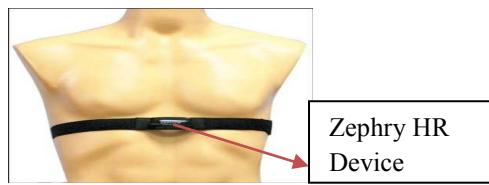


Fig. 5. Zephyr heart rate device worn



Fig. 6. ETS Application

Open ETS-Emergency Tracking System application which is installed in the android mobile as shown in Fig.6. As soon as the ETS is clicked, home page with menus such as Pair HR Device, Pair Controller, Enter Details, Start and Disconnect is opened. In order to get the heart rate, user need to pair the HR device with the android mobile. As soon as the HR device is paired with the mobile, the heart rate starts to get displayed on the obile screen. The entire workflow can be seen in the Fig.7.

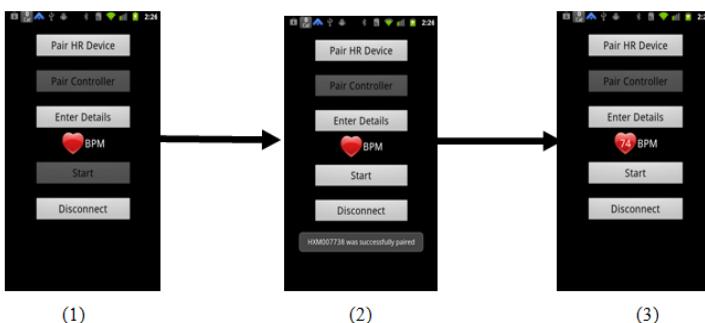


Fig. 7. Entire Workflow of Pairing HR device to Android phone

1. Connects to HR device by tapping on “Pair HR Device”
2. A message shows HR device is successfully paired
3. Heart rate get displayed on the screen “74 BPM (beats per minute)”

The emergency contact number, email address and Facebook account can be integrated in the “Enter Detail page” in the main menu of ETS as shown in Fig.8.



Fig. 8. Enter Details page

When HR device starts to transmit the heart rate to mobile, ETS application performs condition check with respect to $60 < \text{HR} < 120$, if heart rate is less than 60 or more than 120 then the application decides that this condition is critical and starts to track the location of the user using GPS API and simultaneously sends a message to pre-stored emergency contact number and also to a pre-stored Email address. Finally a message containing “This person is under emergency take necessary action” followed by the geographical location of the person is posted on the enabled Facebook wall. Simultaneously an alert signal is send to arduino microcontroller, to acknowledge the risk signal a LED is connected to pin 13 of the microcontroller and that LED blinks under critical situation. The entire work flow is represented in Fig.9.

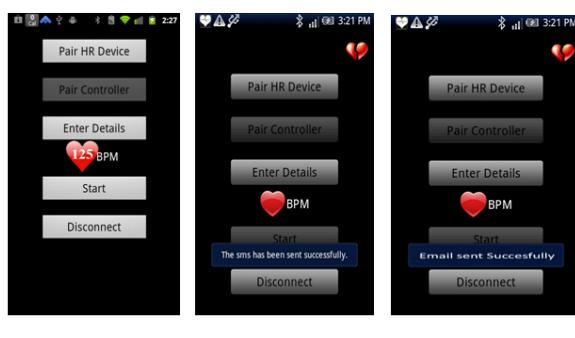


Fig. 9. Abnormal heart rate scenario – workflow

- (1) ETS reads an abnormal heart rate of 125BPM
- (2) SMS was sent to a pre-stored number
- (3) Email was sent to a pre-stored Email address



EMERGENCY!!!

altair8624@gmail.com Add to contacts
To myworld0501@gmail.com

Person is under emergency condition take necessary action
 Latitude : 56.18137
 Longitude : 15.591338
 City: Bergåsa
 Country : sweden

New | Reply | Reply all | Forward | Delete | Junk | Mark as | Move to | Empty |  

Fig. 10. Message sent via SMS, E-Mail and Facebook Wall

Result

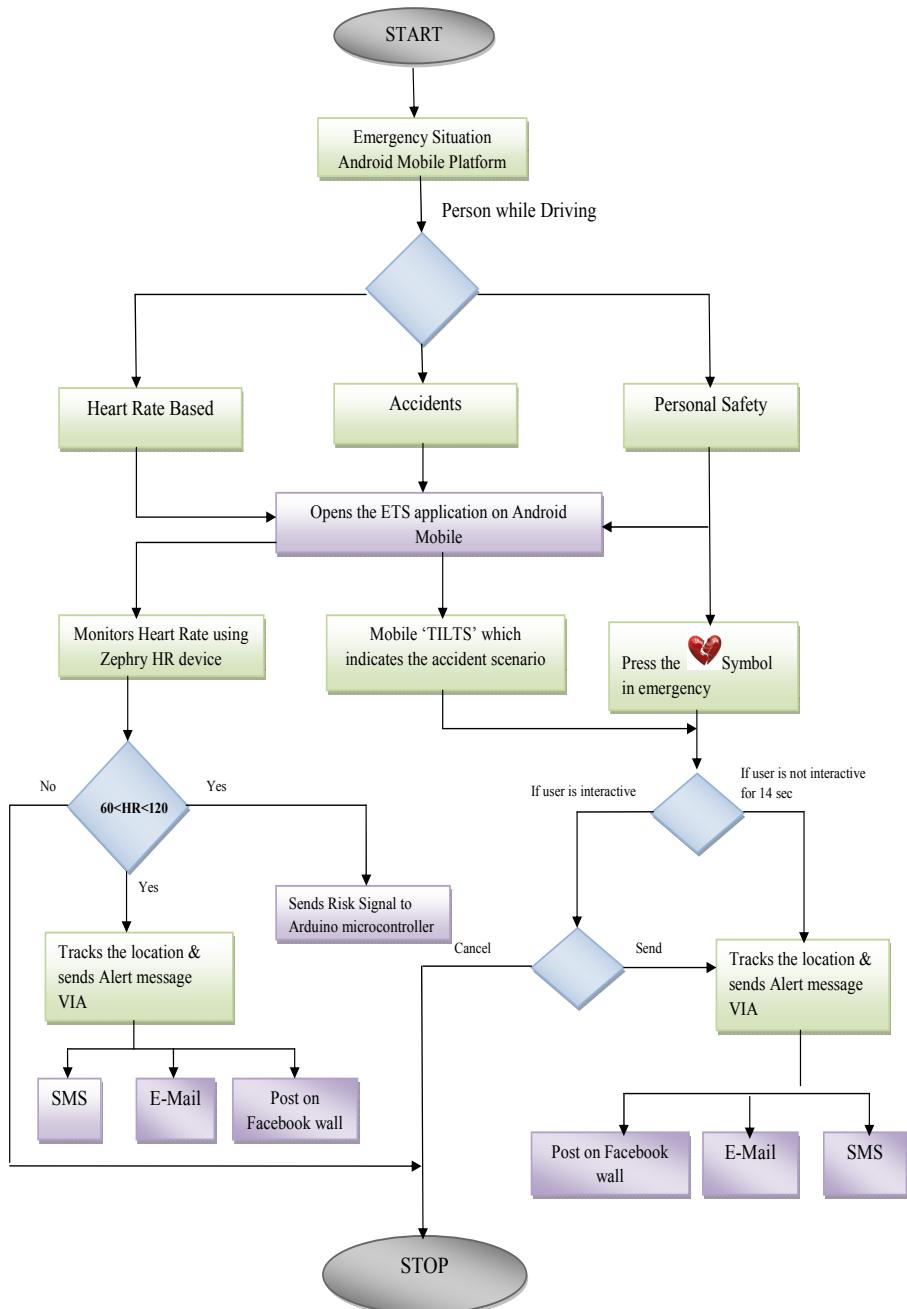


Fig. 11. Working and Result of ETS Application

5 Conclusion

Our research which was based on Emergency Tracking System using Google's Android Mobile platform and application based on that can achieve all the services and process as show in Fig.11. Emergency situation was well sensed by the Android mobile with regards to the heart rate based, accidents and personal safety. Each individual emergency scenario was researched, designed and developed. Person's heart beat was monitored using a specialized HR device which sends the heart beat rate to the mobile which in turn makes a decision with regards to the abnormal heart rate and sends an alert signal to the arduino microcontroller, simultaneously a message was sent to a pre-stored number, email address and was posted on Facebook wall successfully.

Similarly accident based emergency scenario and Personal safety can be incorporated and alert message which contains the GPS location information be sent via SMS, email and message was successfully posted on respective user's Facebook wall. Hence, Android once again proved to be a versatile operating system which allowed us to manipulate various inbuilt features of an Android mobile which made us to develop an intelligent application called as ETS.

References

- [1] Gozalvez, J.: First Google's android phone launched. *IEEE Vehicular Technology Magazine* 3 (September 2008)
- [2] <http://www.xcubelabs.com/evolution-of-mobile-operating-systems.php> (accessed on June 10, 2011)
- [3] Conti, J.P.: The Androids are coming. *IEEE Engineering and Technology* 3(9), 72–77 (2008)
- [4] Open Handset Alliance, Open Handset Announces 14 New Members, http://www.openhandsetalliance.com/android_overview.html (accessed on June 10, 2011)
- [5] Shu, X., Du, Z., Chen, R.: Research on Mobile Location Service Design Based on Android. In: 5th Int. Conf. On Wireless Communication Networking and Mobile Computing, Beijing, pp. 1–4 (2009)
- [6] Whipple, J., Arensman, W., Boler, M.S.: A Public Safety Application of GPS-Enabled smart phones and the Android Operating System. In: IEEE Int. Conf. on System, Man and Cybernetics, San Antonio, pp. 2059–2061 (2009)
- [7] Yang, C.T., Chu, Y.Y., Tsaur, S.C.: Implementation of a medical information service on Android mobile device. In: 4th Int. Conf. on News Trends in Information Science and Service Science, Gyeongju, pp. 72–77 (2010)
- [8] Hou, Q.: Research and implementation of remote heart rate monitoring system based on GSM and MCU. In: 2nd Int. Conf. On Information Science and Engineering, Hangzhou, p. 2293 (2010)
- [9] <http://www.arduino.cc> (accessed on June 10, 2011)
- [10] Neuman, M.R.: Vital Signs: Heart Rate. *IEEE Pulse* 1(3), 51–55 (2010)
- [11] Wen, Y., Yang, R., Chen, Y.: Heart rate monitoring in dynamic movements from a wearable system. In: 5th Intl. Summer School and Symposium on Medical Devices and Biosensors, June 1-3, pp. 272–275 (2008)

- [12] Mahmood, N.H., Uyop, N., Zulkarnain, N., Harun, F.K.C., Kamarudin, M.F., Linoby, A.: LED indicator for heart rate monitoring system in sport application. In: IEEE 7th Int. Colloquium on Signal Processing and its Applications, pp. 64–66, 4–6 (2011)
- [13] <http://www.heartratemonitor-app.co.uk/index.html> (accessed on June 10, 2011)
- [14] <http://www.arduino-toolkit.net> (accessed on June 10, 2011)

Performance Analysis of (AIMM-I46) Addressing, Inter-mobility and Interoperability Management Architecture between IPv4 and IPv6 Networks

Gnana Jayanthi Joseph¹ and S. Albert Rabara²

¹ Dept. of Computer Applications, JJ College of Engineering and Technology,
Tiruchirappalli, Tamilnadu, India-620009

jgnanamtcy@yahoo.com

² Dept. of Computer Science, St. Joseph's College,
Tiruchirappalli, Tamilnadu, India-620002

Abstract. Transition and Mobility management has been a growing concern with numerous problems originating from roaming between IPv6 and IPv4 access networks owing to ever-growing research. Hence the various architectures concerning transition/mobility among IPv6 and IPv4 are studied. It is an observed fact that, there is still a need for more research to be done on the IPv6 transition in order to solve many problems that are not yet resolved. Henceforth, AIMM-I46 architecture has been proposed and designed as an integrated IPv4/IPv6 addressing, mobility and transition mechanism. In this paper, the performance of AIMM-I46 architecture analyzed when different data size, link capacity are used. Adventnet Network Simulator Toolkit 7 is used to evaluate the performance of AIMM-I46, using different performance evaluation metrics such as data loss rate, throughput, packet overhead and latency which have significant implications on any protocol performance. The simulated results shows that the proposed architecture provides inter mobility and interoperability between IPv4 and IPv6 networks.

Keywords: Mobility Architecture, IPv4-IPv6 Interoperability, Address Mapping, Mixed Network, IPv4 to IPv6 Address Mapping.

1 Introduction

The Internet Protocol version 4 (IPv4) has proven to be remarkably popular, robust, easily implemented, interoperable, and has served the Internet well for over 25 years. However, on account of investigated IPv4 address space constraints, security reasons, etc., cellular phone service providers needed to adopt the Internet Engineering Task Force (IETF) proposed IPv6, the next generation of network protocol [1] supporting the deployment of new applications over the Internet and thereby open-up a broad field of technological development [2, 3]. Companies such as Cisco, Microsoft and Nokia have issued white papers on accelerating IPv6 progress [4, 5, 6].

Realizing the need for IPv6 transition and the significant features of IPv6, several countries have started initiating IPv6 deployment in their countries. Countries such as Europe, China, and Japan took the lead for deploying IPv6 in their networks [7].

Several research projects on IPv6 are being carried out all over the world. Europe's 6INIT project [8], US's 6REN/6TAP project [9], Japan's WIDE project [10] and WIDE project [11] are a few of the IPv6 projects. The IPv6 research center was established at Birla Institute of Technology and Science (BITS), Pilani in India.

IPv6 has been deployed in JAPAN in 2001, in CHINA in 2006, in EUROPE in 2007, in FRANCE in 2008, in INDIA and USA in 2009 and in KOREA in 2010. The Chinese government, published the Olympic Games in Beijing, through the IPv6 Internet Protocol address at <http://ipv6.beijing2008.cn/en> (2001:252:0:1::2008:6 and 2001:252:0:1::2008:8) in 2008. US Government drafted a roadmap for IPv6 transition between 2010-2011 and targeted to have its network services to be available over IPv6 by the year 2012. Australian Government has prepared a draft for IPv6 deployment by the end of the year 2013 [12].

Many countries have not yet initiated IPv6 transition and deployment in their countries. Therefore, IPv4 to IPv6 transition will take some time and during the IPv6 transition time, both IPv4 and IPv6 will co-exist [13]. To provide communication between IPv6 and IPv4 networks, while IPv4 and IPv6 networks co-exist, IETF Next Generation TRANSITION (NGTRANS) WG proposed three transition mechanisms such as dual stack, tunneling and translation [13].

Various researches were proposed in implementing the IPv6 transition mechanisms but each one has its limitations. These researches have provided the communication between IPv4 nodes roaming in IPv4 based networks and IPv6 nodes roaming in IPv6 based networks but have not provided interoperability and mobility services. The research work Mobility management based on mobile IP in combined IPv4/IPv6 networks is dual stacked in which IPv4 address must be allocated for all IPv6 nodes. Until now, there is no proper proposal for interoperability, inter-mobility and addressing management between IPv4/IPv6 networks [14].

The nodes mobility is supported with Mobile Internet Protocols at the network layer and it defines MIP operations according to each IP version (i.e. MIP for IPv4 nodes mobility and MIPv6 for IPv6 nodes mobility) [15, 16]. However, MIPv4 and MIPv6 cannot support the mobility of mobile nodes when mobile nodes move between two different IP networks. The mobile IP can be used only when the mobile node moves within the same IP network because the MIPv4 protocol is not compatible with the MIPv6 protocol. The existing IPv4–IPv6 mobility protocols do not provide solutions for the following scenarios like when mobile node roams into different IP network, it should be attached to the new IP visited network by means of address configuration; for Inter-mobility of mobile nodes between different IP networks; and for Interoperability of mobile nodes with other IP networks. This serves as the motivation to design, implement and study the performance of a mechanism AIMM-I46.

This paper is organized as follows: Section 2 reviews a study on the IETF's transition mechanisms and the architectures implemented them. Section 3 discusses the IMM-I46 proposed system. Section 4 describes the implementation details with the algorithm required. Section 5 concludes and spells out the future direction of research developments.

2 Related Work

IETF have proposed three transition mechanisms namely *Dual Stack*, *Tunneling and Translation* have been developed for managing the transition from IPv4 to IPv6 and vice versa. Other researchers have also proposed architectures implementing dual stack mechanism such as Dual Stack Transition Mechanism (DSTM) [17], Dual Stack MIPv4 [18], Dual Stack MIPv6 (DSMIPv6) [19]; architectures implementing tunneling mechanism such as RoamIP [20], Virtual Overlay [21], 64 Translation as Residential Gateway [22]; architectures implementing translation mechanism such as Network Address Translation-Protocol Translator (NAT-PT) [23, 24], Network Address Port Translation-PT (NAPT-PT) protocol [25], Stateless IP/ICMP Translation Algorithm (SIIT) [26], Bump in the Stack (BIS) [27], Bi-directional Mapping System (BDMS) [28], Transport Relay Translator (TRT) [29] and SOCKS64 [30], Bump-In the-API (BIA) [31], Mobile Internet Protocol-Application Level Gateway (MIP-ALG) [32].

These transition architectures do not consider a scenario like IPv6 nodes roaming in IPv4 network; IPv4 nodes roaming in IPv6 network and initiate communications with other nodes irrespective of IP version of network. The proposed system IMM-I46 meets these issues by providing inter-mobility as well as inter-operability while integrating IPv4 and IPv6 networks. The IMM-I46 system permits both IPv4 and IPv6 mobile users to roam freely either into IPv4 based networks or IPv6 based networks, get serviced and connected with internet.

3 AIMM-I46 Architecture

The AIMM-I46 architecture, depicted in Fig. 1, is the integration of the architecture proposed in [33, 34, 35] and aims to permit an IPv6 and IPv4 mobile node to roam into either IPv4 based network or IPv6 based network and get serviced.

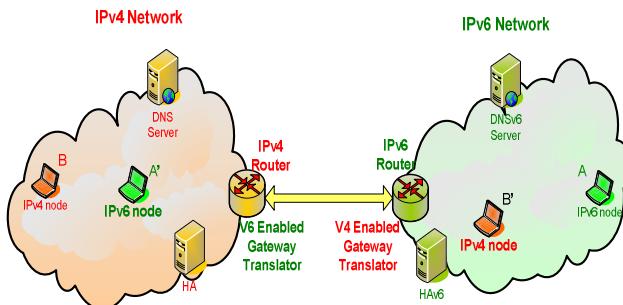


Fig. 1. Proposed AIMM-I46 Architecture

This new architecture is based on the network level translation and depends on the two main components:

- v4 Enabled Gateway Translator in IPv6 border router
- v6 Enabled Gateway Translator in IPv4 border router

4 AIMM-I46: Performance Evaluations

4.1 Simulation Setup

The proposed research has been simulated using Network Simulator Toolkit “Adventnet WebNMS” version 7. IPv4 LAN and IPv6 LAN are created. Both IPv4 and IPv6 routers used are the Cisco 7600. IPv4 Switch used is the Cisco 3750. Windows 2000 is used in all IPv4 nodes. Linux is used in all IPv6 nodes. Command-Line Interface (CLI) is set with 2323 in the IPv4 router, IPv4 switch and in IPv6 router. Port addresses are set with default values in all IPv6 nodes where Telnet port was set to 2424, Trivial File Transfer Protocol (TFTP) is set to 6969. Simple Network Management Protocol (SNMP) is set to 8001 in all IPv4 and IPv6 nodes. All the proposed interfaces have been implemented using Java and Python scripting languages.

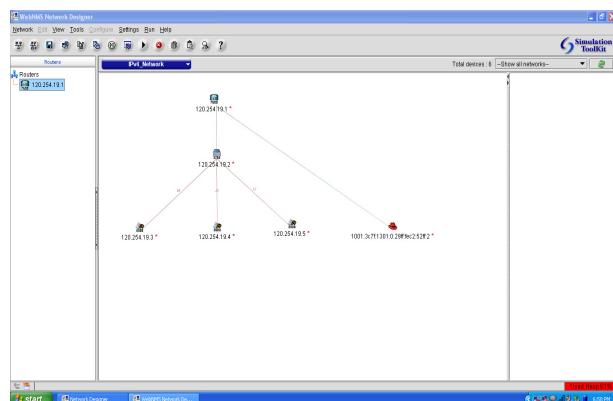


Fig. 2. IPv4 Network Topology in Simulation Environment

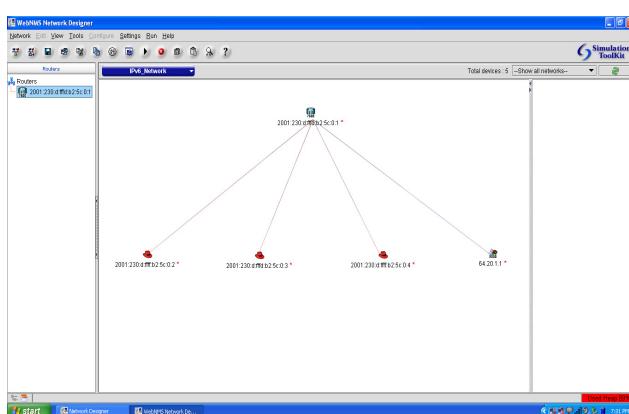


Fig. 3. IPv6 Network Topology in Simulation Environment

The proposed interfaces developed for V6EGT and V4EGT have been incorporated into the IPv4 router and IPv6 router functionalities. The parameters considered for the simulations are (i) packet size (512, 1024, ... 524288 in bytes); (ii) estimated delay; (iii) IPv4 router Maximum Transmission Unit (MTU) (set to 576 bytes); (iv) IPv6 router MTU (set to 1280 bytes).

4.2 Simulation Scenario

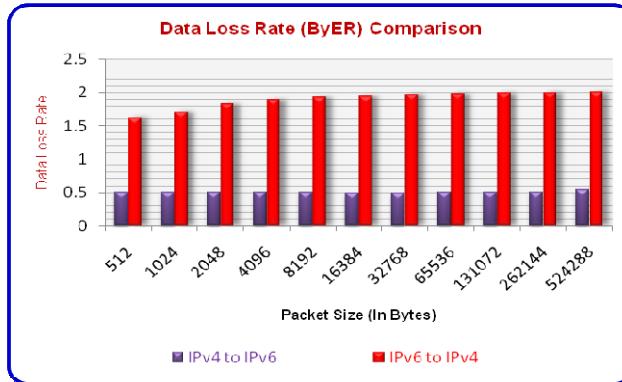
To illustrate the architecture proposed in [43], it is sufficient to create two LAN networks as in figure (2) and figure (3) namely i) IPv4 LAN (*with IPv4 nodes and IPv4 router*) and ii) IPv4 LAN (*with IPv6 nodes and IPv6 router*) with their corresponding boundary router each, shared by all connections. IPv4 LAN created is allowed to have IPv6 node along with IPv4 nodes and similarly IPv6 network is allowed to have also IPv4 node along with IPv6 nodes. The protocols of the v6enabled gateway translator and v4enabled gateway translator, IPv4 and IPv6 router functionalities are included as the scripts in the IPv6 router routine and IPv4 router routine respectively, in order to allow the IPv6 nodes to also roam into IPv4 network, initiate communication with IPv4 nodes and to allow the IPv4 nodes to also roam into IPv6 network, initiate communication with IPv6 nodes.

4.3 Simulation Results and Performance Analysis

The primary goal of this simulation experiment is to investigate the behavior of IPv4 nodes initiated communication with IPv6 destined nodes, and IPv6 nodes initiated communication with IPv4 destined nodes, where the source and the destination nodes roaming in different IP versioned networks. However, the IPv4 initiated communication with IPv4 destined communication and viz. are also simulated, the study and analysis are focused on the communications between the IPv4 and IPv6.

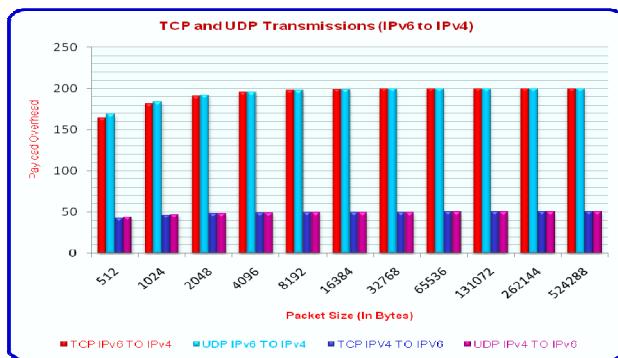
4.3.1 Data Loss in Byte-Per-Rate (ByER) Analysis

Data Loss in Byte-Per-Rate (ByER) is studied from the ratio between the numbers of bytes received at the receiving node to the total number of bytes transferred from the source node whenever the communication takes place from IPv4 node to the IPv6 node and also from IPv6 node to the IPv4 node. Fig. 4 shows the simulated results for the data loss rate analysis. The data loss rate for the communication between IPv4 nodes and the communication between IPv6 nodes were found apparently equal where the data loss rate is 10%. IPv4 packets were successfully sent from IPv4 source nodes to the IPv6 destined nodes and the data loss rate for the communication from IPv4 node to IPv6 node is very less where the data loss rate is 5%. Many IPv6 packets were successfully sent from the IPv6 source node to the IPv4 destined nodes. However, some IPv6 packets were lost while the communication is from the IPv6 source node to the IPv4 destined nodes. The data loss rate for the communication from IPv6 node to IPv4 node is observed with much loss as the IPv6 packet length is higher than the IPv4 packet length; the IPv4 router buffer size is lesser than the IPv6 router buffer size and the data loss rate found is about 18% and when the packet size is increased, the data loss rate was also increased.

**Fig. 4.** Data Loss Comparison

4.3.2 Max Payload Overhead Data Rate Analysis

The simulated results for the Payload Overhead Data Rate for UDP Transmissions analysis are plotted in graph given in fig. 5. They indicate that both TCP and UDP payload overhead data rate increases as the packet size increases for the communications taking place from IPv4 node to IPv6 node; and from IPv6 node to IPv4 node. Both TCP and UDP payload overhead data rates are much less when the communications take place from IPv4 node to IPv6 node than the other communications taking place from IPv4 node to IPv4 node; from IPv6 node to IPv6 node; and from IPv6 node to IPv4 node. A comparison is also made between TCP payload and UDP payload for the communication taking from IPv4 node to IPv6 node and from IPv6 node to IPv4 node. The graphs show that in both TCP and UDP transmissions, the payload overhead is high for the communications from IPv6 node to IPv4 node. This is due to the many fragmentations have to take place for the IPv6 packets than that of IPv4 packets.

**Fig. 5.** Payload Data-Rate Comparison

4.3.3 Throughput Analysis

Factors affecting throughput such as differences between IPv6 and IPv4 stacks, processing overhead and delays are not reflected in the theoretical values. The

throughput generally increases with packet size. The maximum throughput is reached for the largest packet sizes. Simulated results plotted in the graph are presented in fig. 6. Packets transmission from IPv6 node to IPv4 node exhibits the best throughput performance. It is observed that the maximum throughput is reached for the largest packet sizes both in IPv4 and IPv6. However, the results show that the IPv4 to IPv6 communication exhibits the best throughput performance. However, factors affecting throughput such as differences between IPv6 and IPv4 stacks, processing overhead and delays are not reflected in the theoretical values. It is also observed that the maximum throughput is reached for the largest packet sizes.

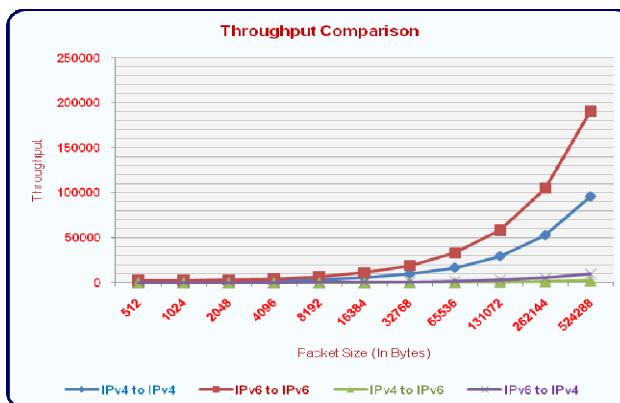


Fig. 6. Throughput Comparison

4.3.4 Latency Analysis

Graph in fig. 7 indicates clearly that packet transmission from IPv4 node to IPv6 node has the least latency and that packet transmission from IPv6 node to IPv4 node has the highest latency. This is because IPv6 router MTU is too big and at the receiving end where the IPv4 router MTU is 576 bytes. Therefore the IPv6 router receives an ICMPv6 “packet too big” message, retransmits the MTU discover packet with a

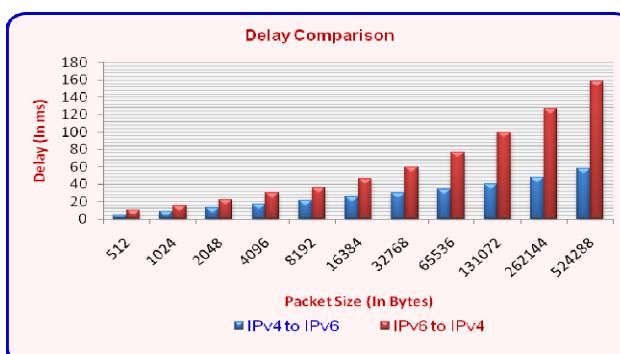


Fig. 7. Delay Comparison

smaller MTU. This process is repeated until the IPv6 node receives a response that the discover packet arrived whole transmitted packet. Hence there was such a delay for the communication from IPv6 node to IPv4 node.

However, all the IPv6 packets were transmitted to the destined IPv4 nodes. But IPv4 packets were easily transmitted to the IPv6 destined node with a negligible amount of delay for the smaller size packets and larger size IPv4 packets were transmitted with the considerable delay.

5 Conclusion

The proposed architecture in this research work is a novel architecture designed for facilitating operability and mobility between IPv6 and IPv4 networks. This architecture is specifically designed to facilitate either IPv6 or IPv4 nodes to roam freely without any network restrictions. The proposed architecture in particular, guarantees for both IPv6 and IPv4 initiated as well as destined communications. The architecture is tested by establishing a test bed with various packet sizes when the communication is from IPv4 nodes roaming into IPv6 networks to IPv6 nodes wherein the communication is established from IPv6 and IPv4 networks. Further, there is no need of any additional hardware equipment to be included in the existing IPv4 networks and IPv6 networks. The interfaces developed need to be installed both in clients as well as in routers which act as the servers. Hence the proposed approach is very economical. Further research is required to integrate the AIMM-I46 proposal with network applications and services for clients. The proposed architecture can be extended for further research with security mechanisms for the interoperability between IPv6 and IPv4 nodes that are roaming in different IP networks. A comparative study may be done with popular architecture “NAT-PT” and recent architecture “BDMS”.

References

1. Bradner, S., Mankin, A.: The Recommendation for the IP Next Generation Protocol. RFC 1752 (January 1995)
2. Esaki, H., Kato, A., Murai, J.: R&D Activities and Testbed Operation in WIDE Project. In: Proceedings of Symposium on Applications and the Internet, pp. 172–177 (January 2003)
3. Allied Telesis, IPv6 White Paper (2007),
http://www.alliedtelesyn.com/media/pdf/ipv6_wp.pdf
4. Cisco, “IPv6 Assessment and Migration Services”, Whitepaper (2005)
5. Microsoft, “IPv6/IPv4 Coexistence and Migration,” White Paper (August 2002)
6. Nokia, “IPv6-Enabling the Mobile Internet,” White Paper (September 2001)
7. IPv6 Task Force Editorial Group, Main Task force Report, Version 1.76, Document no.70 (February 11, 2002)
8. EU fifth frame work, IPv6 Internet Initiative <http://www.6init.com/>
9. US IPv6 Research and Education Network (6REN), <http://www.6ren.net/>
10. Japan’s WIDE project, <http://www.wide.ad.jp/about/index.html>
11. <http://playground.sun.com/pub/ipng/html/ipng-implementations.html>

12. The Business Case and Roadmap for Completing IPv6 Adoption in US Government, Architecture and Infrastructure Committee, Federal Chief Information Officers Council, Version 0.1 (December 22, 2008),
http://osrin.net/docs/DRAFT_Business_Case_&_Roadmap_for_Completing_IPv6_Adoption_in_USG_12242008.pdf
13. Dunn, T.: The IPv6 Transition, Market place. IEEE Internet Computing 6(3), 11–13 (2002)
14. Bi, J., Wu, J., Leng, X.: IPv4/IPv6 Transition Technologies and Univer6 Architecture. International Journal of Computer Science and Network Security 7(1), 232–242 (2007)
15. Perkins, C.: IP Mobility Support, RFC 2002 (October 1996)
16. Perkins, C., Johnson, D., Arkko, J.: Mobility Support in IPv6, RFC 3775 (June 2004)
17. Bound, J., Toutain, L., Richier, J.: Dual Stack IPv6 Dominant Transition Mechanism (DSTM), Internet Draft, draft-bound-dstm-exp-04.txt (October 2005)
18. Tsirtsis, G., Park, V., Soliman, H.: Dual Stack Mobile IPv4, Internet Draft, draft-ietf-mip4-dsmipv4-06, IETF (February 2008)
19. Solima, H.: Mobile IPv6 support for dual stack Hosts and Routers, Internet Draft, draft-ietf-mext-nemo-v4traversal-03, IETF (May 2008)
20. Turanyi, Z.R., Szabo, C.: Global Internet Roaming with RoamIP. ACM Journal on Mobile Computing and Communications Review 4(3), 58–68 (2001)
21. Liu, C.: Support Mobile IPv6 in IPv4 Domains. In: Proceedings of the International Conference on Vehicular Technology (VTC), pp. 2700–2704 (May 2004) ISBN: 7803-8255-2/04
22. Seo, S.-H., Kong, I.-Y.: A Performance Analysis Model of PC-based Software Router Supporting IPv6-IPv4 Translation for Residential Gateway. In: Proceedings of the Fourth Annual ACIS International Conference on Computer and Information Science, ICIS 2005 (2005) 0-7695-2296-3/05
23. Aoun, C., Davies, E.: Reasons to Move the Network Address Translator - Protocol Translator. RFC 4966 (July 2007)
24. Bangnulo, M., Matthews, P., van Beijnum, I.: NAT64/DNS64: Network Address and Protocol Translation from IPv6 Clients to IPv4 Servers", Internet Draft, draft-bagnulo-behave-nat64-00 (June 2008)
25. Srisuresh, P., Egevang, K.: Traditional IP Network Address Translator (Traditional NAT). RFC 3022 (January 2001)
26. Nordmark, E.: Stateless IP/ICMP Translation Algorithm (SIIT), RFC 2765 (February 2000)
27. Mackay, M., Edwards, C., Dunmore, M., Chown, T., Carvalho, G.: A Scenario-Based Review of IPv6 Transition Tools. IEEE Internet Computing, 27–35 (May-June 2003) 1089-7801/03
28. AlJa'afréh, R., Mellor, J., Awan, I.: Implementation of IPv4/IPv6 BDMS Translation Mechanism. In: IEEE Proceedings of the Second UKSIM European Symposium on Computer Modeling and Simulation, pp. 512–517 (2008) ISBN 978-0-7695-3325-4/08
29. Hagino, J., Yamamoto, K.: An IPv6-to-IPv4 Transport Relay Translator, RFC 3142 (June 2001)
30. Kitamura, H.: A SOCKS-based IPv6/IPv4 Gateway Mechanism, RFC 3089 (April 2001)
31. Lee, S., Shin, M.-K., Kim, Y.-J., Nordmark, E., Durand, A.: Dual Stack Hosts Using "Bump-in-the-API" (BIA), RFC 3338 (October 2002)
32. Choi, H.H., Cho, D.H.: Mobility management based on mobile IP in mixed IPv4/IPv6 networks. In: IEEE 58th Proceedings of VTC 2003-Fall, pp. 2048–2052 (October 2003)

33. Gnana Jayanthi, J., Rabara, S.A.: IMM-I46: Inter Mobility Management for an Integrated IPv4 and IPv6 Network. In: Das, V.V., Vijayakumar, R., Debnath, N.C., Stephen, J., Meghanathan, N., Sankaranarayanan, S., Thankachan, P.M., Gaol, F.L., Thankachan, N. (eds.) BAIP 2010. CCIS, vol. 70, pp. 17–21. Springer, Heidelberg (2010) ISBN: 978-3-642-12213-2,
34. Gnana Jayanthi, J., Albert Rabara, S.: IPv6 Addressing Architecture in IPv4 Network. In: IEEE Proceedings of the Second International Conference on Communication Software and Networks (ICCSN 2010), Singapore, pp. 461–465 (2010) ISBN: 978-0-7695-3961-4/10
35. Gnana Jayanthi, J., Albert Rabara, S.: IPv4 Addressing Architecture in IPv6 Network. In: IEEE Proceedings of the Second International Conference on Advanced Computer Control (ICACC 2010), China, pp. 282–287 (2010) ISBN: 978-1-4244-5845-5

Strong Neighborhood Based Stable Connected Dominating Sets for Mobile Ad Hoc Networks

Natarajan Meghanathan^{1,*} and Michael Terrell²

¹ Jackson State University

Jackson, MS, USA

natarajan.meghanathan@jsu.edu

² Grambling State University

Grambling, LA, USA

Abstract. We propose an algorithm to determine stable connected dominating sets (SN-CDS) for mobile ad hoc networks (MANETs) using the notion of a “strong neighborhood,” defined based on a “threshold neighborhood distance ratio” ($TNDR \leq 1$). A node j at a physical Euclidean distance of r from node i is said to be in the strong neighborhood of node i if $r/R \leq TNDR$ where R is the fixed transmission range of all nodes in the network. A non-CDS node is said to be covered if at least one of its neighbor nodes is in the SN-CDS. The proposed algorithm prefers to include a covered node with the maximum number (≥ 1) of uncovered neighbors into the SN-CDS; ties are broken using node ids. The algorithm stops when every node is either in the SN-CDS or has at least one neighbor node in the SN-CDS. If $TNDR = 1$, then SN-CDS corresponds to the maximum-density based CDS (MaxD-CDS) algorithm, a heuristic to approximate a CDS with the minimum number of constituent nodes. We observe that an SN-CDS (with $TNDR < 1$) has a significantly longer lifetime than a MaxD-CDS and for a given condition of network density and node mobility, the difference in the lifetime increases as the value of $TNDR$ decreases. The tradeoff is lower connectivity as well as a larger constituent node size and hop count per path.

Keywords: Strong Neighborhood, Connected Dominating Set, Stability, Mobile Ad hoc Networks, Maximum Density, Graph Theory Algorithm.

1 Introduction

A mobile ad hoc network (MANET) is a dynamic distributed system of arbitrarily moving wireless nodes that operate with limited battery charge and resource constraints (like limited bandwidth, memory and processing capacity). MANET routes are multi-hop in nature due to the limited transmission range of the wireless nodes. Due to node mobility, these paths have a limited lifetime and have to be often reconfigured to continue communication. Due to the resource and mobility constraints, MANET routing protocols prefer to discover a route only when needed rather than proactively

* Corresponding author.

determining and maintaining them [1][2]. On-demand route discovery is conducted through a flooding-based route-request-reply cycle wherein a source node (that has data to send to a destination or to a multicast group) broadcasts a Route Request (RREQ) message to its neighbors [3].

Flooding maximizes the chances that at least one RREQ message reaches the destination (including on the minimum-hop path to the destination). However, flooding is often considered to trigger the “broadcast storm problem” [4] as it is associated with redundant retransmissions; because, each node in the network broadcasts the RREQ message once to its neighborhood. Several strategies have been considered to minimize the broadcast storm problem and the most significant and effective among them is to use a connected dominating set (CDS) of the underlying network graph at the time of route discovery (e.g., [5][6][7]). A CDS of a network of nodes is the subset of nodes such that every node in the network is either in the CDS or is a neighbor node (that is located within the transmission range) of a node in the CDS [8]. A non-CDS node is said to be covered if it has at least one neighbor node in the CDS. If we could find a CDS that has the least number of constituent nodes (called the Minimum CDS, MCDS) to cover all the nodes in the network, we can incur the minimum number of retransmissions if only the nodes constituting the CDS broadcast a message (such as the RREQ message) in the neighborhood. However, the problem of determining a MCDS is considered to be NP-complete [8] and hence several heuristics have been proposed to approximate a MCDS. The Maximum density-based CDS (MaxD-CDS) algorithm [9] that was earlier proposed by us is one such heuristic. The common thread among the heuristics proposed for approximating the MCDS (including the MaxD-CDS algorithm) is to include a covered node that has the largest number of uncovered neighbor nodes into the CDS, during each of the iterations of the algorithm until all the nodes in the network are covered. This way, it has been observed [9] that the approximated MCDS will have the lowest or closer to the lowest number of constituent CDS nodes.

The motivation for the research leading to this paper stemmed from observations in our earlier research (e.g. [9][10]) that an MCDS (like MaxD-CDS) is significantly unstable in the presence of node mobility when used in the context of a MANET. This could be attributed to the requirement to cover all the nodes of a network with the lowest possible number of constituent MCDS nodes and with node mobility, the chances of a non-CDS node not having a MCDS node as its neighbor in the near future is quite high. Also, since the MCDS spans the entire network and is composed of the least number of constituent nodes, the physical Euclidean distance between any two MCDS nodes is often close to the transmission range of the nodes at the time of the CDS construction and is vulnerable to break (i.e., the two MCDS nodes move out of their transmission range) at any time in the near future.

In this paper, we propose the notion of considering only the “Strong Neighborhood” (SN) of a node, rather than considering all neighbor nodes within the transmission range of the node (referred to as the “Open Neighborhood”), as the neighbor nodes that will be covered due to the inclusion of a node into the CDS. Accordingly, our CDS is referred to as the SN-CDS. The Strong Neighborhood of a node is defined based on a parameter called the Threshold Neighbor Distance Ratio (TNDR) that can be at most 1. A node j at a physical Euclidean distance of r from node i is said to be in the strong neighborhood of node i if $r/R \leq TNDR$ where R is the

fixed transmission range of all nodes in the network (i.e., we assume a homogeneous MANET). The physical Euclidean distance between two nodes i and j located at (X_i, Y_i) and (X_j, Y_j) respectively is given by $dis(i, j) = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2}$.

The construction of the SN-CDS starts with the inclusion of the node with the maximum number of strong neighbors into the CDS. As usual for any CDS algorithm, a non-CDS node is said to be covered if at least one of its neighbor nodes is in the SN-CDS. The set of covered nodes are considered to be the candidate nodes for inclusion into the SN-CDS and they are preferred in the decreasing order of the number of uncovered nodes in their strong neighborhood. We continue the inclusion of the covered nodes into the SN-CDS, one node at a time, until all the nodes in the network are covered by at least one node in the SN-CDS. If $TNDR = 1$, then SN-CDS corresponds to the MaxD-CDS. Hence, for purpose of clarity in the rest of the paper, we refer to an SN-CDS as a CDS formed using $TNDR < 1$ and a CDS formed using $TNDR$ of 1 is referred to as MaxD-CDS.

2 Algorithm to Construct Strong Neighborhood-Based Connected Dominating Set

The SN-CDS algorithm uses the following auxiliary variables and functions:

- *SN-CDS-Node-List* – includes all the nodes that are part of the strong neighborhood based CDS, initialized to Φ .
- *Covered-Nodes-List* – includes nodes that are either in *SN-CDS-Node-List* or covered by a node in the *SN-CDS-Node-List*, initialized to Φ .
- *Uncovered-Nodes-List* – includes nodes that are not covered by any node in the *SN-CDS-Node-List*, initialized to the Vertex set of the input static graph.
- *Priority-Queue* – includes nodes that are in the *Covered-Nodes-List* and not in the *SN-CDS-Node-List*. The nodes in the priority queue are probable candidates for inclusion to the *SN-CDS-Node-List*. The list is sorted in the decreasing order of the number of uncovered strong neighbors of the nodes. A dequeue operation returns the node with the largest number of uncovered strong neighbors; if there is a tie, then a node is randomly chosen from among the contending nodes. The priority queue is initialized to Φ .
- *MaxUncoveredStrongNeighbors* – the maximum value for the number of uncovered strong neighbors of any node in the network at the beginning of the algorithm; initialized to $-\infty$ to start with and is computed while determining the strong neighbors of the nodes in the network.
- *UncoveredStrongNeighbors(u)* – the set of all strong neighbors of node u that are not yet covered. Initially, $\forall u, \text{uncoveredStrongNeighbors}(u) = \Phi$.

2.1 Description of the SN-CDS Algorithm, Time Complexity and Validation

The SN-CDS construction algorithm (pseudo code in Figure 3) primarily works as follows: The algorithm inputs a snapshot of the network (referred to as the static graph $G = (V, E)$ with vertex set V and edge set E) at the time instant during which we want to determine the SN-CDS. The algorithm also inputs the Threshold Neighborhood

Distance Ratio (*TNDR*) and transmission range (*R*) for the network. For every node in the vertex set *V*, the nodes constituting the strong neighborhood is constructed by calling the *Compute-Strong-Neighborhood* function (pseudo code in Figure 1). This function also returns the maximum value for the number of uncovered strong neighbors for any node in the network, referred to as *maxUncoveredStrongNeighbors*. The node that has this maximum value of uncovered strong neighbors (i.e., the size of the *Strong Neighborhood* set of the node equals *maxUncoveredStrongNeighbors*) is selected as the *Start Node* (the first node to be included into the *SN-CDS-Node-List*) by calling the *Choose-Start-Node* function (pseudo code in Figure 2). If there is a tie, the *Choose-Start-Node* function randomly returns one among the contending nodes as the *Start Node*.

Input: Vertex set, V , of the network graph; $TNDR$ and R
Output: $maxUncoveredStrongNeighbors$ and the Strong Neighborhood set for every vertex in V

```

Begin Compute-Strong-Neighborhood
  for every vertex  $u \in V$  do
    for every vertex  $v \in V$  do
      if  $dist(u, v) \leq TNDR$  then
         $R = SN_u \cup \{v\}$ 
      end if
    end for // loop for every possible neighbor node  $v$  of  $u$ 
    if  $maxUncoveredStrongNeighbors < |SN_u|$  then
       $maxUncoveredStrongNeighbors = |SN_u|$ 
    end if
  end for // loop for vertex  $u$ 
  return  $maxUncoveredStrongNeighbors$ 
End Compute-Strong-Neighborhood

```

Fig. 1. Pseudo Code to Compute the Strong Neighborhood of the Nodes

Input: Vertex set (V) of the network graph, $maxUncoveredStrongNeighbors$
Output: *Start Node* for inclusion to the *SN-CDS-Node-List*

```

Begin Choose-Start-Node
  Contending-Node-List =  $\emptyset$ 
  for every vertex  $u \in V$  do
    if  $(maxUncoveredStrongNeighbors \text{ equals } |SN_u|)$  then
      Contending-Node-List = Contending-Node-List  $\cup \{u\}$ 
    end if
  end for
  Generate a random integer  $rand$  from the range  $\{1, \dots, |\text{Contending-Node-List}|\}$ 
  Start Node = Contending-Node-List[ $rand$ ]
  return Start Node
End Choose-Start-Node

```

Fig. 2. Pseudo Code to Choose the Starting Node for the SN-CDS

When the *Start Node* is added to the *SN-CDS-Node-List*, all of its strong neighbors are said to be covered; these nodes are removed from the *Uncovered-Nodes-List* and added to the *Covered-Nodes-List* and to the *Priority-Queue*. If both the *Uncovered-Nodes-List* and the *Priority-Queue* are not empty, we dequeue the *Priority-Queue* to extract a node s that is not yet in the *SN-CDS-Node-List* and has the largest number of uncovered strong neighbor nodes. All the uncovered strong neighbor nodes of s are now removed from the *Uncovered-Nodes-List* and added to the *Covered-Nodes-List* as well as to the *Priority-Queue*. The number of uncovered strong neighbors of each node in the network is then updated based on the additional node coverage obtained during the iteration and accordingly, the *Priority-Queue* is re-sorted in the decreasing order of the number of uncovered strong neighbors of the nodes in the queue. The above procedure is repeated for several iterations until the *Uncovered-Nodes-List* becomes empty or the *Priority-Queue* becomes empty.

Note that during an iteration, if the node s extracted from the *Priority-Queue* has all its strong neighbor nodes already covered, then it implies that all the other nodes, if any, in the *Priority-Queue* also have “zero” uncovered strong neighbor nodes. However, we have not yet broken from the while loop (i.e. the *Uncovered-Nodes-List* is not yet empty), indicating that the underlying network based on the strong neighborhood of the nodes is not connected and hence the algorithm returns NULL

(i.e. a SN-CDS for the entire network does not exist). Also, even after exiting from the while loop, if the *Priority-Queue* becomes empty and the *Uncovered-Nodes-List* has at least one node, then the underlying network is considered to be disconnected (based on the strong neighborhood of the nodes) and the algorithm returns NULL. If the underlying network is connected based on the strong neighborhood of the nodes, then the algorithm does not return NULL and returns the *SN-CDS-Node-List* after all the nodes in the network are included to the *Covered-Nodes-List*.

Input: Static Network Graph $G = (V, E)$, $TNDR$ and R

Output: *SN-CDS-Node-List*

Begin SN-CDS Algorithm

```

max UncoveredStrongNeighbors = Compute-Strong-Neighborhood( $V$ )
Start Node = Choose-Start-Node( $V$ , max UncoveredStrongNeighbors)
Uncovered-Nodes-List = Uncovered-Nodes-List - {Start Node}
Covered-Nodes-List = Covered-Nodes-List U {Start Node}
Priority-Queue = Priority-Queue U {Start Node}

 $\forall u \in V$ ,  $uncoveredStrongNeighbors(u) = \{ v \mid v \in SN_u \}$ 

while (Uncovered-Nodes-List  $\neq \Phi$  and Priority-Queue  $\neq \Phi$ ) do
    node  $s = \text{Dequeue}(\text{Priority-Queue})$ 
    if ( $uncoveredStrongNeighbors(s) = \Phi$ ) then
        return NULL; // the underlying network is not connected
    end if

     $SN\text{-CDS-Node-List} = SN\text{-CDS-Node-List} U \{s\}$ 

    for all node  $u \in \text{Neighbors}(s)$  do
        if ( $u \in \text{Uncovered-Nodes-List}$ ) then
             $Uncovered\text{-Nodes-List} = Uncovered\text{-Nodes-List} - \{u\}$ 
             $Covered\text{-Nodes-List} = Covered\text{-Nodes-List} U \{u\}$ 
             $Priority\text{-Queue} = Priority\text{-Queue} U \{u\}$ 
        end if
    end for

     $\forall u \in V$ ,  $uncoveredStrongNeighbors(u)$ 
        =  $\{ v \mid v \in SN_u \text{ AND } v \in \text{Uncovered-Nodes-List} \}$ 

    Re-sort the entries in Priority-Queue in the decreasing order of the number
    of uncovered strong neighbors for the nodes in the queue

end while

if (Uncovered-Nodes-List  $\neq \Phi$  and Priority-Queue  $= \Phi$ ) then
    return NULL; // the underlying network is not connected
end if

return SN-CDS-Node-List
```

End SN-CDS Algorithm

Fig. 3. Pseudo Code for the SN-CDS Construction Algorithm

The time complexity of the SN-CDS algorithm and the MaxD-CDS algorithm (see Section 2.2) depends on whether we maintain the *Priority-Queue* of $|V|$ nodes as an array or a binary heap. When stored as an array, it takes $O(|V|)$ time to extract, insert an element or re-sort the array. On the other hand, all of these operations can be done on a binary heap of $|V|$ elements in $O(\log|V|)$ time. Both the MaxD-CDS and SN-CDS algorithms require all the $O(|V|)$ nodes and the $O(|E|)$ edges in the underlying network to be explored for inclusion into the CDS. Assuming that the *Priority-Queue* is implemented as a binary heap (as is done in our simulations), the overall-time complexity of both the MaxD-CDS and SN-CDS algorithms is $O(|E| + |V|\log|V|)$.

3 Simulations

We conduct our simulations in a discrete-event simulator developed in Java. The dimensions of the network topology are 1000m x 1000m. The fixed transmission range per node is 250m. The number of nodes in the network is varied by conducting simulations with 50, 100 and 150 nodes to represent networks of low, moderate and high density respectively, corresponding to an average neighborhood size of approximately 10, 20 and 30 nodes if the transmission range per node is 250m. The mobility model used in the simulations is the Random Waypoint model [12]; each node randomly chooses a destination location (within the network boundary) to move to with a velocity randomly chosen from the range $[0, \dots, v_{max}]$. Once a node reaches the targeted location, it continues to move to a different randomly chosen destination location with a velocity again randomly chosen from the above range. Each node continues to move like this for the simulation time of 1000 seconds. The movement of a node is independent of the other nodes in the network. The values of v_{max} chosen are 5 m/s and 50 m/s, representing scenarios of low and high node mobility respectively.

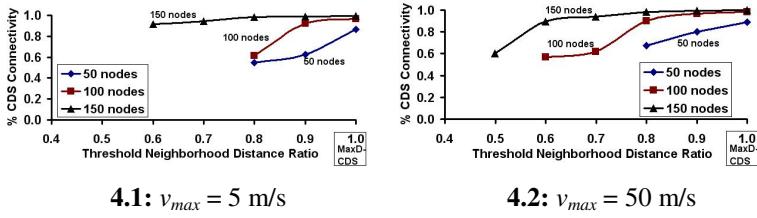


Fig. 4. CDS Connectivity vs. Threshold Neighborhood Distance Ratio

We construct snapshots of the network topology for every 0.25 seconds, starting from time 0 to the simulation time of 1000 seconds. If a CDS is not known or does not exist for the network snapshot at a particular time instant t , we run the appropriate CDS algorithm on that network snapshot. The CDS determined during a particular time instant is validated for existence during the subsequent time instants until the CDS ceases to exist. A CDS exists at a particular time instant if the nodes constituting the CDS stay connected (i.e., reachable from one another directly or through multi-hop paths) and every non-CDS node has at least one CDS node as its neighbor. The above procedure is continued until the end of the simulation time.

The algorithm to construct the MaxD-CDS would be similar to that described in the previous section for SN-CDS. The difference primarily lies in using the open neighborhood, determined based on the fixed transmission range (R) of the nodes, rather than the strong neighborhood. To begin with, the algorithm gives preference to the nodes that have the maximum number of uncovered neighbors for inclusion to the MaxD-CDS. Any tie could be broken randomly or by choosing the node with the lowest or the largest ID among the contending nodes. In this paper, we break the tie by randomly choosing a neighbor node among the contending nodes with the maximum number of uncovered neighbors (the same way ties are broken with the SN-CDS algorithm). During subsequent iterations, the covered node that has the largest number of uncovered neighbors is chosen for inclusion to the MaxD-CDS.

For the SN-CDS, the TNDR values are varied from 0.5 to 0.9; MaxD-CDS is nothing but SN-CDS operated with a TNDR value of 1.0. The number of strong neighbors for a node decreases with decrease in the TNDR value. Since the SN-CDS is constructed on a network graph with constituent edges included based on the strong neighborhood of the nodes, the connectivity of the network gets lower as the TNDR value is lowered (observed in Figure 4). Since a CDS exists as long as the underlying network is connected, we report the network connectivity as the CDS connectivity. The CDS connectivity is the ratio of the number of time instants a CDS exists for the network to that of the total number of time instants considered during the simulation.

In addition to CDS connectivity, we measure the Effective CDS Lifetime (product of the absolute CDS Lifetime and percentage CDS connectivity), CDS Node Size, CDS Edge Size and Hop Count per s - d path. Each data point in Figures 4 through 8 is an average computed over 10 mobility trace files generated for every combination of network density and node mobility values considered in the simulations.

To measure the hop count per path, we run the Breadth First Search algorithm on a CDS-induced sub graph for 15 source-destination (s - d) pairs – the role of the source or destination could be assigned to any node (CDS node or non-CDS node) in the network. The CDS-induced sub graph for a particular time instant comprises of all the nodes in the network and edges that may exist between any two CDS nodes and between a CDS node and a non-CDS node. Two non-CDS nodes have to communicate through one or more CDS nodes as intermediate nodes, even if the two non-CDS nodes are neighbors of each other. However, two CDS nodes can communicate directly if they are neighbors of each other.

For better accuracy, we collected only performance results for network connectivity of 0.5 or above. Also, we observe appreciable SN-CDS connectivity (i.e., 50% connectivity) for all three network densities (50, 100 and 150 nodes) only for TNDR values of 0.8 and 0.9. For the rest of the paper, we will refer to the SN-CDS performance obtained with a TNDR value of 0.9 while comparing with that of MaxD-CDS (TNDR = 1.0). The performance metric values at these two TNDR values are displayed in Figures 5 through 8.

Effective CDS Lifetime: Since the connectivity of the SN-CDS depends on the value of the TNDR values used, we calculate *Effective CDS Lifetime* that takes into consideration not only the stability of a particular CDS, when it exists, but also the percentage of times such CDS can be determined across the entire simulation period. We notice from Figure 5, the effective lifetime of the SN-CDS (@ TNDR = 0.9) is

always significantly larger than that of the MaxD-CDS, especially with increase in network density as well as node mobility. The effective lifetime of SN-CDS is 10% (at $v_{max} = 5$ m/s and 50 nodes) to 170% (at $v_{max} = 50$ m/s and 150 nodes) larger than that of the MaxD-CDS. The relatively high stability of the SN-CDS could be primarily attributed to the TNDR constraint and the resulting side-effect of requiring slightly more nodes (as part of the SN-CDS) to cover the rest of the nodes in the network. With the TNDR restriction, the physical Euclidean distance between the constituent end nodes of the edges included into the SN-CDS is not closer to the transmission range of the nodes; and as a result, the chances that these edges are likely to break in the near future is not high. The tradeoff is a slightly larger number of constituent nodes in the SN-CDS (i.e., the CDS Node Size) and an accompanying increase in the number of edges between the SN-CDS nodes.

The MaxD-CDS algorithm includes the bare minimum number of nodes into the CDS and the edges between these CDS nodes are highly vulnerable to break in the immediate future (due to the physical Euclidean distance between the end nodes of these edges being closer to the transmission range of the nodes). From Figures 5, 6 and 7, one can infer that with at most a 22% increase in the CDS Node Size and at most a 45% increase in the CDS Edge Size, the lifetime of the SN-CDS increases as large as 170% more than that of the MaxD-CDS. Hence, it is worth including slightly fewer nodes into the SN-CDS in order to sustain a significantly longer lifetime.

CDS Node Size: The MaxD-CDS algorithm is 100% density-based and is designed to minimize the number of constituent nodes of the CDS. On the other hand, even though the SN-CDS algorithm gives preference to include (into the CDS) nodes with a relatively larger number of uncovered neighbors, the neighborhood of a node is decided based on the TNDR. The edges constituting the *strong neighborhood* topology on which the SN-CDS is determined are relatively more stable than the edges constituting the *open neighborhood* topology on which the MaxD-CDS is determined. Thus, the SN-CDS algorithm is primarily stability-oriented and the objective of minimizing the number of constituent CDS nodes is only secondary. As a result, we do observe an increase in the SN-CDS Node Size (refer Figure 6); but, it is not a significant increase since the algorithm is still density-based and prefers to include nodes (into the SN-CDS) that have more uncovered (strong) neighbors.

CDS Edge Size: In the case of a MaxD-CDS, with the CDS nodes likely to be far away from each other (to cover all the nodes in the network with the minimal CDS Node Size), the number of edges between the CDS nodes is barely the minimum required to keep the CDS nodes connected. This contributes to the fragile nature of the MaxD-CDS; even if one or two edges (between the CDS nodes) break, the whole MaxD-CDS could get disconnected. On the other hand, the SN-CDS incorporates relatively more edges (attributable to the larger CDS Node Size) between the CDS nodes. However, since the SN-CDS also has a (secondary) objective of minimizing the CDS Node Size, there is not a plethora of nodes and edges constituting the CDS (refer Figures 6 and 7).

Hop Count per Path: On average, the SN-CDS incurs a slightly larger hop count per path (Figure 8) between any two nodes in the network. This could be attributed to the

lower physical Euclidean distance of the edges constituting the SN-CDS, compared to that of a MaxD-CDS. As a result, on average, more intermediate CDS nodes are required to connect any two randomly chosen source and destination nodes in the network. As mentioned earlier, the source and destination nodes could be any two nodes in the network and need not be the CDS nodes. The hop count per path incurred with an SN-CDS is almost the same as that incurred with a MaxD-CDS in low-density networks and the difference between the hop count per path incurred with the two CDSs increase with increase in the network density. The hop count per path determined through the SN-CDS is at most 24% more than that determined using the MaxD-CDS (noticed for network density of 150 nodes). Like the CDS Lifetime vs. CDS Node Size tradeoff observed earlier, the performance of the SN-CDS and the MaxD-CDS vis-à-vis the hop count per path could be accounted as the CDS Lifetime-Hop Count tradeoff. In both cases, the tradeoff is more favorable towards SN-CDS.

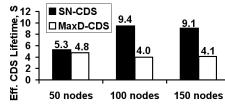
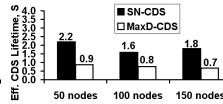
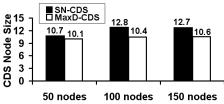
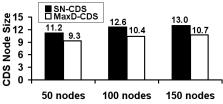
5.1: $v_{max} = 5$ m/s5.2: $v_{max} = 50$ m/s6.1: $v_{max} = 5$ m/s6.2: $v_{max} = 50$ m/s

Fig. 5. Effective CDS Lifetime

Fig. 6. CDS Node Size

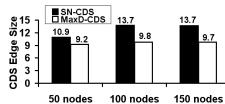
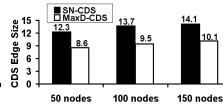
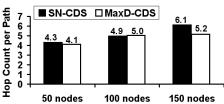
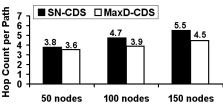
7.1: $v_{max} = 5$ m/s7.2: $v_{max} = 50$ m/s8.1: $v_{max} = 5$ m/s8.2: $v_{max} = 50$ m/s

Fig. 7. CDS Edge Size

Fig. 8. Hop Count per Path

4 Conclusions

In this paper, we have proposed a strong neighborhood-based algorithm to determine long-living stable CDS without incurring a significant increase in the CDS Node Size. The proposed SN-CDS algorithm has the same run-time complexity as the standard maximum density based MaxD-CDS algorithm. We observe the MaxD-CDS to be quite unstable (i.e., incur a lower CDS Lifetime) in the presence of node mobility. The CDS Lifetime vs. CDS Node Size tradeoff and the CDS Lifetime vs. Hop Count per Path tradeoff are both favorable towards the SN-CDS. The SN-CDS lifetime can be as large as 170% more than MaxD-CDS lifetime; with the increase (for the SN-CDS) in the CDS Node Size, CDS Edge Size and the Hop Count per path being at most 22%, 45% and 24% respectively. The tradeoff increases with increase in network density. At the same time, the relatively larger number of constituent nodes and edges (with increase in network density) makes the SN-CDS to be more robust to node mobility and link failures. Since, the hop count per path is not excessively high and there are more nodes that are part of the SN-CDS, we anticipate that by routing through the SN-CDS, we would be able to better balance the forwarding load among

the CDS nodes (compared to routing through the MaxD-CDS) and thereby, improve the fairness of node usage. On the other hand, since the MaxD-CDS induced sub graph has fewer nodes, the nodes that are part of the MaxD-CDS can be significantly over-working compared to the rest of the nodes in the network and this can lead to premature failure of the MaxD-CDS nodes.

Acknowledgments. The work leading to this paper was partly funded through the U.S. National Science Foundation (NSF) grants DUE-0941959 and CNS-0851646. The views and conclusions contained in this paper are those of the authors and do not represent the official policies, either expressed or implied, of the funding agency.

References

1. Broch, J., Maltz, D.A., Johnson, D.B., Hu, Y.C., Jetcheva, J.: A Performance Comparison of Multi-hop Wireless Ad hoc Network Routing Protocols. In: The 4th International Conference on Mobile Computing and Networking, pp. 85–97. ACM, USA (1998)
2. Johansson, P., Larsson, T., Hedman, N., Mielczarek, B., Degermark, M.: Scenario-based Performance Analysis of Routing Protocols for Mobile Ad hoc Networks. In: The 5th International Conference on Mobile Computing and Networking, pp. 195–206. ACM, USA (1999)
3. Siva Ram Murthy, C., Manoj, B. S.: Ad hoc Wireless Networks – Architectures and Protocols. Prentice Hall, USA (2004)
4. Ni, S.-Y., Tsenf, Y.-C., Chen, Y.-S., Sheu, J.-P.: The Broadcast Storm Problem in a Mobile Ad hoc Network. In: The 5th International Conference on Mobile Computing and Networking, pp. 151–162. ACM, USA (1999)
5. Sinha, P., Sivakumar, R., Bhargavan, V.: Enhancing Ad hoc Routing with Dynamic Virtual Infrastructures. In: The 20th International Conference on Computer and Communications Societies, pp. 1763–1772. IEEE, USA (2001)
6. Wang, F., Min, M., Li, Y., Du, D.: On the Construction of Stable Virtual Backbones in Mobile Ad hoc Networks. In: The International Performance Computing and Communications Conference (IPCCC). IEEE, USA (2005)
7. Sakai, K., Sun, M.-T., Ku, W.-S., Okada, H.: Maintaining CDS in Mobile Ad Hoc Networks. In: Li, Y., Huynh, D.T., Das, S.K., Du, D.-Z. (eds.) WASA 2008. LNCS, vol. 5258, pp. 141–153. Springer, Heidelberg (2008)
8. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 3rd edn. MIT Press/McGraw Hill, New York, USA (2009)
9. Meghanathan, N.: An Algorithm to Determine the Sequence of Stable Connected Dominating Sets in Mobile Ad Hoc Networks. In: The 2nd Advanced International Conference on Telecommunications. IARIA, French Caribbean (2006)
10. Meghanathan, N.: An algorithm to determine minimum velocity-based stable connected dominating sets for ad hoc networks. In: Ranka, S., Banerjee, A., Biswas, K.K., Dua, S., Mishra, P., Moona, R., Poon, S.-H., Wang, C.-L. (eds.) IC3 2010. Communications in Computer and Information Science, vol. 94, pp. 206–217. Springer, Heidelberg (2010)
11. Abolhasan, M., Wysocki, T., Dutkiewicz, E.: A Review of Routing Protocols for Mobile Ad hoc Networks. Ad Hoc Networks 2(1), 1–22 (2004)
12. Bettstetter, C., Hartenstein, H., Perez-Costa, X.: Stochastic Properties of the Random-Way Point Mobility Model. Wireless Networks 10(5), 555–567 (2004)

OERD - On Demand and Efficient Replication

Dereplication

Vardhan Manu, Gupta Paras, and Kushwaha Dharmender Singh

Computer Science and Engineering Department, MNNIT Allahabad
Allahabad, India

{rcs1002, cs1006, dsk}@mnnit.ac.in
<http://www.mnnit.ac.in>

Abstract. For many years, file replication and dereplication in distributed computing environment has been researched to enhance and optimize the scalability of the entire system. Although numerous work have been proposed on the issues of file replication, a comprehensive approach still misses out on various fronts. An effort has been made in the present work to propose a reliable and comprehensive file replication and memory aware dereplication mechanism for a trusted private cloud, based on the file threshold. The proposed approach introduces a File Replication Server (FRS) that is responsible for replicating the file on peer FRS, when the file threshold limit is reached. The proposed approach handles file replication, dereplication, access and performance transparency to the system, thereby ensuring the replication and dereplication decisions about the files in a seamless and efficient manner. The approach is simulated on JAVA platform. A comparative study of the proposed approach with the Request Reply Acknowledgement (RRA) and Request Reply (RR) protocol is presented, showing the significant reduction by 37.5% to 58%, in terms of total number of messages exchanged for file replication.

Keywords: Replication, Logical resource (LR), Service, Private cloud, IaaS.

1 Introduction

Cloud computing is the future of technology, with its application distributed in every field. Cloud computing eliminates the need of having efficient hardware resources and infrastructure requirements by providing the resources on Pay as you go basis. All that is required is a machine with an enabled web browser. Cloud can be deployed in three ways viz., Public, Private [15] and Hybrid cloud. When it comes to cloud delivery architecture, there exist three delivery models, viz., SaaS [12], PaaS [12] and IaaS [12]. IaaS is most widely used delivery architecture, as it provides the provision for processing, storage, networks and other fundamental computing resources. It also provides control over the deployed application and limited control of selecting the network components. The delivery of cloud through IaaS is typically platform virtualization. Application programming interface provides accessibility to software, which enables the machine to interact with cloud software, in the same way the user interface facilitates inter-action between a human and a computer.

Distributed Computing is always restricted by the issues like consistency, transparency, reliability that need to be considered for a specific application. The On-demand and Efficient Replication and Dereplication (OERD) approach considers the various issues like loosely coupled and tightly coupled system, granularity, stateless or stateful server, mutable or immutable files and consistency. Granularity refers to the unit of sharing that can be a file or a record. It is of different type's viz., granularity of locking, sharing and data transfer. Granularity needs to be considered, when there are multiple requests for a particular resource. It helps in deciding the level of parallelism that can be achieved by the application and also shows concern about the problem of false sharing. With high level of parallelism, this problem can be minimized, by providing fine grain granularity. To increase the system reliability, some fault tolerant mechanism should be used, so that the system keeps functioning in case of failure. One such method is replication, which replicates the critical software components, so that if one of them fails, the others can be used to continue. Replication means high availability of resources. Resources can be physical or logical. Physical resources include memory and storage capacity, whereas logical resources include file, data and services that need to be replicated or made available on demand, depending upon the application requirement. On-demand and Efficient Replication and Dereplication (OERD), provides an on-demand replication and dereplication of logical resources (files, service), with a view to minimize the network resource utilization by minimizing the message exchange overhead, to speed up the overall system performance. Dereplication is done to delete the file. The rest of the paper is organized as follows. Section 2 presents related work. Section 3 introduces the proposed architecture in a private cloud environment. Section 4 discusses the simulation and results with a case study for OERD approach and then the result is concluded in Section 5 followed by the future work.

2 Related Work

Replication in cloud environment is done to achieve high availability of resources. Resources can be replicated dynamically or on-demand to minimize the overhead of maintaining the consistency of the replicated files, to some extent. Similar kind of work is carried out by the authors in distributed environment considering the various performance issues that can arise and affect the overall system performance. Richard T. Hurley and Soon Aun Yeap [1] have proposed file replication and migration policy, by which the total mean response time for a requested file at a particular site can be reduced. To avoid and restrict the issue of consistency and overhead of maintaining too many copies of the files, in dynamic file replication, the concept of de-replication is proposed. It is based on the concept of least recently used file, where the file is selected for de-replication, if it is not requested for the longest period of time at the storage site.

Sometimes instead of using file replication mechanism, it is preferred to use process migration to achieve better system performance and minimum utilization of network resources. A similar approach has been taken by Anna Hac [3]. Author has proposed the file replication, migration and process migration techniques based on the workload of local and remote host. A file can be replicated on every machine, where a

process can access the replicated file locally. File migration is preferred only when the file size is small, as it will utilize the disk space and use the network resources, thus significantly degrading the system performance. Process migration is a good solution when the file to be accessed is large. Author considers file migration, only for the files having small size, to be as good as process migration.

Concern with cloud computing is, how to deploy the application on cloud and in what manner should the deployed application, be delivered as a service. Pengzhi Xu, et. al. [5] presented a prototype named Posix Cloud, which is designed to deliver general purpose cloud storage via standard POSIX interface and provides support for the traditional applications which are based on standard file system interface. This storage can enhance the performance, if optimized for special purpose storage and customized interface, which can be used by dedicated application and services.

To facilitate logical resource (file, service) replication in cloud environment, a replication mechanism is required which aims at facilitating replication considering the issues related with replication in cloud and distributed computing. Wei-Tek Tsai, et. al. [6] has proposed a replication scheme for services. Whenever there is an increase in number of request a service can handle, additional resources are acquired by replicating the service. Two types of service replication are proposed, Active and Passive Service Level MapReduce Approach (SLMR) for replication. According to authors, the traditional service replication is passive, that does not participate in the decision on when to replicate, where to replicate and number of copies to replicate.

3 Proposed Approach

On-demand replication of logical resources (file or service), is achieved by replicating resources, when the number of requests for a specific resource reaches the threshold value. It increases the performance of the overall system and makes it capable of handling faults in case of failure. On-demand replication and dereplication reduces the utilization of network resources by minimizing the message passing overhead for a particular operation namely replication, thus ensuring the consistent system performance.

Dereplication of files will take place in such a manner that it will fulfill the size requirement of incoming files. While maintaining the space management overhead, file deletion depends upon three criteria. These criteria are last modification time of the file, number of replica available of a file and file size.

Case 1: Last modification time based approach- Last modification time is the time at which the file was last modified. Files are sorted on the usage basis and delete the file which was not requested for largest period of time at the storage site. If only single copy of file exists in the system, in that case there will be a drawback of using this approach. If a file with only one replica is deleted then information of that file will be lost. To overcome from this approach deletion will be done on number of replica available basis.

Case 2: Number of replicas available of a file based approach- Number of replicas available for a file is the count on the number of copies available for a particular file. Whenever a copy of file is created, this increases the number of replicas for a file. Files having many copies or more than one replica can be considered for deletion.

Files with one replica are not deleted to avoid losing information of the file. So, a check is performed, whether or not there are other copies available for that file. If only single copy of file exists in the system, in that case next probable file for deletion, will be selected from the sorted file list on the basis of last modification time.

Case 3: File size based approach- File size is the size of a file required on a disk. This approach is used when time required for deletion considered as important factor. If there exists minute difference in the last modification time of the two files and number of replicas available for both files is more than one, the file with minimum file size among the two will be deleted, so as to minimize the deletion time. The proposed architecture is discussed below:

3.1 Architecture

The architecture proposed, implements, On-demand and Efficient Replication and Dereplication (OERD), which consists of loosely coupled systems, capable of providing various kind of services like replication, storage, I/O specific, computation specific and discovery of resources in the cloud environment. Based on the application requirement, the resources are made available to the client.

Fig. 1 shows the set of peer servers called the File Replication Servers (FRS), responsible for providing the replication service in the cloud environment. Based on the number of request received for a particular file by the FRS, the resource (File) is replicated when the total number of request reaches a threshold value.

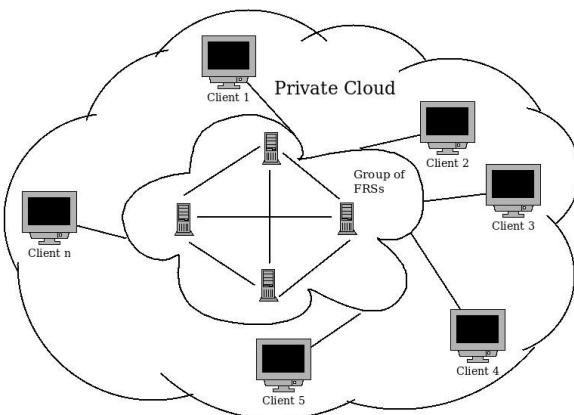


Fig. 1. Architecture of private cloud

3.2 Data Structures

Fig. 1 shows the group of File Replicating Servers (FRSs), which presents the peer servers for the proposed architecture.

Table. 1 shows the format of file request count table, in which the peer server will maintain the information about the files requested by the clients. This table helps in deciding, when to create the replica of the requested file. It is created when the total

number of requests on a server reaches a threshold value. The table has the following fields/attributes:

FileID: uniquely identifies the requested file.

Filename: name of the file requested by the client.

Request Count: increments the count by one, whenever a file request is fulfilled by that node.

Metadata: stores the data about the file to identify various file attributes as shown in Table 2.

Table 1. File request count table

FILE ID	FILENAME	REQUEST COUNT	METADAT
Type: Int Size: 16 bits	Type: String Length:20chars	Type: Integer Size: confined to there request threshold	Type: String Size: 100 chars

Table 2. File Metadata

Attribute Name	Data Type
File Name	String
Last Modification Date	yyyy-mm-dd
Last Modification Time	hh:mm
File Size	Long
File Replica	Integer
Flag Value	Boolean

Table 3 represents the format of peer FRS table, which maintains the IP and PORT, of the peer FRS having the replica of the requested file.

SERVER IP: denotes the IP address of the peer FRS on which the file is replicated.

SERVER PORT: denotes the port address of the peer FRS to which the network messages are forwarded.

Table 3. Peer FRS table

SERVER IP	SERVER PORT
Type: String Length: length of any valid IP address	Type: Short int Size: 16 bits

3.3 Message Definitions for OERD Approach

OERD approach has four types of messages viz., M₁, M₂, M₃, and M₄. Functionality implemented by each message is described below:

M₁ : This message will contain the request for GET/PUT operation. It consists of three tuples which include the following details:

- Source Machine Type(either client or FRS)
- Request Type (either GET or PUT)
- Filename

M₂: This message responds to the request based on the server's current state, as described below:

- It informs the client, if the server is overloaded,
- Checks whether the file exists on server or not,
- If the server is ready to accept the request or not.

M₃: Copy the file from source to destination.

M₄: Provides the details to the client, about the peer server on which the logical resource is being replicated, described as under:

- Peer server IP (i.e. IP of peer server having the replica of requested logical resource)
- Server port (i.e. denotes the port address of the peer server on which the network messages are forwarded)

3.4 RRA and RR Protocol

RRA [16] protocol use three messages for completing a request viz., request message, reply message and acknowledgement message. RR [16] protocol use two messages for completing a request viz., request message and the reply message which also serves as the acknowledgement for the request message.

As shown in Table. 4, for RR protocol, minimum number of messages required to complete the replication mechanism will be, no less than or equal to eight. Similarly, for RRA protocol, when used for replication, minimum number of messages required to complete the replication mechanism reaches to twelve.

3.5 Strengths and Limitations of the Proposed Methodology

Strengths

- OERD approach is designed to provide the following features:
- OERD approach provides access, migration and performance transparency to the system.
- It automatically distributes the system load on the peer servers.
- It is designed to ensure implicit process addressing, thus increases overall transparency of the system.
- Asynchronous communication is used, thus ensuring that the system will keep accepting the requests without blocking its state.
- This approach controls file replication redundancy to certain extent.

Limitations

False sharing can occur with coarse grained granularity, thus reducing the parallelism. But it will not affect the overall performance of the system.

4 Simulation and Results

The proposed model is simulated on JAVA platform. OERD approach is compared with Request Reply Acknowledgement (RRA) [16] and Request Reply (RR) protocol [16]. It outperforms RRA and RR protocol in terms of on-demand replication. Given below are the details and possible cases for better understanding of OERD approach:

4.1 Case 1

The first case is where the client C_j sends a GET request to the server S_o and receive the LR_i from it. Numbers of messages exchanged are described as below:

- $M_1 : C_j \rightarrow S_o$
- $M_2 : S_o \rightarrow C_j$
- $M_3 : S_o \rightarrow C_j$

Total number of messages exchanged: 3

4.2 Case 2

So report the OVERLOAD condition to the client C_j . Number of messages exchanged are described as below:

- $M_1 : C_j \rightarrow S_o$
- $M_2 : S_o \rightarrow C_j$
- $M_4 : S_o \rightarrow C_j$

Total number of messages exchanged: 3

4.3 Case 3

Replication of the LR_i from S_o to and client C_j is informed about the replicated LR_i . Number of messages exchanged is described as below:

- $M_1 : C_j \rightarrow S_o$
- $M_1 : S_o \rightarrow S_r$
- $M_4 : S_o \rightarrow C_j$
- $M_2 : S_o \rightarrow C_j$
- $M_3 : S_o \rightarrow S_r$

Total number of messages exchanged: 5

Table 4 shows the comparison of the OERD approach in terms of messages exchanged per request with the existing RR protocol and RRA protocol. It shows the number of messages exchanged under the scenario viz., GET, PUT, OVERLOAD and REPLICATE for OERD, RR and RRA protocol. In RR and RRA protocol, there is no routine mechanism for getting the IP address of the peer server on which the file is

replicated, as compared to OERD approach. Here ST AT U S message, in case of RR and RRA protocol, will provide the IP address of peer server containing the copy of replicated resource.

Fig. 2 shows the comparison between OERD, RR RRA proto- col. It shows the total number of messages exchanged per resource per client for successfully completing an operation (GET, PUT, OVERLOAD and REPLICATE). It clearly shows that OERD approach runs well for REPLICATE and OVERLOAD operation. It outperforms the other two protocols, when used for these (REPLICATE and OVERLOAD) operations. In terms of total number of messages exchanged, OERD approach shows significant performance improvement, for OVERLOAD and REPLICATE operation, as compared to GET/PUT.

Table 4. Number of messages exchanged per request

Operation	OERD	RR Protocol	RRA Protocol
GET	3	2	3
PUT	2	2	3
OVERLOAD	3	$M_1 + STATUS = 4$	$M_1 + STATUS = 6$
REPLICATE	5	$M_1 + STATUS + M_1 + M_3 = 8$	$M_1 + STATUS + M_1 + M_3 = 12$

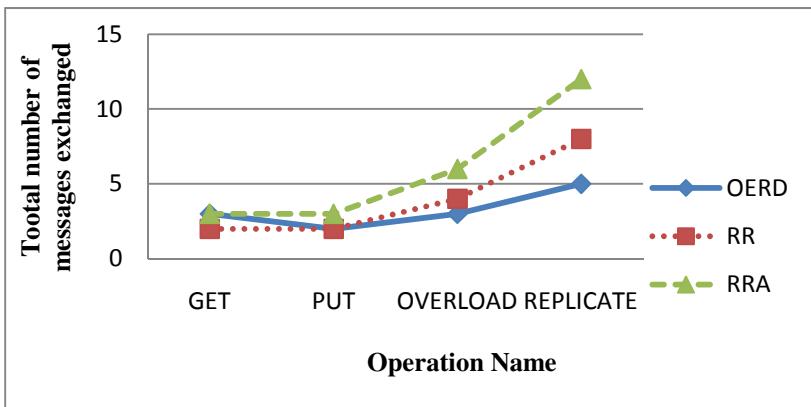


Fig. 2. Line graph showing comparison between OERD, RR and RRA protocol

5 Conclusion

On demand and Efficient Replication and Dereplication approach (OERD), proposed in this paper, aims at implementing the replication and dereplication mechanism, with minimum number of messages required to complete these operation, thus minimizing the time required to replicate or dereplicate a file and at the same time network resource utilization. The proposed OERD approach provides coarse grained granularity because the unit of data transfer across the network is an entire file as compared to a page or a record. OERD approach ensures various types of transparencies viz., migration, access and performance. Migration decisions about the

file movement are made by OERD approach, without any user intervention. It locates the resources and fetches them to fulfill the clients request in a transparent manner. It is also responsible for replicating the file, from one peer server to the other peer server, when the total number of request, on a peer server, for transferring a particular file reaches the threshold value. This enhances the performance of the peer servers, thus enhancing the system performance. Work on proposing a consistency mechanism, increasing reliability and security setup of OERD is under progress.

References

1. Hurley, R.T., Yeap, S.A.: File migration and file replication: a symbiotic relationship. *IEEE Trans. on Parallel and Distributed Systems* 7, 578–586 (1996)
2. Mei, A., Mancini, L.V., Jajodia, S.: Secure Dynamic Fragment and Replica Allocation in Large-Scale Distributed File System. *IEEE Trans. on Parallel and Distributed Systems* 14(9), 885–896 (2003)
3. Hac, A.: A Distributed Algorithm for Performance Improvement Through File Replication, File Migration, and Process Migration. *IEEE Trans. on Software Engg.* 15(2), 1459–1470 (1989)
4. Xiong, K., Perros, H.: Service Performance and Analysis in Cloud Computing. In: World Conference on Services-I, pp. 693–700 (2009)
5. Xu, P., Zheng, W., Wu, Y., Huang, X., Xu, C.: Enabling cloud storage to support traditional applications. In: 5th Annual China Grid Conference, pp. 167–172 (2010)
6. Tsai, W.-T., Zhong, P., Elston, J., Bai, X., Chen, Y.: Service Replication with Map Reduce in Clouds. In: 10th International Symp. on Autonomous Decentralized Systems (ISADS), pp. 381–388 (2011)
7. Sato, H., Matsuoka, S., Endo, T.: File Clustering Based Replication Algorithm in a Grid Environment. In: 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, pp. 204–211 (2009)
8. Panagos, E., Delis, A.: Selective Replication for Content Management Environments. *IEEE Journal on Internet Computing* 9(3), 45–51 (2005)
9. Cheng, H.Y., King, C.T.: File Replication for Enhancing the Availability of Parallel I/O Systems on Clusters. In: 1st IEEE Computer Society International Workshop on Cluster Computing, pp. 137–144 (1999)
10. Cabri, G., Corradi, A., Zambonelli, F.: Experience of Adaptive Replication in Distributed File Systems. In: IEEE Proc. of 22nd EUROMICRO Conf. on Beyond 2000: Hardware and Software Design Strategies, pp. 459–466 (1996)
11. Walters, J.P., Chaudhary, V.: Replication-Based Fault Tolerance for MPI Application. *IEEE Trans. on Parallel and Distributed Systems* 20(7), 997–1010 (2009)
12. Cloud computing - A premier The Internet protocol Journal 12(3)
http://www.cisco.com/web/about/ac123/ac147/archived_issues/i_pj_12-3/123_cloud1.html (accessed on October 22, 2011)
13. Identifying Applications for Public and Private Clouds, Tom Nolle, Searchcloudcomputing,
http://searchcloudcomputing.techtarget.com/tip/0,289483,sid201_gci1358701,00.html?track=NLT-1329&ad=710605&asrc=EM_NLT_7835341&uid=8788654 (accessed on October 22, 2011)

14. Cloud Security Alliance (CSA) <https://cloudsecurityalliance.org/> (accessed on October 22, 2011)
15. Zheng, L., Hu, Y., Yang, C.: Design and Research on Private Cloud Computing Architecture to Support Smart Grid. In: International Conf. on Intelligent Human-Machine Systems and Cybernetics (IHMSC), August 26-27, pp. 159–161 (2011)
16. Spector, A.Z.: Performing remote operation efficiently on a local computer Network. Communications of the ACM 25(4), 246–259 (1982)
17. Venkatasubramanian, N., Talcott, C.L.: A Reflective Framework for Providing Safe QoS-enabled Customizable Middleware. In: Workshop on Reflective Middleware, RM 2000 (2000)
18. Chou, C.-F., Golubchik, L., Lui, J.C.S.: Striping doesn't scale: how to achieve scalability for continuous media servers with replication. In: 20th International Conference on Distributed Computing Systems, pp. 64–71 (2000)
19. Venkatasubramanian, N., Deshpande, M., Mohapatra, S., Gutierrez-Nolasco, S., Wickramasuriya, J.: Design and implementation of a composable reflective middleware framework. In: 21st International Conference on Distributed Computing Systems, pp. 644–653 (April 2001)

A Mathematical Model for Performance Evaluation and Comparison of MAP Selection Schemes in n Layer HMIPv6 Networks

Abhishek Majumder

Department of Computer Science & Engineering, Tripura University, Tripura, India
abhi2012@gmail.com

Abstract. Hierarchical Mobile IPv6 introduced Mobility Anchor Points to reduce the signaling overhead due to frequent change of Access Routers by the Mobile Nodes. In order to increase the scalability and to improve the overall performance of the network multiple layers of MAPs are introduced. This creates the problem of optimal MAP selection by the MNs. Several MAP selection schemes have been proposed for this purpose. This paper considers three MAP selection schemes: Furthest MAP selection scheme, Nearest MAP selection scheme and Mobility based MAP selection scheme. All the three schemes are compared with respect to signaling overhead, handoff latency and tunneling cost in multi layer MAP network architecture. Analytical results show that the furthest MAP selection scheme has minimum inter domain binding update cost, the nearest MAP selection scheme has low cost in other respects except inter domain BU cost but the Mobility based MAP selection scheme offers most optimal performance in all respects.

Keywords: Mobile IPv6 (MIPv6), Hierarchical Mobile IPv6 (HMIPv6), Mobility Anchor Point, Mathematical Model.

1 Introduction

Mobile IPv6 [1] has been proposed to mitigate the challenge of providing internet service to increasing number of mobile internet users. But it has high signaling overhead and long handoff delay. To overcome this problem Hierarchical MIPv6 (HMIPv6) [2] was proposed. In HMIPv6, Mobility Anchor Points (MAPs) were introduced to handle mobility locally.

For selection of an optimal MAP several MAP selection algorithms have been proposed. Out of those, most popular schemes are Distance based MAP selection scheme and Mobility based MAP selection scheme. Distance based schemes are of two types: Furthest MAP selection scheme [2] and Nearest MAP selection scheme [5]. On the other hand, in Mobility based MAP selection scheme [3], [4], [6] the MN selects the MAP based on its speed of movement.

In this paper, a mathematical model of MIPv6 network has been considered. Based on the mathematical model the Furthest MAP selection, Nearest MAP selection and Mobility Based MAP selection schemes are analyzed. Performances of these three

schemes are compared with respect binding update cost, MAP handoff delay and tunneling cost. The effect of varying number of layers in MAP hierarchy is also observed.

The rest of the paper is organized as follows. Section 2 introduces the three MAP selection schemes. Section 3 and 4 presents the assumptions made for the mathematical model and the analysis of the schemes. The numerical results are described in section 5. Finally in section 6 conclusion of this paper is drawn.

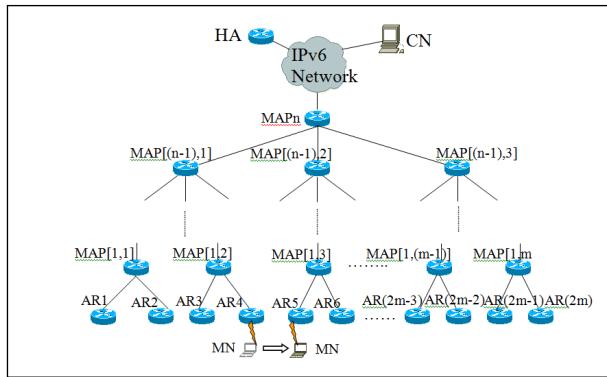


Fig. 1. Hierarchical architecture of HMIPv6

2 MAP Selection Schemes

Fig. 1 shows a hierarchical architecture of HMIPv6. MAP[i,j] specifies jth MAP of layer i. In Fig. 1 AR1 is in the domain of MAP[1,1] and all its parent MAPs. All the ARs are in the domain of MAP_n. When a MN enters into the vicinity of an AR it has the option of registering its LCoA (on-link care-of-address) in any of the MAP belonging to different layers. In Fig. 1 as the MN moves from AR4 to AR5, it has to register its LCoA to a MAP. If the MN's LCoA was registered in MAP[1,2], after the handoff the MN selects a MAP and sends BU to the HA. On the other hand, if the MN's LCoA was registered to MAP_n, the MN will only update its LCoA in MAP_n.

2.1 Distance Based MAP Selection Scheme

The Distance based MAP selection scheme was proposed in the specification of HMIPv6 [2]. In this scheme the MN selects the Furthest MAP to register its LCoA. It can be safely assumed that the further node covers more number of ARs than the nearer MAPs. So, the more distanced MAP the MN selects, the less number of MAP domain changes take place. Every MAP domain change by the MN results in BU to the HA and CN, thus burdening the global network. In Fig. 1 the MN will select the MAP_n as its serving MAP .The advantage of using Furthest MAP is that it reduces the number of BUs to the HA.

There are several problems associated with Furthest MAP selection scheme. In this scheme, the Furthest MAP is the bottleneck point for the network traffic. On the other

hand, intra-domain BU cost and handoff latency becomes more as the hop count from the MN to the Furthest MAP is higher compared to the nearer MAPs. Selection of nearer MAP [5] reduces the intra-domain handoff cost as well as delay.

2.2 Mobility Based MAP Selection Scheme

This MAP selection mechanism is based on the MN's velocity. In the velocity-based MAP selection scheme [3], [4], [6]-[8], there are mainly two steps. First the velocity of the MN is estimated then the MN selects the MAP to register. To estimate the velocity, each MAP records not only the binding update of the MN, but also the time when the binding is updated. When the MN moves into a new MAP domain, it sends BU to the new MAP. The residence time of the MN in the previous MAP can be calculated by subtracting binding update time of old MAP from that of new MAP. After obtaining the MAP residence time the velocity of the MN will be calculated. Highly mobile nodes and less mobile nodes select the higher layer MAP and lower layer MAP respectively.

3 Modeling and Assumption

This section describes the mathematical model as well as assumptions for analysis of the network performance under different MAP selection schemes. To evaluate the performance of each of the schemes mathematical expressions for BU cost, handoff delay and tunneling are formulated [9-12]. Fig. 1 shows the n layered MAP network architecture considered for analysis. Table 1 describes the abbreviation and definition used in this modeling.

Table 1. Symbols and definitions

Symbol	Definition
BU_{Total}	Total Binding Update cost
BU_{AR}	Binding update cost due to AR change
BU_{MAP}	Binding update cost due to MAP change
N_{AR}	Number of AR change
D_{MAP-AR}	Hop count between an AR and its corresponding MAP
N_{MAP}	Number of MAP change
D_{HA-MAP}	Hop count between HA and MAP of a MN
D_{HA-AR}	Hop count between HA and AR of a MN
$BUSIZE$	Binding Update Message Size
$BACKSIZE$	Binding Acknowledgement Message Size
TC_{Total}	Total Tunneling Cost
TC_I	Tunneling Cost when packets destined to the MN reach its currently serving MAP via HA
TC_D	Tunneling Cost when packets destined to the MN reach its currently serving MAP directly
A	Number of packets transmitted by the CN via HA in a time unit
H	Size of the tunneling header

Table 1. (continued)

Symbol	Definition
N	Number of MAP layers in the network
LA _w	Per hop delay in wired network
LA _{wl}	Per hop delay in wireless network
LA _{AR}	Latency for AR change in the same MAP domain
LA _{MAP}	Latency for MAP domain change
F	MN follows Furthest MAP Selection Scheme
N	MN follows Nearest MAP Selection Scheme
M	MN follows Mobility Based MAP Selection Scheme
Δ	Decrease in mobility rate with increase in number of MAP layers in mobility based MAP selection scheme.

In addition the following assumptions are made:

- i. The hop distance from the l^{th} layer MAP to AR follows poisson distribution [13] with mean λ_l , $\lambda_l > 0$. The probability mass function is given by:
 $p(i) = p(x=i) = e^{-\lambda_l} (\lambda_l)^i / i!$ where $l=1,2,\dots,n$ and $i=0,1,2,\dots$
- ii. The residence time of the MN in an AR follows exponential distribution [13]-[16] with rate $\lambda_{\text{AR}} > 0$. So the probability density function is given by:
 $p_{\text{AR}}(t) = \lambda_{\text{AR}} e^{-\lambda_{\text{AR}} t}$ where $t > 0$
- iii. The residence time of the MN in a l^{th} MAP domain follows exponential distribution [13]-[16] with rate λ_{ML} . So the probability density function is:
 $p_{\text{ML}}(t) = \lambda_{\text{ML}} e^{-\lambda_{\text{ML}} t}$ where $t > 0$
- iv. The residence time of the MN in a MAP domain when it uses the mobility based MAP selection scheme follows exponential distribution [13]-[16] with rate λ_M . So the probability density function is given by:
 $p_M(t) = \lambda_M e^{-\lambda_M t}$ where $t > 0$
The value of λ_M decreases by δ as the number of MAP layer in the network increases by 1.
- v. The hop count from HA to a MAP belonging to any layer is assumed to be constant.
- vi. The packet rate follows poisson distribution [13] with rate λ_p , $\lambda_p > 0$. The probability mass function given by:
 $p_p(j) = e^{-\lambda_p} (\lambda_p)^j / j!$ where $j > 0$
Expected number of packets in a time unit $E[N_p] = \lambda_p$

4 Analysis

4.1 Binding Update Cost

The BU cost of a MN in a time unit can be calculated as

$$\text{BU}_{\text{Total}} = \text{BU}_{\text{AR}} + \text{BU}_{\text{MAP}} + \text{BACK}_{\text{MAP}} + \text{BACK}_{\text{AR}} \quad (1)$$

$$\text{BU}_{\text{AR}} = (E[N_{\text{AR}}] + 1) \cdot E[D_{\text{MAP-AR}}] \cdot \text{BUSIZE} \quad (2)$$

$$\text{BACK}_{\text{AR}} = (E[N_{\text{AR}}] + 1) \cdot E[D_{\text{MAP-AR}}] \cdot \text{BACKSIZE} \quad (3)$$

$$BU_{MAP} = (E[N_{MAP}] + 1)D_{HA-AR} \text{BUSIZE} = (E[N_{MAP}] + 1)(D_{HA-MAP} + E[D_{MAP-AR}])\text{BUSIZE} \quad (4)$$

$$BACK_{MAP} = (E[N_{MAP}] + 1)D_{HA-AR} \text{BACKSIZE} = (E[N_{MAP}] + 1)(D_{HA-MAP} + E[D_{MAP-AR}])\text{BACKSIZE} \quad (5)$$

In furthest and nearest MAP selection scheme the MN selects the furthest MAP (n^{th} layer MAP) and nearest MAP (1st layer MAP) respectively. So expected hop distance from its AR to MAP is given by

$$E_F[D_{MAP-AR}] = \lambda_n \quad (6)$$

$$\text{and } E_N[D_{MAP-AR}] = \lambda_1 \quad (7)$$

On the other hand, in case of MMAP the MN selects the MAP based on its velocity. Let the residence time of the MN be divided into following segments

$$\infty > t_1 > t_2 > t_3 > \dots > t_{n-2} > t_{n-1} > t_n$$

If the MN's residence time is between ∞ and t_1 , the MN selects the layer 1 MAP to register its LCoA. Similarly, if the MN's residence time is between t_1 and t_2 , the MN selects the layer2 MAP. The highly mobile MNs having residence time between t_{n-1} and t_n selects the n^{th} layer MAP. The expected distance from its serving AR to MAP is calculated as

$$\begin{aligned} E_M[D_{MAP-AR}] \\ = & E[\lambda_1] \\ = & [p(x=1)\lambda_1 + p(x=2)\lambda_2 + \dots + p(x=n-1)\lambda_{n-1} + p(x=n)\lambda_n] \\ = & \int_{t_1}^{\infty} \lambda_M e^{-\lambda_M t} dt \cdot \lambda_1 + \int_{t_2}^{t_1} \lambda_M e^{-\lambda_M t} dt \cdot \lambda_2 + \dots + \int_{t_{n-1}}^{t_{n-2}} \lambda_M e^{-\lambda_M t} dt \cdot \lambda_{n-1} + \int_{t_n}^{t_{n-1}} \lambda_M e^{-\lambda_M t} dt \cdot \lambda_n \\ = & (\lambda_1 + \lambda_2)e^{-\lambda_M t_1} + (\lambda_2 + \lambda_3)e^{-\lambda_M t_2} + \dots + (\lambda_{n-1} + \lambda_n)e^{-\lambda_M t_{n-1}} + \lambda_n e^{-\lambda_M t_n} \end{aligned} \quad (8)$$

Since the residence time of the MN in an AR follows exponential distribution with rate λ_{AR} , the expected residence time of the MN in an AR is given as

$$E_{AR}[t] = 1/\lambda_{AR} \quad (9)$$

The number of AR changes by the MN in a time unit for furthest, nearest and mobility based MAP selection scheme is obtained as

$$E_F[N_{AR}] = E_N[N_{AR}] = E_M[N_{AR}] = 1/E_{AR}[t] = \lambda_{AR} \quad (10)$$

Expected number of MAP changes by MN in furthest, nearest and mobility based MAP selection scheme can be computed as

$$E_F[N_{MAP}] = 1/(1/\lambda_{Mn}) = \lambda_{Mn} \quad (11)$$

$$E_N[N_{MAP}] = 1/(1/\lambda_{M1}) = \lambda_{M1} \quad (12)$$

$$E_M[N_{MAP}] = 1/(1/\lambda_M) = \lambda_M \quad (13)$$

So, total BU cost per time unit for a MN using furthest, nearest and mobility based MAP selection scheme is derived from Eq. (1)-(13) as

$$\begin{aligned} BU_{FTotal} &= BU_{FAR} + BU_{FMAP} + BACK_{FMAP} + BACK_{FAR} \\ &= (\lambda_{AR} + 1) \cdot \lambda_n \cdot (BUSIZE + BACKSIZE) + (\lambda_{Mn} + 1) \cdot (D_{HA-MAP} + \lambda_n) \cdot (BUSIZE + BACKSIZE) \end{aligned} \quad (14)$$

$$\begin{aligned} BU_{NTotal} &= BU_{NAR} + BU_{NMAP} + BACK_{NMAP} + BACK_{NAR} \\ &= (\lambda_{AR} + 1) \cdot \lambda_1 \cdot (BUSIZE + BACKSIZE) + (\lambda_{M1} + 1) \cdot (D_{HA-MAP} + \lambda_1) \cdot (BUSIZE + BACKSIZE) \end{aligned} \quad (15)$$

$$\begin{aligned} BU_{MTOTAL} &= BU_{MAR} + BU_{MMAP} + BACK_{MMAP} + BACK_{MAR} \\ &= (BUSIZE + BACKSIZE) [\{(\lambda_1 + \lambda_2)e^{-\lambda_{Mt1}} + (\lambda_2 + \lambda_3)e^{-\lambda_{Mt2}} + \dots + (\lambda_{n-1} + \lambda_n)e^{-\lambda_{Mtn-1}} + \lambda_n e^{-\lambda_{Mtn}}\} \cdot (\lambda_{AR} + 1) + (D_{HA-MAP} + \{(\lambda_1 + \lambda_2)e^{-\lambda_{Mt1}} + (\lambda_2 + \lambda_3)e^{-\lambda_{Mt2}} + \dots + (\lambda_{n-1} + \lambda_n)e^{-\lambda_{Mtn-1}} + \lambda_n e^{-\lambda_{Mtn}}\}) \cdot (\lambda_{M1} + 1)] \end{aligned} \quad (16)$$

4.2 Handoff Latency

Handoffs occur when a MN either changes its serving AR or MAP. There are two types of handoff: intra-domain handoff and inter-domain handoff. In intra-domain handoff, MN moves from one AR to another within the vicinity of the same MAP and MN sends BU to its MAP only. On the other hand, in case of inter-domain handoff the MN moves from one MAP domain to another and BU is sent to MN's HA. So, intra-domain and inter-domain handoff latency of a MN is given as

$$LA_{AR} = 2 \cdot \{E[D_{MAP-AR}] \cdot LA_w + LA_{wl}\} \quad (21)$$

$$LA_{MAP} = 2 \cdot \{(E[D_{MAP-AR}] + D_{HA-MAP}) \cdot LA_w + LA_{wl}\} \quad (22)$$

Using Eq. 6, 21, 22 the handoff latency for furthest MAP selection scheme can be computed as

$$LA_{FAR} = 2 \cdot \{\lambda_n \cdot LA_w + LA_{wl}\} \quad (23)$$

$$LA_{FMAP} = 2 \cdot \{(\lambda_n + D_{HA-MAP}) \cdot LA_w + LA_{wl}\} \quad (24)$$

Using Eq. 7, 21, 22 the handoff latency for nearest MAP selection scheme can be written as

$$LA_{NAR} = 2 \cdot \{\lambda_1 \cdot LA_w + LA_{wl}\} \quad (25)$$

$$LA_{NMAP} = 2 \cdot \{(\lambda_1 + D_{HA-MAP}) \cdot LA_w + LA_{wl}\} \quad (26)$$

The handoff latency for mobility based MAP selection scheme can be computed using Eq. 8, 21, 22 and is given as

$$LA_{MAR} = 2 \cdot [(\lambda_1 + \lambda_2)e^{-\lambda_{Mt1}} + (\lambda_2 + \lambda_3)e^{-\lambda_{Mt2}} + \dots + (\lambda_{n-1} + \lambda_n)e^{-\lambda_{Mtn-1}} + \lambda_n e^{-\lambda_{Mtn}}] \cdot LA_w + LA_{wl} \quad (27)$$

$$LA_{MMAP} = 2 \cdot [(\lambda_1 + \lambda_2)e^{-\lambda_{Mt1}} + (\lambda_2 + \lambda_3)e^{-\lambda_{Mt2}} + \dots + (\lambda_{n-1} + \lambda_n)e^{-\lambda_{Mtn-1}} + \lambda_n e^{-\lambda_{Mtn}} + D_{HA-MAP}] \cdot LA_w + LA_{wl} \quad (28)$$

4.3 Tunneling Cost

In HMIPv6 the CN sends the packets destined to the MN to its HA. The HA receives the packets. If MN is in its home network, HA forwards the packets to the MN. But when the MN moves in a foreign network, the HA appends a tunneling header to the packet and tunnels it to MN's currently serving MAP. On receiving the packet, the MAP again tunnels the packet to the MN after adding another tunneling header. Thus

two tunneling headers are required. As soon as the CN receives the binding update from the MN, it stops sending the packets via HA; rather it uses an optimal route to the MAP. In this case only the MAP appends a tunneling header. The tunneling cost can be calculated as

$$TC_{Total} = \alpha \cdot TC_I + (\lambda_p - \alpha) \cdot TC_D \quad (29)$$

$$\alpha = E[N_{MAP}] \cdot LA_{MAP} \cdot E[N_p] \quad (30)$$

$$TC_I = h \cdot D_{HA-MAP} + 2h \cdot E[D_{MAP-AR}] \quad (31)$$

$$TC_D = h \cdot E[D_{MAP-AR}] \quad (32)$$

Using Eq. 6,7,8 and 22,24,26 and 29,30,31,32 tunneling cost of the MN for furthest, nearest and mobility based MAP selection scheme can be obtained as

$$TC_{FI} = h \cdot D_{HA-MAP} + 2h \cdot \lambda_n, \quad TC_{FD} = h \cdot \lambda_n, \quad \alpha = \lambda_{Mn} \cdot LA_{FMAP} \cdot \lambda_p \\ TC_{FTotal} = [\lambda_{Mn} 2 \{(\lambda_n + D_{HA-MAP})LA_w + LA_{wl}\} \lambda_p] (hD_{HA-MAP} + 2h\lambda_n) + [\lambda_p - \lambda_{Mn} 2 \{(\lambda_n + D_{HA-MAP})LA_w + LA_{wl}\} \lambda_p] h\lambda_n \quad (33)$$

$$TC_{NI} = h \cdot D_{HA-MAP} + 2h \cdot \lambda_1, \quad TC_{ND} = h \cdot \lambda_1, \quad \alpha = \lambda_{M1} \cdot LA_{NMAP} \cdot \lambda_p \\ TC_{NTotal} = [\lambda_{M1} 2 \{(\lambda_1 + D_{HA-MAP})LA_w + LA_{wl}\} \lambda_p] (hD_{HA-MAP} + 2h\lambda_1) + [\lambda_p - \lambda_{M1} 2 \{(\lambda_1 + D_{HA-MAP})LA_w + LA_{wl}\} \lambda_p] h\lambda_1 \quad (34)$$

$$TC_{MI} = h \cdot D_{HA-MAP} + 2h \cdot \{(\lambda_1 + \lambda_2)e^{-\lambda_M t_1} + (\lambda_2 + \lambda_3)e^{-\lambda_M t_2} + \dots + (\lambda_{n-1} + \lambda_n)e^{-\lambda_M t_{n-1}} + \lambda_n e^{-\lambda_M t_n}\}$$

$$TC_{MD} = h \cdot \{(\lambda_1 + \lambda_2)e^{-\lambda_M t_1} + (\lambda_2 + \lambda_3)e^{-\lambda_M t_2} + \dots + (\lambda_{n-1} + \lambda_n)e^{-\lambda_M t_{n-1}} + \lambda_n e^{-\lambda_M t_n}\}$$

$$\alpha = \lambda_M \cdot LA_{MMAP} \cdot \lambda_p$$

$$TC_{MTotal} \\ = \lambda_M 2 \{(\lambda_1 + \lambda_2)e^{-\lambda_M t_1} + (\lambda_2 + \lambda_3)e^{-\lambda_M t_2} + \dots + (\lambda_{n-1} + \lambda_n)e^{-\lambda_M t_{n-1}} + \lambda_n e^{-\lambda_M t_n}\} [hD_{HAMAP} + 2h \{(\lambda_1 + \lambda_2)e^{-\lambda_M t_1} + (\lambda_2 + \lambda_3)e^{-\lambda_M t_2} + \dots + (\lambda_{n-1} + \lambda_n)e^{-\lambda_M t_{n-1}} + \lambda_n e^{-\lambda_M t_n}\}] \\ + [\lambda_p - \lambda_M 2 \{(\lambda_1 + \lambda_2)e^{-\lambda_M t_1} + (\lambda_2 + \lambda_3)e^{-\lambda_M t_2} + \dots + (\lambda_{n-1} + \lambda_n)e^{-\lambda_M t_{n-1}} + \lambda_n e^{-\lambda_M t_n}\}] [h \{(\lambda_1 + \lambda_2)e^{-\lambda_M t_1} + (\lambda_2 + \lambda_3)e^{-\lambda_M t_2} + \dots + (\lambda_{n-1} + \lambda_n)e^{-\lambda_M t_{n-1}} + \lambda_n e^{-\lambda_M t_n}\}] + [hD_{HAMAP} + 2h \{(\lambda_1 + \lambda_2)e^{-\lambda_M t_1} + (\lambda_2 + \lambda_3)e^{-\lambda_M t_2} + \dots + (\lambda_{n-1} + \lambda_n)e^{-\lambda_M t_{n-1}} + \lambda_n e^{-\lambda_M t_n}\}] \quad (35)$$

5 Numerical Results

The formulas derived in section 4 are used to evaluate and compare the performance of furthest MAP selection scheme, nearest MAP selection scheme and mobility based MAP selection scheme. For numerical analysis the typical values of default parameters are shown in table 2. The size of BU and BACK messages follows the specification of MIPv6 [1] and HMIPv6 [2]. In the numerical analysis number of MAP layers varies from 3 to 7 and number of nodes in the network is assumed to be 1,00,000.

Table 2. Typical values of parameters

BU		72 bytes	$l=7$	λ_{MI}	0.00001
BACK		52 bytes	LA_w	1×10^{-11}	
H		40 bytes	LA_{wl}	1×10^{-10}	
1=l	λ_1	2	λ_M	δ	3
	λ_{MI}	10			0.2

Table 2. (continued)

l=2	λ_l	4	λ_{AR}	100
	λ_{MI}	1	D _{HA-MAP}	20
l=3	λ_l	6	n	3,4,5,6,7
	λ_{MI}	0.1	m	100000
l=4	λ_l	8	λ_p	100000
	λ_{MI}	0.01	n=3 t ₁ , t ₂ , t ₃	0.05, 0.01, 0
l=5	λ_l	10	n=4 t ₁ , t ₂ , t ₃ , t ₄	0.1, 0.05, 0.01, 0
	λ_{MI}	0.001	n=5 t ₁ , t ₂ , t ₃ , t ₄ , t ₅	0.5, 0.1, 0.05, 0.01, 0
l=6	λ_l	12	n=6 t ₁ , t ₂ , t ₃ , t ₄ , t ₅ , t ₆	1.0, 0.5, 0.1, 0.05, 0.01, 0
	λ_{MI}	0.0001	n=7 t ₁ , t ₂ , t ₃ , t ₄ , t ₅ , t ₆ , t ₇	5, 1, 0.5, 0.1, 0.05, 0.01, 0
l=7	λ_l	14		

5.1 Binding Update Cost

In this section, number of MAP layers in the network (n) is varied from 3 to 7 to realize its effect of on BU cost per time unit. Fig. 2 shows the variation of intra domain BU cost per time as the number of MAP layers in the network changes. It can be noticed that if the MN follows furthest MAP selection scheme, the intra domain BU cost is the highest among the three MAP selection schemes. Since the hop distance of the higher layer MAP is greater than that of lower layer MAP, the intra domain BU message has to traverse larger number of hops. Thus with the increase in number of MAP layers the intra domain BU cost increases. On the other hand, when the MN follows the nearest MAP selection scheme intra domain BU cost in the minimum compared to that of rest two. This is because in this scheme the MN always selects the layer 1 MAP. The BU cost remains the same although the number of MAP layer in the network increases. When the MN follows the mobility based MAP selection scheme the intra domain BU cost is higher than that of nearest MAP selection scheme but very much lower than furthest MAP selection scheme. This is because the MN selects the MAP based on its speed and most of the time intermediate MAPs gets selected. As the number of MAP layers increases in the network the intra domain BU cost also increases.

In furthest MAP selection scheme MN selects highest layer MAP. As a result very few inter domain BUs are performed and traffic on the global network reduces significantly. With the increase in number of MAP layers in the network, the distance of the MAP from AR increases but inter domain BU frequency decreases. The inter domain BU cost has a tendency to increase as MAP distance increases but at the same time has a tendency to decrease as inter domain BU frequency decreases. Fig. 3 shows that furthest MAP selection scheme has the minimum inter domain BU cost compared to rest two. In case of nearest MAP selection scheme MN selects the lowest layer MAP. The frequency of inter domain BU is maximum although distance of the MN from MAP is minimum. Nearest MAP selection scheme has maximum inter domain BU cost among the three. From Fig. 3 it can be observed that the inter domain BU cost for mobility bases MAP selection scheme is greater than furthest MAP selection scheme but smaller than nearest MAP selection scheme. Since intermediate

MAPs mostly get selected the two variables: BU frequency and MAP distance are optimized. As the number of MAP layer increases the highly mobile MN gets the opportunity to select higher layer MAP having larger domain. So the inter domain BU cost does not change much with the increase in number of MAP layers. Fig. 4 shows the variation of total BU cost with the increase in number of MAP layers.

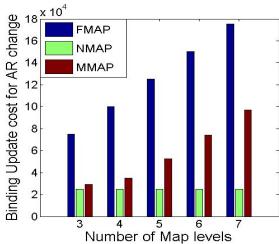


Fig. 2. Intra domain BU cost vs. number of MAP layers

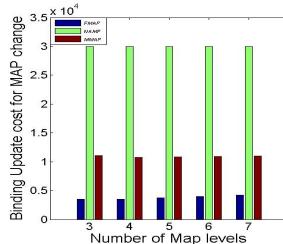


Fig. 3. Inter domain BU cost vs. number of MAP layers

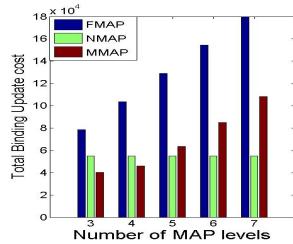


Fig. 4. Total BU cost vs. number of MAP layers

5.2 Handoff Latency

Fig. 5 shows the intra domain handoff latency for varying number of MAP layers in the network. In furthest MAP selection scheme the intra domain handoff latency is the highest compared to nearest and mobility based MAP selection scheme. This is because the MN selects the furthest MAP with respect to hop distance. With the increase in number of MAP layers intra domain BU latency increases. In nearest MAP selection scheme the MN selects the MAP having least hop distance among all available MAPs. So, intra domain handoff latency is lowest. The intra domain latency is constant in varying number of MAP layers in the network. In mobility based MAP selection scheme the MN selects its MAP from all the layers depending on its velocity. The intra domain handoff latency is slightly higher than that of nearest MAP selection scheme when the number of MAP layers is 3. As the number of MAP layer increases intra domain handoff latency increases and the difference between mobility based and nearest MAP selection scheme also increases. This is because the average distance of AR from MAP increases with the increase in number of MAP layers. The distance of AR from MAP is variable and is only responsible for variation of inter domain MAP delay with increasing number of MAP layers. Since in the furthest MAP selection scheme the distance of AR from MAP is highest, the inter domain handoff latency is also highest among all. Nearest MAP selection scheme enables the MN to select the nearest MAP, so the inter domain handoff latency is the lowest. In case of mobility based MAP selection scheme on an average MN selects the intermediate MAP. This results in a inter domain handoff latency which is less than furthest MAP selection scheme but greater than nearest MAP selection scheme. Figure 6 shows the comparison of the inter domain handoff latency among the three schemes. From the Fig. 6 it is clear that with the increase in number of MAP layers

the inter domain handoff latency of mobility based and furthest MAP selection scheme increases. It is also observed that the inter domain handoff latency increases at a higher rate if the number of MAP layers is greater than 4.

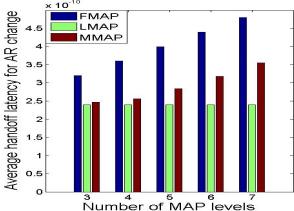


Fig. 5. Average handoff latency for AR change vs number of MAP layers

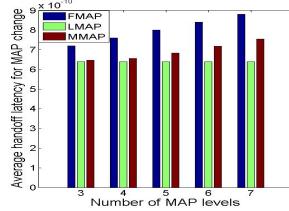


Fig. 6. Average handoff latency for MAP change vs number of MAP layers

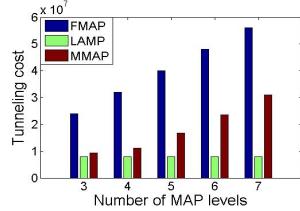


Fig. 7. Tunneling cost vs number of MAP layers

5.3 Tunneling Cost

Fig. 7 shows the tunneling cost per time unit in varying number of MAP layers. In furthest MAP selection scheme though the number of inter domain BUs are minimum but inter domain handoff latency is the highest. So the furthest MAP selection scheme has the highest tunneling cost. As the number of MAP layers increase tunneling cost also increases. Nearest MAP selection scheme has the smallest inter domain handoff latency but the highest number of inter domain BUs. This leads to the lowest tunneling cost. The tunneling cost does not change with the increase in number of MAP layers because of the assumption of the nearest MAP distance to be constant. The tunneling cost of mobility bases scheme is higher than nearest MAP selection scheme and lower than furthest MAP selection scheme. The tunneling cost increases as number of MAP layers is increased. Similar to inter domain handoff latency tunneling cost of mobility based MAP selection scheme also increases drastically if number of MAP layers are higher than 4.

6 Conclusion

In this paper, furthest, nearest and mobility based MAP are compared in multi layer MAP environment. From the comparison and analysis it is clear that the nearest MAP selection scheme performs better than rest two with respect to total BU cost, handoff latency and tunneling cost. It can also be concluded that there exist an optimal number of MAP layers in which the network following mobility based MAP selection scheme performs the best. For the scenario considered in this paper optimal number of MAP layer is found to be 4. It may vary in different network scenarios. Therefore mobility based scheme performs most optimally in all respects compared to furthest and nearest MAP selection scheme when the number of MAP layers is increased up to certain extent.

References

1. Perkins, C., Johnson, D., Arkko, J.: Mobility Support in IPv6. Internet Engineering Task Force Request for Comments 6275 (2011)
2. Soliman, H., Castelluccia, C., ElMalki, K., Bellier, L.: Hierarchical Mobile IPv6 (HMIPv6) Mobility Management. Network Working Group Request for Comments 5380 (2008)
3. Kawano, K., Kinoshita, K., Murakami, K.: A multilayer hierarchical distributed IP mobility management scheme for wide area networks. In: International Conference on Computer Communication Networks, pp. 480–484 (2002)
4. Kawano, K., Kinoshita, K., Murakami, K.: Multilayer hierarchical mobility management scheme in complicated structured networks. In: IEEE International Conference on Local Computer Networks (LCN), pp. 34–41 (2004)
5. Xu, Y., Lee, H.C.J., Thing, V.L.L.: A local mobility agent selection algorithm for mobile networks. In: IEEE International Conference on Communication, vol. 2, pp. 1074–1079 (2003)
6. Silva, P., Sirisena, H.: A mobility management protocol for IP-based cellular Networks. In: International Conference on Computer Communication Networks, pp. 476–482 (2001)
7. Kawano, K., Kinoshita, K., Murakami, K.: A mobility-based terminal management in IPv6 networks. IEICE Transactions on Communications E85-B(10), 2090–2099 (2002)
8. Thing, V., Lee, H., Xu, Y.: Designs and analysis of local mobility agents discovery, selection and failure detection for Mobile IPv6. In: IEEE International Conference on Mobile and Wireless Communication Networks, pp. 465–469 (2002)
9. Pack, S., Kwon, T., Choi, Y.: A performance comparison of mobility anchor point selection schemes in Hierarchical Mobile IPv6 networks. Computer Networks 51(6), 630–1642 (2007)
10. Lei, Y.X., Kuo, G.S.: Impact of MAP selection on handover performance for multimedia services in multi-layer HMIPv6 networks. In: IEEE Wireless Communications and Networking Conference, pp. 3904–3909 (2007)
11. Makaya, C., Pierre, S.: An Analytical Framework for Performance Evaluation of IPv6-Based Mobility Management Protocols. IEEE Transactions on Wireless Communications 7(3), 972–983 (2008)
12. Kong, K.S., Roh, S.J., Hwang, C.S.: A comparative analysis study on the performance of IP mobility protocols: mobile IPv6 and hierarchical mobile IPv6. In: International Conference on Advances in Mobile Computing & Multimedia, pp. 437–446 (2004)
13. Ross, S.M.: Introduction to probability and statistics for engineers and scientist, 3rd edn. Elsevier academic press
14. Ortigoza-Guerrero, L., Aghvami, A.H.: A prioritized handoff dynamic channel allocation strategy for PCS. IEEE Transactions on Vehicular Technology 48(4), 1203–1215 (1999)
15. Zhuang, W., Bensaou, B., Chua, K.C.: Adaptive quality of service handoff priority scheme for mobile multimedia networks. IEEE Transactions on Vehicular Technology 49(2), 494–505 (2000)
16. Haung, Y.: Determining the optimal buffer size for short message transfer in a heterogeneous GPRS/UMTS Network. IEEE Transactions on Vehicular Technology 52(1), 16–225 (2003)

Utilizing Genetic Algorithm in a Sink Driven, Energy Aware Routing Protocol for Wireless Sensor Networks

Hosny M. Ibrahim, Nagwa M. Omar, and Ali H. Ahmed

Department of Information Technology,
Faculty of Computers and Information
Assuit University, Egypt

hibrahim@uencom.com, nagwa_omar@hotmail.com,
alihussin.it@gmail.com

Abstract. Wireless Sensor Network (WSN) is a self-organized wireless ad-hoc network comprising large number of resource constrained devices called sensors. Usually sensors battery drainage is the main constraint in developing powerful WSN applications. Accordingly, a power conserving strategy must be implemented in all WSN layers. This paper focuses on the network layer which includes routing techniques as a main participant in power conserving applications. The main goal of the present work is to develop a routing technique based on genetic algorithm which aims to minimize total consumed power per round; hence lifetime is maximized compared to other techniques. The proposed technique enables sensor network to continue its operation during the continuous sensor failure without introducing additional control packets. Genetic algorithm is used in the proposed technique to find the minimum power ring which passes through all sensors and the base station. The algorithm operates on the base station only to save sensor's memory, processing resources, and indeed the power consumption.

Keywords: Wireless Sensor Networks, Ring Topology, Genetic Algorithm.

1 Introduction

Wireless sensor networks (WSNs) originally motivated in 1987 by the Defence Advanced Research Projects Agency (DARPA) [1, 2, 3]. Nowadays WSN is used in many industrial and civilian application areas [4, 5, 6]. WSN is subject to a variety of constraints which impact the design of WSN applications. Sensors' energy is the main constraint in WSN because sensors are powered through batteries, which must be recharged when drained. Battery drainage is considered failure in WSN and must be acted upon. Energy conserving strategies varies according to the layer where the designed protocol is located in [7, 8]. This paper focuses on the network layer, where paths from data sources to sink devices are found. If the multi-hop communication model is used, the network layer has to identify good paths from the source sensor to the sink across multiple sensors acting as relays. Designing multi-hop routing protocol for WSNs is challenging due to its resource scarcity and the unreliable wireless medium. In this regard there are different routing protocols for WSN that can be

classified according to network organization, route discovery and protocol operation [1]. Also network topology highly affects the routing protocol complexity [9] and indeed power consumption. Problems in fully connected, star, bus and mesh topologies [9] led us to use the ring topology in the proposed work. In ring topologies, there is no leader node. Messages generally travel around the ring in a single direction. However, if the ring is cut the self-healing ring network (SHR) which is proposed in [9] enables to use the reverse path instead so, SHR has two rings and more fault tolerant.

The objective of this paper is developing a proactive multi-hop routing technique based on the ring topology which handles the power constraint. The proposed technique combines all the sensors and the base station in a single power saving ring chosen by the genetic algorithm via an evolutionary process. Then, the routing and power information is broadcasted from the base station to the sensors to be stored. Every sensor uses its received routing table in delivering data; hence negotiation is minimized.

The remainder of this paper is organized as follows: section 2 illustrates the proposed technique, section 3 shows the simulation results compared to related work. Conclusion is given in section 4.

2 The Proposed Routing Technique

The proposed technique is an energy aware routing technique based on genetic algorithm which makes use from the ring topologies. The simulation results show that this protocol extends the sensor network lifetime. Genetic algorithm is used in the ring selection process because of its high performance in the rapid global search. The limitation in sensors resources is also considered in our work as sensor stores small routing table contains only four entries. Sensor network can continue to operate even if a number of failed nodes exist. Finally we assumed that every node knows its position and sends it to the base station during the initialization phase via a direct communication [10,11]. The proposed technique is described in details in the following subsections.

2.1 Network Topology

In multi-hop transmission communication network, topologies can be classified as fully connected, mesh, bus, or ring. The first three are inconsistent with WSNs [9, 12] accordingly; the ring topology is used in the proposed work. During the rest of this paper we will refer to ring as virtual ring because constructed rings may have intersections in the actual topology. There are two main advantages of the ring topology:

1. Each sensor communicates only with two sensors.
2. All sensors are connected via a single ring without loops.

Concerning the first advantage, a sensor communicates only with two sensors; previous and next hop sensor which enables routing table to hold only two entries. The Next hop sensor is located on the shortest-hop path to the base station while previous hop is located on the reverse longer path. Figure1, shows a simple network of 30 sensors deployed randomly and the routing table of some sensors. The number above circles is the sensor's ID. One should notice that the ring is not always simple as in the figure.

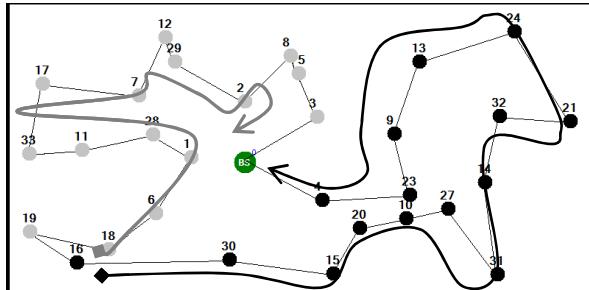
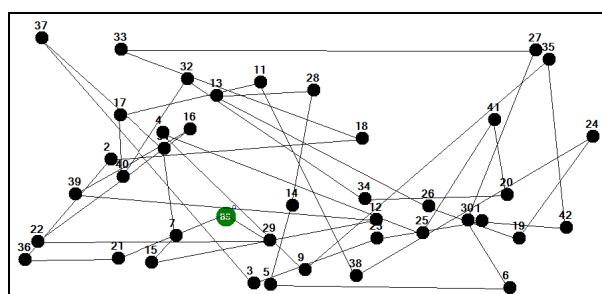
**Fig. 1.** WSN Virtual Ring and Routing

Figure 1 shows sensors with two colors sets to indicate the direction of the default next hops in every set (as indicated by arrows). Additional tuning is added to this advantage which is; two next hops and two previous hops are used as primary and alternate hops. This tuning doubles routing table size but the benefit beyond this is that if the primary next hop fails, source sensor increases transmission power by a prespecified amount (later in this paper we will quantify this power amount) and use the alternate next hop instead of routing data via previous hop.

Table 1. Routing Table of Some Sensors of the Network in Fig.3.

Sensor ID	Next Hop	Previous Hop
3	Base Station	5
23	4	9
4	Base Station	23

The benefit beyond the second advantage is that live lock can't occur. Later, the operation of routing data from source sensor to the base station is illustrated. Virtual rings must be constructed carefully to minimize total consumed power. For this purpose, Genetic algorithm [13] is used to connect all sensors together in a single power efficient ring. Figure 2 shows an example of power consuming virtual ring topology.

**Fig. 2.** Power Consuming Ring.

In [19], it is found that longer multi-hop routes consumes less power than single-hop routes so in the proposed technique multi-hop routing is used. In the present work the average number of hops from source to base station is actually more than the hops involved in the shortest path techniques but from other perspective we have saved sensors memory and processing required in the negotiation for shortest path.

2.2 Wireless Energy Transmission Model

Efficient routing protocols must offload sensors which maximizes the network lifetime. The sensor's transmission and receiving power which depends on distance must be minimized [14, 15, 16]. Genetic algorithm as well as other evolutionary techniques [13, 17, 18] are known to find the optimum solution for problems. The simplicity and the efficiency of the Genetic algorithm motivate us to use it in constructing a virtual ring containing all network sensors with minimum link distances between sensors. Power is directly proportional to the square of distance for constant received power. Accordingly, in order to construct power saving virtual rings, one parameter is needed to be adjusted for optimality that's total distance between sensors. Eqs.(1) and (2) illustrate the model for the radio hardware energy dissipation [20, 21, 22, 23]. For the experiments described here, both the free space power loss (d^2) and the multipath fading power loss(d^4) channel models were used, depending on the distance from the transmitter to the receiver.

$$E_{TX}(K, d) = \begin{cases} KE_{elec} + K\varepsilon_{fs}d^2 & d \leq d_0 \\ KE_{elec} + K\varepsilon_{mp}d^4 & d > d_0 \end{cases} \quad (1)$$

$$E_{RX}(K, d) = KE_{elec} \quad (2)$$

Where E_{TX} and E_{RX} are the transmission and receiving power respectively and K is the packet length in bits, d is the Euclidian distance from source to destination in meters, $\varepsilon_{mp} = 0.0013\text{pJ/bit/m}^4$, $\varepsilon_{fs} = 10\text{pJ/bit/m}^2$, $E_{elec} = 50\text{nJ/bit}$ And $d_0 = \sqrt{\varepsilon_{fs}/\varepsilon_{mp}}$.

2.3 Genetic Algorithm

Genetic algorithms [14, 16, 17] generate solutions to optimization problems using techniques inspired by natural evolution, such as, mutation, selection, and crossover on chromosomes. As stated earlier, Genetic algorithm is used in the present work for its simplicity and hardware consistency. Chromosome coding is a formulation to the problem which an optimum solution is required. In order to start the algorithm, the following operations must be fulfilled: (1) Gene selection and chromosome coding (2) Cost function formulation. These operations are illustrated in detail as follows.

Chromosome Coding. Eq.3 describes the chromosome which represents a virtual ring. Every sensor has a unique (ID). The virtual ring is represented by the sequence of sensor ids in any permutation without repeating. If there is a sensor network of N sensors then a virtual ring that contains all sensors is represented by the following design variable (DV):

$$DV = [S_0, S_1, S_2, S_3, \dots, S_{N-1}] \quad (3)$$

This represents a ring that connects sensor S_i by sensor S_{i+1} , and S_{N-1} by sensor S_0 .

Cost Function Formulation. Distance between a sensor, i , and any other sensor, j , is calculated as:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (4)$$

Where x_i, y_i and x_j, y_j are sensor i and j coordinates respectively. A vector LD_i for each sensor, i , is obtained with the following format:

$$LD_i = [X_{i0}, X_{i1}, X_{i2}, \dots, X_{ij}, \dots, X_{iN-1}], j \neq i, i, j \in [0, N-1] \quad (5)$$

Where the segment X_{ij} is either d_{ij}^2 or d_{ij}^4 according to the distance between the sensor i and sensor j and value of d_0 that is mentioned in eq.1. Genetic algorithm chooses only one segment from every LD_i and sets the other segments to zero so that the energy cost is the sum of all the contents of every LD_i . Eq.6 shows the change in LD_i values.

$$X_{ij} = \begin{cases} 0 & \text{segment not chosen} \\ d_{ij}^2 \text{ or } d_{ij}^4 & \text{segment chosen} \end{cases} \quad (6)$$

A ring will contain N non zero segments chosen from the N^2 segments from every LD_i for every sensor. The energy cost (EC_K) for the K^{th} constructed virtual ring is:

$$EC_K = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} X_{ij}, X_{ij} \in LD_i \quad (7)$$

Eq.7 attempts to add $N \times N$ segments but only N of them segments have a non zero values according to eq.6.

2.4 The Proposed Technique in Action

This section collects the previously discussed steps together to build the proposed routing technique. The proposed routing technique operates on the base station and on any sensor as follows:

Proposed Routing Technique Operation on the Base Station. The technique operation on base station is divided into three main operations:(1)Computing the minimum length virtual ring.(2)Routing information distribution.(3) Processing of The sensor's received data. In the first operation, the process of mutation and crossover is performed on DV for obtaining the minimum EC virtual ring. The second operation is the distribution of routing information to all the sensors in the topology; this is achieved via a high power message (beacon message) from base station to every sensor.The third operation is handling the data received from sensors. A packet received by the base station holds information about the monitored phenomena and information about the path taken from the source. The following pseudocode illustrates the steps held by the base station to complete this operation:

1. *Read active sensor list location information*
2. *Calculate the minimum power virtual Ring*
3. *Distribute routing information and the power amount required to reach any hop in the routing table to every sensor the network.*
4. *Set up a timer for resending new route updates*
5. *If the timer interval elapsed*
 6. *Go to step 1*
 7. *If a packet arrived*
 8. *Examine packet header and extract path information*
 9. *If a packets comes from an alternate previous hop*
 10. *Mark the indicated failed sensors*
 11. *Process data*
 12. *Go to step 1*
 13. *Mark the indicated failed sensors (if any) and update active sensor list*
 14. *Process packets data*
 15. *Go to step 4.*
 16. *End.*

The timer is used to prevent the oscillation that occurs if a total network failure occurs, the total failure occurs when all the next and previous hops occurs in this case any packet transmitted from this sensor will continue to oscillate back and forth, this is the main reason of using a timer.

Proposed Routing Technique Operation on any Sensor. Every sensor in the network has to:(1)Negotiate with the base station to establish the location information. (2)Listen for the route updates and power information sent by the base station.(3)Use its routing table in forwarding data to base station.Operations 1 and 2 are trivial; the following pseudocode illustrates operation 3:

- Assuming a data is required needed to be transmitted to the base station:*
1. *Prepare and send the packet contains send it –but keep it in memory - via the next hop*
 2. *If acknowledgment not received from the receiving hop*
 3. *Encapsulate next hop ID in the packet header and sent it via alternate next hop*
 4. *If acknowledgment not received from the receiving hop*
 5. *Add alternate next hop ID in the packet header and sent it via alternate next hop*
 6. *If acknowledgment not received from the receiving hop*
 7. *Add alternate next hop ID in the stored packet header and sent it via previous hop*
 8. *If acknowledgment not received from the receiving hop*
 9. *Add previous hop ID in the stored packet header and sent it via alternate previous hop*
 10. *If acknowledgment not received from the receiving hop*
 11. *Wait for the new route and power information*
 12. *End*

3 Simulation Results

In order to test the proposed routing technique, genetic algorithm performance is tested via a special simulator built using C# .Net . Routing and power information obtained from the previous simulator is supplied as input to ns-2 simulator. The network has following characteristics : (1) Fixed; i.e., sensors and the base station are all non-moving after deployment. (2) All nodes initially have equal energy; the base station has infinite power supply, powerful computation and processing abilities. (3) A round is defined as the process of gathering all the data from nodes to the sink, regardless of how much time it takes. Most of the simulations results are obtained using the following parameters unless otherwise mentioned: (1) Sensor deployment is random on a given area of 100x100m. (2) The sink node is located and fixed at a (100,100) position. (3) Initial sensor energy is chosen to be 0.0005 J. (4) In a single round every sensor transmits ten 500-byte packets. Simulation results are divided into two sets, the first set is concerned with monitoring the genetic algorithm performance of selecting the best virtual ring and how this affects the lifetime of a sensor network. The second set is dedicated for measuring the performance of the proposed routing algorithm compared with the related work.

Genetic Algorithm Performance. In order to ensure that genetic algorithm operates as expected i.e. the least cost virtual ring is obtained, Fig.3 is plotted to show the cost associated with every constructed virtual ring versus ring cost, in a simulation to a network with 50 sensors.

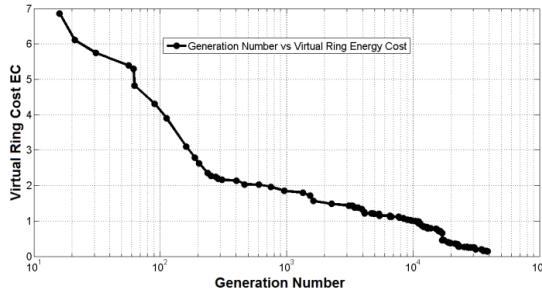


Fig. 3. The Cost/Generation Number Relationship

Figure 4 shows the results of the second simulation for testing genetic performance. In this simulation the relationship between network lifetime and generation number is illustrated. In order to hold this experiment, sensors routing table and power information was derived from a set of an early phases virtual rings (actually they are power consuming) in addition to later phases (power saving). This routing information is used in the ns-2 simulation then the lifetime is recorded for every virtual ring. The previous plot proves that the final computed virtual ring is the one which gives the maximum network lifetime where early constructed virtual rings is power consuming and negatively affects lifetime. Note that in this experiment lifetime is expressed in seconds so that we can handle incomplete rounds, also note that sensor's initial power is .0005 J.

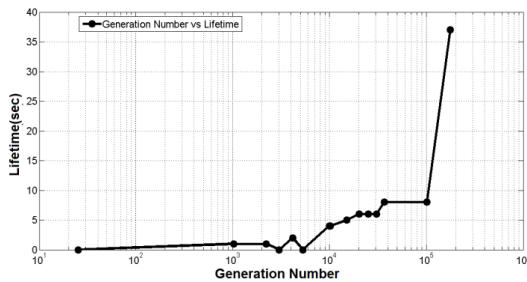


Fig. 4. The Network Lifetime versus Ring Generation Number

From the previous experiments, we can ensure that genetic algorithm behaves as expected and finally the used routing information is derived from the best ring, the next experiments tests the proposed routing protocol lifetime and power residual compared with the related work.

Proposed Routing Algorithm Performance. Lifetime is the most important metric of a routing technique, so, figure 6 compares the proposed technique lifetime against the *ELGA*, *EBGA* and *Direct* techniques [24]. In this simulation initial sensors power is set to 1J so that we can compare the proposed technique lifetime with other techniques. As shown from figure 5, the proposed technique gives the maximum network lifetime versus other techniques. This persuades readers by the aptness of ring topologies in wireless sensor networks if and only if they carefully managed. The next simulation is held to show power residual amount of 20 sensors when the first failure occurs versus the power residual of *ELGA* [24] where the genetic algorithm is used to keep the least average energy consumption, *EBGA* [24] where genetic algorithm is used in addition to taking energy balance into consideration, and the *Direct* technique [25] where direct communication between the node and the sink node is used. The results are shown in figure 6. As shown from the figure few sensors are highly affected and their power is near depletion such as node 10 while others are not. The sensors with small available power involved in a relatively long route compared to other sensors. One should note that the average lifetime of the proposed technique greater than other compared routing techniques. One can conclude from figure 5 and 6 that; while

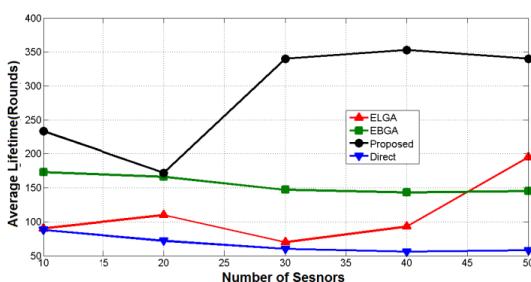


Fig. 5. Proposed Routing Techniqueagainst EBGA, ELGA and Direct

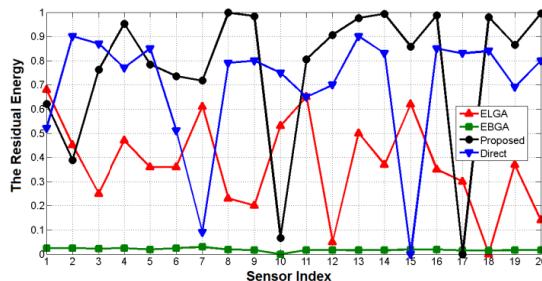


Fig. 6. Power Residual when the First Failure Occurs

the proposed technique's power residual gives greater variance than other, it gives the maximum lifetime. Other near fully powered sensors may be involved as alternate hops which indeed extends the network lifetime.

4 Conclusion

In this paper, we proposed a new centralized routing algorithm for WSN with the aid of the genetic algorithm. Compared with ELGA, EBGA and Direct techniques simulation results show that our protocol can reduce the total energy consumption which prolongs the network lifetime hence more data is delivered to base station. Using the ring topology made the proposed technique fault tolerant, simple and applicable in wireless sensor network.

References

1. Dargie, W., Poellabauer, C.: Fundamentals of Wireless Sensor Networks Theory and Practice. Wiley (2010)
2. Xu, N.: A survey of sensor network applications. IEEE Communications Magazine 40(8), 102–114 (2002)
3. Kumar, S.P.: Sensor networks: Evolution, opportunities, and challenges. Proceedings of the IEEE 91(8), 1247–1256 (2003)
4. Akkaya, K., Younis, M.: A survey on routing protocols for wireless sensor networks. Ad Hoc Networks 3(3), 325–349 (2005)
5. Khedo, K.K., Perseedoss, R., Mungur, A.: A Wireless Sensor Network Monitoring System. International Journal of Wireless & Mobile Networks (IJWMN) 2(2), 31–45 (2010)
6. Zorzi, M.: Wireless sensor networks. IEEE Wireless Communications 11(6), 2–2 (2004)
7. Dalvi, S., Sahoo, A., Deo, A.: A MAC-Aware Energy Efficient Reliable Transport Protocol for Wireless Sensor Networks. In: IEEE Wireless Communications and Networking Conference, pp. 1–6 (2009)
8. Mansouri, V.S.: Transport Protocol for Wireless Sensor Networks. In: 2nd IEEE International Conference on Computer Science and Information Technology, pp. 464–468 (2009)
9. Lewis, F.: Wireless Sensor Network, Technologies, Protocols, and Applications. Computer 38(4), 393–422 (2002)

10. Patwari, N., Hero III, A.O.: Signal strength localization bounds in ad hoc & sensor networks when transmit powers are random. In: Fourth IEEE Workshop on Sensor Array and Multichannel Processing, Waltham (2006)
11. Zhuang, W., Song, G., Tan, J., Song, A.: Localization for hybrid sensor networks in unknown environments using received signal strength indicator. In: Proc. of the IEEE International (2008)
12. Nanda, A., Kumar, A.: Node Sensing & Dynamic Discovering Routes for Wireless Sensor Networks. *International Journal of Computer Science and Information Security* 7(3), 122–131 (2010)
13. Daniels, W.: An Introduction to Numerical Methods and Optimization Techniques. Elsevier (1978)
14. Sivanandam, S.N., Deepa, S.N.: Introduction to Genetic Algorithms. Springer (2008)
15. Callaway, E.H.: Wireless Sensor Networks: Architectures and Protocols. CRC Press (2004)
16. Nallusamy, R., Duraiswamy, K.: Energy efficient dynamic shortest path routing in wireless Ad hoc sensor networks using genetic algorithm. In: ICWCSC, pp. 1–5 (2010)
17. Chi, L., Fan, D.: Genetic Algorithm. *Biotechnology and applied Biochemistry* 58(3), 175–184 (2011)
18. Goldberg, Y.K.S.N.D., Karp, B., Seshan, S.: Genetic algorithms in search, optimization and machine learning. Addison-Wesley (1989)
19. Fedor, S., Collier, M.: On the problem of energy efficiency of multi-hop vs one-hop routing in Wireless Sensor Networks. In: AINAW 2007, vol. 2, pp. 380–385 (2007)
20. Heinzelman, B., Chandrakasan, P., Balakrishnan, H.: An application-specific protocol architecture for wireless Microsensor networks. *IEEE Transaction on Wireless Communication* 1(4), 660–670 (2002)
21. Dasgupta, K., Kalpakis, K., Namjoshi, P.: An efficient clustering-based heuristic for data gathering and aggregation in sensor networks. *IEEE Wireless Communications and Networking*, 1948–1953 (2003)
22. Heinzelman, W., Chandrakasan, P., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. *IEEE Transactions on Wireless Communications* 1(4), 660–670 (2002)
23. Rappaport, T., et al.: Wireless communications: principles and practice. Prentice Hall PTR, New Jersey (1996)
24. Wenliang, G., Huichang, S., Jun, Y., Yifei, Z.: Application of Genetic Algorithm in Energy-Efficient Routing. In: China-Japan Joint Microwave Conference, pp. 737–740 (2008)

Load Balancing with Reduced Unnecessary Handoff in Hierarchical Macro/Femto-cell WiMAX Networks

Prasun Chowdhury, Anindita Kundu, Iti Saha Misra, and Salil K. Sanyal

Department of Electronics and Telecommunication Engineering

Jadavpur University, Kolkata-700032, India

prasun.jucal@ieee.org, kundu.anindita@gmail.com,

iti@etce.jdvu.ac.in, s_sanyal@ieee.org

Abstract. The hierarchical macro/femto cell WiMAX networks are observed to be quite promising for mobile operators as it improves their network coverage and capacity at the outskirt of the macro cell. However, this new technology introduces increased number of macro/femto handoff which inturn may affect the system performance. Users moving with high velocity or undergoing real-time transmission suffers degraded performace due to huge number of unnecessary macro/femto handoff. Our proposed handoff decision algorithm eliminates the unnecessary handoff while balancing the load of the macro and femto cells. The performance of the proposed algorithm is analyzed using Continuous Time Markov Chain (CTMC) Model. In addition, we have also contributed a method to determine the balanced threshold level of the received signal strength (RSS) from macro base station (BS). The balanced threshold level provides equal load distribution to the macro and femto BSs. The balanced threshold level is evaluated based on the distant location of the femto cells for small scaled networks. Numerical analysis shows that threshold level above the balanced threhold results in higher load distribution to the femto BSs.

Keywords: Hierarchical WiMAX Networks, Handoff, Continuous Time Markov Chain, QoS Management, Load Balancing.

1 Introduction

WiMAX (Worldwide Interoperability for Microwave Access) has been widely accepted as the next generation wireless technology to facilitate broadband wireless communication in metropoliton area networks [1]. The recent development of hierarchical macro/femto cell networks is a realistic way to provide better service quality for indoor end users. In WiMAX, femtocells are a cost effective means to provide ubiquitous connectivity. The femto cellular base station is a miniaturized low-cost and low-power Base Station (BS) which uses a general broadband access network as its backhaul [2]. With the introduction of femto cells the total number of active users in the service area increases. However, the mobility of these active users lead to handoff in the hierarchical cell structures.

In recent literatures like [3], [4] authors have presented WiMAX femto cell system architectures and evaluate its performance in terms of network coverage, system capacity and performance of mobile station in indoor environment. On the other hand, authors of [5], [6] have compared the performance in private and public access method in WiMAX femto cell environment. However, the process of handoff and QoS requirement of the mobile stations have not been considered in any of the aforementioned papers.

A variety of handoff algorithms based on received signal strength (RSS) have been considered in [7], [8], [9]. Velocity of the mobile node have been considered in [8], [9] as a parameter for handoff decision. However, QoS guarantee has not been considered in these papers too.

In [10] authors have proposed a new handoff algorithm but QoS profile and network load balancing are not taken into account. Moreover, no relation between the femto cells and the whole system has been considered in their simulation which does not reflect a real mobile WiMAX architecture.

In this paper, we have considered a WiMAX system where a mobile station is moving from a macro cell to a femto cell. We have assumed that the mobile station gives higher priority to the femto BSs than the macro BS. Thus a mobile station selects the femto BS as its serving BS when it receives signal from both the macro and femto BSs as well as the RSS from macro BS falls below its threshold level. In this paper, we have also determined the balanced threshold level of RSS from macro BS based on the distant location of the femto cells for small scaled network. Balanced threshold level provides equal load distribution to the macro and femto BSs. Numerical analysis shows that threshold level above the balanced threshold results in higher load distribution to the femto BSs.

To achieve a QoS aware hierarchical networks, our handoff decision algorithm is based on two main factors viz. the velocity of mobile station and the service type of the mobile station. When a user moves with a ongoing call at a very high velocity (above velocity threshold) from one end of the hierarchical cell to the other end, it is expected that the user will experience huge number of macro/femto handoff in a short period of time. This burdens the overhead of the macro BS. A considerable amount of packet loss may also be encountered which degrades the call quality. On the other hand, though a user moves with a velocity lower than the velocity threshold it experiences comparatively lower handoff rate but handoff still happens. In this case, if the ongoing call is a real-time service then packet loss will hamper the call quality.

To avoid this quality degradation, in our handoff decision algorithm we have considered a velocity threshold such that any user moving with a velocity higher than the velocity threshold or undergoing real-time transmission will not undergo any macro/femto handoff. Thereby, we eliminate unnecessary handoff and provide improved system performance.

The remainder of the paper is organized as follows: Section 2 discusses system model and the detailed description of the proposed handoff decision algorithm of WiMAX macro/femto-cell networks. Section 3 shows analytical model and QoS performance evaluation parameters. Numerical results are discussed in section 4. Finally, Section 5 concludes the paper.

2 System Model and Proposed Handoff Decision Algorithm

A WiMAX macro cell of 1.2 km radius is considered along with multiple femto cells deployed randomly at a distance of atleast ‘R’ meter from the macro BS as shown in Fig. 1. No femto cell is considered within ‘R’ meter of radius of the macro BS because the RSS of the mobile nodes residing within this area is assumed to be quite high. A number of mobile users are deployed randomly under the coverage of the macro BS with varied velocity and undergoing calls of varied service type. The rest of the system model parameters are shown in Table 1.

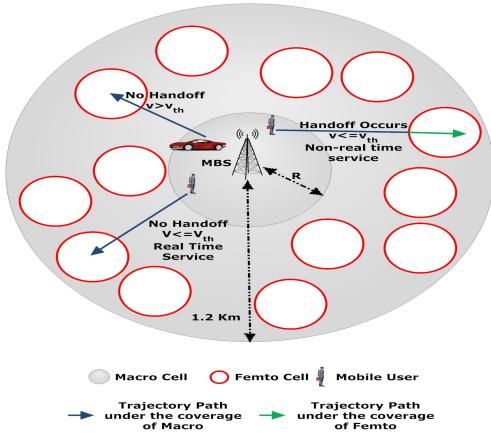


Table 1. System model parameters

Parameters	Value
Macro cell radius	1.2 Km
Femto cell radius	30m
Real-time service type	UGS, rtPS
Non real-time service type	nrtPS, BE

Fig. 1. Handoff decisions in the hierarchical system model

Macro/femto cellular handoff comprises of two main phases – handoff strategy and handoff decision algorithm. The first phase deals with the RSS comparison while in the second phase the system decides when to trigger the handoff. In this paper, we contribute an efficient algorithm for the second phase. The handoff strategy proposed in [7] has been considered for the first phase.

Let RSS_m and RSS_f denote the received signal strength from the macro BS and femto BS experienced by a mobile node at any instant of time. As a mobile node moves in a straight line from the macro BS to femto BS with constant low velocity as shown in Fig. 1, conventional handoff occurs. The conventional handoff algorithm with the RSS comparison [7] can be expressed as in equation (1).

$$RSS_m < RSS_{m,th} \quad \text{and} \quad RSS_f > RSS_m + \Delta \quad (1)$$

where $RSS_{m,th}$ and Δ denotes the minimum RSS threshold level from the macro BS and the value of hysteresis respectively.

The pathloss encountered by a mobile node as it moves away from the macro BS diminishes the RSS_m . As the distance from the macro BS increases, this pathloss triggers the handoff situation where RSS_f becomes higher than RSS_m . In our scenario,

we have considered the ITU pathloss model in slow fading channel [11] as shown in equations (2) and (3).

$$PL_m = 15.3 + 37.6 \log_{10}(D) + PL_{hw} \quad \text{where, } PL_{hw} = 10 \quad (2)$$

$$PL_f = 38.46 + 20 \log_{10}(d) + 0.7d \quad (3)$$

where PL_m and PL_f denotes the pathloss from macro and femto BS respectively. ‘D’ and ‘d’ are the corresponding distance of the mobile user from macro and femto BS.

Thus the resulting RSS_m and RSS_f encountered by the mobile node is shown in equations (4) and (5) respectively.

$$RSS_m = P_{m,tx} - PL_m \quad (4)$$

$$RSS_f = P_{f,tx} - PL_f \quad (5)$$

where $P_{m,tx}$ and $P_{f,tx}$ denotes the transmit power of macro BS and femto BS respectively.

However, in our proposed handoff decision phase the conventional handoff is not adopted in the following two cases.

Case I: When the mobile user is moving at a very high velocity

Conventional handoff is not applicable when a mobile user is moving with a very high velocity. As a user moves with a very high velocity it undergoes huge number of macro/femto handoff within a very short period of time. The overhead of the macro BS thus increases unnecessarily. Hence, in this paper we have considered a velocity threshold ‘ V_{th} ’ of 10 Kmph (non motor vehicular speed) and simulated our scenario accordingly. If a user moves with a velocity ‘V’ such that $V > V_{th}$, unlike conventional scenario the user will not undergo handoff. Thus the unnecessary handoff is eliminated and improved QoS is guaranteed.

Case II: When the mobile user is undergoing a UGS or rtPS i.e. real-time call

When a user is moving with a real-time connection the number of handoff encountered degrades the call quality proportionately. Hence, in our scenario, no handoff is triggered for them in order to maintain the call quality. Thereby unnecessary handoff count decreases and improved QoS is assured to the real-time users.

In our scenario, the QoS guarantee achieved by considering only case I is referred to as soft QoS guarantee while QoS guarantee achieved by considering both case I and case II is called hard QoS guarantee. Soft QoS guarantee only will reduce the overhead of the network. On the other hand, hard QoS guarantee will reduce the network overhead as well as increase user satisfaction.

3 Analytical Model and Performance Evaluation Parameters

In this paper, the performance evaluation of the WiMAX macro/femto-cell networks is obtained by using Continuous Time Markov Chain (CTMC) Model [12]. In addition,

we have considered Pareto distribution for the arrival process of the priority service types, so the network model experiences a flow of service requests in a continuous time domain. The network undergoes a continuous change in its current state due to the occurrence of events (i.e. arrival and service of priority calls). It is necessary to observe the short-lived states of the network in order to analyze its performance more accurately. This is only possible if the network is modeled with CTMC.

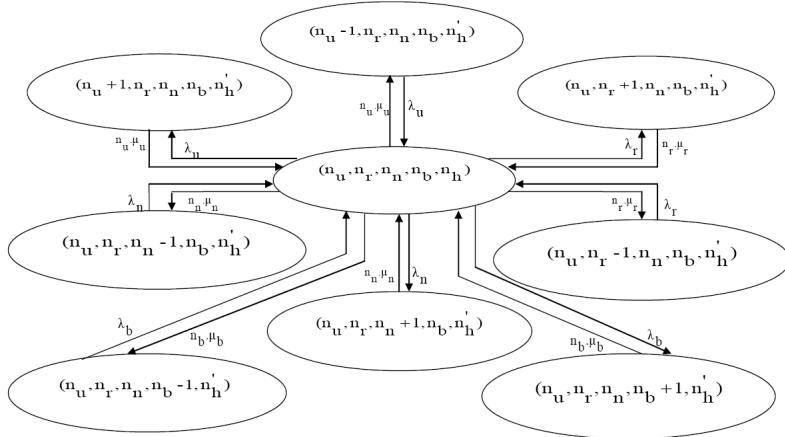


Fig. 2. State transition diagram of the hierarchical WiMAX networks

A hierarchical WiMAX networks consisting of single macro BS along with multiple femto BS is considered. The macro BS will receive the handoff requests from the users directly. Four types of services i.e. UGS, rtPS, nrtPS and BE need QoS guarantees and request for a handoff whenever it find any suitable femto BS in the near vicinity. The hierarchical networks change state from one to another upon the admission or termination of a service type. Further, it is assumed that the hierarchical networks either admit or terminate only one service type at a particular instance of time. So the next state of the hierarchical networks depends only on the present state of the hierarchical networks but does not depend on the previous states of the hierarchical networks. Therefore, the states of the hierarchical networks form a Markov Chain and accordingly the hierarchical networks can be analytically modeled as shown in Fig. 2. In this scenario the hierarchical networks can uniquely be represented in the form of a five dimensional Markov Chain $(n_u, n_r, n_n, n_b, n_h)$ based on the number of services residing within the hierarchical networks and the total number of macro/femto handoff occurred in the network.

State $s = (n_u, n_r, n_n, n_b, n_h)$ represents that the hierarchical networks have currently admitted ' n_u ', ' n_r ', ' n_n ' and ' n_b ' number of UGS, rtPS, nrtPS and BE service respectively. ' n_h ' represents the total number of macro/femto handoff occurred in that state of the hierarchical networks. We have assumed that initially no users are present under the coverage of femto cells. Hence, the total number of users present under the coverage area of femto BSs is also indicated by the parameter ' n_h' '. In Fig. 2, n_h' is

the modified values of the variable ' n_h ' after state transition. Pareto distribution is considered for the arrival process of the newly originated UGS, rtPS, nrtPS and BE with rates of λ_u , λ_r , λ_n and λ_b respectively. This is because Pareto distribution supports more practical traffic model [13]. However, Poisson distribution is an ideal model, which is not practical in real WiMAX networks. The service times of UGS, rtPS, nrtPS and BE connections are exponentially distributed with mean $1/\mu_u$, $1/\mu_r$, $1/\mu_n$ and $1/\mu_b$ respectively.

Let the steady state probability of the state $s = (n_u, n_r, n_n, n_b, n_h)$ be represented by $\pi_{(n_u, n_r, n_n, n_b, n_h)}(s)$. As the Markov chain is irreducible, thereby observing the outgoing and incoming states for a given state s , the steady state probabilities of all states of the hierarchical networks have been evaluated.

From the steady state probabilities we can determine various QoS performance parameters of the system as given below.

A. Handoff Probability (HO_Prob)

Number of handoff occurred in a particular state of the hierarchical networks multiplied with the steady state probability of that state will give the handoff probability of that particular state. Thereby, handoff probability of the hierarchical networks is obtained by summing the handoff probabilities of all the states of the hierarchical networks.

Hence, the probability of handoff occurred in hierarchical macro/femto networks can be calculated as follows.

$$\text{HO_Prob} = \sum_{\forall s} n_h * \pi_{(n_u, n_r, n_n, n_b, n_h)}(s) \quad (6)$$

B. Macro Load (ML)

The Macro load is defined as the ratio of the number of users residing under macro BS to the total number of users present within the macro/femto hierarchical networks. ML can be calculated as follows.

$$\text{ML} = \sum_{\forall s} \frac{(n_u + n_r + n_n + n_b - n_h) * \pi_{(n_u, n_r, n_n, n_b, n_h)}(s)}{n_u + n_r + n_n + n_b} \quad (7)$$

C. Femto Load (FL)

The Femto load is defined as the ratio of the number of handoff users residing under femto BSs to the total number of users present within the macro/femto hierarchical networks. FL can be calculated as follows.

$$\text{FL} = \sum_{\forall s} \frac{n_h * \pi_{(n_u, n_r, n_n, n_b, n_h)}(s)}{n_u + n_r + n_n + n_b} \quad (8)$$

4 Numerical Results and Discussions

The contribution of this paper lies in balancing the network load between the macro BS and femto BSs and the reduction of the unnecessary handoff. Exhaustive simulations have been carried out under MATLAB version 7.3. Since our main aim is to reduce unnecessary handoff not the handoff latency, in our simulation the value of hysteresis has been taken as zero. The arrival rates of all the connections are assumed to be same i.e. $\lambda_u = \lambda_r = \lambda_n = \lambda_b$. The values of the rest of the simulation parameters are shown in Table 2. In Table 2 the macro and femto transmit power has been taken from [14]. The results associated with the load balancing are shown in Fig. 3a, 3b, 4 and 5 while reduction of the unnecessary handoff is exhibited in Fig. 6. Justifications behind all the numerical results have also been provided.

Table 2. Simulation parameters

Parameters	Value
Macro transmit power ($P_{m,tx}$)	46 dBm
Femto transmit power ($P_{f,tx}$)	20 dBm
Velocity threshold (V_{th})	10 kmph
Traffic ratio of UGS, rtPS, nrtPS, BE	1:1:1:1
$\mu_u = \mu_r = \mu_n = \mu_b$	0.2
Femto cell deployment	Random

As the threshold level of RSS_m or $RSS_{m,th}$ increases, the load of the macro BS decreases and the load of the femto BSs increases. This is revealed from Fig. 3a and 3b respectively. With increase in the value of $RSS_{m,th}$ the macro/femto handoff count increases. Hence, the macro load decreases while increasing the femto BSs load.

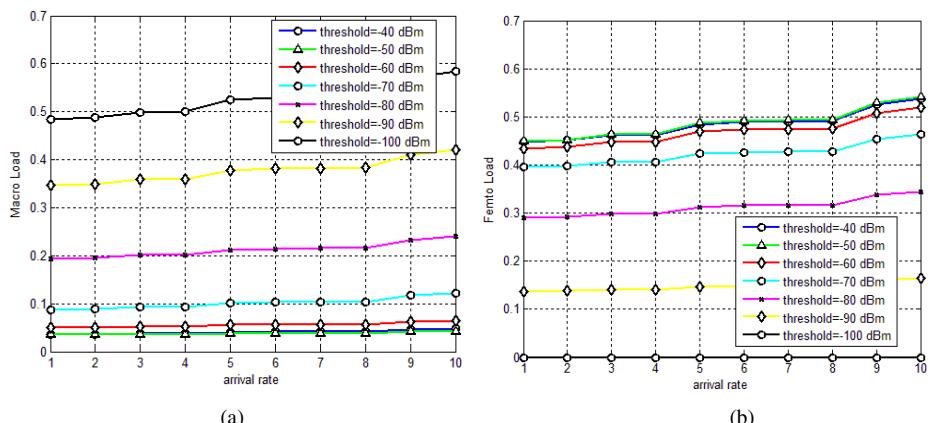


Fig. 3. (a) Macro Load for various $RSS_{m,th}$ when ' R '=100m and (b) Femto Load for various $RSS_{m,th}$ when ' R '=100m

A saturation is observed in the load of both macro and femto BSs when $RSS_{m,th}$ reaches -50 dBm. No change is observed when value of $RSS_{m,th}$ increased further. It happens due to absence of femto cells within ' R '=100 m radius of macro BS as shown in Fig. 1. Hence, when $RSS_{m,th}$ goes above -50 dBm, the mobile nodes residing within ' R '=100 m of macro BS do not find any femto BS. So no handoff is triggered and the load remains unchanged. Again, femto load is found to be zero at $RSS_{m,th}$ = -100dBm due to macro cell outage. At this level of RSS threshold no femto cells are present to trigger handoff.

To balance the load of the macro BS and femto BSs, a study of the variation of their load has been performed with respect to the $RSS_{m,th}$. This variation is performed when femto cells are located at ' R '=100m apart from macro BS and is shown in Fig. 4. The point of intersection of these two variations provides the value of $RSS_{m,th}$ at which the macro and femto load are observed to be same. It is observed that at $RSS_{m,th}$ = -83.4 dBm a balance between the load of the macro BS and femto BSs is achieved. Henceforth, Macro threshold level ($RSS_{m,th}$) at which load balancing is achieved is referred to as balanced threshold level. Load distribution to the femto BSs increases as the $RSS_{m,th}$ goes above the balanced threshold level i.e. -83.4 dBm.

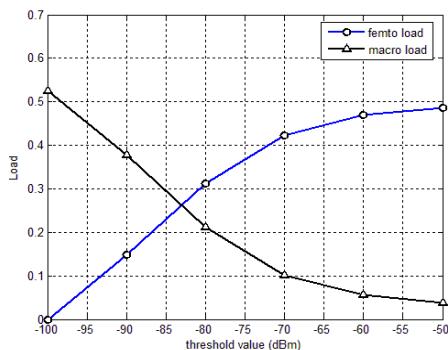


Fig. 4. Comparison of macro/femto load for various $RSS_{m,th}$ when ' R '=100m

Table 3. Threshold level for load balancing

Variation of ' R ' (meter)	Balanced threshold level (dBm)
200	-85.7
300	-87.1
400	-88.6
500	-89.8
600	-90.9
700	-92.1
800	-93.8
900	-95
1000	-95
1100	-95

The above mentioned scenario has been generalized by varying the parameter ' R '. The corresponding balanced threshold level is evaluated in the similar way and is shown in Table 3. Thus, in order to have higher load distribution to the femto cells the macro BS should set the macro threshold level above the balanced threshold level with respect to the value of ' R '. The method of determining the balanced threshold level has been assessed for small scaled networks as the simulations are very time-consuming for broad scaled networks. This method can also be applied for broad scaled networks conceptually to calculate the corresponding balanced threshold level.

In Table 3, as the value of ' R ' increases, the balanced threshold level is observed to decrease gradually unless ' R ' reaches 900 meters. From 900 meters onwards the macro threshold level is observed to remain constant at -95 dBm. The reason behind this is elaborated in Fig. 5. Fig. 5 shows the variation of the RSS of the mobile station with respect to its distance ' D ' from the macro BS as obtained from equation (4).

From Fig. 5 we see that as the mobile station reaches the outskirt of the macro cell edge i.e when $D = 1.2$ km, the RSS encountered is -95 dBm. Thus combining Fig. 5 and Table 3 we conclude that even if the distance of the femto cells from the macro BS is 900 meters and beyond, the macro threshold has to be above -95 dBm to achieve higher load distribution to the femto cells.

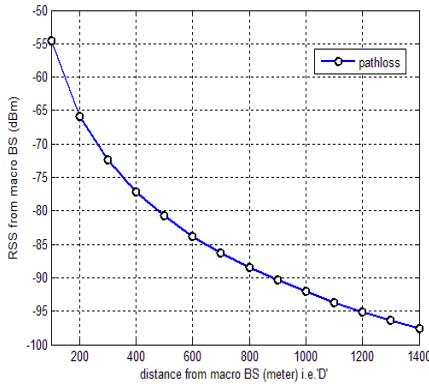


Fig. 5. Pathloss from macro BS

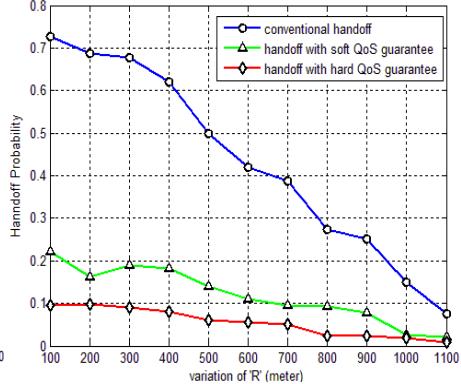


Fig. 6. Handoff probability at $RSS_{m.th} = -70$ dBm

Considering the above fact, we have observed the handoff probability for various kind of handoff decision discussed in this paper keeping $RSS_{m.th}$ at -70 dBm which is above the balanced threshold level for any ' R' . The result is shown in Fig. 6.

Fig. 6 reveals that the probability of handoff decreases to a considerable amount when unnecessary handoff is eliminated from the conventional handoff scheme. Again, the handoff probability in each case is also observed to fall gradually as ' R' increases. With increase in ' R' ', the outskirt region to be covered by the femto cells decreases. This in turn lowers the number of femto cells and thereby the handoff probability decreases. The handoff probability is observed to be much lower in case of hard QoS guarantee than the case of soft QoS guarantee. In a hierarchical cell scenario as shown in Fig. 1, if a user moves with very high velocity or undergoes real-time call while moving from one end of the cell to the other end, reduction in the number of handoff is going to improve the call quality considerably. Thus soft QoS guarantee will ensure better call quality while hard QoS guarantee will provide much better call quality compared to conventional handoff scheme.

5 Conclusion

In this paper, we have proposed a handoff decision algorithm for reducing unnecessary handoff in macro/femto hierarchical networks while balancing the load of macro and the femto BSs. The performance of the proposed algorithm is also analyzed using CTMC model. Balanced threshold level of RSS from macro BS have been evaluated with respect to the distant location of the femto cells. Macro threshold

level ($\text{RSS}_{m,\text{th}}$) set above the balanced threshold level results in higher load distribution to the femto cells. Hard QoS and soft QoS guarantee – the two decision of handoff reduction technique proposed in this paper shows how unnecessary handoff has been reduced while balancing the load of the macro and femto BSs. Soft QoS guarantee only will reduce the overhead of the network. On the other hand, hard QoS guarantee will reduce the network overhead as well as increase user satisfaction. So a service provider can choose any one of the QoS guarantee level depending upon their requirement.

Since simulations are very time-consuming for broad scaled networks, the method of determining the balanced threshold level has been assessed for small scaled networks. However, this method can also be applied for broad scaled networks conceptually and the corresponding balanced threshold level can be calculated accordingly.

Acknowledgments. The authors deeply acknowledge the support from DST, Govt. of India for this work in the form of FIST 2007 Project on “Broadband Wireless Communications” in the Department of ETCE, Jadavpur University.

References

1. IEEE 802.16 Standard- Local and Metropoliton Area Networks-Part 16. IEEE Draft P802.16/D3-(2001)
2. Shu-ping, Y., Talwar, S., Seong-choon, L., Heechang, K.: WiMAX Femtocells: A Perspective on Network Architecture, Capacity, and Coverage. *IEEE Communications Magazine* 46(10), 58–65 (2008)
3. Kim, R.Y., Kwak, J.S., Etemad, K.: WiMAX femtocell: requirements, challenges, and solutions. *IEEE Communication Magazine* 47, 84–91 (2009)
4. Zeng, H., Zhu, C., Chen, W.: System performance of selforganizingnetwork algorithm in WiMAX femtocells. In: Proceedings of the 4th ACM International Conference Proceeding Series, pp. 1–9. ICST, Brussels (2008)
5. Lopez-Perez, D., Valcarce, A., De La Roche, G., Liu, E., Zhang, J.: Access Methods to WiMAX Femtocells: A downlink system-level case study. In: 11th IEEE International Conference on Communication Systems, pp. 1657–1662 (2008)
6. Claussen, H.: Performance of macro- and co-channel femtocells in a hierarchical cell structure. In: IEEE 18th International Symposium on PIMRC 2007, Athens, Greece, pp. 1–5 (2007)
7. Halgamuge, M., et al.: Signal-based evaluation of handoff algorithms. *IEEE Communication Letters* 9(9), 790–792 (2005)
8. Hsin-Piao, L., Rong-Terng, J., Ding-Bing, L.: Validation of an improved location-based handover algorithm using GSM measurement data. *IEEE Transactions on Mobile Computing* 4, 530–536 (2005)
9. Denko, M.K.: A mobility management scheme for hybrid wired and wireless networks. In: Proceedings of the 20th International Conference on Advanced Information Networking and Applications, vol. 02, pp. 366–372 (2006)
10. Moon, J., Cho, D.: Efficient handoff algorithm for inbound mobility in hierarchical macro/femto cell networks. *IEEE Communications Letters* 13(10), 755–757 (2009)

11. Oh, D.C., Lee, H.C., Lee, Y.H.: Cognitive Radio Based Femtocell Resource Allocation. In: International Conference on Information and Communication Technology Convergence (ICTC), pp. 274–279 (2010)
12. Ross, S.M.: Probability Models for Computer Science. Elseveir (June 2001)
13. Baugh, C.R., Huang, J.: Traffic Model for 802.16 TG3 Mac/PHY Simulations. IEEE 802.16 working group document (2001)
14. Chandrasekhar, V., Andrews, J., Gatherer, A.: Femtocell Networks: A Survey. IEEE Communication Magazine 46(9), 59–67 (2008)

A Study on Transmission-Control Middleware on an Android Terminal in a WLAN Environment

Hiromi Hirai, Kaori Miki, Saneyasu Yamaguchi, and Masato Oguchi

Ochanomizu University, Department of Information Sciences,
Tokyo, Japan

hiromi,kaori@ogl.is.ocha.ac.jp,
sane@cc.kogakuin.ac.jp,
oguchi@computer.org
<http://ogl.is.ocha.ac.jp/>

Abstract. In this study, we present a transmission-control middleware, which enables an Android terminal to select a suitable TCP in a WLAN Environment. Various approaches toward developing a congestion control algorithm of TCP have been proposed to prevent congestion. Some of these approaches are loss based whereas others are delay based to predict network traffic, and their hybrid type also exists such as Compound-TCP[9] and TCPIllinois[10]. However, all of these approaches are designed to allow each terminal to run independently. Moreover, in the case of a mobile terminal, its TCP is limited to behave modestly to avoid filling the bandwidth. In this paper, we suggest a middleware that exchanges communication conditions to predict traffic on the basis of the number of communication terminals connected to the same access point. In the future, we will improve the middleware to predict the values of the Congestion Window of the other terminals.

Keywords: Middleware, TCP, Congestion control, Android.

1 Introduction

In recent years, the digital convergence era has arrived. Computers are required to be convenient anywhere and anytime. The huge demand for portable computers produced the smartphone, a mobile computer. However, the architecture of the smartphone is different from that of the general-purpose PC because of the poor function of the I/O interface and the limited hardware of the mobile phone. To overcome this disadvantage, the smartphone is constantly connected to the Internet and obtains information through the Cloud. The proliferation of SNS services has also stimulated users to publish information with their smartphones. For instance, users can upload their videos to Youtube and synchronize their calendars or address books in the Cloud. Thus, the performance of the smartphone is measured not only by its processing power but also by its network availability. We focus on the networking system of the smartphone to improve upon it.

We adopt Android as the OS to be researched because it has the highest share in the international smartphone market. Moreover, Android's source code is open, and developers can freely customize or remodel the code.

The smartphone is a useful device that enables users to connect to the Internet, regardless of time and location. If a user uses a smartphone while traveling, the device connects the user from access point to access point. However, if a large number of smartphones connect to an access point at the same time, congestion will frequently occur.

In this study, we attempt to develop a middleware through which each internet device can share its own Congestion Window (CWND) to predict network traffic. To date, a collection of various TCPs have been developed. The previous TCPs predicted the traffic via packet loss or delay. All of them are designed for general-purpose PC to control the data flow independently. Android is one of OSs developed for mobile device. It is adopted TCP-cubic[1] and the highest value is limited so as not to interrupt own working system. TCP-cubic has loss-based congestion control, with which CWND is decreased drastically by a packet loss. In WLAN environment, packet loss does not always indicate congestion because of noise. Actually packet loss rate in WLAN is much higher than that in wired LAN. Nowadays, the technology of cloud is also growing as rapidly as that of smartphone. Smartphone often depends on cloud for processing. In this study, the case is assumed in which smartphone transmits huge packets to cloud server as shown Figure 1. In recent years, smartphone tends to enjoy mass rapid communications. Cloud service supplies reliable WAN to be accessed from any part of the world. Therefore, most of the packet loss takes place between smartphone and an access point. Moreover the value of CWND is important because RTT is expected high.

If only one Android terminal monopolizes an access point or base station without any obstacles, the terminal is supposed to maintain the highest communication throughput. Nevertheless, the terminals are scrambling against each other for bandwidth while a large number of nodes usually connect to an access

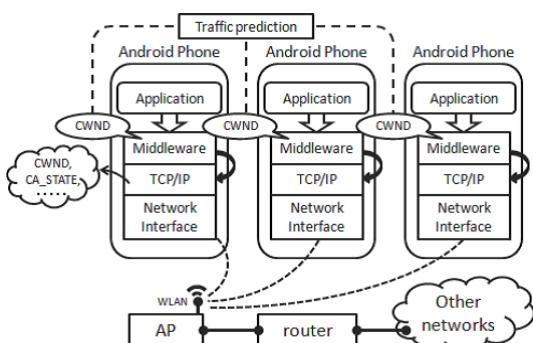


Fig. 1. The concept of our proposal

point. In this case, the transport layer plays a critical role because it controls the data flow. CWND also indirectly influences the communication throughput in an Android terminal, as demonstrated by [14].

Additionally, in a crowded situation, CWND becomes insecure and is excessively restricted. CWND, which controls the amount of data segments to flow, is expected to increase as much as possible. However, a huge packet from a terminal whose CWND is excessively fixed may disturb the whole network and even interrupt the transmission with frequent packet losses and retransmissions [13]. Furthermore, no TCP algorithm always behaves appropriately [8], as any TCP has advantages and disadvantages.

Thus, TCP switching is important for each terminal because the terminals need to be able to adapt to their environments. Because the conventional transport layer for mobile terminals is designed to behave modestly to avoid jamming the network traffic, transmissions by mobile terminals have more latency. However, in recent years, mobile terminals have also come to enjoy rapid communication on a massive scale (i.e., an adequate mechanism is required to control mobile communication.) Therefore, we propose a transmission-control middleware to flexibly control TCP.

2 Android OS

2.1 Architecture

Android is a software stack for mobile devices that includes an operating system, middleware and key applications[2]. It was developed by the Open Handset Alliance (OHA), which is mainly composed of Google. As shown in Figure 2, Android, whose basement is Linux 2.6, has extra components for the smartphone and the tablet. Android typically has an original virtual machine or Dalvik VM.

As a Linux distribution, Android has the C library and is able to execute native code. However, the Dalvik binary written in Java is also useful on Android. Dalvic VM provides high-quality portability and security. Moreover, with the Dalvik library, which has various user interfaces and frameworks, users can enjoy the intuitive functions.

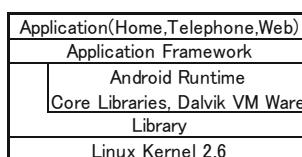


Fig. 2. Architecture of Android

2.2 Application

The number of Android applications has increased rapidly for several reasons. First, both the package of the Android OS and the Android SDK used to develop applications are free. Furthermore, although they are free, their quality is still high. Second, the Android Market has enabled developers to spread Applications easily. Once an application is registered in the Android Market, it can be downloaded all over the world. The Android Market has also enabled users to obtain applications easily. Because of Dalvik VM, users only have to download a byte-codes to execute an application. Various applications can be easily installed to improve the usability of the smartphone itself. In this study, we understand the importance of applications on Android, although we focus on Android as a system platform. We improve the communication performance of Android through Dalvik VM.

3 Congestion Control of Android

It is reported that initial CWND should be more than 10 to finish transmission quickly and that it interrupts no other communication[4]. We adopted this idea and set it 10.

3.1 Congestion Control Algorithm

As is widely known, the transport layer on the Linux OS plays a critical role because it controls the data flow. CWND is a parameter that limits the data flow to control network congestion. In the exact definition, CWND denotes the number of maximum packets that can be sent continuously without receiving acknowledgement from a data receiver. CWND also indirectly influences the communication throughput in the Android terminal, as demonstrated in a 2010 study[14]. In particular, in a high-delay environment, the impact is high.

CWND is directly related to the sender's transmission rate and should be set on the basis of the available bandwidth of an end-to-end connection so as to prevent network congestion. The default TCP of Android is TCP-cubic, which is an enhanced version of TCP-bic. Both of them have loss-based control to detect congestion. In the control, CWND is increased gradually per an Acknowledgement and halved each time a single packet loss is experienced. There are various error events of transmission, such as Local device congestion, Duplicate Acknowledgment, Selective Acknowledgment Options and Timeout. Especially in wireless LAN environment, packet loss is frequently caused by noise. Loss-based TCP assumes that packet loss happens only due to overflow of data segments. However, Packet loss does not always indicate congestion in wireless LAN environment. Actually delay-based TCPs, such as TCP-westwood and TCP-vegas, can fill more bandwidth than loss-based ones[5][6][7].

Previous study[13] improved TCP-cubic not to reduce CWND too much but to fill traffic with more segments. In this paper, we made experiment in switching

TCP between TCP-cubic as default and original TCP. The original TCP is designed to adjust the sender's CWND high and transmit packets aggressively. It is argued that aggressive TCP might interrupt the others' transmission. However, it should be allowed for a smartphone to use it when no other devices are communicating.

3.2 How to Obtain the Parameters in Kernel

CWND is a parameter in Kernel. Kernel is a special software program that is different from other applications. Because it cannot accept normal debug methods, the behavior of kernel during communication is difficult to observe, even in the case of the general-purpose PC. This problem can be solved by using the Kernel Monitor[12].

In this paper, we applied the Kernel Monitor to Android, an embedded system. Kernel Monitor is our original system tool that can record the value of each parameter in the kernel during the communication process. CWND, timestamps, the size of the socket buffer queue, error events and the other parameters are recordable. An overview of the Kernel Monitor is shown in Figure 3. The Kernel Monitor can leave a log of TCP behavior by inserting the monitor function into the source code of TCP and rebuilding the kernel.

Because Android is an embedded terminal, the amount of resources in Android terminals, such as storage and memory, are not sufficient to output the log that is input at the same time. To obtain a real-time log, we customized the Kernel Monitor to leave only the latest log and embedded it into the Android Kernel.

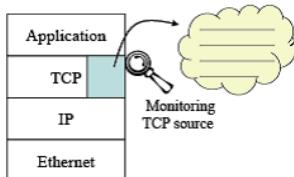


Fig. 3. Kernel monitor

3.3 Congestion in Coexisting Environment

If the network is crowded, packet loss often occurs. In this case, CWND is decreased drastically to prevent any more congestion. The transmission rate is also reduced at the same time. However, the present congestion control is running independently. If terminals increase their CWND and lose some packets at the same time, all of the terminals will decrease their CWND only to leave the traffic light. Furthermore, the TCPs of mobile terminals are designed to be modest in transmission. However, if only one terminal monopolizes an access

point, it should be allowed to use more bandwidth. Therefore, synchronized congestion control must fill the bandwidth. In this study, we will offer not only TCP-friendliness but also fairness and effectiveness by providing synchronized congestion control in the future.

3.4 The Design of Transmission-Control Middleware

We propose a middleware that shares a communication condition among the other mobile terminals connected to an access point and selects a suitable TCP for the environment. Currently, common applications on smartphones are supported by Cloud services. Cloud services are protected in such a reliable network that the bottleneck of an end-to-end network lies around the WLAN access point. Therefore, traffic can be predicted more accurately by the other communication condition.

The middleware we present in this paper is composed of 3 parts. first one is an agent program written in C between application and Kernel Monitor. It is crosscompiled by GNU ARM tool chain[3]. It lets proc interface output the log left by Kernel Monitor and inform application of timestamp and CWND. Second one is a resident service activity that executes the agent program and get information about TCP to share among the other Android terminals during communication. Third one is a manager to predict the traffic and select suitable TCP. Android device is mobile and might be used in various environments. Thus we developed a middleware so as to keep more bandwidth and higher communication throughput.

4 The Policy of Switching TCP

4.1 Suitable Congestion Control Algorithm

TCP is designed for general purposes and does not always behave properly. Even if only one or two Android terminals are connected to an access point, the default TCP behaves modestly. In other words, it takes an accidental packet loss

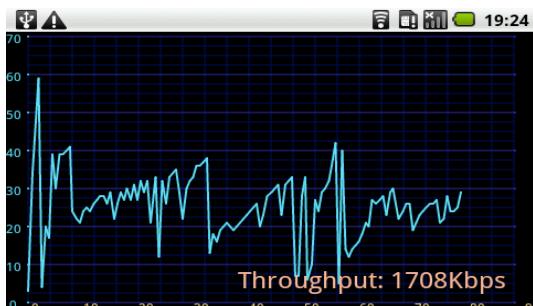


Fig. 4. Congestion control of default TCP (x-axis:time,y-axis:CWND)

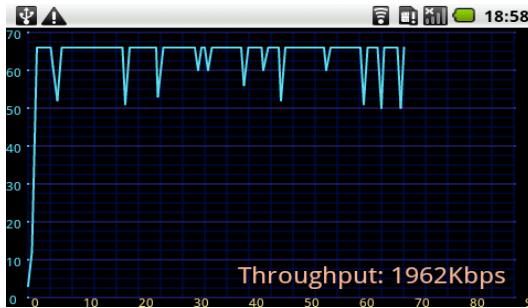


Fig. 5. Congestion control of original TCP (x-axis:time,y-axis:CWND)

as congestion and decreases the CWND drastically if there is enough bandwidth. Default TCP means that TCP-cubic is initially embedded in the Android kernel.

In this study, the middleware observes the number of Android terminals connected to the same access point. If it notices that there is no other terminal, the middleware switches its TCP from default to original. If the middleware notices that some terminals are connecting to the access point, it switches its TCP from original to default to avoid negatively affecting the other terminal's transmission.

In Figure 4 and 5, the experiment is conducted in Round Trip Time 128 ms and artificial packet loss 1%. These graphs are drawn on a TCP visualization tool as the author's graduation thesis. The figures show the behavior of the default TCP and that of the original one. In these figures, the horizontal axis and vertical axis denote time (seconds) and CWND, respectively. These congestion controls are observed if each terminal executes an application that transports a 16 Mbyte packet. Additionally, they are examined separately to monopolize the network.

Compared with the default TCP, the original TCP rapidly increases the CWND and decreases little even if several packets are lost. Moreover, low CWND is also recovered from quickly. It is reported that the original TCP is effective if RTT is longer than 64 ms. Therefore, the default TCP tends to excessively limit the CWND to retain high throughput.

5 Sharing of Communication Condition

Currently, most of the programs used to switch TCPs are working on Dalvik VM. We will embed them into the library as a middleware in the future.

5.1 An Application on Terminal

This application measures the latency taken to transmit some packets to measure the communication throughput and visualizes the congestion control.

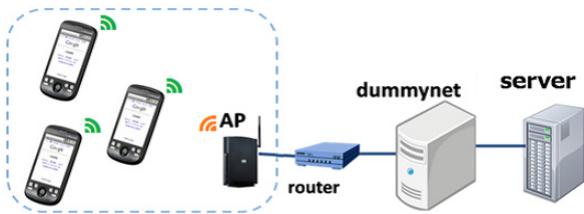


Fig. 6. The model of experiment

5.2 How the Application Obtains Parameter in Kernel

In this experiment, we used two server programs on a PC. One is a receiver of packets from an experimental Android device. The other is an echo server that receives information about the communication condition and broadcasts it throughout the access point. If the receiver receives the whole packet, it replies with a sign. By comparing the time to start with the time of this sign, the application calculates the throughput. As shown in the figures of this application, the communication throughput of the transmission appears at the bottom.

However, while the connection is being established, the agent program is receiving its own TCP information left by the Kernel Monitor to analyze the TCP once per 800 ms in parallel. The function of accessing the Kernel Monitor is in an infinite loop with *sleep()*.

5.3 Exchanging CWND

The analysis of own TCP behavior is shown in section 5.2. We explain the sharing system in the next section.

Each terminal sends its own CWND by UDP after analysis. UDP is better to exchange each infomation because previous data are not required to retransmit in the case of packet loss. The information broadcasted from a terminal is caught by the other terminals' middlewares. Afterward, the middleware of the terminals considers whether it should switch TCPs. The visualization tool is also listening to the broadcasted information and drawing the graph on the basis of the data.

5.4 Experimental Environment

The experimental environment is shown in table 1 and figure 6. FreeBSD was inserted between the access point and the server machine to work as Dummynet. The artificial delay is set at 128 seconds. The TCP behaviors of all of the terminals in an access point are visualized by using the original tool.

5.5 The Detail about the Experiment

In these experiments, three Android devices connect to the server through a WLAN access point. Whereas one terminal is transmitting 16 Mbyte of data,

Table 1. The experimental environment

Android	Hardware	HT-03a
	Model number	AOSP on Sapphire(US)
	Firmware version	2.1-update1
	Baseband version	62.50S.20.17H_2.22.19.26I
	Kernel version	2.6.29-00481-ga8089eb-dirty
	Build number	aosp_sapphire_us-eng 2.1-update1
Server	CPU	Intel Pentium M 1.30GHz
	Main Memory	256MB
	OS	Linux2.6.32-31-generic
AP	Maker	BUFFALO
	Product name	WHR-G301N/U AirStation
	Mode	IEEE 802.11g

the other terminals start transmitting 4 Mbyte of data to interrupt the transmission. We conducted two experiments related to the transmission to compare non switching-TCP and switching-TCP.

In experiment 1, all of the Android terminals are transmitting with the default TCP. In contrast, in experiment 2, the middlewares are running to control the TCP in each section 4.1. That is, the main terminal transfers 16 Mbyte of data with the original TCP when it monopolizes the access point. Then, the main one switches its TCP from the original to the default when one of the other terminals begins transmitting. Finally, the main terminal changes its TCP from the default to the original when all of the others finish transmitting.

5.6 Experimental Evaluation

The results of these experiments are shown in Figure 7 and 8. In experiment 1, the main terminal behaved modestly, even though the other terminals did not communicate. The value of CWND was fluctuating during the whole time. In experiment 2, because of the switching TCP, the main terminal could transfer

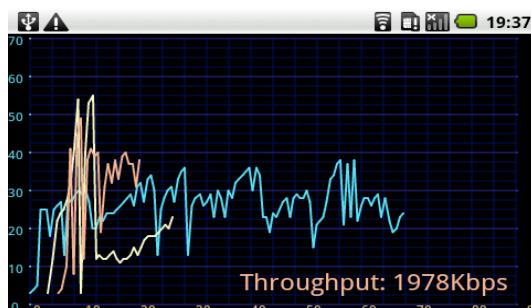
**Fig. 7.** The result of experiment 1 (x-axis:time,y-axis:CWND)



Fig. 8. The result of experiment 2 (x-axis:time,y-axis:CWND)

more packets to fill the rest of the bandwidth than it could in experiment 1 before the others had started transmitting and after they completed their transmissions.

The typical results are presented. No terminal showed the same behavior or communication throughput.

6 Conclusion and Future Works

In this paper, we showed the middleware that exchanges communication conditions in an access point. Communication throughput on the Android terminal is improved with the original TCP. Switching TCP is effective method because it allows the original TCP to run only if no other terminals are communicating. Overflow in traffic often has a negative impact on the other terminals or the network itself. Because the original TCP is aggressively designed, we must take care when using it.

We would like to improve this transmission-control middleware to fill the network traffic to the extent that congestion occurs. The number of communicating terminals is cared by the middleware in this paper. The middleware will be designed to consider the CWNDS of the other terminals in the future.

References

1. Ha, S., Rhee, I., Xu, L.: CUBIC: A New TCP-Friendly High-Speed TCP Variant. *SIGOPS Operating Systems Review* 42(5), 64–74 (2008)
2. android developers, <http://developer.android.com>
3. Sourcery G++ Lite for ARM GNU/Linux, <http://www.codesourcery.com/>
4. Dukkipati, N., Rece, T., Cheng, Y., Chu, J., Herbert, T., Agarwal, A., Jain, A., Sutin, N.: An Argument for Increasing TCP's Initial Congestion Window. In: Proc. ACM SIGCOMM Computer Communications Review, vol. 40(3), pp. 27–33 (July 2010)
5. Mascolo, S., Casetti, C., Gerla, M., Sanadidi, M.Y., Wang, R.: TCP Westwood: Bandwidth Estimation for Enhanced Transport over Wireless Links. In: Proc. ACM SIGMOBILE 7, 2001, Rome, Italy (July 2001)

6. Grieco, L.A., Mascolo, S.: Performance Evaluation and Comparison of Westwood+, New Reno, and Vegas TCP Congestion Control. Proc. Computer Communication Review 34(2), 25–38 (2004)
7. Capone, A., Fratta, L., Martignon, F.: Bandwidth Estimation Schemes for TCP over Wireless Networks. IEEE Transactions on Mobile Computing 3(2) (April-June 2004)
8. Katto, J., Ogura, K., Fujikawa, T., Kaneko, K., Su, Z.: On Hybrid TCP Congestion Control. In: Proc. ICCCNS, TCP(2008)
9. Blanc, A., Avrachenkov, K., Collange, D., Neglia, G.: Compound TCP with Random Losses. In: Fratta, L., Schulzrinne, H., Takahashi, Y., Spaniol, O. (eds.) NETWORKING 2009. LNCS, vol. 5550, pp. 482–494. Springer, Heidelberg (2009)
10. Liu, S., Car, T.B., Srikant, R.: TCP Illinois: A Loss and Delay Based Congestion Control Algorithm for High Speed Networks. In: Proc. VALUETOOLS, Pisa, Italy (October 2006)
11. Leith, D.J., Andrew, L.L.H., Quetchenbach, T., Shorten, R.N., Lavi, K.: Experimental Evaluation of Delay/Loss-based TCP Congestion Control Algorithms. In: Proc. the 6th International Workshop on Protocols for Fast Long-Distance Networks (PFLDnet 2008), Manchester, UK, March 5-7 (2008)
12. Miki, K., Yamaguchi, S., Oguchi, M.: Kernel Monitor of Transport Layer Developed for Android Working on Mobile Phone Terminals. In: Proc. the Tenth International Conference on Networks (ICN 2011), St. Seattle, USA (April 2010)
13. Miki, K., Yamaguchi, S., Oguchi, M.: A Study about Performance Improvement and Analysis of Communication on Android in a Wireless LAN with Kernel Monitor. In: Multimedia, Distributed, Cooperative, and Mobile Symposium, DICOMO 2011, 7H-2 (July 2011)
14. Hirai, H., Miki, K., Yamaguchi, S., Oguchi, M.: A Visualization Tool for Communication Performance of Android Terminal. In: The 73rd National Convention of IPSJ, 5V-9 (March 2011)

A Study of Location-Based Data Replication Techniques and Location Services for MANETs

C.B. Chandrakala¹, K.V. Prema², and K.S. Hareesha³

¹ Dept. of Information and Communication Technology,
Manipal Institute of Technology, Manipal University, Manipal, India

² Department of CS&E, MITS University, Rajasthan, India

³ Dept. Of Computer Science and Engineering, Manipal Institute of Technology,
Manipal University, Manipal, India
{chandrakala.cb@manipal.edu}

Abstract. A special class of mobile application has been made feasible by Mobile ad-hoc networks. They benefit from the fast deployment and reconfiguration of the networks, are mainly characterized by the need to support many-to-many interaction schema within groups of cooperating mobile hosts and are likely to use replication of data objects to achieve performances and high data availability. This strong group orientation requires specialized solutions that combine adaptation to the fully mobile environment and provide the adequate level of data and service availability with certain amount of fault tolerance. In this paper performance analysis of existing Location-based MANETs techniques with respect to various issues they address such as Data availability and data consistency, Partition detection, etc., is carried out.

Keywords: Wireless Network, Mobile Ad-Hoc Network, Location-based routing, Data replication, Location services.

1 Introduction

Mobile ad hoc networks (MANETs) are networks composed of a set of communicating devices able to spontaneously interconnect without any pre-existing infrastructure. This systems consist of a set of mobile hosts, connected to the network through wireless links. They differ from traditional and nomadic distributed systems in that they have no fixed infrastructure. mobile hosts can isolate themselves completely and groups may evolve independently. Connectivity may be asymmetric or symmetric depending, for instance, on the radio frequency of the transmission used by the hosts. Radio connectivity is, by default, not transitive. However, ad-hoc routing protocols have been defined [1] in order to overcome this limitation and allow routing of packets through mobile hosts. Pure ad-hoc networks have encountered so far limited applications that range from small ad-hoc groups to share information in meetings for a short time, to military applications on battlefields, and discovery or emergency networks in disaster areas.

In ad-hoc networks, the lack of infrastructure imposes a constraint that no hosts can be referred to as server. The distinction between the role of servers and clients in mobile ad-hoc networks is in fact blurred. The challenges of this kind of networks are:

Heterogeneity: heterogeneity issues for ad-hoc networks are because of the lack of a fixed infrastructure, it is more likely that hosts equipped with different network adaptors meet and are willing to communicate.

Openness: The fact that no servers are available on a core network implies that new behavior cannot be downloaded from a specific and known location, but need to be retrieved, if at all existing. Moreover, due to higher dynamicity than in fixed or nomadic networks, the chances that a host will come into contact with a service never met before, and for which it has no specific interaction behaviour, or that it will encounter a new environment, for which new functionalities are needed, increase.

Scalability: coordination of many hosts in a completely unstructured environment is very challenging. Decentralized solutions for service discovery and advertisement need to be put in place and exploited in order to obtain the best results.

Resource-sharing: The lack of structure implies that security is much more difficult to obtain. The existence of specific locations to register authentication data of hosts cannot be assumed; links are wireless, allowing for more risks of eves-dropping. Concurrency can be controlled using transactions; however, this becomes more difficult due to mobility, scarce resources and connectivity. Moreover, in ad-hoc networks, a host can just go out of range in the middle of a transaction.

In order to exchange information, a node requires multiple hops in the manets due to restricted transmission power in the network. Therefore, routing discovery and maintenance are the major issues to consider in manets [1]. Due to features like unpredictability of environment, unreliability of wireless medium, resource constrained nodes and dynamic topology, manets are susceptible to various types of faults. These faults can be grouped as follows : *Transmission errors, Node or link failures, Route breakages, Congested nodes or link*. For the fault tolerance, the replication is a means to ensure continuity of service for data sharing in the network.

The rest of the paper is organized as follows: The existing techniques/approaches for achieving fault tolerance to various above mentioned faults and improving of data availability are discussed in Section 2. In Section 3, we have the comparison of various location-based replication techniques and few special location services by detailing on the various quantitative metrics like replication cost, data availability, storage cost . Finally, we conclude and give perspectives of our comparative study.

2 Location-Services for MANETs

Location services [14] are used in mobile ad hoc networks either to locate the geographic position of a given node in the network(position location service) or for locating a data item (content location service). The Location services although used for different purpose, have a significant amount of similarity in terms of implementation and are therefore studied together.

Yet, despite their similarity, there are a number of differences between position and content location services:

Geographical knowledge: Position location services usually utilize the geographical position of nodes (inherently available to it) in order to implement the service itself, while data location services may not necessarily possess such an information.

Sensitivity to mobility: In position location services, the service has to be responsive to nodes mobility. When a node moves, its position changes and the service must be updated according to the mobility pattern. On the other hand, data location services do not necessarily depend on nodes mobility, but rather on nodes arrivals and departures, as well as on new data advertisements and disposals (due to limited memory consumption the application may dispose some of the data it possess, and in such cases the service has to be updated as well).

A need for additional routing: Position location services usually provide the position of the requested node, while in order to communicate with this node geographical routing should be employed. On the other hand, data location services may not only provide the location of the data, but actually the data itself (indeed, if only the location of the data is provided by the service, than additional routing should be employed to fetch the content).

Context of the study: Content location services have been primary studied in the context of sensor networks and service/resource discovery, while position location services have been primary studied in the context of geographic routing. Additionally, there is a significant amount of similarity, in terms of problem space and design criteria, between content location services and publish subscribe systems.

2.1 Location Based Data Replication Techniques

Location-based data replication protocols aim at achieving improvement with respect to data availability. They use a location based routing protocol to deliver update and query packets to replica servers. A node uses the location information to send the query packets. Using a geographic routing protocol, update and query packets can be sent toward fixed regions or positions instead sending to the nodes. Some or all nodes within the region or in the vicinity of location can act as servers. The following sub-section contains a description of various location based data replication technique.

2.1.1 Rendezvous Regions(RR)[4]

The proposed technique RR in divides the network into geographical regions. Each region is responsible for a set of keys representing the services or data. Hash-table like mapping scheme is used where each key is mapped to a region. Few nodes in each region are elected based on local election mechanism as data provider for maintaining the mapped information. The information requesting node retrieve it from them. A node wishing to update or query a key obtains the home region responsible for that key through the mapping, then uses a geographic-aided routing to send a message to the region.

2.1.2 Geography-Based Content Location Protocol(GCLP)[2]

This is proposed by Tchakarov and Vaidya. GCLP uses an approach propagating the advertisements and queries in cross-shaped trajectories, thus guaranteeing two intersections. The nodes at the intersection of the advertising and query trajectories

answer the Queries. These nodes are called content location servers (CLS). Periodically the content advertisement is performed by the update messages. These update messages are sent through the network in four geographical directions (north, south, east, west).

The content location server receives the advertisement messages. If a server which maintains the location information receives multiple advertisements for a specific resource, it will only forward updates from the content server closest to it. A query message similar to a update message is sent from the client through the network to locate the content.

2.1.3 Geographic Hash Table for Data-Centric Storage(GHT)[3]

The Geographic Hash Table system for Data-Centric Storage has been proposed for sensor networks. The design of the system considers a data object which is associated with a key. The key is hashed with geographic coordinates, and stores a key-value pair at the sensor node geographically nearest the hash of its key. The main required aspect is that the node requires to know their exact geographic location and uses the GPSR routing protocol to identify and to reach a packet's home node (the node closest to the geographic destination).

2.2 Location Services: A Special Case of Location Based Data Replication Techniques

Location services are special cases of location-based data replication protocols when the data items are location information, and a node sends an update packet to server nodes when only it changes its positions. The main objective of the location service is to associate the ID of a node to its geographical position. Each location service performs two basic operations: the *location update* and the *location query*. The location update is responsible for replicating information about the current location of a given node D to a set of nodes called *location servers*. If a node S wants to know the location of node D , it sends a location query message to some or all D 's location servers. survey on location services is available at [7].

Yu et al. [8] have analyzed the scalability of SLURP, SLALoM, GLS, DLM, and HIGH-GRADE. The scalability of location services depends on three parameters (i) the size of the network, denoted by the number of nodes n , (ii) the moving speed of a node, denoted as v , and (iii) the traffic pattern: uniform, and localized. Uniform traffic pattern means that a node chooses its queried nodes with equal probability. Localized uniform pattern means that nodes are more likely to query the location of those nodes that are nearby than those that are faraway.

The most important metrics that are used to evaluate the performance of the location services are:

Location update cost: The average number of forwarding operations each node needs to perform in a second to maintain fresh location information on the location servers.

Location query cost: The average number of packet forwarding operations due to location queries each node needs to perform in a second.

Total storage cost: The number of location records all the location servers store.

Table 2 provides a comparison of few location services proposed for manets. The performance results can be found in [9].

Table 1. Comparison of location-based data replication techniques and Location services

	Location based data replication technique		
	RR[4]	GHT[3]	GCLP[2]
Probability to access a data item when needed	Medium High	Low	Low
Replication decision on information from nodes within constant hop distance	Yes	Yes	Yes
Faults handled	Node failure due to depletion of node energy.	None	None
Effect of topological changes (Robustness)	Affected by node mobility and to some extent by node failures	Affected by both node mobility and node failures	Affected by both node mobility and node failures
Number of replicas deployed in network	$O(n)$	$O(n)$	$O(n)$
Update cost in hops	$O(\sqrt{n})$	$O(\sqrt{n})$	$O(\sqrt{n})$
Storage cost	$O(\sqrt{n})$	$O(1)$	(n)
Query cost in hops	$O(\sqrt{n})$	$O(\sqrt{n})$	$O(\sqrt{n})$
Probability of most recent data value access	Medium High	Low	Low

Table 2. Comparison of Location services

	Location services					
	HIGH-GRADE [5]	CRLS [6]	GLS [10]	DLM [11]	SLURP [12]	SLALoM [13]
Update cost in hops	$O(v \log n)$	$O(v \sqrt{n})$	$O(v\sqrt{n})$	$O(v \sqrt[3]{n})$	$O(v\sqrt{n})$	$O(v \sqrt[3]{n})$
Storage cost	$O(n \log n)$	$O(n\sqrt{n})$	$O(n \log n)$	$O(n\sqrt[3]{n^2})$	$O(n\sqrt{n})$	$O(v \sqrt[3]{n})$
Query cost in hops-uniform	$O(\sqrt{n})$	$O(\sqrt{n})$	$O(\sqrt{n})$	$O(v \sqrt[3]{n})$	$O(\sqrt{n})$	$O(v \sqrt[3]{n})$
Query cost in hops-localized	$O(\log n)$	$O(\sqrt{n})$	$O(\log n)$	$O(v \sqrt[3]{n})$	$O(\sqrt{n})$	$O(v \sqrt[3]{n})$
Probability of most recent data value access	Very Low	Medium Low	Very Low	Very Low	Low	Low

3 Conclusion

The location based data replication techniques for MANETs have been analysed. The analyzed protocols support replication of data with fewer resources overhead. So, the techniques could be a suitable method for scalable ad hoc networks. Compared to all the techniques discussed above RR technique is the most efficient and effective approach to deal with replication of files in an adhoc network. The RR algorithm is effective in comparison with the GCLP because it reduces the update and query cost and it requires approximate location information than GHT that depends on exact location information. The location-based location protocols have the highest scalability because it ignores the effect of topological changes . Except update and query cost, they do not incur any additional communication cost. However, it has the lowest data availability since it does not deal with the issue of network portioning. The data availability of location-based protocols can be easily improved by adding a mechanism that address the issue of data survivability. This mechanism can predict when a server node will drain out of battery power, and replicate the data items before node failure.

References

1. Perkins, C.: Ad-hoc Networking. Addison-Wesley (2001)
2. Tchakarov, J.B., Vaidya, N.H.: Efficient content location in wireless ad hoc networks. In: Mobile Data Management, MDM 2004 (2004)
3. Ratnasamy, S., Karp, B., Yin, L., Yu, F., Estrin, D., Govindan, R., Shenker, S.: Ght: A geographic hash table for data-centric storage in sensornets. In: Proc. First ACM International Workshop on Wireless Sensor Networks and Applications, WSNA (2002)
4. Seada, K., Helmy, A.: Rendezvous regions: A scalable architecture for service location and data-centric storage in large-scale wireless networks. In: IPDPS (2004)
5. Yu, Y., Lu, G.-H., Zhang, Z.-L.: Enhancing location service scalability with high-grade. In: IEEE International Conference on Mobile Ad-hoc and Sensor Systems, MASS 2004 (October 2004)
6. Stojmenovic: A scalable quorum based location update scheme for routing in ad hoc wireless networks. Technical Report TR-99-09, University of Ottawa (September 1999)
7. Stojmenovic, I.: Location updates for efficient routing in ad hoc networks. In: Handbook of Wireless Networks and Mobile Computing, pp. 451–471 (2002)
8. Yu, Y., Lu, G.-H., Zhang, Z.-L.: Location service in ad-hoc networks: Modeling and analysis. In: International Workshop on Theoretical Aspects of Wireless Ad hoc, Sensor and Peer-to-Peer Networks (June 2004)
9. Das, S.M., Pucha, H., Hu, C.: Performance comparison of scalable location services for geographic ad hoc routing. In: Proc. IEEE INFOCOM 2005 (March 2005)
10. Li, J., Jannotti, J., De Couto, D., Karger, D., Morris, R.: A scalable location service for geographic ad hoc routing. In: Proc. ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM), pp. 120–130 (2000)
11. Xue, Y., Li, B., Nahrstedt, K.: A scalable location management scheme in mobile ad- hoc networks. In: Proc. IEEE Conference on Local Computer Networks (LCN 2001), Tampa, Florida (November 2001)

12. Woo, S.-C., Singh, S.: Scalable routing protocol for ad hoc networks. *ACM Wireless Networks* 7(5), 513–529 (2001)
13. Cheng, C.T., Lemberg, H.L., Philip, S.J., van den Berg, E., Zhang, T.: Slalom: A scalable location management scheme for large mobile adhoc networks. In: Proc. IEEE Wireless Communications and Networking Conference (March 2002)
14. Friedman, R., Kliot, G.: Location services in wireless ad hoc and hybrid networks: A survey. Tech. rep. CS-2006-10, Technion, Haifa, Israel (2006),
<http://www.cs.technion.ac.il/users/wwwb/cgi-bin/tr-info.cgi?2006/CS/CS-2006-10>

A Comparative Analysis of Modern Day Network Simulators

Debajyoti Pal

Camellia Institute of Technology
Kolkata 700129, India
debajyoti.pal@gmail.com

Abstract. It is a very common practice to use network simulators for testing different network performance parameters before the real-life deployment of such a network. Apart from ns-2, few other recent network simulators have come into existence today and are gaining in more popularity. In this paper, we survey some of the widespread network simulators that are in use today, and try to evaluate their performance over certain parameters by setting up identical network simulation scenarios.

Keywords: network simulators, scalability, network-loss, topology.

I Introduction

Network simulators help us to reduce the cost, overhead and time that is involved with setting up a real test-bed containing multiple computers, routers and data-links. They allow researchers to implement practical test scenarios that otherwise might be difficult or expensive to emulate in real hardware-for example while experimenting with a new routing protocol. In fact they are particularly useful in allowing researchers to test new network protocols or modify an existing protocol in a controlled and reproducible environment. Majority of the network simulators that are available are based upon the discrete event-based simulation type [1] in which a list of pending events is stored, and those events are processed in order, with some events triggering future events- such as the event of arrival of a packet at one node triggering the event of the arrival of that packet at a downstream node.

The first instance where discrete event-based simulation was applied towards simulating a computer network was published in [2, 3]. ns-2 [4] was a direct successor of all those early efforts. In fact ns-2 is so widely popular that virtually it has become a benchmark for network simulations. The popularity of ns-2 amongst the research fraternity is primarily due to its ability to support a wide variety of protocol models and its support for both wired and wireless networks. However, studies [5, 6] reveal that with the growing number of nodes in a given simulated network ns-2 suffers from scalability problems. Under extreme conditions the problem of efficient memory usage and simulation run time become prominent. New Research domains like Wireless Sensor Networks (WSN), Wireless Multimedia Sensor Networks (WMSN), grid architectures, etc. require simulating a large number of nodes wherein the limitations of ns-2 become prominent. Some modifications have already been

incorporated in the latest version of ns-2 (ns-2.35 released on 4th November, 2011) like the inclusion of parallelization [7]. Other network specific features are still due that has ultimately paved its way to the next generation of the ns-2 simulator, namely ns-3[8]. In fact, the main goal behind introducing ns-3 is to provide a newer and wider simulation platform that supports the latest networking technologies (both wired as well as wireless) and to improve the overall simulation performance when the number of network nodes becomes large.

Although simulators model the real world, but the way they simulate the reality varies significantly across different simulators. Thus the only way to judge which simulator is the best is to go for a comparative analysis of the different available simulators which is the basic theme of this paper. Apart from ns-2 and ns-3 several other network simulators are available in the market targeted both towards the research fraternity and commercial customers. Prominent among them are OMNeT++ [9], Java in Simulation Time (JiST) [10], GloMoSim [11], and TOSSIM [12].

The outline of the remaining paper is as follows. Section 3 provides a brief overview of the different network simulators that are available today and popular in the academic world. To be precise we particularly provide an overall view of ns-2, ns-3, OMNeT++ and JiST. Section 3 discusses about the methodology that we employ to evaluate these network simulators. In fact the same network test bed that we define is put to test across the four different simulators under consideration and we record the various types of simulation results obtained, thereby trying to provide some sort of comparative performance evaluation. It also provides a detailed in-depth discussion about the results that are obtained from the experiment. Finally Section 4 provides a conclusion to the experiment that we perform and tries to compare our work with other related works.

2 Comparison of Different Network Simulators

In this section we review the network simulators under consideration. ns-2 is included in the comparison list because it has been popular among all sections of academics and industry for a very long period of time, and hence serves as the basis for our comparison. ns-3, OMNeT++ and JiST are gaining in popularity these days and have their own advantages and disadvantages.

A. ns-2

ns-2 is a discreet event simulator that supports a wide variety of protocols and can be used for both wired as well as wireless networks. It was built in C++ and provides a simulation interface through OTcl (an object oriented dialect of Tcl). Users describe a network topology by writing OTcl scripts, and then the main ns-2 program simulates that topology with specified parameters. However, the problems those are associated with ns-2 while running a large simulation is well known, with the main constraints being that of start-up time, memory and CPU usage [13]. This problem is of prime importance these days because the simulations are often run for a scalable network size.

B. ns-3

Like its predecessor, ns-3 also uses C++ for the implementation of the simulation models. However, instead of using OTcl as the scripting language, ns-3 uses Python. It should be noted that ns-3 is not an extension of ns-2. ns-3 integrates architectural concepts and code from GTNetS [14], a simulator that has good scalability characteristics. The design decisions for ns-3 were made at the expense of compatibility. In fact, ns-2 models have to be ported to ns-3 in a manual way. ns-3 is a new software development effort focused on improving upon the core architecture, software integration, models, and educational components of ns-2. It provides support for the integration of code by providing standard API's, like Berkley sockets or POSIX threads, which can be easily mapped to the simulation [15].

C. OMNeT++

In contrast to ns-2 and ns-3, which primarily focuses on simulation of computer networks, OMNeT++ can be used for simulating any general network. In fact it is an extensible, modular, component-based C++ simulation library and framework, primarily for building network simulators. "Network" is meant in a broader sense that includes wired and wireless communication networks, on-chip networks, queuing networks, and so on. Domain-specific functionality such as support for sensor networks, wireless ad-hoc networks, Internet protocols, performance modeling, photonic networks, etc., is provided by model frameworks, developed as independent projects [16]. For example, packages such as OMNeT++ Mobility Framework and Castalia [17] facilitate the simulation of mobile ad-hoc networks and wireless sensor networks. OMNeT++ offers an Eclipse-based IDE, a graphical runtime environment, and a host of other tools. There are extensions for real-time simulation, network emulation, alternative programming languages (Java, C#), database integration, SystemC integration, and several other functions.

OMNeT++ simulations use the so-called simple modules. Modules can be connected with each other via gates and combined to form compound modules. Modules communicate through message passing, where messages may carry arbitrary data structures. Modules may have parameters that can be used to customize module behaviour and/or to parameterize the model's topology.

Like ns-2 and ns-3 OMNeT++ rests upon C++ for the implementation of simple modules. However, the combination of simple modules into compound modules, and hence the setup of the entire network topology takes place in a language called Network Descriptor (NED). NED enables us to assign values to the various network parameters, for example the total number of nodes in a network in a dynamic fashion or can later be configured during run-time. In fact OMNeT++ follows a strict object-oriented design philosophy.

D. JiST

JiST is a high-performance discrete event simulation engine that runs over a standard Java virtual machine. It is a prototype of a new general-purpose approach to building discrete event simulators, called virtual machine-based simulation that unifies the traditional systems and language-based simulator designs [18].

Simulation in JiST is made up of entities that actually represent the network elements, for example the nodes, while the simulation events being invoked by some method calls among those entities. Embedding the simulation semantics within the Java language enables JiST to inherit all the properties and libraries of java including the existing compilers. JiST benefits from the automatic garbage collection, type-safety, reflection and many other properties of the Java language. The use of a standard virtual machine provides an efficient, highly-optimized and portable execution platform and allows for important cross-layer optimization between the simulation kernel and running simulation.

Unfortunately, the official development of JiST has stalled, and it is no longer maintained by its author, Rimon Barr. Still, we include it in our paper due to the numerous advantages that it provides.

3 Methodology and Experimental Test Bed

In this section we define our experimental test bed that is used to conduct the simulation experiments across the four different simulators under consideration. The results that are obtained are analyzed thereafter, and conclusions are drawn based upon the same.

With the aim to compare the performance of the different simulator toolkits, we implemented the same network simulation model in all of them. Our simulation does not depend on any particular simulation model tied down to a specific simulator, but is generic in nature that is applicable to all the simulators under consideration. Creating our own network model to evaluate the performance comparatives was preferred because we did not want to be tied down to the already existing network specific models as they tend to influence the efficiency and complexity of the entire simulation scenario in general.

Our simulation models a very basic network which consists of different nodes that are arranged in a square topology as illustrated by figure 1 in next page.

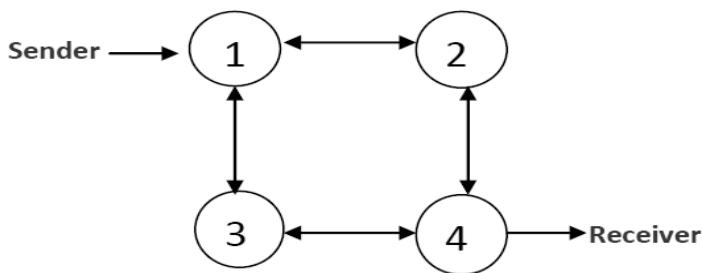


Fig. 1.

It should be clearly noted that we show 4 nodes for illustration purposes only. In reality while carrying out the experiment, the actual number of nodes is varied from a minimum of 16 to 2048. The communication process starts at the sending node which

generates one packet every second that is broadcasted to all its immediate neighbors. The neighboring nodes in turn broadcast the unprocessed packets after a delay of 1 second, to their immediate neighbors respectively, thereby flooding the entire network. The propagation delay of 1s is implemented directly by delaying the simulation event's execution. Also we assume that the nodes do not implement any specific queuing policy. The probability of packet drop is assumed to be the same across every link. We chose this model due to its simplicity, without going into different network specific details.

The simulations were carried on a Intel core-i3-2310M CPU based system having 4 GB RAM, 500 GB hard disk and running dual operating systems (Windows 7 Ultimate and Ubuntu 10.04 LTS). The different observations were done on the latest simulator versions viz. ns-2.35, ns-3.12.1, OMNeT++4.1 and JiST-1.0.6. Java version 6.0.290 provides the run time environment for OMNeT++ and JiST.

The above mentioned network topology along with the constraints are run in the four simulators under consideration for network loss ranging between 0 to 1. The simulation is carried out for 800s in all the cases. The total number of network nodes is increased from 0 to 1400 gradually. Figures 2, 3 and 4 depict the graph of the total number of packets that are dropped in the network v/s the total number of network nodes for network loss corresponding to 0.2, 0.5 and 1.0 respectively.

From the figures it is evident that keeping the network loss constant, with increasing the number of network nodes, the total number of packets dropped also increases. This is quite obvious, and as seen with increase probability of network loss, the number of packets dropped also increases. In fact, corresponding to network loss = 1.0, almost 100% of the packets are dropped in all the simulation scenarios. The percentage of packet dropped for different values of network loss is consistent across all the simulation platforms, although ns-2 in particular is more prone to the effect, however still within the safe limits of tolerance. Thus, it can be concluded that the reference simulation scenario that we use in all the four cases produce almost equivalent results.

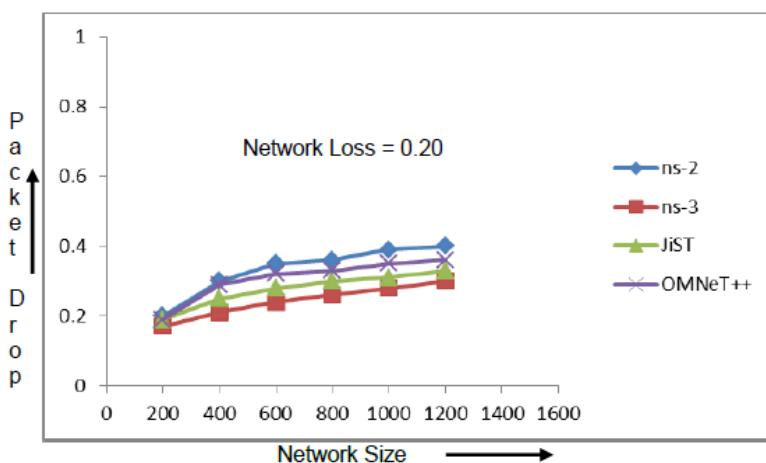


Fig. 2.

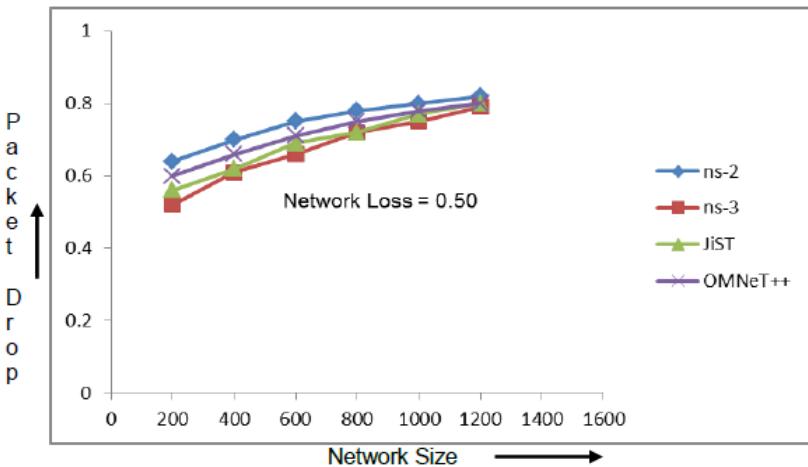


Fig. 3.

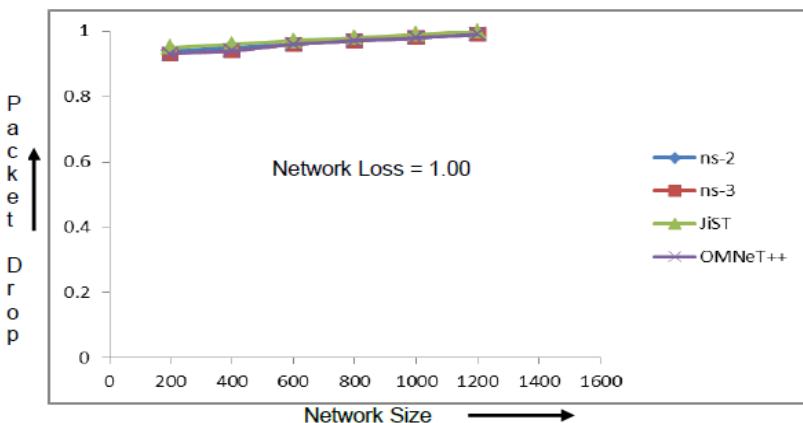


Fig. 4.

Now, we compare the different simulation tools with respect to three very important performance metrics, namely, simulation run-time, memory usage and computation time. For this purpose we consider 2 cases:

Case 1: The network loss is kept constant at 0.02 and the total number of nodes is gradually increased from 0 to 3000.

Case 2: The network loss is varied from 0.0 to 1.0 keeping the total number of nodes constant at 3000.

In both the cases, the simulation time is fixed to 800s.

Figure 5 in the next page shows the graph for simulation run-time v/s the network size.

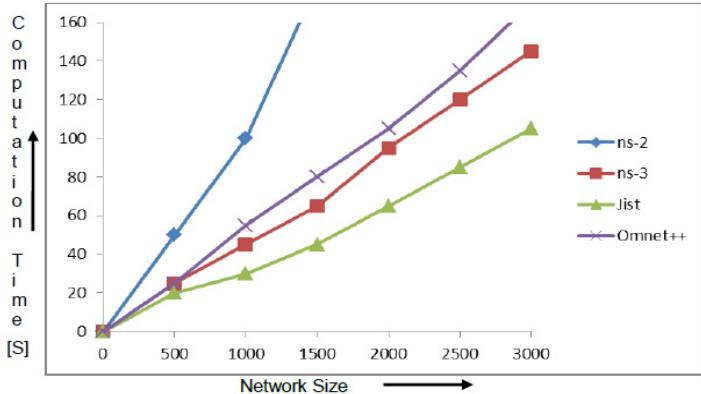


Fig. 5.

It is evident that for simulating the same number of nodes the time taken by ns-2 is maximum, while that taken by JiST is minimum. For a network size of 3000 nodes, while the average time taken by ns-2 is about 440s, that taken by JiST is about 80s. Hence, the simulation run-time for JiST is about 5.5 times faster as compared to ns-2. In fact it is quite surprising that JiST despite being based on a Java platform outperforms all other simulation tools. Primarily, this is due to the JiST architecture. Apart from parallel execution of different entities, JiST also has provisions of various run-time optimizations based upon the analysis of the executed byte-code [19]. Thus, the slowness of Java is purely a myth as compared to a C++ platform [20]. The simulation run-time for ns-3(about 110s) is also far better as compared to ns-2. This is primarily due to the abolishment of oTCL/C++ duality in ns-3. The performance of OMNeT++ is more or less similar to that of ns-3. Thus, apart from ns-3 all the other 3 simulators are equally scalable with respect to the simulation run-time.

Figure 6 below shows the graph for different values of network loss v/s the simulation run-time keeping the total number of nodes fixed at 3000.

As it is evident from the figure, with an increase in network loss, the simulation run-time also decreases. This is quite obvious, because as we increase the network loss, more and more number of packets are removed from the simulation scenario, thereby causing the simulation run-time to drop. For lower order range of network loss, the computation time for ns-2 is maximum while that of JiST is minimum.

Similar to the analysis of the simulation run-time we also investigate the memory usage of the different network simulator tools. Figure 7 shows the graph between network size v/s memory usages keeping the network loss constant.

For network size up to 500 the memory usage of the different simulator tools are comparable, but thereafter and especially as the number of network nodes becomes large the memory consumption of JiST increases to a great deal. In fact it exhibits the

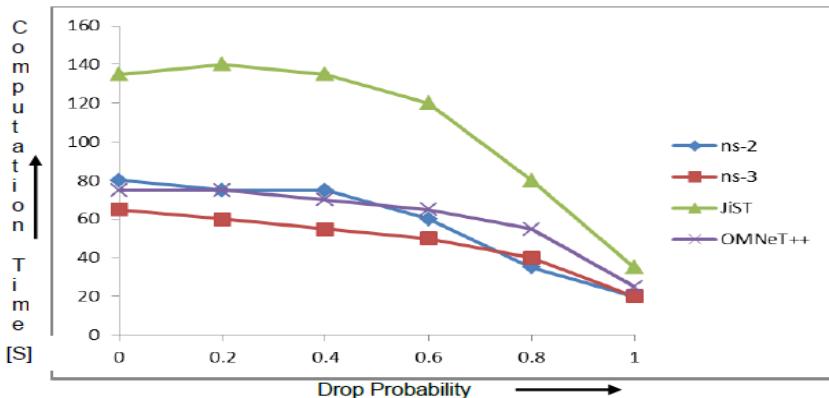


Fig. 6.

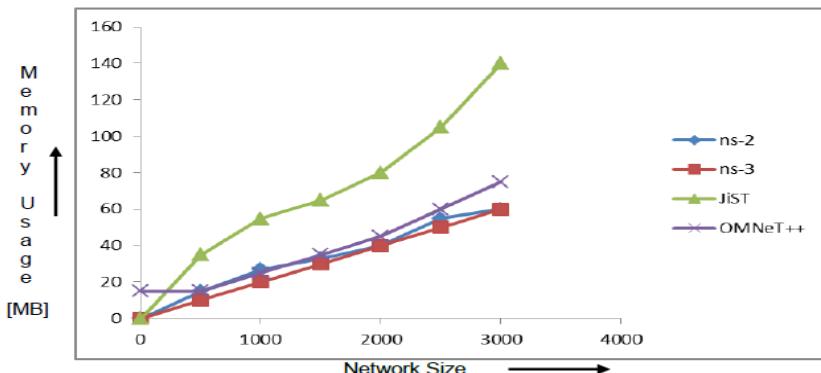


Fig. 7.

worst performance with respect to memory usage. This is quite surprising due to the fact that Java supports automatic garbage collection mechanism. Memory usage of the remaining simulator tools exhibit more or less the same pattern (increasing in a linear fashion with the number of network nodes) with ns-3 being the most efficient amongst all of them in this regard.

Figure 8 below shows the same memory usage for case 2 i.e. a graph between network loss and memory usage keeping the total, number of nodes fixed at 3000.

As expected, with an increase in network loss, since more and more packets are being removed from the simulation scenario, hence the memory utilization of the simulator tools also keeps on decreasing. Again JiST shows the worst memory usage characteristics among all the others. It should also be noted that even for network loss equal to 100% some residual memory is used by all the simulators. This can be attributed to the simulator process itself that keeps on running in the background at all points of time.

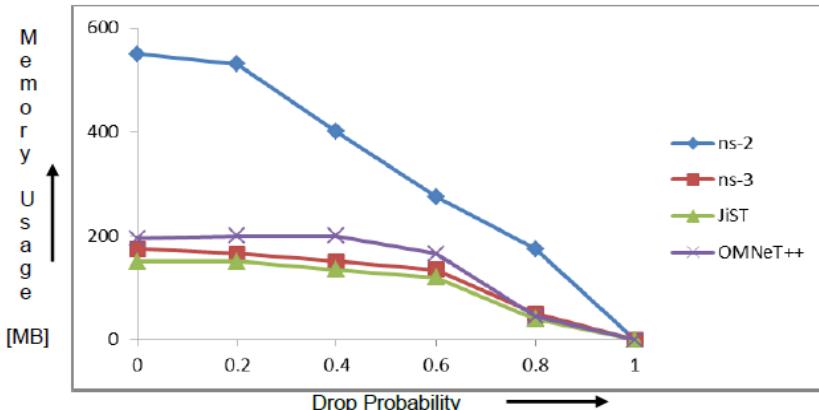


Fig. 8.

4 Conclusion and Other Related Work

A few papers have been found on performance comparison of network simulators. One example is the work provided in [21], where the performance of a TCP-based simulation is implemented in ns-2, SSFNet and J-Sim. Similarly, another paper [22] analyzes the run time performance of ns-2, OMNeT++ and SimPy coming to the same conclusion as ours. Another example is that of [23] which uses some more network simulators.

After conducting an in depth studies of the different network simulators we come to a conclusion that ns-3, OMNeT++ and JiST are all capable of carrying out large scale simulation scenarios. In fact the execution speed of JiST is the best although it is worst with respect to memory utilization. Thus, a tradeoff has to be done between simulation run-time and memory usage and often it depends upon the circumstances and the environment in which we are going to run the simulation. Also the official development and maintenance of JiST has been stopped. The overall performance of ns-2 is worst among all.

Thus we can conclude that the answer to the question which simulator is the best is a very difficult one to be answered and is heavily dependent upon the specific use case. However, if memory requirement is not a constraint and scalability is the primary issue, then ns-3, OMNeT++ and JiST are the obvious choices.

References

1. Fisherman, G.S.: Principles of Discrete Event Simulation. John Wiley & Sons, Inc., New York (1978)
2. Schwartz, D.J., Yemini, Y., Bacon, D.: NEST: A network simulation and prototyping testbed. Commun. ACM 33(10), 63–74 (1990)

3. Keshav, S.: Real: A network simulator, Technical report. University of California at Berkley, CA, USA (1998)
4. The Network Simulator ns-2, <http://isi.edu/nsnam/ns>
5. Xue, Y., Lee, H.S., Yang, M., Kumarawadu, P., Ghenniwa, H., Shen, W.: Performance evaluation of ns-2 simulator for wireless sensor networks. In: Proceedings of the Canadian Conference on Electrical and Computer Engineering (CCECE 2007), pp. 1372–1375 (April 2007)
6. Henderson, T.R., Roy, S., Floyd, S., Riley, G.F.: ns-3 project goals. In: WNS2 2006: Proceeding From the 2006 Workshop on ns-2: the IP Network Simulator, p. 13. ACM, New York (2006)
7. Riley, G.: PDNS project website,
<http://www.cc.gatech.edu/computing/compass/pdns/>
8. Henderson, T.R., Roy, S., Floyd, S., Riley, G.F.: ns-3 project goals. In: WNS2 2006: Proceeding From the 2006 Workshop on ns-2: the IP Network Simulator, p. 13. ACM Press, New York (2006)
9. Varga, A., Hornig, R.: An overview of the OMNeT++ simulation environment. In: Proceedings of the First International Conference on Simulation Tools and Techniques for Communications, Networks and Systems, SIMUTools 2008 (March 2008)
10. Barr, R., Haas, Z.J., van Renesse, R.: JiST: an efficient approach to simulation using virtual machines. *Softw. Pract. Exper.* 35(6), 539–576 (2005)
11. Bagrodia, R., Chen, Y.A., et al.: PARSEC: A Parallel Simulation Environment for Complex System. ULCA technical report (1997)
12. Levis, P., Lee, N., Welsh, M., Culler, D.: TOSSIM: accurate and scalable simulation of entire TinyOS applications. In: Proceedings of the 1st ACM Conference on Embedded Networked Sensor Systems, SenSys 2003 (2003)
13. Huang, P., Heidemann, J.: Minimizing Routing State for Light-weight Network simulation. In: Proceedings of the International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems. IEEE, Cincinnati (August 2001)
14. Riley, G.: Large scale network simulations with GTNetS. In: Proceedings of the 2003 Winter Simulation Conference (2003)
15. ns-3 Overview, <http://www.nsnam.org/docs/release/3.12/manual/ns-3-manual.pdf>
16. OMNet++ overview, <http://www.omnetpp.org/>
17. Pham, H.N., Pediaditakis, D., Boulis, A.: From simulation to real deployments in wsn and back. In: Proceedings of the 2007 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM 2007), pp. 1–6 (June 2007)
18. JiST overview, <http://jist.ece.cornell.edu/>
19. Barr, R., Haas, Z.J., van Renesse, R.: An efficient, unifying approach to simulation using virtual machines. PhD dissertation, Cornell University (May 2004)
20. Lewis, J., Neumann, U.: Performance of Java versus C++,
http://www.idiom.com/_zilla/Computer/javaCbenchmark.html
21. Nicol, D.M.: Scalability of network simulators revisited. In: Proceedings of the Communication Networks and Distributed Systems Modeling and Simulation Conference, Orlando, FL (February 2003)
22. Albeseder, D., Fuegger, M.: Small PC-Network Simulation a comprehensive performance case study. Research Report 77/2005, TU Wien, Institut für Technische Informatik (2005)
23. Weingartner, E., vom Lehn, H., Wehrle, K.: A Performance Comparison of recent network simulators. RWTH Aachen University, Germany

Design and Implementation of a RFID Based Prototype SmArt LibRARY (SALARY) System Using Wireless Sensor Networks

K.S. Kushal¹, H.K. Muttanna Kadal², S. Chetan³, and Shivaputra⁴

^{1, 2} Dept. of M.Tech[VLSI Design & Embedded System],
Dr. Ambedkar Institute of Technology
Bangalore, India

^{3, 4} Dept. of Electronics & Communication,
Dr. Ambedkar Institute of Technology
Bangalore, India
ksk261188@gmail.com

Abstract. "There is a great deal of difference between an eager man who wants to read a book and a tired man who wants a book to read" - G.K. Chesterton

With the colossal collection of traditional and digital including books, journals, audio materials, photographs, e-journals, e-books, web resources and more in recent years, finding an anonymous item is becoming more and more difficult, resulting in a number of practical conflicts. The accessing and the procurement of the details pertaining to a publication is becoming an ubiquitous problem. Widespread use of wireless technologies paired with the recent advances in the wireless applications, manifests that digital data dissemination could be the key to solve emerging problems. Wireless Sensor Network technology has attracted increased attention and is rapidly emerging due to their enormous application potential in diverse fields. This buoyant field is expected to provide an efficient and cost-effective solution to the effluent problems. This paper proposes a SmArt LibRARY (SALARY) System based on wireless sensor network technology which provides advanced features like automatic update in the addition/deletion of a publication, automated guidance, and item reservation mechanism. The paper describes the overall system architecture of SALARY from hardware to software implementation in the view point of sensor networks. We implemented a full-fledged prototype system for library management to realize the design functionalities and features mentioned. Our preliminary test results show that the performance of this WSN based system can effectively satisfy the needs and requirements of existing integrated library system hassles thereby minimizing the time consumed to find the slot of thou publication, real-time information rendering, and smart reservation mechanisms.

Keywords: Radio Frequency Identification(RFID), Wireless Sensor Networks(WSN), Java ME(J2ME), Context Aware.

1 Introduction

"Books serve to show a man that those original thoughts of his aren't very new after all" - Abraham Lincoln

In recent years, with the advancement of mobile and ubiquitous technologies, invisible micro-computers are embedded in to our surrounding environment. These invisible computers interconnect to our home appliances and mobile devices to provide ubiquitous services to the user. Various applications are developed using ubiquitous technologies and one such application is ubiquitous SmArt LibrARY (SALARY) system. This system is an advanced stage of library systems that enjoy the benefits of a browser-based environment, that incorporates all modules including Circulation, Cataloguing, Acquisitions, Serials and Reporting with visually appealing and customizable interface, where the user enjoys convenient access to thou traditional and digital items, books, journals and more, via the handheld device to access the library management content anywhere and anytime.

Providing the right content to the right person at the right place and guiding them to the exact location of the thou publication, avoiding the derange is a real challenge for any ubiquitous library management system application developer. Many researchers have developed many ubiquitous models and prototypes for various applications. One such model is a TANGO System [1], which uses RFID tags for vocabulary learning, RFID based context aware Ubiquitous Learning System[2], Smart PARKing (SPARK) System using Wireless Sensor Networks [3], which makes use of the WSN with advanced features like automated guidance and parking reservation mechanism, which forms the base to this system and supports in a similar manner.

In this ubiquitous system, each user is provided with a Bluetooth enabled handheld device such as a mobile phone integrated with a J2ME application. JSR-82 APIs are used to develop these mobile client applications. We also require a Bluetooth enabled RFID reader. Whenever the student enters into the library and comes nearby/within the range of the RFID reader with his/her RFID enabled mobile device, the client application running on the mobile device captures the object ID and sends it to the application server for the processing. After the processing is done, the information related to the object are sent to the server application via WLAN network , which determines the user details like name, mobile number, semester, materials relative to the user.

Information regarding the publications, books, journals, and more relative to the thou is composed and sent as an SMS to the user mobile phone. Here it also contains information like the availability of the book/journal/publication (i.e. the status), and other details pertaining to the book/publication/journal. It is assumed that all the users who are using this system are registered to the application. Even if they are not registered, users can register to the system through a simple SMS.

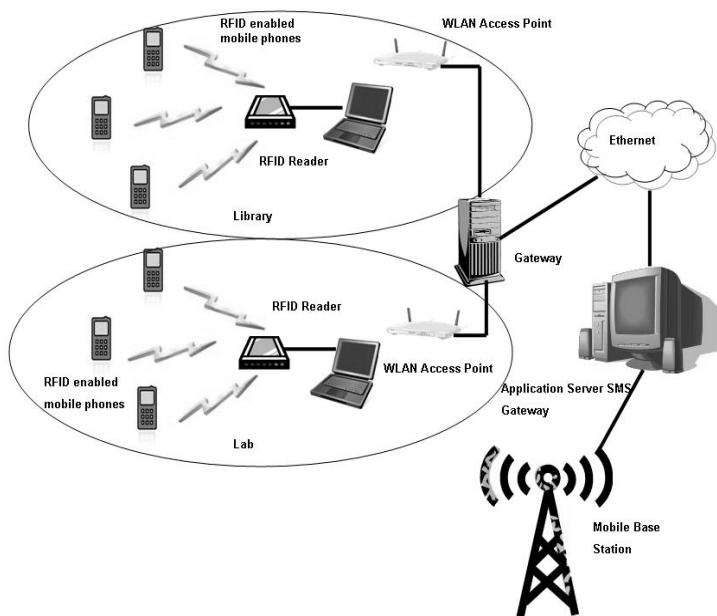
The user can reserve the book/journal/publication via WLAN network through a simple SMS, and also can procure the same in a similar way. This ubiquitous system also contains a WSN which also provides advanced features like automated guidance to the particular slot in which the thus elected book/journal/publication is located in real-time.

2 Related Works

2.1 RFID Based Context Aware Ubiquitous Learning System

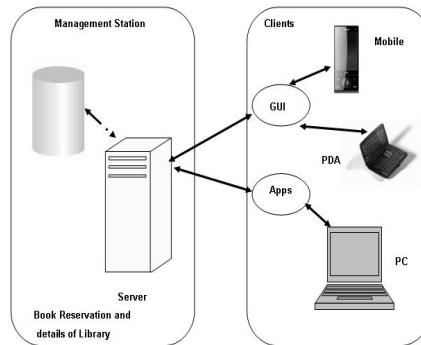
Electronic learning is used to access the learning content using the desktop computers. Mobile learning is an advanced stage of e-learning where in the user is equipped with

the handheld device such as a mobile phone, to access the learning content using various wireless technologies. A new mode of learning mechanism called ubiquitous learning (u-learning) is context aware and also provides anywhere, anytime learning using the handheld devices and sensors. The real challenge is to deliver the learning content to the mobile devices depending upon the surrounding context. Here context plays an important role in delivering the right content to the right device, right person, in the right place and at the right time. This context can be the person who is using the system, surrounding the environment in which the system is being used or the device itself.



2.2 Design and Implementation of a Prototype, Smart PARKing (SPARK) System Using Wireless Sensors

Most of the existing parking management systems rarely address the issues of the parking space management, vehicle guidance, parking lot reservation etc. These systems majorly have control at the entrance and the exit and use vehicle detectors as an essential element to provide smart parking. The widespread use of the WSN technologies paired with the advancements in wireless applications implies that digital data dissemination could be the key for the growing parking challenges. WSN has a great potential towards providing an easy and a cost effective solution to this credible application for various reasons, of which ease of deployment and flexibility to couple with sophisticated but cheap sensors are attributed majorly.



3 System Architecture and Technology

3.1 Radio Frequency Identification

Radio Frequency Identification Technology(RFID), is one of the short range wireless communication technologies used to capture remote object ID using radio waves. Basically it consists of a RF interrogator or reader used to interrogate RF transponder or tag within its radio range. Radio transponders are used to store and retrieve the object details embedded inside the memory of the tag. There are various air-interface protocols used to retrieve the object ID from a distant location. Frequency standards like Low Frequency(LF), High Frequency(HF), Ultra High Frequency(UHF), Microwave Frequency(MF) are being used for developing various RFID applications.

The read range depends on the chosen frequency, orientation of the tag and the reader, working environment and the kind of application being used. RFID can also be used to locate a person or get the contextual information using the electronic sensors and radio tags embedded inside the surrounding environment. This contextual data can be used to provide various services depending upon the person, location, time and the device being used.

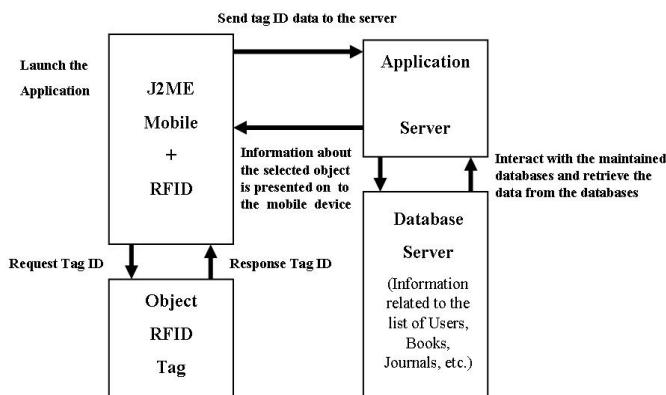


Fig. 1. System architecture of SALARY based Content Presentation

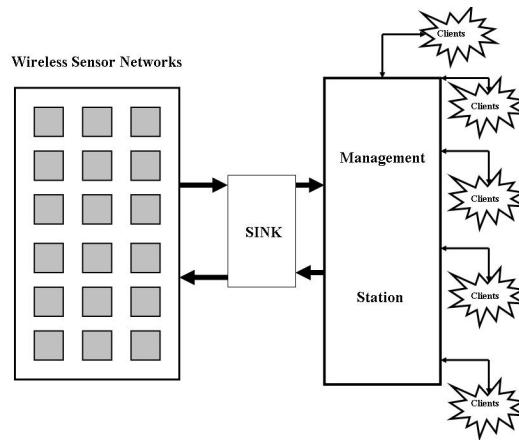


Fig. 2. System architecture for SALARY in guiding the client to the exact location of the item

3.2 Java Platform, Micro Edition

Java Platform, Micro Edition or Java ME is a Java platform designed for mobile internet and embedded systems. It consists of various profiles and configurations like Mobile Information Device Profile (MIDP)[4] and Connected Limited device Configuration (CLDC)[5] as apart of the core architecture. Other optional APIs like JSR-82(Bluetooth) [6], JSR-172(XML Parsing) [7] etc. can be used to develop mobile client applications. Java 2 Enterprise Edition(J2EE) can be used to develop and host server side components like XML files and Servlets.

4 Implementation Details

Mobile client application is developed on J2ME[7] platform. Server application is developed using JAVA Servlets and SMS data is composed from SAX Parsers which is implemented in J2EE. Servlets are deployed in APACHE Tomcat and the user and the objects database is designed using the MySQL[8]. SMS server on Linux platform. Management server is also rigged up for the management of WSN including, Sensor nodes, Sink Node, Guiding Nodes, Status LEDs and the GSM devices.

[#] When a student/ person enters the library, his presence is captured using the RFID Reader located at the entrance of the library block and the RFID Tag ID along with the information related to the student/ person are sent to the server application via WLAN network. At server, user details like the name, mobile number, and the registration number corresponding to the Tag ID are fetched from the database and the contextual information is composed and sent to the user mobile phone as an SMS. Here the contextual information can be his semester, textbooks, journals, reference books and cd & e-journals, also might be the information regarding the attest journals and books, that is relevant to the user in that surrounding context.

5 Results

In this section we present the results of our project with an Application Startup snapshot, obtained after a successful coding and linking of the database and all the

other inter-related systems as mentioned earlier. This result will finally create a .jar/.jad file, which can be loaded onto a mobile device and can be run to have a successful communication with the system and the interaction can be done. The snapshot is that of an emulator of NetBeans 7.0.1 IDE/Eclipse IDE.

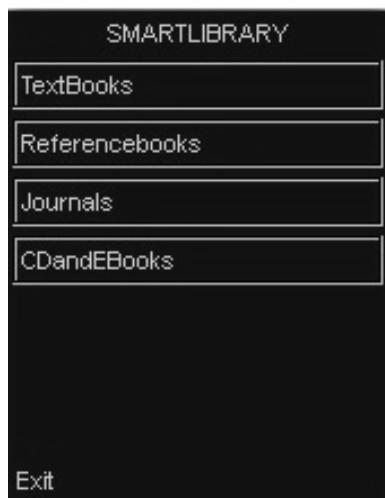


Fig. 3. Startup Application screen snapshot

6 Conclusion

The implementation of the ubiquitous SALARY System is a classic example of using the ubiquitous technologies for a proper management system. Due to the rapid advances in the ubiquitous computing technologies like RFID, large scale deployment of sensor network technologies, the management systems can be made anywhere, anytime, and real-time providing the suitable and necessary services to the client based on the preferences and the context, relatively as given in the table. This is related to RFID based Learning Services Model[2]. The below table describes the future cases in various scenarios of its implementation and the services that are vulnerable with the design of the same.

Table 1. Various learning services corresponding to various surrounding environments

Location	Services
Classroom	Exam Alerts, Marks, Assignments
Laboratory	List of experiments to be done for that day
Canteen	Information about the College Cultural Activities

Acknowledgement. The authors would like to thank Mr. Chetan S, and Mr. Shivaputra for their assistance, implementations, insight and valuable discussion over the course of this project.

Parts of this work have been supported by Dr.Ambedkar Institute of Technology, Bangalore. We would also like to thank our beloved colleagues from M/s Aeronautical Development Agency, DRDO, Bangalore, Karnataka, INDIA and also our beloved friends who lent a helping hand in the success of this project.

References

- [1] Ogata, H., Akamatsu, R., Mitsuhashi, H., Yano, Y., Matsuura, K., Kanenishi, K., Miyoshi, Y., Morikawa, T.: TANGO: Supporting Vocabulary Learning with the RFID tags, Tokushima University,
<http://www.sirc.kyushu-u.ac.jp/rfid-workshop/ogata-paper.pdf>
- [2] Kumar, M., Nava Jyothi, K.: RFID Based Context Aware Ubiquitous Learning System. In: Proceedings of the Wokshop on Ubiquitous Computing, UbiComp INDIA 2011, C-DAC, Hyderabad (2011), <http://www.cdac.in>
- [3] Srikanth, S.V., Pramod, P.J., Dileep, K.P., Tapas, P., Patil, M.U., Sarat Chandra Babu, N.: Design and Implementation of a prototype Smart Parking (SAPRK) System Using Wireless Sensor Networks, CDAC <http://www.cdac.in>
- [4] Mobile Information Device Profile (MIDP),
<http://www.oracle.com/technetwork/java/index-jsp-138820.html>
- [5] Connected Limited Device Configuration (CLDC),
<http://java.sun.com/products/cldc/>
- [6] J2ME Bluetooth APIs,
<http://java.sun.com/javame/reference/apis/jsr082>
- [7] J2ME XML APIs,
<http://java.sun.com/javame/reference/apis/jsr172/>
- [8] MySQL Database & Open source SMS gateway, <http://www.mysql.com/>,
<http://www.kannel.org/>
- [9] Irisnet: Internet-scale Resource Intensive Sensor Network Service,
<http://www.intel-iris.net>
- [10] Tang, V.W.S., Zheng, Y., Cao, J.: An Intelligent Car Park Management System based on Wireless Sensor Networks. In: Proceedings of the 1st International Symposium on Pervasive Computing and Applications, Urumchi, Xinjiang, China

Optimal Route Life Time Prediction of Dynamic Mobile Nodes in Manets

Ajay Kumar, Shany Jophin, M.S. Sheethal, and Priya Philip

Dept of Computer .Science
Adi Shankara Institute of Engineering
And Technology, Kalady
sheethalbasil@gmail.com

Abstract. One of the important and challenging problems in the design of ad hoc networks is the development of an efficient routing protocol that can provide high-quality communications among mobile hosts for that proposing new protocol to evaluate the node lifetime and the link lifetime utilizing the dynamic nature, such as the energy drain rate and the relative mobility estimation rate of nodes. Integrating these two performance metrics by using the proposed route lifetime-prediction algorithm select the least dynamic route with the longest lifetime for persistent data forwarding and based on quadrant. our proposed route Lifetime-prediction protocol in a exploring dynamic nature routing for large scale network (LEDNR) protocol environment based on Ad hoc on demand distance vector routing (AODV).

Keywords: Lifetime prediction, link lifetime (LLT), mobile ad hoc networks (MANETs), node lifetime, route discovery, routing protocol.

1 Introduction

A Mobile ad hoc network (MANET) consists of many mobile nodes that can communicate with each other directly or through intermediate nodes. Often, hosts in a MANET operate with batteries and can roam freely, and thus, a host may exhaust its power or move away, giving no notice to its neighboring nodes, causing changes in network topology. One of the important and challenging problems in the design of ad hoc networks is the development of an efficient routing protocol that can provide high-quality communications among mobile hosts. These studies often attempt to find a stable route that has a long lifetime. We can classify these solutions into two main groups: node lifetime routing algorithms and link lifetime (LLT) routing algorithms.

2 Related Work

By considering the energy state of nodes, such as residual energy and energy drain rate, the node lifetime routing algorithms often select a path consisting of nodes that may survive for the longest time among multiple paths. Shrestha and Mans [3] mentioned that the energy drain rate of a node is affected not only by its own but by

its neighboring data flows as well. Marbukh and Subbarao [4] aimed to preserve network connectivity by choosing a route according to the remaining battery life of nodes along the route. Toh [5] proposed selecting a path with minimum total transmission power when there exist some possible paths, and all nodes through these paths have sufficient residual battery power. Misra and Banerjee [6] proposed selecting a path that has the largest packet transmission capacity (the residual energy divided by the expected energy spent in reliably forwarding a packet) at a “critical” node among multiple paths. The critical node is the node that has the smallest packet transmission capacity in a path. The LLT routing algorithms are used to estimate the lifetime of wireless links between every two adjacent nodes and then to select an optimal path. In the associativity-based routing algorithm, a link is considered to be stable when its lifetime exceeds a specific threshold that depends on the relative speed of mobile hosts.

3 System Architecture

3.1 Description of System Architecture

Since DSR [16] is one of the most popular routing protocols in MANETs and it is easy to extend the routing control message format of DSR, we implement the proposed route lifetime-prediction algorithm in the DSR protocol. The proposed algorithm consists of the following three phases: route discovery, data forwarding, and route maintenance. There are three main differences between the EDNR and the DSR.

First, in the EDNR protocol, every node saves the received signal strength and the received time of the RREQ packet in its local memory, and adds this information into the RREP packet header in a piggyback manner when it receives the RREP for the corresponding RREQ packet to meet the requirement of the connection lifetime-prediction algorithm. Second, node agents need to update their predicted node lifetime during every period.

Finally, the node-lifetime information in the RREP packet is updated when the RREP packet is returned from a destination node to the source node. At every EDNR node agent, a variable NLT, which represents the node lifetime, is added to represent the estimated lifetime of this node, and it is updated by the algorithm in Section 2-A. For the lifetime of a link C_i , there are two sample packets exchanged between nodes N_{i-1} and N_i (packet 1: N_{i-1} RREQ $\xrightarrow{\text{-----}}$ N_i ; packet 2: N_{i-1} RREP $\xleftarrow{\text{-----}}$ N_i). To implement this, every node agent needs to maintain a data structure called RREQ_Info table in its local memory. This structure includes the RREQ id, the forwarding RREQ time, and the RREQ received signal strength. For a path sequence $S, \dots, N_{i-1}, N_i, N_{i+1}, \dots, D$, when an intermediate node N_i receives an RREQ packet from N_{i-1} , it adds this RREQ id, the current time, and the received signal strength to its RREQ_Info table before it continues to forward this RREQ packet.

Similarly, node N_{i+1} also saves the RREQ_Info from node N_i in its local memory. In the returning RREP period, when node N_i receives an RREP packet from node N_{i+1} , the RREQ_Info from N_i (information of N_i RREQ $\xrightarrow{\text{-----}}$ N_{i+1}) has been added to the RREP header by N_{i+1} before node N_{i+1} sends an RREP packet to node

N_i . Simultaneously, node N_i knows the RREP time and the RREP received signal strength from node N_{i+1} (information of N_i RREP $\leftarrow\!\!\!---- N_{i+1}$). Thus, it can obtain the second sample packet that is delivered between the corresponding two nodes (N_i, N_{i+1}), and, thus, we can calculate the connection time TC_i using the connection lifetime-prediction algorithm and then update the local LLT value.

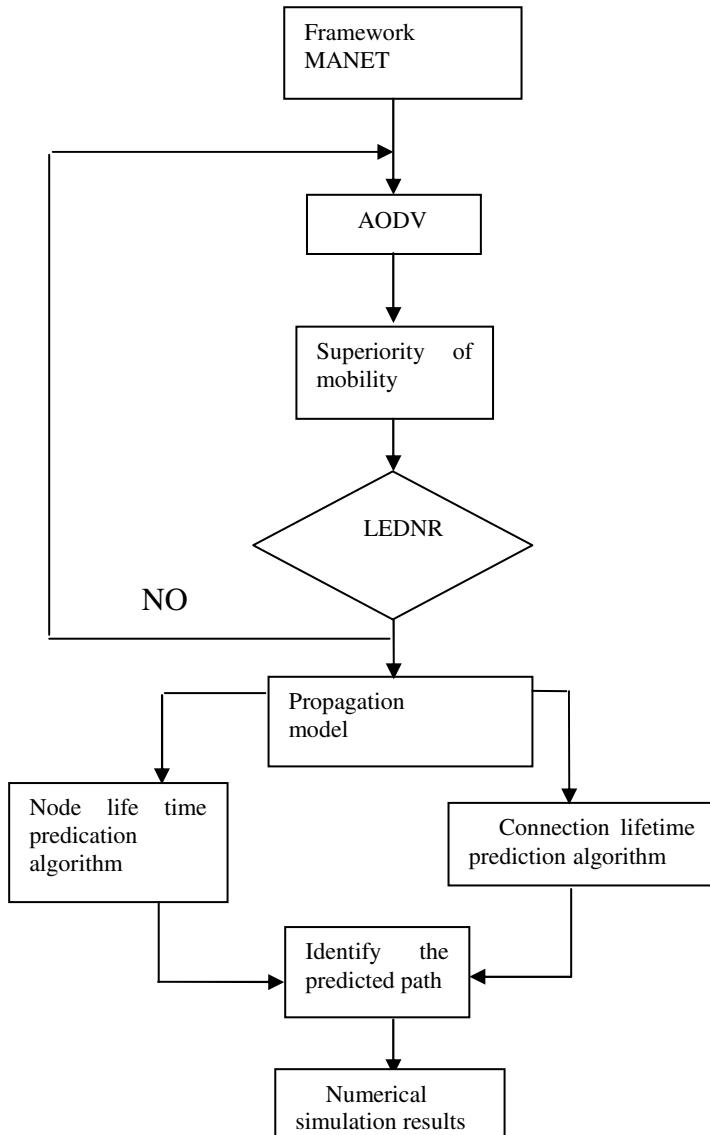


Fig. 1. Flow Diagram

Similarly, node N_i should add the RREQ_Info entry that is received from node N_{i-1} to the RREP header before sending the RREP to node N_{i-1} , and then node N_{i-1} calculates the LLT between nodes (N_{i-1}, N_i) . Three new entries, i.e., *path lifetime* (PLT), *RREQ time*, and *RREQ signal strength*, are added to the common header of an RREP packet. The PLT represents the predicted lifetime of the source route in this packet header and can be updated when RREP packets are forwarded from the destination node to the source node in the route-discovery phase. The *RREQ time* and the *RREQ signal strength* represent the RREQ_Info of the previous RREQ node.

The EDNR node agent only updates the PLT value in the common header of the RREP packet with a local NLT value or LLT value, if $NLT < PLT$ or $LLT < PLT$, before forwarding this RREP packet. When this RREP packet reaches the source node, the PLT becomes the minimum value of the estimated lifetime of all nodes and links through the route from the source node to the destination node, as described in (2). In the persistent data forwarding period, a source node tends to select the path with the longest lifetime (the path with the maximum PLT value) from multiple paths as a source route for data forwarding.

4 Algorithm of Intermediate Node

```

Predict Its lifetime
If its lifetime < Min-lifetime
Replace Min-lifetime with its lifetime
If Sequence Number exists
Compare Min-lifetime of current RREQ with Min lifetime of existing one.
If new Min-lifetime <= old Min-lifetime
Discard new RREQ
If new Min-Lifetime >old Min-lifetime
Replace old Min-Lifetime with new Min lifetime
Forward new RREQ
If Sequence Number does not exist
Save this Min-lifetime
Forward RREQ

```

4.1 Description of Algorithm

4.1.1 Node Lifetime Prediction

If there are two nodes that have the same residual energy level, an active node that is used in many data-forwarding paths consumes energy more quickly, and thus, it has a shorter lifetime than the remaining inactive node. the node lifetime that is based on its current residual energy and its past activity solution that does not need to calculate the predicted node lifetime from each data packet .We use an exponentially weighted moving average method to estimate the energy drain rate evi . E_i represents the current residual energy of node i , and evi is the rate of energy depletion. Ei can simply be obtained online from a battery management instrument, and evi is the statistical value that is obtained from recent history. the estimated energy drain rate in the n th period,

and $ev(n-1)$ i is the estimated energy drain rate in the previous $(n - 1)$ th period. α denotes the coefficient that reflects the relation between ev_n and ev_{n-1} , and it is a constant value with a range of $[0, 1]$.

4.1.2 Connection Life Time Prediction

We are only concerned with the minimum node lifetime or the connection lifetime in a route from two nodes of a stable connection are within the communication range of each other, the connection lifetime may last longer, and they are not a bottleneck from the route to which they belong. It is easier to model the mobility of nodes in a short period during which unstable connections last. Reasonably and simply that the nodes move at a constant speed toward the same direction in such a short period. Easy to measure the distance between nodes N_i and N_{i-1} when we use Global-Positioning-System-based location information. Senders transmit packets with the same power level a receiver can measure the received signal power strength when receiving a packet and then calculates the distance by directly applying the radio propagation model.

If the received signal power strength is lower than a threshold value, we regard this link as an unstable state and then calculate the connection time. Our proposed method requires only two sample packets, and we implement piggyback information on route-request (RREQ) and route-reply (RREP) packets during a route-discovery procedure with no other control message overhead, and thus, it does not increase time complexity

5 Simulation Results

To evaluate the performance of the EDNR, we compare the performance of the EDNR with those of the following three routing protocols: 1) the original DSR, in terms of network throughput, routing failures, and control packet overhead. The original DSR tends to find the shortest path from the source node to the destination node, ignoring the node lifetime and wireless LLT.

Fig. 4 shows the throughput performance in terms of the number of packets for the four routing protocols. The proposed EDNR protocol outperforms the remaining three protocols in varying node velocity environments. Its throughput enhancement is achieved by approximately 79.2%, 14.2%, and 13.8%, compared with that of the original DSR.

Fig. 5 shows the advantage of the EDNR protocol in terms of the number of routing failures. To adapt to dynamically varying network topology environments, the EDNR, protocols do their best to find a more stable route, reducing the number of routing failures by 21.2%, 15.6%, and 14.2%, respectively, compared with that of the original DSR.

Routing overhead is defined as the amount of routing control packets, including RREQ and RREP. Fig. 5 shows the routing overhead of the four routing protocols. The EDNR protocol yields a significant improvement with the help of our proposed route lifetime-prediction algorithm, and its overhead is reduced by 25.6%, 9.4%, and 6.3% However, the length of RREP packets is 3×4 B longer than that of the DSR.

5.1 Simulation Parameters

Table 1. Simulation Parameters

Simulation Time	1000s
Topology Size	1000m x 15000m
Number Of Nodes	100
MAC Type	MAC 802.11
Radio Propogation Model	Two Ray Model
Radio Propogation Range	250m
Pause Time	0s
Max Speed	4m/sec-24m/sec
Energy Model	Energy Model
Initial Energy	100J
Transmit Power	0.4W
Receive Power	0.3W
Traffic Type	CBR
CBR Rate	512 bytes x 6 per second
Number of Connections	50

5.2 Result Analysis

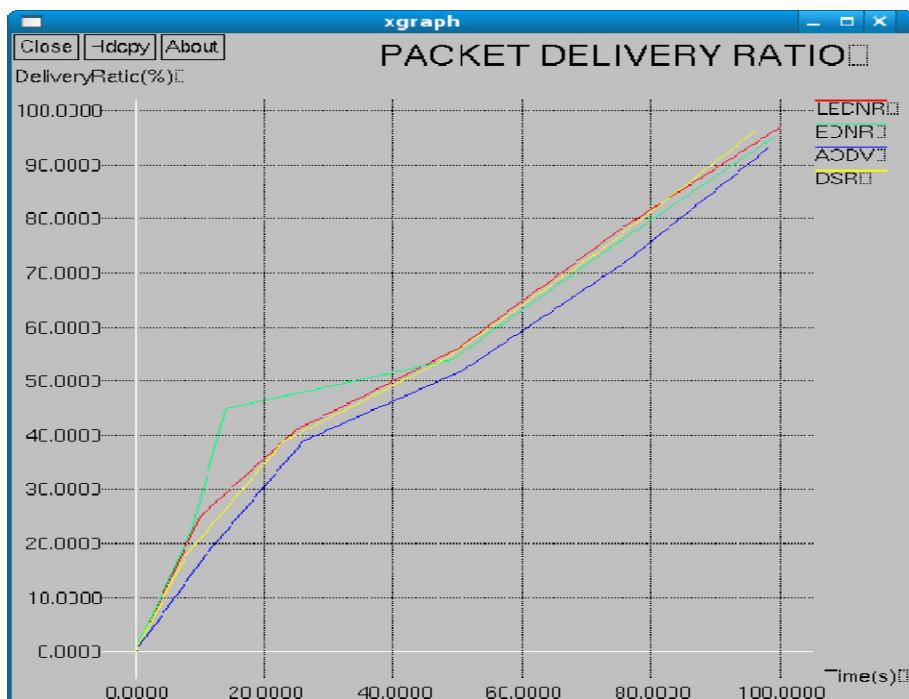
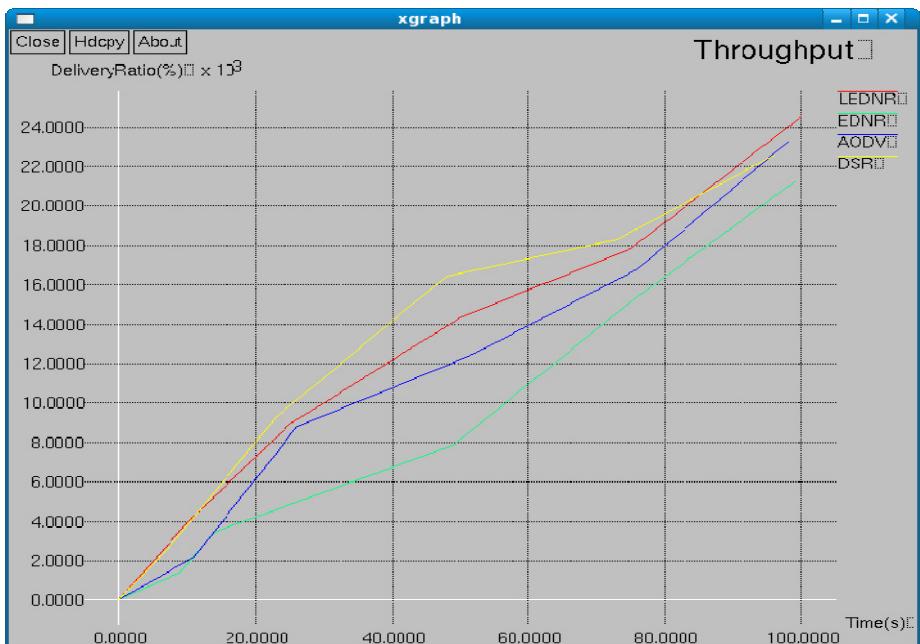
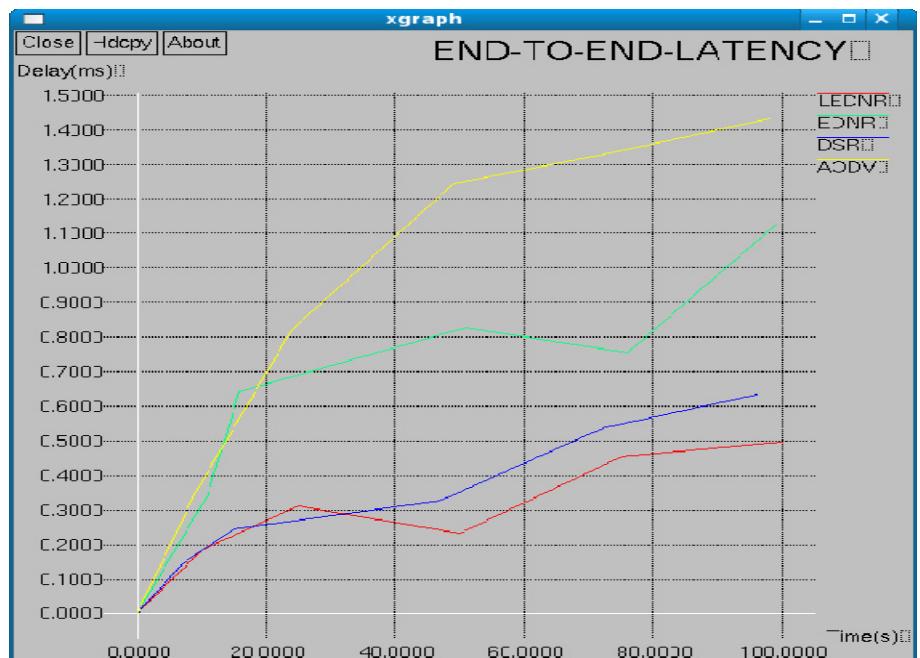
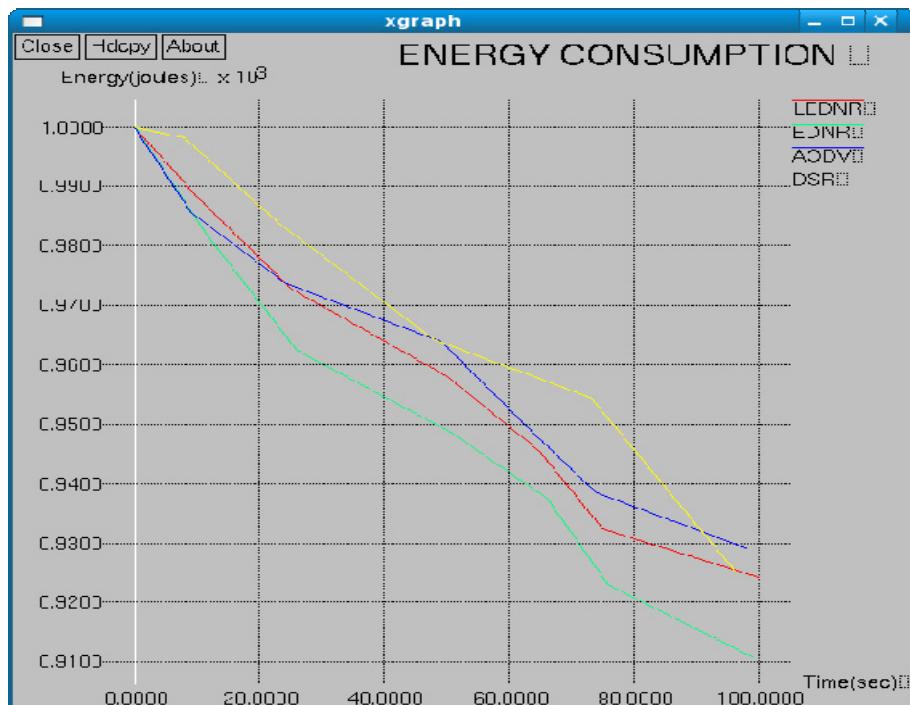


Fig. 2. Packet Delivery Ratio

**Fig. 3.** Throughput**Fig. 4.** End to End Latency

**Fig. 5.** Energy Consumption

5.3 Large Scale-Routing Protocol Performance Evaluation

Table 2. Comparison of performance of protocols

Parameters	Routing Protocol	
	Min	Max
Throughput	EDNR	LEDNR
Energy Consumption	EDNR	DSR
End to End Latency	LEDNR	AODV
Packet Delivery Ratio	AODV	LEDNR

6 Conclusion

In MANETs, a link is formed by two adjacent mobile nodes, which have limited battery energy and can roam freely, and the link is said to be broken if any of the

nodes dies because they run out of energy or they move out of each other's communication range. In this paper, we have considered both the node lifetime and the LLT to predict the route lifetime and have proposed a new algorithm that explores the dynamic nature of mobile nodes, such as the energy drain rate and the relative motion estimation rate of nodes, to evaluate the node lifetime and the LLT. Combining these two metrics by using our proposed route lifetime-prediction algorithm, we can select the least dynamic route with the longest lifetime for persistent data forwarding .Finally, we have evaluated the performance of the proposed LEDNR protocol based on the AODV. Simulation results show that the AODV protocol implemented with LEDNR outperforms the DSR protocol implemented with EDNR mechanisms.

References

1. Zhang, X.M., Zou, F.F.: Exploring the Dynamic Nature of Mobile Nodes for Predicting Route Lifetime in MANETS. *IEEE Transactions on Vehicular Technology* 59(3) (March 2010)
2. Wei, X.H., Chen, G.L., Wan, Y.Y., Zhang, X.M.: Longest lifetime path in mobile ad hoc networks. *J. Softw.* 17(3), 498–508 (2006)
3. Shrestha, N., Mans, B.: Exploiting overhearing: Flow-aware routing for improved lifetime in ad hoc networks. In: Proc. IEEE Int. Conf. Mobile Ad-hoc Sens. Syst., pp. 1–5 (2007)
4. Marbukh, V., Subbarao, M.: Framework for maximum survivability routing for a MANET. In: Proc. MILCOM 2000, pp. 282–286 (2000)
5. Toh, C.-K.: Maximum battery life routing to support ubiquitous mobile computing in wireless ad hoc networks. *IEEE Commun. Mag.* 39(6), 138–147 (2001)
6. Misra, A., Banerjee, S.: MRPC: Maximizing network lifetime for reliable routing in wireless environments. In: Proc. IEEE WCNC, pp. 800–806 (2002)
7. Maleki, M., Dantu, K., Pedram, M.: Lifetime prediction routing in mobile ad hoc networks. In: Proc. IEEE WCNC, pp. 1185–1190 (2003)
8. Toh, C.K.: Associativity-based routing for ad hoc mobile networks. *Wirel. Pers. Commun.—Special Issue on Mobile Networking and Computing Systems* 4(2), 103–139 (1997)
9. Dube, R., Rais, C.D., Wang, K.Y., Tipathi, S.K.: Signal stabilitybased adaptive routing (SSA) for ad hoc mobile networks. *IEEE Pers. Commun.* 4(1), 36–45 (1997)
10. Tickoo, O., Raghunath, S., Kalyanaraman, S.: Route fragility: A novel metric for route selection in mobile ad hoc networks. In: Proc. IEEE ICON, pp. 537–542 (2003)
11. Gerharz, M., de Waal, C., Frank, M., Martini, P.: Link stability in mobile wireless ad hoc networks. In: Proc. 27th Annu. IEEE Conf. Local Comput. Netw., pp. 30–42 (2002)
12. Qin, L., Kunz, T.: Pro-active route maintenance in DSR. *ACM SIGMOBILE Mobile Comput. Commun. Rev.* 6(3), 79–89 (2002)
13. Su, W., Lee, S.J., Gerla, M.: Mobility prediction and routing in ad hoc wireless networks. *Int. J. Netw. Manage.* 11(1), 3–30 (2001)
14. Samar, P., Wicker, S.B.: On the behavior of communication links of a node in a multi-hop mobile environment. In: Proc. Int. Symp. Mobile Ad Hoc Netw. Comput., pp. 145–156 (2004)
15. Wu, X., Sadjadpour, H.R., Garcia-Luna-Aceves, J.J.: An analytical framework for the characterization of link dynamics in MANETs. In: Proc. IEEE Mil. Commun. Conf., pp. 1–7 (2006)

16. Johnson, D., Hu, Y., Maltz, D.: DSR: RFC 4728,
<http://www.ietf.org/rfc/rfc4728.txt>
17. Sarkar, T.K., Ji, Z., Kim, K., Medouri, A., Salazar-Palma, M.: A survey of various propagation models for mobile communication. *IEEE Antennas Propag. Mag.* 45(3), 51–82 (2003)
18. <http://www.isi.edu/nsnam/ns>
19. Bettstetter, C., Resta, G., Santi, P.: The node distribution of the random waypoint mobility model for wireless ad hoc networks. *IEEE Trans. Mobile Comput.* 2(3), 257–269 (2003)
20. Le Boudec, J.-Y., Vojnovic, M.: Perfect simulation and stationarity of a class of mobility models. In: Proc. IEEE INFOCOM, pp. 2743–2754 (2005)

Reachability Analysis of Mobility Models under Idealistic and Realistic Environments

Chirag Kumar¹, C.K. Nagpal², Bharat Bhushan³, and Shailender Gupta⁴

^{1,3,4} YMCA University of Science and Technology,

² Echleon Institute of Technology

{Chiragarora35@gmail.com, Nagpalckumar@rediffmail.com,
bhrts@yahoo.com, Shailender81@gmail.com}

Abstract. The mobility models are used to represent the unpredictable movement pattern of the nodes in Mobile Ad-hoc Network (MANET) and give us an idea regarding their location, velocity and acceleration change over time. These models are used for simulation purpose in standard software tools such as QualNet, ns-2 etc. This paper evaluates the performance of routing protocols for mobility models such as Random Way Point (RWP), Random Walk (RW) and Random Direction (RD) under idealistic and realistic conditions based on a parameter termed as Probability of Reachability (POR). The POR is defined as the fraction of reachable routes to all possible routes between all pairs of sources and destinations. For this purpose a simulator is designed in MATLAB. We observe a marked difference in value of POR under idealistic and realistic conditions.

1 Introduction

MANET [1] is formed by the set of mobile nodes that are connected via wireless links without using any fixed infrastructure. Due to absence of centralized routers, each node in MANET has to act as a gateway, transmitter and receiver, making the routing task even more challenging than other conventional wireless networks. In addition to above several other factors such as areas shape where the network is to be deployed [2], limited bandwidth, processing capability, memory, battery power, and unpredictable movement of the nodes also affect the reachability significantly. The unpredictable behavior of nodes in MANET results in their random organization which alters the topology of the network rapidly and unpredictably. This unpredictable movement pattern of nodes is presented by various researchers by presenting an idea regarding their location, velocity and acceleration change over time and is termed as mobility model [3] [4].

To evaluate the performance of mobility models on routing protocols, network simulator are designed which makes simulation modeling an invaluable tool for understanding the operation of these networks. Once the nodes in these networks are placed, the mobility model determines how the nodes move within the network. A variety of mobility models have been proposed for MANET [5, 6, 7, 8, 9, 10], and a survey of many is presented in [3, 4, 11]. These models vary widely in their movement characteristics. All these models play a significant role in determining the reachability of routing protocols.

Literature study [4, 11, 12, 13] has shown that the mobility model in use can significantly impact the performance of ad hoc routing protocols, based on packet delivery ratio, the control overhead, and the data packet delay. Hence, it is important to use mobility models that accurately represent the intended scenarios in which the protocol is likely to be used. In this way the performance of the protocol can be more accurately predicted. In this paper, we propose to create more realistic movement models by incorporating obstacles in the simulation area. The obstacles [14] are placed within a network area to model the location of buildings within an environment, i.e. a college campus. This paper studies the impact of mobility models such as Random Way Point (RWP), Random Walk (RW) and Random Direction (RD) under realistic conditions on reachability of routing protocols.

The paper has been organized as follows: Section 2 provides the literature survey on mobility models used in our simulation, section 3 provides the simulator design and experimental setup parameters, section 4 describes the simulation and results followed by concluding remarks.

2 Mobility Models Used in MANET

The various mobility models used to evaluate the performance (POR) of routing protocols are as follows

2.1 Random Way Point Mobility Model

The RWP model was proposed by John and Maltz[15] in which all the nodes randomly select different locations as their destinations within the simulation area [3][4][16]. With the start of simulation the nodes start moving towards the selected destinations from their existing locations with uniform velocities selected randomly from the uniformly distributed array $[0, V_{\max}]$. Once the node reaches at the destination, it stays there for some time known as pause time before moving to a new destination. The pause time is selected from the array $[0, T_{\text{pause}}]$. The above process is repeated until the simulation time is over. In RWP model the behavior of the mobile nodes is completely described by the maximum velocity (V_{\max}) and the Pause Time (T_{pause}). Fig 1 shows the movement of a node using RWP Mobility Model.

2.2 Random Walk Mobility Model

This mobility model [4][16][17] was developed and described mathematically by Einstein in 1926 to emulate the unpredictable movements of the particles known as Brownian motion . In this model a node starts its motion by selecting a direction with speed from the pre-specified ranges $[0, 2*\pi]$ and $[0, V_{\max}]$. The node moves for a fixed time interval t or moves for a fixed distance d . After distance d or time t , new direction and speed are selected from the pre-specified ranges. If the specified time or distance is very small then the node's movement pattern will be restricted to a small portion of the simulation area and vice versa. If a particular node reaches to the boundary of simulation area it is bounced back with π -incoming angle and is termed as border effect. This model is a memory-less. Therefore, the current speed and the

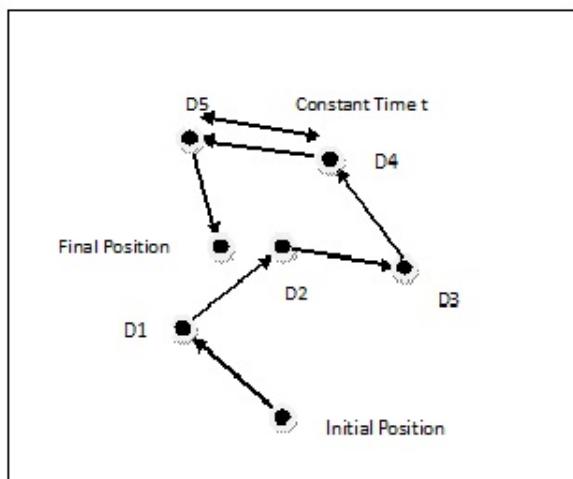


Fig. 1. Node movement in Random Way Point Mobility Model

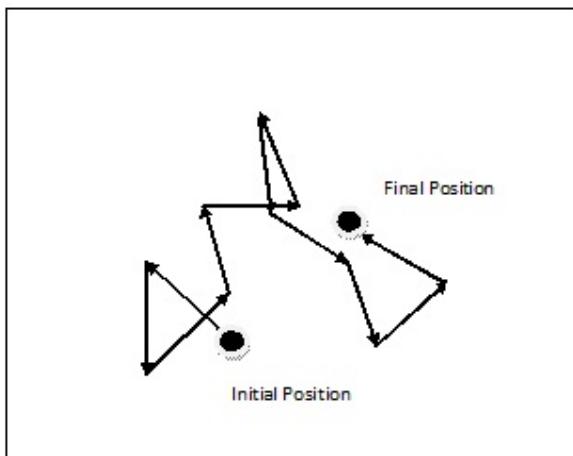


Fig. 2. Node movement in Random Walk Mobility Model

direction of the node is independent of its past speed and direction. Fig 2 shows the movement of a node in case of RW Mobility Model.

2.3 Random Direction Mobility Model (RDM)

The random direction mobility model [3][16][18] was developed in order to overcome the flaws discovered in the RWM Model. The problem in the RWM Model is that the probability of a node to choose a new destination located at the center of the simulation area, or a destination that requires path of the node through the center of the simulation area is high. This results in clustering of mobile nodes near the center of

simulation area. In RDM model the node chooses random direction and velocity from the specified range $[0, 2\pi]$ and $[0, V_{\max}]$. As the node reaches at the border of the simulation area, it waits there for pause time. After expiry of pause time, node chooses a new direction from $[0, \pi]$ and starts moving again towards the simulation border in a new direction. This process continues until the simulation is over. This mobility model is similar to the RWM Model, with a small difference in motion of a node to the border of simulation in RD model instead of motion for constant time or distance as in RWM. Fig 3 shows the movement of a node in case of RD Mobility Model.

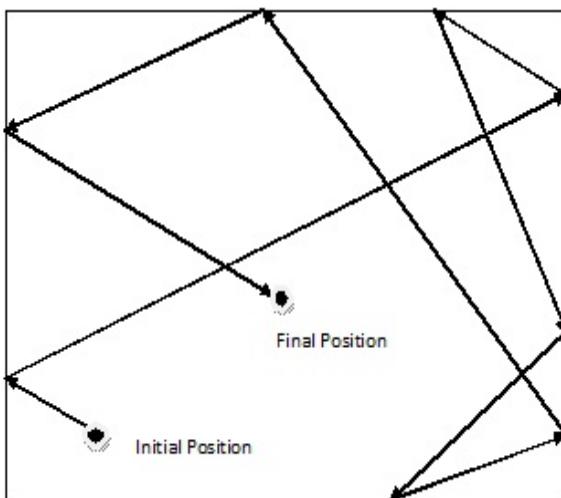


Fig. 3. Node movement in Random Direction Mobility Model

3 Simulation and Experimental Setup

3.1 Metric Used

The Probability of Reachability (POR) [19] is used to evaluate the performance of different mobility models in idealistic and realistic conditions and is defined as “The fraction of possible reachable routes to all possible routes between all different sources to all different destinations”. Mathematically it can be expressed as

$$\text{POR} = \frac{\text{# REACHABLE PATHS BETWEEN SOURCE and DESTINATION}}{\binom{n}{2}}$$

where n is number of nodes.

The POR is calculated by checking the path existence between all pairs of sources and destinations. For this purpose a count variable is taken and initialized to zero value. If the path exists, variable count is incremented. In this way all combinations of source destination pairs are checked. The POR is calculated by using the above

equation. To ensure complete randomness the process is repeated 25 times and average POR is calculated as shown in Fig. 4.

3.2 Simulation Setup Parameter

Various Parameters used for the simulation process are as given in Table-1.

Table 1. Simulation setup Parameters

Parameter	Value
Size of Region	2250000 sq. unit
Mobility Model Used	RWP,RWM,RDM
Number of Nodes Deployed	30
Transmission Range	300 to 450
Routing Algorithm Used	Dijkstra's Shortest Path Routing Algorithm
Routing Strategies Used	Maximum & Minimum Hop Routing
Placement of Nodes	Random
Number of iterations	25

3.3 Simulation Process

In Fig 4 the simulation process is explained using flowchart. Fig 5 and Fig. 6 shows the snapshot of the simulation area, where MANET nodes are to be deployed under idealistic and realistic conditions respectively using MATLAB as simulator. The path in red color shows the Minimum hop routing between the nodes and the path shown by cyan color is for maximum hop routing. The three green rectangles display the obstacles in the given area representing realistic condition.

3.4 Results

3.4.1 Impact of RWP Mobility Model on POR

Fig. 7 shows the impact of varying transmission range on Minimum Hop and Maximum Hop routings protocol under idealistic and realistic environment on RWP mobility model. After comparison of the results it is observed that at very low transmission range there is negligible impact of obstacles for both the routing strategies. As the transmission range increases there is a marginal increase in POR values for both the routing protocols. The POR value approaches to 50% in case of realistic environment while it approaches to nearly 100% in case of idealistic environment as the transmission range value approaches to nearly 450 units for both the protocols.

3.4.2 Impact of RWM Model on POR

Fig. 8 shows the impact of varying transmission range on Minimum Hop and Maximum Hop routings protocol under idealistic and realistic environment on RWM mobility model. It can be easily observed from the graph that even at very low transmission ranges the impact of obstacles on POR is significant. As the transmission range increases further the impact goes on increasing as shown in the graph.

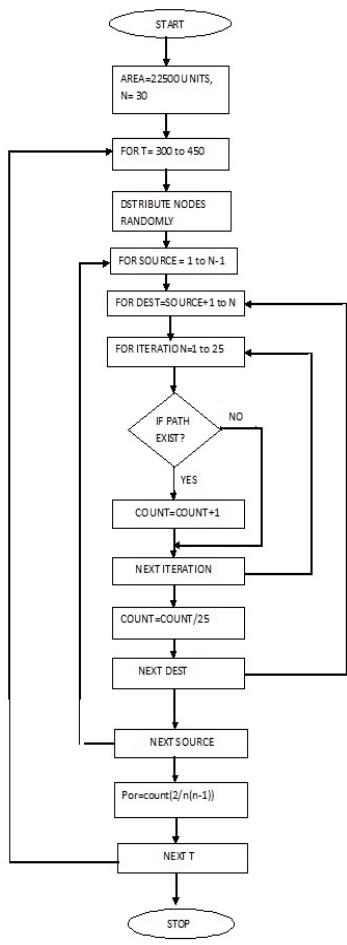


Fig. 4. Flow Chart of Simulation process

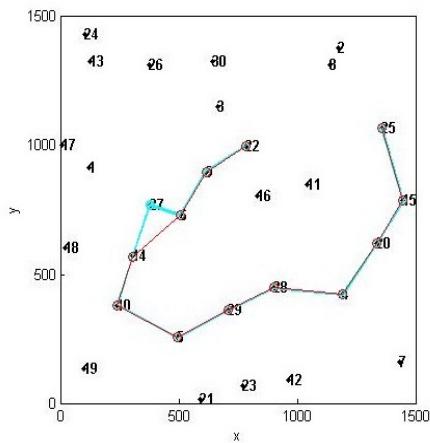


Fig.5. Routing in MANETs in idealistic Conditions

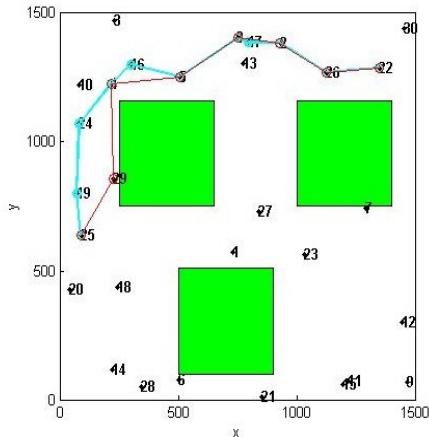


Fig. 6. Routing in MANETs in realistic Conditions

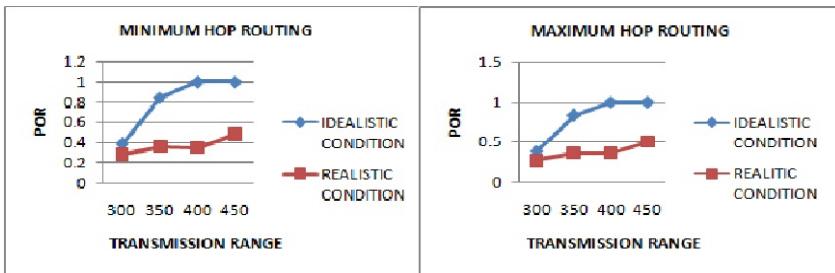


Fig. 7. Performance of RWP Mobility Model using max and min hop routing in different conditions

3.4.3 Impact of RDM Model on POR

Fig. 9 shows the impact of varying transmission range on Minimum Hop and Maximum Hop routings protocol under idealistic and realistic environment on RDM mobility model. It can be easily observed from the graph that even at very low transmission ranges the impact of obstacles on POR is remarkable. As the transmission range increases further the impact goes on increasing as shown in the graph.

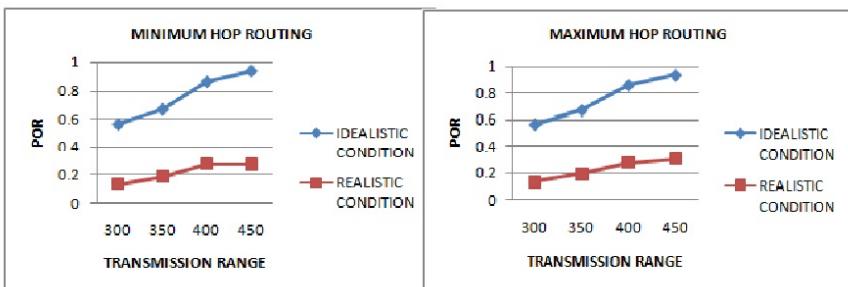


Fig. 8. Performance of RWM Model using max and min hop routing in different conditions

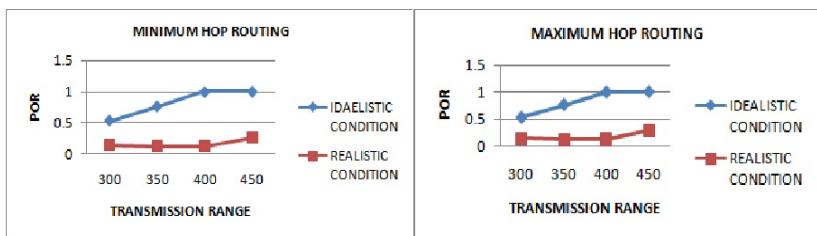


Fig. 9. Performance of RDM Model using max and min hop routing in different conditions

3.4.4 Comparison of Various Mobility Models

Fig 10 and 11 compares the performance of the three mobility models using two routing strategies under idealistic and realistic environmental conditions. The random

way point mobility model is having higher performance in the idealistic as well in realistic environmental conditions in terms of POR thus ensuring the reliable communication in comparison with the other two mobility models. The performance of all the mobility models is very much influenced in the presence of obstacle as can be seen from all the above mentioned results.

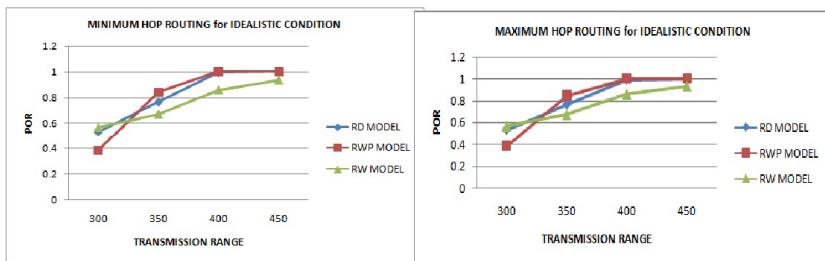


Fig. 10. Performance of Mobility Models utilizing max and min hop routing in idealistic conditions

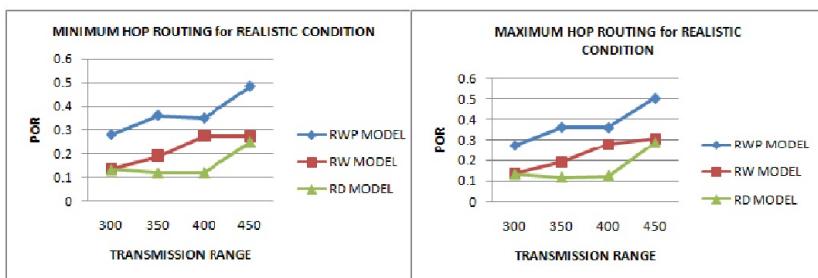


Fig. 11. Performance of Mobility Models utilizing max and min hop routing in Realistic conditions

4 Conclusion

From the various simulation results the concluding remarks can be made as follows:

- The RWM mobility model is having higher POR values in both the environmental conditions (idealistic as well in realistic) for both the protocols.
- At lower value of transmission range the POR value of all the three models is significantly low.
- For idealistic conditions at higher value of transmission range the POR values approaches nearly unity for all the three mobility models for both the protocols.
- The maximum hop routing shows higher value of POR compared to the minimum hop for all the three mobility models.

It can be concluded that the performance of all the mobility models is very much influenced in the presence of obstacles as can be seen from all the above mentioned results.

References

- [1] Macker, J., Corson, S.: Mobile ad-hoc networks, MANET (December 2001), <http://www.ietf.org/proceedings/01dec/183.htm>
- [2] Nagpal, C.K., Gupta, S., Bhushan, B.: Impact of Area's Shape on MANET Performance. In: IEEE Conference WICT (2011)
- [3] Camp, T., Boleng, J., Davies, V.: A Survey of Mobility Models for Ad Hoc Network Research. In: Wireless Communication & Mobile Computing (WCMC): Special issue on Mobile Ad Hoc Networking:Research, Trends and Applications, vol. 2(5), pp. 483–502 (2002)
- [4] Cooper, N., Meghanathan, N.: Impact of Mobility Models in Multi-Path Routing in Mobile Ad Hoc Networks. International Journal of Computer Networks & Communications (IJCNC) 2(1) (January 2010)
- [5] Broch, J., Maltz, D.A., Johnson, D., Hu, Y.-C., Jetcheva, J.: A Performance Comparison of Multi-Hop Wireless Ad Hoc Network Routing Protocols. In: Proceedings of the 4th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom), Dallas, Texas, pp. 85–97 (October 1998)
- [6] Davies, V.: Evaluating Mobility Models Within an Adhoc Network. Master's thesis, Colorado School of Mines (2000)
- [7] Haas, Z.: A New Routing Protocol for Reconfigurable Wireless Networks. In: Proceedings of the IEEE International Conference on Universal Personal Communications (ICUPC), pp. 562–565 (October 1997)
- [8] Hong, X., Gerla, M., Pei, G., Chiang, C.-C.: A Group Mobility Model for Ad hoc Wireless Networks. In: Proceedings of the ACM/IEEE MSWIM 1999, Seattle, WA (August 1999)
- [9] Liang, B., Haas, Z.: Predictive Distance-based Mobility Management for PCS Networks. In: Proceedings of the IEEE Conference on Computer Communication (INFOCOM), New York, NY (March 1999)
- [10] Royer, E.M., Melliar-Smith, P.M., Moser, L.E.: An Analysis of the Optimum Node Density for Ad hoc Mobile Networks. In: Proceedings of the IEEE International Conference on Communications, Helsinki, Finland, pp. 857–861 (March 2001)
- [11] Kumar, S., Sharma, S.C., Suman, B.: Classification and Evaluation of Mobility Metrics for Mobility Model Movement Patterns in Mobile Ad-Hoc Networks. International Journal on Applications of Graph Theory in Wireless Ad Hoc and Sensor Networks (GRAPH-HOC) 3(3) (September 2011)
- [12] Divecha, B., Abraham, A., Grosan, C., Sanyal, S.: Impact of node mobility on MANET routing protocols models. Journal of Digital Information Management (February 1, 2007)
- [13] Venkateswaran, A., Sarangan, V., Gautam, N., Acharya, R.: Impact of mobility prediction on the temporal stability of MANET clustering algorithms. In: Proceedings of the 2nd ACM International Workshop on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks (2005)
- [14] Jardosh, A., Belding Royer, E.M., Almeroth, K.C., Suri, S.: Towards Realistic Mobility Models For Mobile Ad hoc Networks. In: MobiCom 2003, San Diego, California, USA, September 14-19 (2003)

- [15] Random Waypoint Model,
<http://www.netlab.tkk.fi/~esa/java/rwp/rwpmode1.shtml>
- [16] Saad, M.I.M., Zukarnain, Z.A.: Performance Analysis of Random-Based Mobility Models in MANET Routing Protocol. European Journal of Scientific Research 32(4), 444–454 (2009) ISSN:1450-216X
- [17] Zonoozi, M., Dassanayake, P.: User Mobility Modeling and Characterization of Mobility Pattern. IEEE Journal on Selected Areas in Communications 15(7), 1239–1252 (1997)
- [18] Buruhanudeen, S., Othman, M., Othman, M., Ali, B.M.: “Mobility Models, Broadcasting Methods and Factors Contributing Towards the Efficiency of the MANET Routing Protocols: Overview”, paper id-123
- [19] Kumar, C., Gupta, S., Bhushan, B.: Impact of Various Factors on Probability of Reachability in Manet: A Survey. International Journal on Applications of Graph Theory in Wireless Ad Hoc Networks and Sensor Networks (GRAPH-HOC) 3(3) (September 2011)

Chaotic Cipher Using Arnolds and Duffings Map

Mina Mishra¹ and V.H. Mankar²

¹ Nagpur University, Nagpur, Maharashtra, India
minamishraetc@gmail.com

² Department of Electronics Engineering, Government Polytechnic,
Nagpur, Maharashtra, India
vhmarkar@gmail.com

Abstract. This paper deals with the application of concept of identifiability based on output equality approach on chaotic ciphers developed using 2-D chaotic maps, Duffings and Arnolds Cat map and they are named as Duffings and Arnold's Cat, according to the map used. Due to the less key space generally many chaotic cryptosystem developed are found to be weak against Brute force attack which is an essential issue to be solved. Thus, concept of identifiability proved to be a necessary condition to be fulfilled by the designed chaotic cipher to resist brute force attack, which is an exhaustive search. As 2-D chaotic maps provide more key space than 1-D maps thus they are considered to be more suitable. This work is accompanied with analysis results obtained from these developed cipher. Moreover, identifiable keys are searched for different input texts at various key values. The ciphers are also analyzed for plaintext sensitivity and key sensitivity for its validity to provide security.

Keywords: Arnolds Cat map, Duffings map, Identifiability, Brute force attack, linear cryptanalysis, Differential cryptanalysis, Output-Equality approach.

1 Introduction

Chaotic systems [1] have many interesting features, such as the sensitivity to the initial condition and control parameter and mixing property, which have relationships with the requirements of pseudo-random coding and cryptography. For example, the sensitivity to the initial condition and the mixing property can be connected with confusion and diffusion property of a good cryptosystem [2] [3]. Thus, it is a natural idea to use chaos as a source to construct new encryption systems. Discretized chaotic maps [4] produce pseudorandom sequence of good auto-correlations, complexity, and random-looking from many aspects. Therefore, it has been widely used in chaos based secure communication systems.

The aim of this paper is to crypt analyze two of the stream symmetric chaotic ciphers constructed using one of the promising chaotic scheme known as message-embedded scheme [6] [7] in which 2-D chaotic maps, Duffings and Arnolds Cat map are used. 2-D chaotic system response depends on two parameters that act as secret key in the ciphers due to which complexity and key space is increased compared to

1-D chaotic map for example, Logistic. Both of the ciphers are analyzed for key space, avalanche effect and strength against Brute-force and Known-plaintext attack.

A cryptanalytic method, known as output equality based on the identifiability concept, is the solution to the problem of less key space in chaotic ciphers. It is possible to test in prior about the cipher strength against Brute-force attack using it. Both of the mentioned ciphers are concluded to provide security against the Brute-force attack. Identifiability concept fulfills the necessary condition but not sufficient as the developed cryptosystems must be tested for sensitivity and other statistical tests to result in a robust cipher. Thus both the ciphers are tested for sensitivity and it is concluded that some of the keys selected from domain of key space of the ciphers seem to have good key sensitivity and resist known plaintext attack for the available first two characters of plaintext.

This paper is organized into five sections as follows. Section 2, presents the background and in section 3 algorithm for encryption used in developing ciphers is provided. Then in section 4, analysis result in tabulated form and discussions are presented. Section 5, discusses about the conclusions derived.

2 Background

Message-Embedded Scheme: According to this scheme at the transmitter side, the plain text is encrypted by an encryption rule which uses non-linear function and the state generated by the chaotic system in the transmitter. The scrambled output signal is used further to drive the chaotic system such that the chaotic dynamics is changed continuously in a very complex way. Then another state variable of the chaotic system in the transmitter is transmitted through the channel.

At the receiver side, the reconstruction of the plaintext is done by decrypting the input by using the reverse of encryption method.

Arnold's Cat Map: Arnolds Cat map is a 2-D discrete-time dynamical system, which takes a point(x, y) in the plane and maps it to a new point using equations:

$$x(k+1) = (a - 1) \bmod [2x(k) + y(k), N];$$

$$y(k+1) = \bmod[x(k) + (1 - b)y(k), N];$$

a, b and N are parameters on which the map depends. At a=0.3, b=0.345, map exhibits chaotic nature.

Duffings map: Duffings map is a 2-D discrete-time dynamical system, which takes a point(x, y) in the plane and maps it to a new point using equations:

$$x(k+1) = y(k);$$

$$y(k+1) = -bx(k) + ay(k) - y(k)^3$$

a and b are parameters on which the map depends. At a = 2.75, b = 0.2, map exhibits chaotic nature.

Non-Linear Function: Modular function is used as non-linear function in the construction of ciphers.

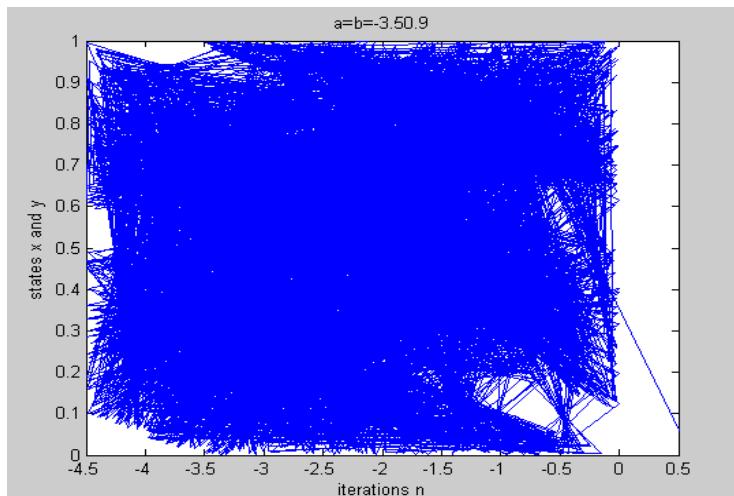


Fig. 1. Plot of Arnolds Cat Map at $x(0) = 0.5$, $y(0) = 0.06$, $a = -3.5$, $b = 0.9$, $n=5000$

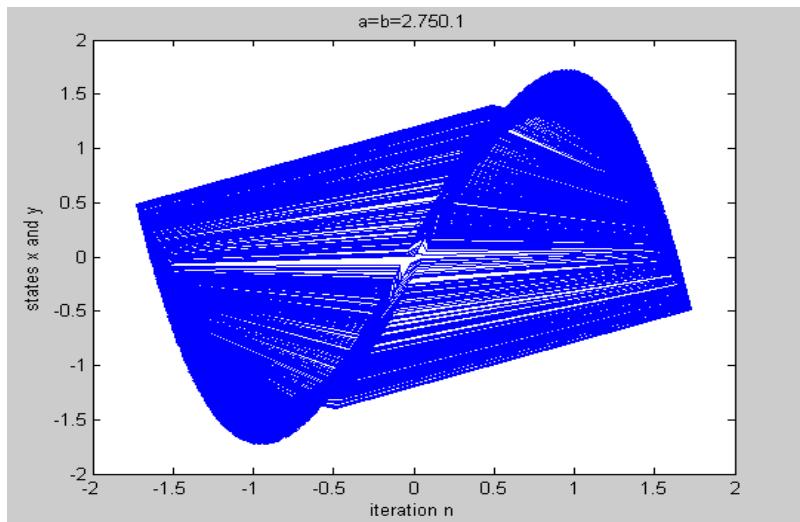


Fig. 2. Plot of Duffings Map at $x(0) = -0.04$, $y(0) = 0.2$, $a= 2.75$, $b=0.1$, $n=5000$

The Mod numeric function returns the remainder when the dividend is divided by the divisor. The result is negative only if the dividend is negative. Both the numbers

must be integers. The function returns an integer. If any number is NULL, the result is NULL. For example:

Mod (5, 3) returns 2.
Mod (-5, 3) returns -2.

Cryptanalysis: Cryptanalysis is the study of attacks against cryptographic schemes to disclose its possible weakness. During crypt analyzing a ciphering algorithm, the general assumption made is that the cryptanalyst knows exactly the design and working of the cryptosystem under study, i.e., he/she knows everything about the cryptosystem except the secret key. It is possible to differentiate between different levels of attacks on cryptosystems. They are briefly explained as follows:

1. **Cipher text-only attack:** The attacker possesses a string of cipher text.
2. **Known plain text:** The attacker possesses some portion of plain text and the corresponding cipher text.
3. **Chosen plain text:** The attacker has obtained temporary access to the encryption machinery. Hence he/she can choose a plain text string, p, and construct the corresponding cipher text string.
4. **Chosen cipher text:** The attacker has obtained temporary access to the decryption machinery. Hence he/she can choose a cipher text string, c, and construct the corresponding plain text string.
5. **Brute Force Attack:** A brute force attack is the method of breaking a cipher by trying every possible key. The brute force attack is the most expensive one, owing to the exhaustive search.

In addition to the five general attacks described above, there are some other specialized attacks, like, differential and linear attacks.

Differential cryptanalysis is a kind of chosen-plaintext attack aimed at finding the secret key in a cipher. It analyzes the effect of particular differences in chosen plaintext pairs on the differences of the resultant cipher text pairs. These differences can be used to assign probabilities to the possible keys and to locate the most probable key.

Linear cryptanalysis is a type of known-plaintext attack, whose purpose is to construct a linear approximate expression of the cipher under study. It is a method of finding a linear approximation expression or linear path between plaintext and cipher text bits and then extends it to the entire algorithm and finally reaches a linear approximate expression without intermediate value.

Security Analysis: Various cryptanalytic procedures are developed to test the validity of newly constructed ciphers and they are as follows:

- (a) **Key Space Analysis:** Key space belongs to the chaotic region of the system in case of chaotic ciphers. The total key space is a product of all the parameters involved. Once the key has been defined and key space has been properly characterized, the good key is chosen randomly from the large key domain.

- (b) **Identifiability Test method:** From the crypto graphical point of view, the size of the key space should not be smaller than 2^{100} to provide a high level security so that it can resist all kind of Brute force attack. A fundamental issue of all kinds of cryptosystem is the key. No matter how strong and how well designed the encryption algorithm might be, if the key is poorly chosen or the key space is too small, the cryptosystem will be easily broken. Unfortunately, chaotic cryptosystem has a small key space region and it is non-linear because all the keys are not equally strong. The keys should be chosen from the chaotic region. To solve the problem of small key space and weakness against brute force attack, identifiability concept is quite advantageous.

A cryptanalytic procedure, known as output equality based on the identifiability concept, is carried out on the developed ciphers. It is found that in chaotic ciphers, there exists a unique solution for a particular input for certain domain of values of parameters. The response of any system to a particular input is the solution of that particular system and it contains all the information about the parameters of system. In the discussed ciphers, system parameters are acting as a secret key. This type of analysis is also known as parametric analysis.

The output equality method is explained as follows:

“For the same inputs and initial condition, transmitter system is parameterized at different values of parameter taken from the existing domain of parameter space, if the output response of the system obtained after some value of iteration, parameterized at a particular value coincides with the output response of the same system parameterized at some other value of parameter within the domain for the same number of iteration, then both the parameters are said to be equal and identifiable. The system is said to possess unique solution at that particular value of parameter and the system is said to be structurally identifiable.”

If parameter of the transmitter is identifiable, it is more difficult for the eavesdropper to find it by a brute force attack. Consequently, this parameter can play the role of the secret key against brute force attack. If parameter is not identifiable, the eavesdropper has a higher favorable chance to find it by a brute force attack and thus, the parameter vector cannot play the role of the secret key against brute force attack.

- (c) **Plaintext sensitivity Test Method:** The percentage of change in bits of cipher text obtained after encryption of plaintext, which is derived by changing single bit from the original plaintext from the bits of cipher text obtained after encryption of original plaintext. With the change in single bit of plaintext, there must be ideally 50% change in bits of cipher text to resist differential cryptanalysis (chosen-plaintext attack) and statistical analysis.
- (d) **Key sensitivity Test Method:** The percentage of change in bits of cipher text obtained after encryption of plaintext using key, which is flipped by single bit from the original key, from bits of cipher text obtained after encryption of plaintext using original key, which requires ideally 50% change in cipher text bits to resist Linear and statistical attacks.

- (e) **Known plaintext attack Method:** For observing this attack on developed cryptosystem it is assumed that the opponent knows everything about the algorithm, he/she has the corresponding cipher text of plaintext and some portion of plaintext. With this much information, the opponent tries to find out the secret key.

3 Algorithm for the Developed Ciphers

Encryption Algorithm:

Step-1: Read plaintext and key vector.

Step-2: Convert plaintext into its ASCII values.

Step-3: Each value of ASCII values are transformed using following steps:

- (a) The chaotic map is iterated for a number of times to output a random state.
- (b) ASCII value is mixed with non-Linear function and the output state of chaotic system obtained after a fixed value of iteration.
- (c) Again chaotic system is iterated for a fixed number of times and an output state is obtained.
- (d) The response obtained in (b) is mixed with the output state obtained in (d) and output values are obtained as cipher text.

Step-4: Convert cipher text into characters.

Step-5: Read the cipher text.

Decryption algorithm is reverse of encryption process and the original information is retained using the same secret key using which encryption is being done and it is kept secret between authenticated sender and receiver only.

4 Results and Analysis

Table 1 and 2 summarizes the simulated result data produced after analyzing both the ciphers using above discussed (section 2) cryptanalytic procedures. Ten different values of keys are chosen from key space of respective ciphers and are analyzed for its security. From both the observation tables, it can be seen that **plaintext sensitivity** of Arnold's cat cipher ranges from 0.5 to 2.5 % and Duffings cipher ranges from 0.5 to 2 %, which is not sufficient. **Key sensitivity** for each of the cipher ranges from 0 to 36 % and from 0 to 51 %, respectively. Thus key sensitivity property of some keys from both the ciphers shows satisfactory values. Both ciphers are robust against **known plaintext attack** for the available first two characters of plaintext. Key space of ciphers shows lesser range than compared to the required limit i.e. 2^{100} to resist **Brute-force attack** but **identifiable keys** conclude that the developed ciphers can resist Brute-force attack.

- a. **Arnold's Cat cipher:** Key space is from [-5 0.4] to [-0.9 1.5] = 5×10^{16}

Table 1. Analysis Table for Arnold's Cat

Sl. No .	Plain text	Key value	Cipher text	Plain text sensitivty (in %)	Key sensitivty (in %)	Domain for key With increment = 0.0001	Identifiability of key for iteration value =2 or 3	Robustness against known plaintext attack for p=[p1 p2].	Whether key can act as secret key against Brute Force attack?
1.	What is your name?	[0.0034 0.0013]	k ³ ¬k ³ 4k ^o À ¹ 2k ¹ ¬ ^o	1.9737	24.3421	[0 0]to[0.005 0.004]	I	R	YES
2.	I am going to market.	[-4.4977 0.2034]	kk¬k ² °12k ^o k _o ¬ ¹ 2y ^o y ^o	1.7045	21.5909	[-4.5 0.2]to[-4.495 0.204]	I	R	YES
3.	My college name is s.s.c.e.t .	[-2.9982 -0.6981]	kÀk [®] ° ..° ² °k ¹ ¬ ^o k [°] 3/4k ³ 4y ³ y [®] y [°] y ^o	1.2500	25.4167	[-3 - 0.7]to[-2.995 - 0.696]	I	R	YES
4.	Hello! how are you?	[-2.992 -0.694]	i@μμj ±,Ài [®] ®iÀ, ³ 4	1.3158	0	[-2.995 - 0.696]to[-2.99 - 0.692]	NI	R	NO
5.	Sita is singing very well.	[-2.99 - 0.692]	h± ¹ © h±>h» ±¶±¶ -h ³ 4- °Ah ^o - v	0.4630	36.1111	[-2.99 - 0.692]to[-2.985 - 0.688]	NI	R	NO
6.	Ram scored 98 marks in Maths.	[-2.985 -0.688]	lμÀ- lμl ^o μ ^o μ ^o À ± ³ Àl À±,z	0.4630	24.0741	[-2.985 - 0.688] to[-2.98 - 0.684]	NI	R	NO
7.	Jaycee publication.	[-0.9957 0.0101]	l- Å- ±±1 ¼À [®] , μ ⁻ - Àμ»°z U	0.5952	23.8095	[-1 0.01]to[-0.995 0.014]	I	R	YES
8.	Thank you,sir.	[-0.99 0.018]	k ³ ¬k ³ À ^o Aw ³ 4 ¹ 2y	2.5000	26.6667	[-0.995 0.014]to[-0.99 0.018]	NI	R	NO

Table 1. (continued)

9.	The match was very exciting.	[-0.985 3 0.021 3]	$k^3\bar{k}^{-\zeta}\bar{\zeta}^{(3)}$ $\bar{k}\bar{A}^{-3/4}\bar{k}\bar{A}^{\circ}$ $1/2\bar{A}\bar{k}^{\circ}\bar{A}^{(3)}$ $\zeta^{-12}y$	1.2931	0	[-0.99 0.018]to[-0.985 0.022]	I	R	YES
10.	I will be leaving at 9p.m.	[0.203 5 1.400 9]	$ll\bar{\mu}_{\pm}^{1\otimes\pm}$ $l_{\pm}\bar{\mu}^{(3)}l-$ $\bar{A}^{1/4}z^1z$	0.4630	28.2407	[0.2 1.4]to[0. 205 1.404]	I	R	YES

NI – Non – Identifiable; I – Identifiable; R- Robust; p[p1 p2...p n]- First ‘n’ characters of available plaintext string.]

b. **Duffings cipher:** Key space is from [1.8 -0.59] to [2.9 0.2] = 9×10^{14}

Table 2. Analysis Table for Duffings

Sl. No .	Plainext	Key value	Ciphertext	Plaintext sensitivity (in %)	Key sensitivity (in %)	Domain for key With increment t = 0.0001	Identifiability of key for iteration value =2 or 3	Robustness against known plaintext attack for p=[p1 p2].	Whether key can act as secret key against Brute Force attack?
1.	What is your name?	[1.803 -0.584]	$k\zeta^3\bar{\zeta}k^{-3/4}k\bar{A}^{\circ}\bar{A}^{1/2}k^1\bar{\zeta}^{\circ}$	1.9737	28.9474	[1.8 -0.59]to[1.805 -0.586]	NI	R	NO
2.	I am going to market .	[1.8995 0.0068]	J!bn!hpjoh! up!nbslfu/	1.2987	0	[1.895 0.006]to[1.9 0.01]	I	R	YES
3.	My college name is s.s.c.e.t.	[1.9015 0.0100]	!Nz!dpmmfh f!obnf!jt!t/t/ d/f/u/	0.9524	0	[1.897 0.008]to[1.902 0.012]	I	R	YES
4.	Hello! how are you?	[1.8127 - 0.5804]	!Ifmmp"ipx! bsf!zpv@	1.5038	21.0526	[1.81 -0.582]to[1.815 -0.578]	I	R	YES

Table 2. (continued)

5.	Sita is singing very well.	[1.7547 0.1521]	!Tjub!jt!johj oh!wfsz!xfm m/	1.0582	0	[1.75 0.15]to[1. 755 0.154]	I	R	YES
6.	Ram scored 98 marks in Maths.	[1.4045 0.0020]	!Sbn!tdpsfe!: 9!nbslt!jo!N buit/	0.9524	0	[1.4 2.15]to[1. 405 2.154]	I	R	YES
7.	Jaycee publication.	[1.7544 0.1540]	!Kbzdff!qv cmjdbujpo/	1.4286	0	[1.75 0.15]to[1. 755 0.154]	I	R	YES
8.	Thank you,sir .	[1.6247 0.0937]	!Uibol!zpv-tjs/	1.9048	0	[1.62 0.09]to[1. 625 0.094]	I	R	YES
9.	The match was very excitin g.	[2.802 - 0.094]	"Vjg"ocvej "ycu"xgt{" gzekvkpi0	0.4926	51.2931	[2.8 - 0.1]to[2.8 05 - 0.096]	NI	R	NO
10.	I will be leaving at 9p.m.	[2.0045 0.1509]	!J!xjmm!cf! mfbwjoh!b u!:q/n/	1.0582	0	[2 0.15]to[2. 005 0.154]	I	R	YES

5 Conclusions

A generalized algorithm for chaotic ciphers developed using one of the promising scheme known as message-embedded are discussed. Ciphers are named according to the chaotic map used in it, known as Arnold's Cat and Duffings respectively and parameter of chaotic map acts as secret key. Due to the less key space generally many chaotic cryptosystem developed are found to be weak against Brute force attack which is an essential issue to be solved. Thus, concept of identifiability proved to be a necessary condition to be fulfilled by the designed chaotic cipher to resist brute force attack, which is an exhaustive search. As 2-D chaotic maps provide more key space than 1-D maps thus they are considered to be more suitable. This work is accompanied with analysis results obtained from these developed cipher. Moreover, identifiable keys are searched for different input texts at various key values. The ciphers are also analyzed for plaintext sensitivity and key sensitivity for its validity to provide security.

A comparison table no.3 shows that both ciphers are found to resist Brute-force attack as they consist of identifiable keys. Key sensitivity property is also good for some of the keys selected from domain of key space. Ciphers are determined to resist known plaintext attack for available first two characters of plaintext. If available characters are not the starting characters of plaintext then ciphers shows robustness against the attack for available any number of plaintext characters. Both the ciphers are more secure than one – time pad ciphers. Algorithm consists of simple structure, complex and random response.

Table 3. Comparison between the two ciphers

Name of Cipher	key Space	Range of plaintext sensitivity	Range of key sensitivity	Identifiable key	Robust against known plaintext attack	Whether key space $> 2^{100}$
Duffing's	9×10^{14}	0.5 to 2 %	0 to 51 %	Yes	Yes	No
Arnold's Cat map	5×10^{16}	0.5 to 2.5 %	0 to 36 %	Yes	Yes	No

References

1. Jakimoski, G., Kocarev, L.: Chaos and Cryptography: Block Encryption Ciphers Based on Chaotic Maps. *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications* 48(2), 163–169 (2001)
2. Anstett, F., Millerioux, G., Bloch, G.: Global adaptive synchronization based upon polytopic observers. In: Proc. IEEE Int. Symp. Circuits Syst., Vancouver, BC, Canada (May 2004)
3. Yang, T.: A survey of chaotic secure communication systems. *Int. J. Comput. Cogn.* 2(2) (2004)
4. de Oliveira, L.P., Sobottka, M.: Cryptography with chaotic mixing. *Chaos, Solitons and Fractals* 35, 466–471 (2008)
5. Alvarez, G., Li, S.: Some Basic Cryptographic Requirements for Chaos-based Cryptosystems. *Int. J. Bifurc. Chaos* (2006)
6. Millérioux, G., Hernandez, A., Amigó, J.: Conventional cryptography and message-embedding. In: Proc. 2005 Int. Symp. Nonlinear Theory and its Applications (NOLTA 2005), Bruges, Belgium, October 18-21 (2005)
7. Anstett, F., Millerioux, G., Bloch, G.: Message-embedded cryptosystems: Cryptanalysis and identifiability. In: Proc. 44th IEEE Conf. Decision and Control, Sevilla, Spain, December 12-15 (2005)
8. Yin, R., Yuan, J., Yang, Q., Shan, X., Wang, X.: Linear Cryptanalysis for a Chaos-based Stream Cipher. *World Academy of Science, Engineering and Technology* 60 (2009)
9. Masuda, N., Aihara, K.: Cryptosystems with Discretized chaotic maps. *IEEE Trans. Circuits and Syst. I* 49, 28–40 (2002)

10. Anstett, F., Milleroux, G., Bloch, G.: Chaotic Cryptosystems: Cryptanalysis and Identifiability. *IEEE Transactions on Circuits and Systems—I* 53(12) (December 2006)
11. Alvarez, G., Montoya, F., Romera, M., Pastor, G.: Cryptanalysis of a chaotic encryption system. *Physics Letters A* 276, 191–196 (2000)
12. Beth, T., Lazic, D.E., Mathias, A.: Cryptanalysis of Cryptosystems Based on Remote Chaos Replication. Springer, New York (1994)

Effect of Sound Speed on Localization Algorithm for Underwater Sensor Networks

Samedha S. Naik¹ and Manisha J. Nene²

¹ Department of Computer Engineering,

² Department of Applied Mathematics and Computer Engineering

Defence Institute of Advance Technology, Pune, India 411 025

{cse10samedha, mjnene}@diat.ac.in

Abstract. Autonomous Under Water Sensor Networks UWSNs form distributed amorphous computing environments. Efficient resolution for an unreachable UWSN which includes failure-prone nodes will require strategies that are as simple as possible in computations and local communications, to facilitate self-organization. Localization of Under Water Sensor Networks UWSNs is the most challenging and essential task. In this paper we propose a localization technique for UWSNs which is implemented using a self-organizing localizing algorithm. When acoustic waves propagate through a medium, it travels with varying speed. This change in speed of sound wave is highly influenced by ocean parameters. In our proposed work we study the effect of sound speed on localization algorithm for Underwater Sensor Networks. The results show that our proposed localization technique performs better.

Keywords: Underwater Sensor Network, Distributed, Self-Organizing, Localization, Distance estimation, Sound Speed.

1 Introduction

The vastness of the ocean, covering 3/4th of the earth's surface, is still not been observed and explored largely. Nowadays ocean monitoring systems gather information from ocean surface, coastlines and island, but a very few work is done for seabed monitoring and surveillance. The main advantage of using underwater sensor networks is that conventional large, expensive, individual ocean monitoring equipment units can be replaced by relatively small and less expensive underwater sensor nodes that are able to communicate with each other via acoustic signals.

In many situations, wireless sensor nodes are expected to be deployed in an ad-hoc fashion (i.e. air-dropped over an area). With ad-hoc deployment however, one cannot accurately predict or plan a-priori the location of each sensor. Precise knowledge of node location in ad-hoc deployed networks yields a wide variety of profound advantages. Knowledge of location can be used to report the geographical origin of events, data tagging, to assist in target tracking, geographic aware routing, and node tracking, to administer the sensor network and evaluate its coverage.

Localization of UWSN is an active area of research and is one of the most challenging and essential tasks for UWSN. Many UWSN enabled potential applications

like environment monitoring, review how human activities affect the marine ecosystem, undersea explorations, detect underwater oil fields, disaster prevention, monitoring ocean currents and winds-tsunamis, assisted navigation, locate dangerous rocks in shallow waters, distributed tactical surveillance and intrusion detection are of great benefit to the mankind and resource conservation of the planet. All these applications demand the need of efficient self organizing algorithms to make the sensor nodes location-aware.

In this paper we focus on the distributed localization techniques to compute accurate location, even when errors in communication are induced. We propose a new trilateration method to compute the unknown location of a sensor using three Reference Nodes RN's. Our proposed localization algorithm is completely decentralized and distributed in nature. Also our proposed localization technique reduces localization errors for UWSN caused due to varying sound speed.

2 Background

Sensor nodes which are placed underwater use acoustic signals to communicate with each other. Communication using acoustic waves under harsh physical layer environments is very challenging. The variable speed of sound and the long propagation delays under water pose a unique set of challenges for localization in UWSN [1]. Radio Frequencies RF can work at the most on the ocean surface but fails for underwater. Following are the reasons why acoustic communication is preferred over RF and optical waves: RF waves can travel in sea only at extra low frequencies (30-300 Hz). Hence large antenna and high transmission power is required. Other reasons are limited bandwidth, propagation delay (5 orders of magnitude greater than on terrestrial), very high bit error rates and temporary loss of connectivity. The underwater channel is severely impaired, especially due to multi-path and fading. Battery power is limited and usually batteries cannot be re-charged, also because solar energy cannot be exploited. Underwater sensors are prone to failures because of fouling and corrosion.

2.1 System Architecture

In our proposal we consider the requirement for underwater sensor networks to be self-organizing which implies that there is no fine control over the positioning of the sensor nodes when the network is deployed in the ocean. Consequently, we assume that nodes are randomly distributed across the environment. For simplicity and ease of simulation we limit the underwater environment to two dimensions and can be easily extended for 3 dimensional scenarios.

Nodes are dropped into the ocean either by plane or ship. Once they settle on the sea floor they start communicating to each other. The sensors must then estimate their position using different positioning algorithm. The proposed algorithm does not rely on any existence of previous infrastructure. The ordinary nodes are assumed to be grounded on the ocean bed. At present we assume that there are no water currents, hence this technique can be applied for terrestrial as well as underwater environment. Anchor nodes are able to detect their position by means of GPS satellite system attained before diving into the ocean. In order to perform collaborative sensing tasks the sensor nodes must

estimate their position by means of a distributed positioning algorithm. The communication architecture of underwater acoustic sensor networks constitute of sensor nodes that are anchored to the bottom of the ocean as shown in Fig. 1. Underwater sensor nodes are interconnected to one another by means of wireless acoustic links. They can relay data from the ocean bottom network to a surface station.

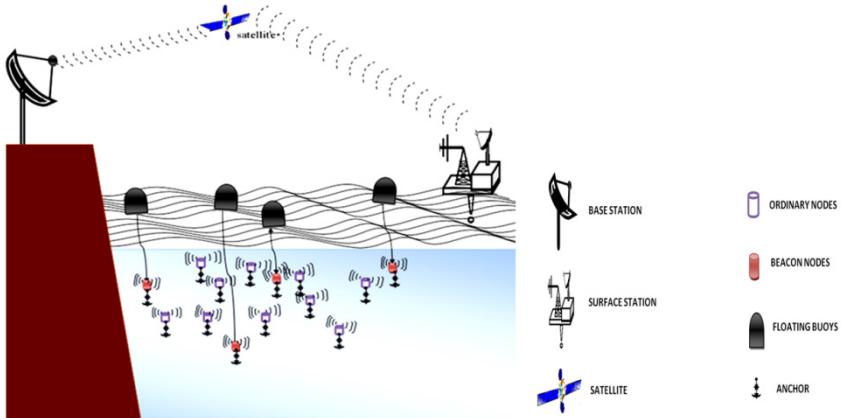


Fig. 1. System Architecture for Underwater Sensor Network

Uw-sensors are equipped with two acoustic transceivers, namely vertical and horizontal transceiver. The horizontal transceiver is used to send commands and configuration data to the other sensors and the vertical link is used to relay data to a surface station. Two types of sensor nodes are deployed (i). Anchor nodes or Beacons, these are the nodes which know its location. (ii). Ordinary nodes whose locations are yet to be decided using the proposed trilateration method.

2.2 Deployment

Various deployment techniques are used to deploy the anchor nodes underwater [2][3][4][5]. The attributes of a sensor node are Node ID, Network ID, Beacon flag, list of reference nodes, x and y position and depth at which node is place. Node ID is a unique number which identifies a node. Network ID tells to which network this node belongs. Beacon Flag stores the status of a node i.e whether it is a anchor or a ordinary node. If beacon flag is set to '1' it implies that the node is a beacon/anchor node and knows its location. If beacon flag is set to '0' the node is a ordinary node whose location is yet to be found out. Before deployment all the ordinary nodes are set to '0' and anchor nodes are set to '1'. List of Reference Nodes RN consists of all the reference nodes in its communication range and their distance from itself.

2.3 Distance Estimation

In order to find the position of sensor node, minimum of three RN are needed. The distance between these RN are used to calculate the exact position of unknown node.

Distance can be measured by various range based distance measuring techniques as discussed in [6]. Some of the techniques are received signal strengths (RSS), time of arrival (TOA) or time difference of arrival (TDOA) and Angle of Arrival AOA. Amongst the above mentioned distance estimation techniques TOA is most suitable for underwater scenario [7]. ToA algorithm can be used in underwater environments measuring arrival time by using acoustic signal only. Hence we are going to use TOA to calculate distance in our algorithm. For TOA-based systems, the one-way propagation time is measured, and the distance between measuring unit and signal transmitter is calculated. This algorithm estimates the distance between nodes by measuring the propagation time of a signal. So it requires precise time synchronization between two nodes. In this case the distance between two nodes is directly proportional to the time the signal takes to propagate from one point to another. If signal is sent at time t_1 and reached the receiver node at time t_2 , the distance between two nodes can be defined as in equation (1). Where S_r is the propagation speed of acoustic signal (1500 m/s). From this method we get the list of all possible RN's in the communication range.

$$d = S_r (t_1 - t_2) \quad (1)$$

3 Error Introduced in Distance Estimation

Lateration is the process of determining absolute or relative locations of points by measurement of distances, using the geometry of circles, spheres or triangles. Lateration computes the position of an object by measuring its distance from multiple reference positions. Calculating a nodes position in two dimensions requires distance measurements from atleast 3 RN's as shown in Fig. 2. Radius 1, 2 and 3 are determined by using TOA method as explained in the previous section [7]. Lo-calization algorithm will work very fine when there are no Distance measurement errors. Distance measurement errors are errors in the distance estimates between the non localized node and references. Following are 2 cases that can be encountered during position estimation.

Case 1: For ideal condition, without errors in distance calculated.

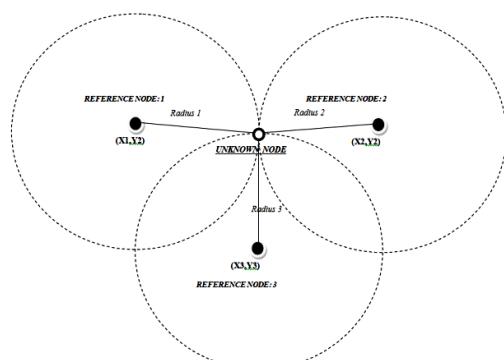


Fig. 2. Determining position using lateration using 3 reference positions

In Fig. 2, we consider that there are no distance estimation errors. In such a case the position of the unknown node can be directly found out as the intersection of circles drawn from each reference by taking the distance measured by TOA as the radius.

Case 2: For conditions, when errors are introduced in distance calculated.

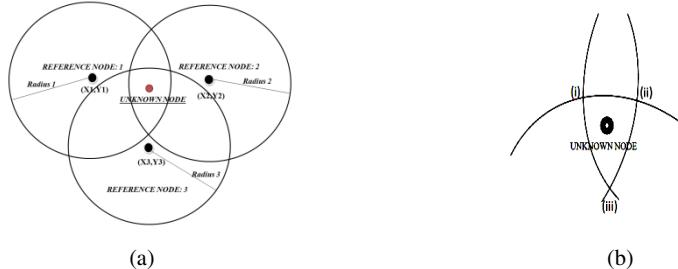


Fig. 3. Determining position using lateration using 3 reference positions

In this case while calculating the position of the unknown node, there will not be a common intersection point amongst the three circles, as shown in Fig. 3(a). In fact the task will be now to calculate the exact location shown in the region enclosed by the three circles. Fig. 3(b) [9]. In order to overcome with such situation, Zenon et al [11], proposed a solution to compute the approximate position of the unknown node.

3.1 Effect of Temperature, Salinity and Depth on Sound Speed

Sound speed depends on temperature, salinity and depth. Hence distance calculated is directly influenced by speed of sound, as in equation(1)[8].

Temperature: Sea water temperature decreases from the surface to the seabed. The time and space variability is maximal in the shallower layers, but decreases with depth. Beyond a typical depth(around 1000m in open oceans, but shallower in closed seas: e.g., 100-200m in Mediterranean), the average temperature remains stable, decreasing very slowly with depth and varying very little from one place to the other.

Depth: Hydrostatic pressure makes the sound velocity increase with depth. This increase is linear as a first approximately of around 0.017m/s per meter down.

Salinity: Sea water is made of a mix of pure water and dissolved salts. The salts mass percentage defines salinity. Salinity in the large ocean basins is on average 35 p.s.u. Fig. 4 shows that temperature and salinity profiles have both spatial and temporal variation [12].

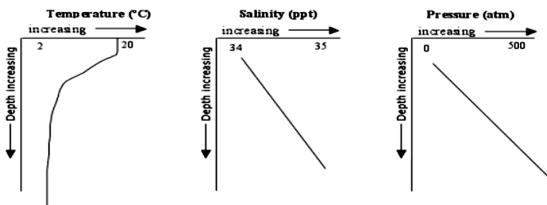


Fig. 4. Depth profiles from the open ocean of temperature, salinity and depth

3.2 Calculating Speed of Sound

We will be using a new equation proposed by Leroy et al[10] for the calculation of sound speed in seawater which is a function of temperature, salinity, depth, and latitude in all oceans and open seas. Equation (2) shows the relation between underwater parameters and the speed. Where T : Temperature, S : Salinity, Z : Depth and ϕ : Latitude.

$$\text{Speed of sound } 'c' = 1402.5 + 5T - 5.44 \times 10^{-2} T^2 + 2.1 \times 10^{-4} T^3 + 1.33S - 1.23 \times 10^{-2} ST + 8.7 \times 10^{-5} ST^2 + 1.56 \times 10^{-2} Z + 2.55 \times 10^{-7} Z^2 - 7.3 \times 10^{-12} Z^3 + 1.2 \times 10^{-6} Z(\phi-45) - 9.5 \times 10^{-13} TZ^3 + 3 \times 10^{-7} T^2 Z + 1.43 \times 10^{-5} SZ \quad (2)$$

3.3 Our Proposed Equation for Estimating Errors

When we calculate the distance between the RN and the Ordinary Node ON, we assume that the speed of acoustic signal will be 1500m/s. As in equation (3)

$$d_1 = S_1 * t \quad (3)$$

But as we have seen earlier that the speed always varies with the change in oceanographic parameters like temperature, Salinity and depth at which the node is placed. The new distance with errors can now be calculated as in equation (4).

$$d_2 = S_2 * t \quad (4)$$

Where, S_2 is the speed of sound calculated for varying temperature, salinity and Depth values. t will remain same as the time taken will be calculated using TOA method. When network consist of N sensors then any nodes constellation can be fully described as N by N matrix. Elements of this matrix d_{ij} equal to distance between neighbor nodes i and j ($i, j = 1 \dots N$), -1 if nodes i and j are too far to communicate and 0 if $j=i$. To find node i position it is necessary to know at least three $d_{ij} > 0$ elements where $j = 1 \dots N$ with $j \neq i$. Communication range R has to be greater than the distance d_{ij} . For simulation purpose we propose an equation (5) to calculate the erroneous distance.

$$d_{ij}^* = \begin{cases} d_{ij} ((d_2/d_1) - 1) & \text{when } d_{ij} \leq R \\ -1 & \text{when } d_{ij} > R \end{cases} \quad (5)$$

4 Our Proposed Algorithm for Position Estimation

Before we start the localization algorithm it is necessary to select appropriate RN's. All the three selected RN's should adhere to the following 2 conditions: no two RN's should have the same position and the three nodes should not lie on a same line.

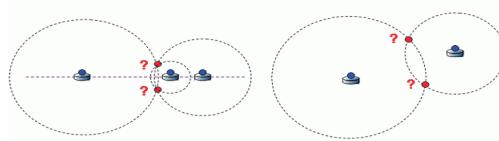


Fig. 5. Exceptions in position computation step.

Fig. 5 depicts the two exceptions. In these two cases, it is not possible to compute unknown node. To avoid such a situation two parameters ϵ and γ are introduced. ϵ is the minimum distance between each RN's that can be considered while selecting the reference. γ is the minimum angle of a triangle formed between the three reference. Introduction of these parameters also help in the refinement of the positions calculated.

Algorithm: By using this distance as radius and each RN's location as a centre we get a diagram as in Fig. 3(a). From this situation we first calculate six intersection points using, from overlapping of three pair of circles. We propose to use the Clever Area-Based method to find points of intersection between two circles. The proposed localization technique is depicted in Fig. 6.

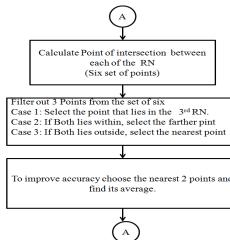


Fig. 6. Determining position using lateration using 3 reference positions

5 Performance Analysis

5.1 Simulation Scenario

The proposed localization scheme was simulated in C. In our simulation 250 nodes were deployed over an area of 1000 x 1000 meters. The communication range of every sensor node 'R' was taken between 250 and 500 meters. The initial RN were placed at (0,0),(0,500) and (500,0) respectively. The erroneous distance calculation is as shown in equation (5). All transmitters and receivers in the system have to be precisely synchronized. Nodes with beacon flag as '1' can act as RN's. Considering a two dimensional Architecture, all nodes are placed at a same depth in the ocean. The simulations were run for 50 simulations to take average results for different underwater parameters.

5.2 Results

Fig. 7, shows the result for actual position and localized position of sensor nodes using our proposed technique. These results were computed while keeping $T=2, S=30, Z=4000, \phi=60, \varepsilon=50, \gamma=10$ and $R=500$.

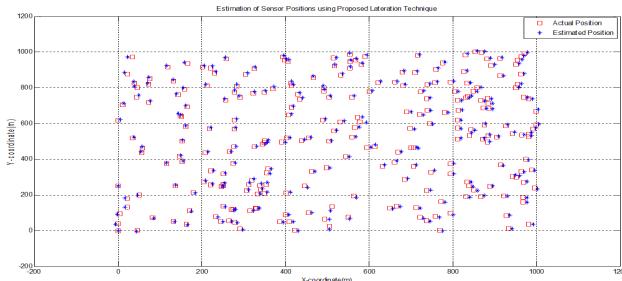


Fig. 7. Simulation results for proposed technique

Table 1 shows the average error calculated for different values of Temperature, Salinity and latitudes. From Fig. 8 we can infer that temperature is a dominating factor which influences the average error during localization. In this result S was set to 35, Altitude to 60 , ε to 50 and γ to 10. The depth was considered to be 4000mtr. The results of our simulation shows that the positioning algorithm performs better for different sound speed in UWSN. Since no such results are available in literature for the sake of comparison, to the best of our knowledge we are the first to propose such a technique for UWSN scenario.

Table 1. Average error estimation for underwater parameters at depth of 4000m.

TEMPERATURE($T^{\circ}\text{C}$)	2	5	8	11	20	23
SALINITY (S %)						
LATITUDE(ϕ°)						
$\phi=0, S=30$	9.767	11.991	16.096	17.020	26.272	27.602
$\phi=0, S=35$	9.506	12.248	16.263	19.644	27.19	31.145
$\phi=0, S=38$	11.718	13.564	18.184	20.082	28.262	31.459
$\phi=60, S=30$	9.544	13.482	14.209	16.885	27.133	27.236
$\phi=60, S=35$	9.062	12.200	16.852	21.014	27.473	32.038
$\phi=60, S=38$	10.165	14.632	17.423	19.602	28.421	30.885

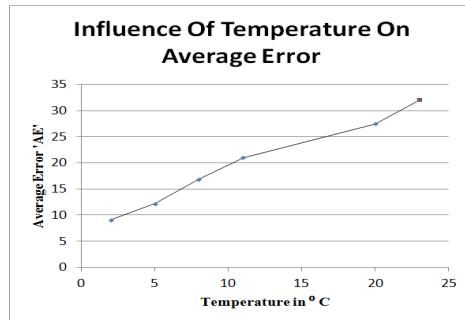


Fig. 8. Simulation results for proposed technique

5 Conclusion

We have proposed and simulated a localization technique which can accommodate the errors encountered during propagation of acoustic waves between the deployed underwater sensor nodes. Our proposed self-organizing algorithm for position estimation was simulated and results were documented. The simulation was carried out for varying ocean parameters such as salinity, temperature, depth and latitude found at different layers in the ocean. Our proposed technique is efficient and calculates Under Water Sensor locations with minimum localization errors in harsh underwater environments. To the best of our knowledge we are the first to propose such a localization technique for underwater sensor network scenario.

References

1. Akyildiz, I.F., Pompili, D., Melodia, T.: Underwater acoustic sensor networks research challenges. *Science Direct* (2005)
2. Erol, M., Filipe, L.: AUV-Aided Localization for Underwater Sensor Networks. In: *WASA* (2007)
3. Erol, M., Luiz, F.M., Geria, M.: Localization with Dive’N’Rise (DNR) Beacons for Underwater Acoustic Sensor Networks. In: *WUWNet 2007* (2007)
4. Langendoen, K., Reijers, N.: Distributed localization in wireless sensor networks: a quantitative comparison. *Computer Networks* 43, 499–518 (2003)
5. Waldmeyer, M., Tan, H.-P., Winston, K.G.: Multi-stage AUV-aided Localization for Underwater Wireless Sensor Networks (2011)
6. Liu, H.: Survey of Wireless Indoor Positioning techniques and systems. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews* 37(6) (2007)
7. Lee, K.H., Yu, C.H., Choi, J.W., Seo, Y.B.: ToA based Sensor Localization in Underwater Wireless Sensor Networks. In: *SICE Annual Conference* (2008)
8. Lurton, X.: *An Introduction to Underwater Acoustics, Principles and Applications*. Springer Publication (2010)
9. Yu, C.H., Lee, K.H., Moon, H.P., Choi, J.W., Seo, Y.B.: Sensor Localization Algorithms in Underwater Wireless Sensor Networks. In: *ICROS-SICE International Joint Conference* (2009)

10. Leroy, C.C., Stephen, P.: A new equation for the accurate calculation of sound speed in all Oceans. *Journal of the Acoustical Society of America*, 2774–2782 (2008)
11. Chaczko, Z., Klempous, R., Nikodem, J., Nikodem, M.: Methods of Sensors Localization in Wireless Sensor Networks. In: 2007 Proceedings of the 14th Annual IEEE Conference (2007)
12. Ali, M.M., Jain, S., Radhika, R.: Effect of Temperature and Salinity on Sound Speed in the Central Arabian Sea. *The open Ocean Engineering Journal* 4, 71–76 (2011)

An Analytical Model for Power Control B-MAC Protocol in WSN

V. Ramchand and D.K. Lobiyal

School of Computer and Systems Sciences,
Jawaharlal Nehru University, New Delhi, India
rchand.jnu@gmail.com,
lobiyal@gmail.com

Abstract. This paper presents an analytical model for estimating throughput and energy consumption in B-MAC protocol for Wireless Sensor Networks. The design includes transmission power control and multi-hop transmission of frames through adjusted transmitted power level. Proposed model reduces collision with contention level notification. The proposed model has been simulated using MATLAB. The simulations reveal better results for throughput and energy consumption of the proposed model as compared to B-MAC protocol.

Keywords: Wireless Sensor Networks, Possion distribution and Energy efficiency.

1 Introduction

Wireless communication is the most exciting area in the field of communication research. Over five decades it has been a major topic of research. Wireless Sensor Networks (WSNs) are an emerging technology that has become one of the fastest growing areas in the communication industry. They consist of sensor nodes that use low power consumption which are powered by small replaceable batteries that collect real world data, process it, and transmit the data to their destination nodes or a sink node or a server. WSN based applications usually have comfortable bandwidth requirement, the demand for using this medium is increasing with wide range of deployment for monitoring and surveillance systems as well as for military, Internet and scientific purposes. Wireless sensor networks will play an important role in future generation for multimedia applications such as video surveillance systems.

Transmission power control is provoked from potential benefits. The benefit is a more efficient use of the network resources. Allow a large number of simultaneous transmissions, power control increases the whole network capacity. Secondly energy saving is achieved by minimizing the average transmission power. The transmission power level is directly related to the power consumption of the wireless network interface. The lifetime of node's battery is becoming an important issue to the manufacturers and consumers, as devices are being used more frequently for transmission of signals\data packets\frames. It is becoming great interest to control the transmission power level of every node so that the lifetime of the wireless sensor network will be maximized.

Multiple access-based collision avoidance MAC protocols have made that a sender-receiver pair should first ensure exclusive access to the channel in the sender and receiver neighborhood before initiating a data packet transmission. Acquiring the floor allows the sender-receiver pair to avoid collisions due to hidden and exposed stations in shared channel wireless networks. The protocol mechanism used to achieve such collision avoidance typically involves preceding a data packet transmission with the exchange of a RTS/CTS (request-to-send/clear-to-send) control packet handshake between the sender and receiver. This handshake allows any station that either hears a control packet or senses a busy carrier to avoid a collision by deferring its own transmissions while the ongoing data transmission is in progress.

2 Related Work

MAC layer has a vast impact on the energy consumption of sensor nodes. Communication is a major source of power consumption and the MAC layer design manages the transmission and reception of data over the wireless medium using the radio. The MAC layer is responsible for access to the shared medium. MAC protocols assist nodes in deciding when to access the channel. Pattern-MAC (P-MAC) for sensor networks adaptively determines the sleep-wake up schedules for a node based on its own traffic, and the traffic patterns of its neighbors. This protocol achieves a better throughput at high loads, and conserves more energy at light loads. In P-MAC, the sleep-wake up times of the sensor nodes is adaptively determined. The schedules are decided based on a node's own traffic and that of its neighbors. The improved performance of P-MAC suggests that 'pattern exchange' is a promising framework for improving the energy efficiency of the MAC protocols used in sensor networks [5].

S-MAC protocol achieves energy conserving through three basic techniques [9]. Nodes sleep periodically instead of listening continuously to an idle channel. Transceivers are turned off for the time the shared medium is used for transmission by other nodes overhearing is avoided, and a message passing scheme is used with the help of store-and-forward technique based on the buffer capacity. Each of the nodes has a fixed duty cycle. It can be used to tradeoff bandwidth and latency for energy conserving, but it does not allow adapting to network traffic. However, S-MAC protocol allows transmitting large messages by fragmenting them. It mitigates problems with higher delays and requires large storage buffers.

3 Proposed Analytical of B-MAC

3.1 Analytical Model for Power Consumption

The network function begins when a node senses an event and starts transmitting the sensed event in the form of message, data, frame or packet etc. The task of a node is to sense for events, transmit \ receive the data with other nodes, forward the data to a head node or sink node when ever required until the battery power drains. On a given time, either a node or few nodes may transmit out of N number of nodes deployed in the field.

The probability of transmitting nodes varies over time. Nodes active time, sleep time idle time as followed as per analytical model for power control T-MAC protocol [3]. At the initial stage of the network all the nodes are equipped with equal energy. Therefore, more number of nodes may employ themselves in sensing the events. Those nodes which have sensed some events will involved in forwarding of events as frames or signal, it leads to increase in high contention level among neighboring nodes. Any node before initiating a transmission estimates contention level to avoid collision with others

$$C_L = (A_N - \sum_{i=1}^n \pi r^2) * n_t / N_A \quad (1)$$

C_L is the contention level, A_N is any node that measures the current contention level among neighbors, $\sum_{i=1}^n$ are the nodes which are in contention to communicate, πr^2 is the circular area where all the contending nodes reside, n_t is the number of nodes contending at time t and N_A is total number of nodes in a given area. A node that wins the contention starts transmitting. Once the transmission is over the node goes to listen state. Further, nodes which are in the backoff mode wake up once the timers expire.

3.1.1 Contention Notification

Contention Notification (CN) messages alert the neighbor nodes not to act as hidden terminals when contention is high. Every node makes a local decision to send a CN message based on its local estimates of the contention level. Estimating contention level is either by receiving acknowledgment from the one-hop receiver or by measuring the carrier to noise plus interference ratio between the source and destination. Other way of estimating contention is by measuring the noise level of the channel. Any node in the network that has a frame to transmit senses the channel with the Clear Channel Assessment algorithm before initiating the transmission. When the noise level of the channel is higher than CCA threshold, the node takes random backoff. A node starts transmitting only when the noise level of the channel is smaller than CCA threshold. Noise level of a channel is measured by carrier to noise density ratio (CNDR),

$$CNDR = (E_f/n) * (N/C_A) * (R/B) \quad (2)$$

where E_f is the energy consumption in one frame transmission, n is the noise level of current frame transmission, N/C_A is number of nodes in given coverage area, R is the rate at which a frame is transmitted and B is the channel bandwidth.

3.1.2 Power Estimation of Nodes

Source node transmit a frame to a destination node,

$$P_r = P_t (\lambda / 4\pi d)^n \quad (3)$$

Given P_t is the transmit power of source and P_r is the power when the frame reaches the receiver. λ is the average arrival rate, and d is the distance from source to destination. n is the noise level of the channel. Source node sends RTS to destination node with the power level P_t , and the destination node receives the RTS package with the power level P_r then,

$$P_r = P_{\text{frame}} (\lambda / 4\pi d)^n$$

$$\text{where } P_{\text{frame}} = (P_t * R_s) / P_r$$

where R_s is the sensitivity of received signal, sensitivity in a receiver is normally defined as signal s produce noise ratio at the transceiver / receiver node.

$$S_i = K (N_s + N_r) * B * \text{SNR}$$

S_i is the Sensitivity, K is the Boltzman constant, N_s equivalent noise at source node, N_r equivalent noise at receiver node, B is bandwidth, SNR signal to noise ratio

3.1.3 Adjusted Transmission Power

Let P_{\max} is the max transmission power of nodes, and P is the current transmission power of node. E_{\max} is the maximum energy level of nodes at the initial stage of the network operation begins, with P_{res} is the residual power of a node. Optimal degree of a node is N , current degree of node is n . To adjust the transmission power according to suitable degree, the adjusted transmission power P_{adj} is shown below,

$$P_{\text{adj}} = P + [N - n/N] * P$$

Further transmission power of node will be based on its residual power. Therefore, power available P_{AVAIL}

$$P_{\text{AVAIL}} = (P_{\text{res}} / E_{\max}) * P_{\max}$$

According to the residual power, improved adjusted power IP_{adj} is given as,

$$IP_{\text{adj}} = P_{\text{adj}} + [(P_{\text{res}} / E_{\max}) * P_{\max}] - P \quad (4)$$

3.1.4 Probabilities of Control Packet and Frame Exchange

When a node acquires frames for transmission at the rate λ_r and arrival follows Poisson distribution. On a frame to transmit it initiates the transmission with control packet exchange. To transmit an RTS packet a node takes μ_R time, therefore the average arrival rate of receiving an RTS packet $\lambda = 1/\mu_R$. Therefore, probability of transmitting an RTS packet is calculated as

$$P(X=1) = e^{-(1/\mu_R)} * (1/\mu_R)^1 / 1! \quad (5)$$

On arrival of RTS packet the receiver calculate the power level of the transmitted control packet and reply the CTS packet with required power level. The probability of replying the CTS packet is represented as

$$P(X=1) = e^{-(1/\mu_C)} * (1/\mu_C)^1 / 1! \quad (6)$$

The node on receiving the reply it forwards the acquired frames to its one hop neighbor node. The frames are forwarded at the rate λ_r . Therefore, the power consumption of transmitting node is

$$P_t = p_r * \lambda r + (e^{-(1/f)} * (1/f)^1) / 1! \quad (7)$$

where $(e^{-(1/f)} * (1/f)^1) / 1!$ is the probability of receiving a frame and p_r is the power consumed for receiving a frame.

On arrival of frame the receiver calculate the power level of the transmitted frame, it reply the ACK with required power level. The probability of replying the ACK is represented as

$$P(X=1) = e^{-(1/\lambda_{ACK}) * (1/\lambda_{ACK})^1} / 1! \quad (8)$$

On arrival of ACK from the receiver the transmitting node send the remaining frames with adjusted transmit power. Power consumption of a node transmitting frames with the adjusted power is given by

$$P_{t(IP_{adj})} = p_r \lambda_r + (e^{-(1/f) * (1/f)^n}) / n! \quad (9)$$

where $(e^{-(1/f) * (1/f)^n}) / n!$ is the probability of receiving n number of frames.

4 Simulation Results

The proposed model for estimating energy consumption in B-MAC protocol is implemented in mat lab. Proposed methodology has been designed with fewer numbers of mobile nodes which are placed equidistance to each other. All the nodes are mobile by nature they relocated their positions subject to requirements of the tasks of their own or with the request of nodes with one hop communication range. The proposed protocol has been tested for unicast and broadcast communication with multi-hop transmission of frames. For unicast communication control packets like RTS-CTS are used. Adjusted transmission power level is used for broadcast communication. After broadcasting a frame, all nodes in the coverage area should refrain themselves from transmitting until one frame time has elapsed to allow transmitting the other node initiating a transmission is more efficient than control packet exchange for broadcast traffic. Frame length is varied as per the requirement of the application. The Simulation parameters used in the work are listed in the tables below.

Table 1a. Simulation parameter

Parameter	Value
Number of nodes	20
Contention window per slot duration	400 μ s
Communication bandwidth	15 Kbps
Transmission Range	2 meters
Transmitting and Receiving antenna gain	Gt=1, Gr=1
Transmission power	0.031622777W
Carrier Sense Power	5.011872e- 12W
Received Power Threshold	5.82587e-09W
Traffic type	VBR
Initial Energy	500 Joule

Table 1b. Frame parameters

Length (Bytes)	
Preamble	8
Synchronization	2
Header	5
Footer (CRC)	2
Frame length	Variable

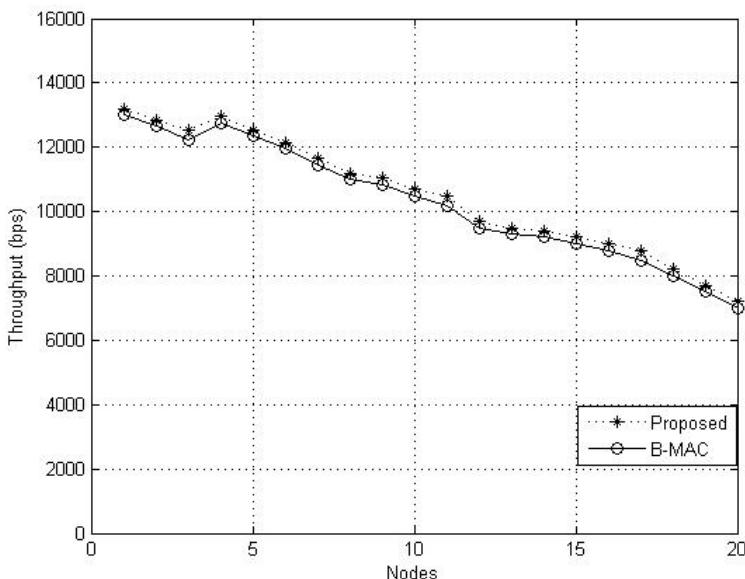
**Fig. 1.** Throughput

Fig 1 shows the throughput of existing B-MAC protocol and the proposed model. Under low transmission rate unicast messages are exchanged with the use of control packet transmissions. Adjusted transmission power level is used for broadcasting frames. Proposed model deliver frames and achieves marginally better throughput for multi-hop transmission of frames than the existing work.

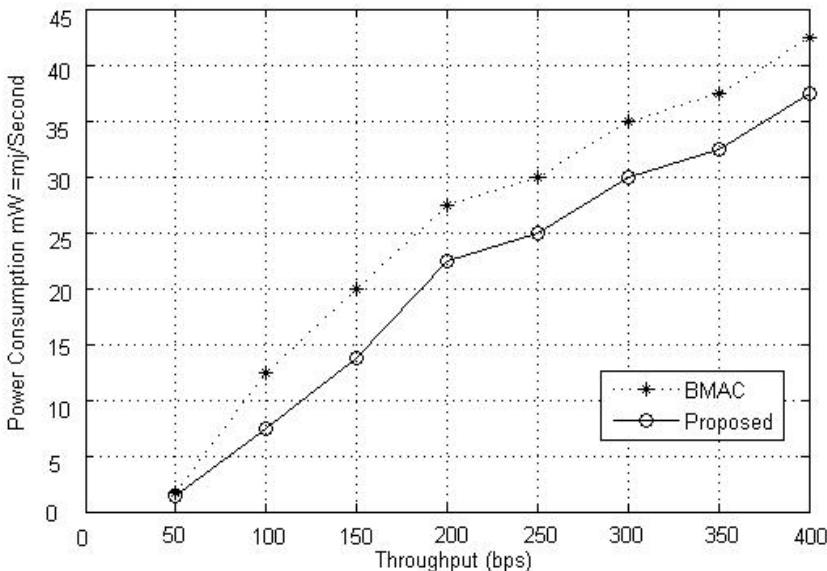


Fig. 2. Power consumption

The above figure shows the comparison of energy consumption between B-MAC and proposed model. Energy efficiency is measured based on unicast, broadcast and multi-hop transmission of frames. While measuring the efficiency, sending rates are varied. The above figure presents the energy consumption of nodes involved in different duty cycles. As we observe in the multi-hop throughput, under low data rates, existing MAC has slightly lower throughput. The figure also shows the impact of the bits transmitted per second and power consumption of nodes in milli-watts. It is quite clear from the figure that the proposed work out performs the existing model in energy consumption.

5 Conclusion and Future Work

In this work, we have proposed analytical model for estimating throughput of multi-hops and energy consumption in B-MAC protocol for Wireless Sensor Networks. The power consumption for an individual node is calculated for multi-hop communication. A node in the network saves its energy by changing its mode periodically. The proposed protocol shows better results than B-MAC protocol in terms of energy consumption. While designing the methodology for B-MAC, utilization variation in synchronization errors and transmission fairness and border nodes going away from the transmission are not focused. This can be explored in the future work as an extension of the current work.

References

1. Goldsmith, A.: *Wireless Communication*. Cambridge University press (2007)
2. Siva Ram Murthy, C., Manoj, B.S.: *Ad hoc Wireless Sensor Networks* Department of Computer Science & Engineering, IIT Madras. Pearson Education(2005)
3. Ramchand, V., Lobiyal, D.K.: An Analytic model for Power Control T-MAC protocol. *International Journal of Computer Applications(0975-8887) 12(1)* (December 2010)
4. Hamid, Y.M.A., Lobiyal, D.K.: IPCM/COMPOW: An Efficient Power Saving Scheme for Multi-Hop Wireless Ad Hoc Networks. In: ICWN 2008, pp. 452–458 (2008)
5. Kim, Y., Shin, H., Cha, H.: Y-MAC An Energy Efficient Multi-channel MAC Protocol for Dense Wireless Sensor Networks. In: International Conference of Information Processing in Sensor Networks (2008)
6. Rhee, I., Warrier, A., Aia, M., Min, J., Sichitiu, M.L.: Z-MAC: a hybrid MAC for wireless sensor networks. *IEEE/ACM Transactions on Networking* 16(3), 511–524 (2008)
7. Rajendran, V., Obraczka, K., Garcia-Luna-Aceves, J.J.: Energy-efficient, collision-free medium access control for wireless sensor networks. *Wireless Networks* 12, 63–78 (2006)
8. Ji, P., Wu, C., Zhang, Y., Jia, Z.: Research of an energy-aware MAC protocol in wireless sensor network. In: Control and Decision Conference, pp. 4686–4690 (2008) (in Chinese)
9. Ye, W., Heidehann, J., Estrin, D.: An Energy Efficient MAC protocol for Wireless Sensor Networks. In: INFOCOM 2002, IEEE Computer and Communication Societies Proceedings, vol. 3 (2002)

Enterprise Mobility – A Future Transformation Strategy for Organizations

Jitendra Maan

Tata Consultancy Services, TCS Towers, 249 D & E,
Udyog Vihar Phase IV, Gurgaon, Haryana, India – 122001
Jitendra.maan@tcs.com

Abstract. In the changing business environment, where enterprise users are expected to handle critical tasks and decision-making in real-time, it has become a business imperative to stay competitive and agile by adopting mobility to handle business needs. In a typical enterprise, the entire business value chain is geographically fragmented which drives the need to mobile-enable the existing enterprise applications.

This paper highlights the key customer pain areas in Enterprise Mobility adoption across the enterprises. It, not only addresses the Critical Success Factors for enterprise mobility enablement but also outlines the business needs to rapidly create enterprise mobile solutions across all lines of businesses.

The paper enlightens the value impact of enterprise mobility on workplace, organizations and technology and discusses the critical growth factors, key enablers and transformational strategies along with enterprise mobile application components.

Moreover, the paper outlines key design considerations, recommendations and enterprise mobility value proposition to potential organizations in terms of tangible and financially oriented results.

Keywords: Mobile computing, Enterprise Mobile Applications, Mobile Internet, Enterprise Mobility, Mobile Solutions, Rich Mobile Applications, Mobile web experience, Mobility Services.

1 Introduction

Today, enterprises are faced with increased global competition in an environment where customers are demanding faster delivery, better service and also want to gain significant and immediate business value by increasing productivity and reducing operational cost. Most enterprises today face a common set of challenges when it comes to integrating their mobile workers into their enterprise processes. Adopting Mobility to the changing business needs is a sustained challenge.

Enterprise Mobility is the next logical transition in mobile technology evolution which will continue to gain more prominence in the enterprises not just to improve the return on investment, but also to expand global reach and improve operational efficiency of the mobile workforce.

In such a highly competitive environment, Mobile applications not only offers a compelling total cost of ownership (TCO) advantage but it will also continue to gain more traction among the enterprises to create rich mobile applications as they will be easier to maintain and enhance as the mobile device market continues to rapidly change with continuous innovation in mobile technologies, platforms and devices.

2 Enterprise Mobility – Customer Pain Areas

There is an ongoing challenge in front of CTOs/CIOs/IT Heads to develop an understanding what to mobilize, when to mobilize, as well as how to develop and execute mobility strategy in the context of their overall business eco-system. Mobile paradigm is different from the normal client-server based application development. The bottom line is that mobile application development is new to most organizations and comes with unique challenges.

The key customer challenges/pain areas include the following:

- Real Time access to critical information is not available when needed
- Increasing customer expectation of immediate response to problem resolution
- Lack of adoption of globally available mobility solutions
- Re-use/utilize existing web-oriented infrastructure
- Scalable and flexible Mobility solutions with changing business needs
- Lack of standardization in Mobile offerings from Mobile platform vendors.
- Fragmented mobile platforms and diversity issues. Eg Multiplicity of –
 - Connected devices,
 - Operating Systems,
 - Browsers,
 - Form factors,
 - Input Methods.

3 Critical Enablers of Enterprise Mobility

Mobile Applications are one of the key enablers of Enterprise Mobility. Business users are expected to handle critical tasks and decision making in real-time, no matter where they are. Driven by explosive growth in smartphone and tablet sales, mobility deployment in enterprises is going to take-off in a big way. Consumerization of IT drives enterprises to enable consumer mobile devices and applications in the workplace. Today, enterprises are much more concerned about the security of sensitive and critical corporate data. The success of enterprise mobility is largely dependent on providing adequate security levels, given the inherent vulnerability of mobile devices and the data that resides on them.

Most companies, regardless of their size and location, today face a common set of challenges when it comes to integrating their mobile workers into the enterprise's processes. They have invested a lot in implementing ERP and CRM systems. Most companies will adopt Mobile technologies gradually, without ripping and replacing their core enterprise resource planning (ERP) Platforms.

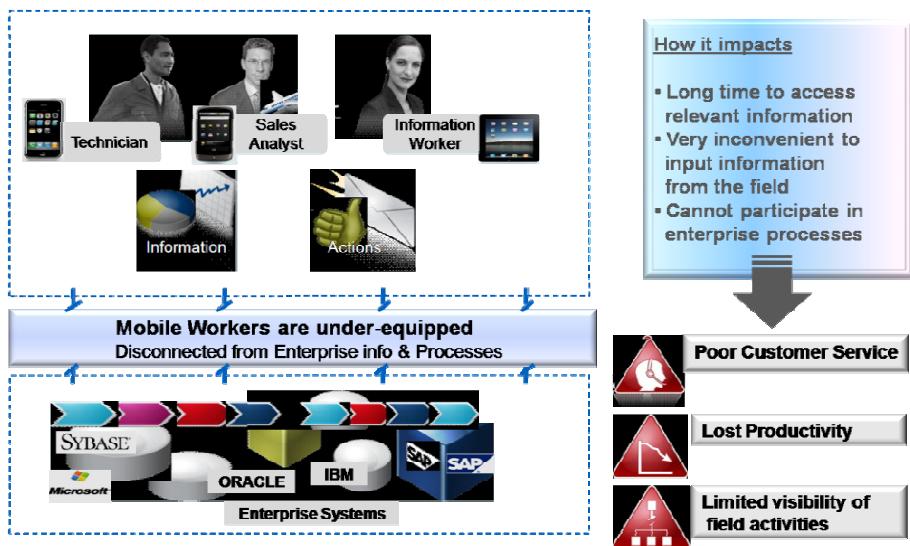


Fig. 1. Impact of Enterprise Mobility across organizations

However, these applications are rendered useless as soon as a mobile worker steps out of the office, because much of the activity of customer-facing or mobile professionals is dependent on timely and accurate access to the enterprise information and processes. For example, CRM systems have been implemented to support the activities of sales and services people, but because these people are mostly mobile, the value of the CRM implementations is dramatically reduced without the necessary mobility support.

If the mobile workers are equipped with information on their mobile devices, customers can reap a lot of benefits in terms of costs and opportunity. It becomes the exponential benefits for the entire organization in the terms of,

- Enhance Productivity
- Flexibility And Speed
- Optimize Field Resources

3.1 Enterprise Mobility – Critical Success Factors

The critical success factors for Enterprise Mobility are as follows -

- Well articulated enterprise mobility strategy
- Well defined enterprise mobility architecture
- Tablet-enabling only applications with clear mobile use case and ROI
- Focus on User Experience Design
- Automated testing and analytics
- Well managed resourcing plan
- Direct or indirect device vendor partnerships
- Development Partners who can provide Comprehensive mobility support

3.2 Enterprise Mobility – Mobile Web Paradigm

Mobility is one of the top three most disruptive technologies making its way into the enterprises. It is disruptive because it has the potential to radically change how business is done.

Adoption of Mobile Web applications continue to grow with increased penetration of emerging mobile platforms, devices (like larger form factor Smart-phones, tablets) with continuous richness of micro browsers. Using mobile web technology, organizations are getting more out of their existing web-oriented systems by delivering applications and web content directly to mobile users.

4 Enterprise Mobility – Key Business imperatives

A few key business imperatives are given below –

- Evolving the customer experience through mobile self service applications
- Need for making the work force more productive and efficient. There are numerous examples where mobility enables introduction of a connected computer in places in business processes where none exists today such as paper-and-pen based processes. This increases the velocity and accuracy of business processes. A few of them are as follows
 - Approvals for various business workflows across industries
 - Enabling Sales force and Field service technicians across industries
 - Operations and Plant management in Manufacturing
 - Inventory Management across industries
- There is a dire need for businesses to be more agile by having business critical alerts and information at their fingertips enabling users to respond to business critical events to a whole new level by making decisions on the go. Tablets can enable this by performing on the spot “what-if” analysis on receiving push-notifications

5 Mobile Application Models

Majority of Mobile Applications contain some level of common functionality that spans across layers and tiers. It is customary to examine the mobility functions required in each layer, and then abstract the functionality into common components to address cross-cutting concerns. And, such common components can be configured depending on the specific requirements of each layer of the mobile application.

5.1 Native v/s Mobile Web v/s Hybrid Applications

It is imperative for enterprises to have a strong mobility platform that would address strong enterprise mobility business cases. It's critical to evangelize the right mobility strategy before selecting a mobile platform which would help in correctly examining

each of the layers in mobile platforms and how do they fit into an organization's IT architecture landscape.

The figure below shows different Mobile Application Models – Native Mobile Applications, Mobile Web Application and Hybrid Mobile Applications. Such Model helps architects and developers to take right design decisions for cross platform strategy.

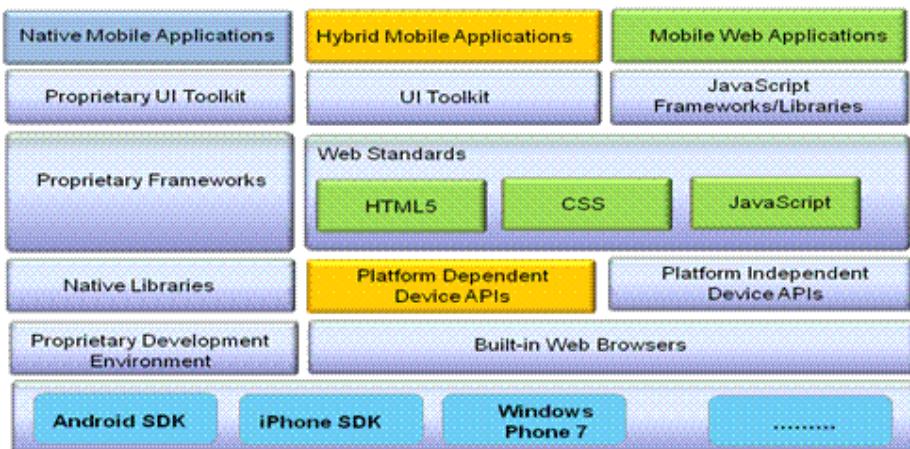


Fig. 2. Mobile Application Models- Mobile Web v/s Native v/s Hybrid Apps

5.2 Enterprise Mobile Application Components

Any enterprise mobile applications, be it a thin client or thick client or be it an application for enterprise user or an end customer, there is a need to connect to the enterprise network. Figure 3 depicts the various blocks involved in hosting and delivering an enterprise mobile application. The recommended mode of data capture from an Enterprise is web services. Information required for the application from the enterprise is made available through the web services through the Enterprise Firewall. In this process Mobile Middleware plays a vital role.

It is important to utilise and build on the established principles and best practices. These are consolidated and are available as a reference for enterprise mobile architecture engagements.

Mobile middle-ware architecture supports various services and components. A few salient features of middle-ware services as depicted below -

- **Data Access Manager Service** – Acts as a service layer to fetch data from backend systems/databases.
- **Authentication Service** – Authenticate user credentials on the server
- **User Management Service** – Authenticates the validity of user/group
- **Composition Service** – Helps to compose the UI components.
- **Synchronization Service**– Synchronization of data between the Server and Client

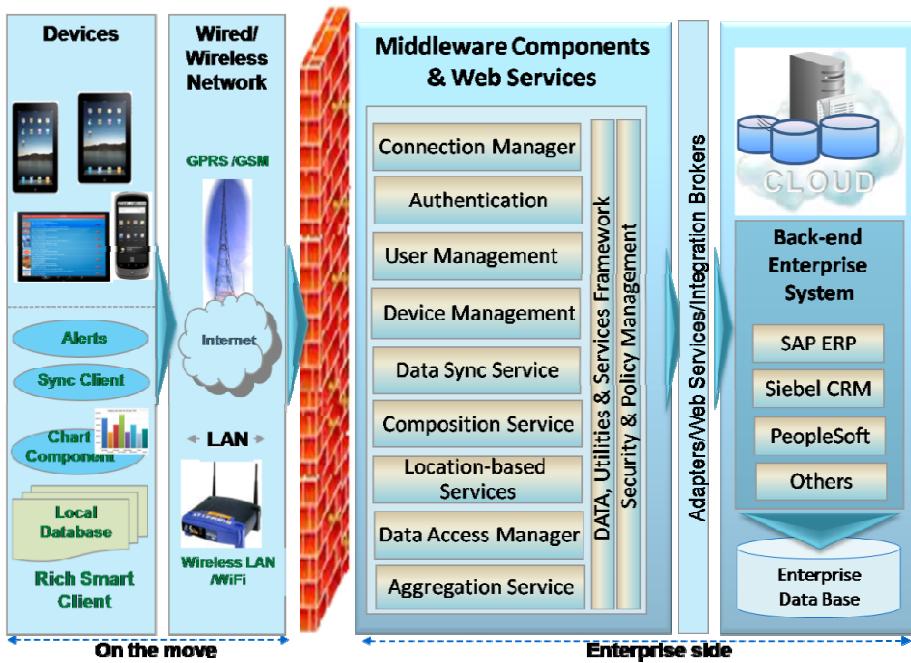


Fig. 3. Key Components of Enterprise Mobile Applications

Key cross cutting concern is depicted below:

- **Security and Policy Management** – To ensure the data Security, encrypt data, manage device loss considerations. This module will implement the already laid-out security standards & policies practiced in the organization.

In addition it serves as the same integration conduit for all new mobile application needing to interface to other existing backend applications and systems. Organizations will need to manage mobile solution integration to on-premise and cloud-based enterprise systems and data, as well as improve data transmission to mobile devices based on bandwidth consumption. This includes management of variability in network connectivity and performance – including tools for enabling off-line mode and ways to deal with frequent air-interface switching between 4G/3G and Wi-Fi networks of varying signal strengths.

6 Enterprise Mobility – Key Design considerations, Design Recommendations and Results

6.1 Design Considerations

Enterprise Mobility adoption would help to improve business user experience by adding value by simultaneously improving workforce productivity and identifying where mobility can be utilized in overall value chain.

Here are the few factors that would be considered while designing enterprise focused mobile application –

- **Design from the ground up** to take advantage of the new opportunities that mobile offers
- **Understand context of Mobile users** - One of the essential things while developing/designing for the rich Mobile application is to understand the context of the user. Mobile context is different and is fundamental to the mobile UI strategy.
- **Understand changing user behavior** – Define viewpoint for the Mobile user
- **Simplified User Interface** - Keep the user interface as simple and intuitive, as possible.
- **Content** - Issues like navigation, image sizes, page weight and scripts all need to be considered when thinking about web application on mobile devices.
- **Full Web Browsing** – Delivering a full desktop like browsing experience to the user
- **Authentication and Authorization strategy** - Every design decision must take into account an effective authentication and authorization strategy for all possible mobility scenarios – Fully connected and occasionally connected modes.
- **Caching to improve response time** - Mobile App design should factor-in caching as a mechanism to improve performance and responsiveness of mobile application. A right caching strategy should include cache expiration and flushing policies.
- **Always design for mobile first** – Don't just re-purpose your web application
- **Adaptive rendering** – Browsers supporting most of the mark-up languages and mobile devices
- **Interactive and fluid UI interface** – Reduced page reloads using right-mix of AJAX technology by retrieving only the data that user wants from the server.
- **Content adaptation or multi-serving Strategies** -Content adaptation strategies would help to deliver content as per user device capabilities. For example, if the device is a newer XHTML MP-compliant device, then an XHTML MP version is delivered & customized according to the screen size of the device.
- **Progressive enhancement philosophy** – Mobile browser uses all those layers (like HTML, CSS and JavaScript layers) which it can handle easily. A few decisions play a vital role to decide on which feature to support and/or leave out for which browser.
- **Support for Open APIs** - Open APIs are being exposed, ideally suited for mobile and cloud applications, but they require appropriate monitoring and management
- **Over-The-Air (OTA) deploy-ability** - Mobility environment is highly heterogeneous, Over-The-Air provisioning/delivery of applications plays an important role as it enables easy deployment and upgrades to mobile applications.

6.2 Key Design Recommendations

The Key recommendations, that may be considered while developing the Mobile applications are given below –

- Understand the customer mobile platform preferences
- Select the platforms that are used by key target segments

- Have a clear business case for multi-platform mobile development
- Take into account the unique nature of mobile application development
- Keep as much of the logic in the network-side as possible
- Mobility solutions must utilize the existing Web-oriented infrastructure by reusing Web Services architecture as much as possible.
- Test the mobile application on actual devices, with real customers, using it in the typical and expected real-world contexts/situations.
- Develop rich enterprise mobile applications while minimizing costs using -
 - Right architectural approach
 - Deep knowledge of mobile as a medium
 - Competitive delivery models

6.3 Key Result Area

Organizations need a strong mobility platform to gain the benefits from enterprise mobility. Enterprise Mobility will not only open up new channels for collaboration but provides the means for accessing the critical customer data on the go - Anytime, anywhere. Based on my study with customers from different industry clusters, I came up with the fact that a sharp distinction has to be drawn between creating a technology-focused **mobile strategy**, which refers to helping employees find the right devices, platform, and applications, and a business-focused **strategy for mobility**, which analyzes how mobility will affect various stakeholders. Based on the informal study with several enterprises, the following conclusions are drawn on the possible impact of enterprise mobility and value proposition to potential organizations in terms of tangible and financially oriented results (Refer the figure '4' below).



Fig. 4. Enterprise Mobility value proposition: Tangible and Financial Results

7 Conclusion

It is now clear that mobilizing the enterprise has become a key imperative for next generation businesses where mobility enablement would help them to manage both on-the-move workers and overall business efficiency. Realizing the importance of enterprise mobility as a strategic priority, enterprises will empower their enterprise business processes with the enterprise-wide deployment of mobile applications. Enterprise Mobility has the potential to fundamentally transform enterprises, their business value chains and markets. Organization will have a clear vision to explore how to operate in ubiquitous computing eco-system when location constraints are obliterated and mobility becomes the key endpoint for delivery.

Besides the benefits of a mobility platform, a full-fledged platform, sometimes may not be addressing the need for every organization. In such a scenario, opting for enterprise mobility services from a hosted/managed services vendor is a viable choice. Now that several service providers are moving to the cloud for better service penetration and cost optimization, this option should certainly be explored by enterprises to improve overall customer satisfaction and set themselves apart from their competition.

References

1. Saeteras, J.: Mobile Web vs. Native Apps, Revisited (2010)
2. Moore, D., Budd, R., Benson, E.: Professional Rich Internet Applications: AJAX and Beyond. Wrox Press (© 2007 Citation)
3. Designing for the Mobile Web, Sitepoint,
<http://www.sitepoint.com/designing-for-mobile-web>
4. Barnes, S.J.: International Journal of Mobile Communications 1(4)(October 2003)
5. Basole, R.C.: Enterprise mobility: Researching a new paradigm. Information Knowledge Systems Management 7, 1–7 (2008)
6. Barnes, S.J.: Enterprise Mobility: Concept and Examples. International Journal of Mobile Communications 1(4), 341–359 (2003)
7. Mobilizing Applications for the Enterprise. Mobile Enterprise Magazine (June 2011),
<http://tiny.cc/mznfc>

Allocation of Guard Channels for QoS in Hierarchical Cellular Network

Kashish Parwani* and G.N. Purohit

Banasthali University, AIM & ACT, Banasthali-304022, India
kparwani1@yahoo.com, gn_purohitjaipur@yahoo.co.in

Abstract. This paper present a dynamic guard channel assignment technique based on a two low layer of hierarchical cellular architecture which evaluates the QoS of (LMST) low speed and (HSMT) high speed moving terminals in an indoor area. The lower layer of the proposed architecture is based on a femto cellular solution for absorbing the traffic loads of LSMT. The higher layer is based on a picocell solution for absorbing the traffic load of the HSMT. The result show that using the optimum number of channels and adjusting dynamically the number of guard channels in each layer, the QoS of LSMT and HSMT are evaluated, having a small negative effect on the QoS of LSMT.

Keywords: Quality of Service (QoS), Picocell, Femtocell, Guard Channel, Hierarchical Cellular Network (HCN).

1 Introduction

In the past few years, all of the femtocell vendors have made their own efforts with different structures and methods to fix the femtocell into the current network (UMTS, CDMA) or Wireless Interoperability for Microwave Access (Wi-Max) etc. Continuing on this path would result in over hyped technology solutions that would make it difficult to inter-operate with each other and keep the cost of femtocell at a reasonable level. An industry wide standardization becomes essential to enable the wide spread adoption and deployment of femtocells by telecom operators around the world. Mobile communications systems experience a rapid increase in the number of subscribers, which places extra demands on their capacity. One of the major challenges in such networks is the utilization of limited resources effectively in order to provide high availability of service. To solve this problem, the geographical area is to divide into small cells and reuse the channels in the system and this type of network is called the hierarchical cellular network (HCN) and has four types of layers i.e. macro cell, micro cell, picocell and femtocell. But more serious problem that arises in this architecture is the handoff issue which occurs when a mobile user moves from one cell to neighboring one. [1] The problem of handoff issue becomes more serious, for high speed moving terminals where the handoff rate increases and the probability that an ongoing call will be dropped due to the lack of a free traffic channel.

* Corresponding author.

The handoff blocking probability is to be more important than the blocking probability of new calls because the call is already active and the QoS is more sensitive for handoff calls.

In this paper our concentration is only on two low layers i.e. picocell and femtocell of HCN and subscribers are divided into two types of class i.e. low speed moving terminal (LSMT) which are connected to the femtocell layer and high speed moving terminal (HSMT) which are connected to the picocell layer and using an Erlang model with prioritized handoff procedure is analyzed and the dynamic adjustment of guard channel into technique based on a multi layer architecture.

2 Related Work

Many models have been developed in wireless cellular networks which provide the best performance and efficiency. In [6], developed the flexible hierarchical cellular systems with a bandwidth efficient handoff schemes. Chandrasekhar. V, et al, [2] gave femtocell networks a survey which described the features of femtocell. In July 2007, the Femto Forum [18] was founded to promote femtocells standardization and deployment worldwide. Jie Zhang et al, [8] studied femtocell technology and deployment. Wu.X, et. al. [3] compared the performance of two tier cellular networks based on queuing handoff calls. K. Parwani, et.al. [9], proposed performance measures of mobile communication networks with hierarchical cellular networks. In [7], Jung and et al, studied the performance of a combined guard channel and channel reservation with queuing resource management scheme for efficient handing of handoff calls in macro/femtocell hierarchical cellular network. K.Ioannou,et.al [12] proposed a dynamic guard channel assignment technique based on a two layer cellular network. In [11], K. Parwani and Purohit compared the blocking probability of two low layer of hierarchical cellular network with a queue and without a queue.

3 System Description

In cellular system, the handoff calls are given higher priorities using prioritized networks. To provide prioritization of the handoff calls is to reserve some channels exclusively for handoff calls, also known as guard channel. ‘C’ is the available channels in every femtocell and priority is given for handoff calls and guard channels (C_h) exclusively are assigning for handoff calls. The remaining ($C-C_h$) channels are shared both by new and handoff calls and mean holding time T_h are considered in all femtocells.

New and handoff calls of LSMT are generated in the area of femtocell according to Poisson point process, with mean arrival rates of λ_n^L and λ_h^L respectively and new calls and handoff calls of HSMT are generated with mean rates of λ_n^H and λ_h^H per cell. The relative motilities are defined as

$$a_L = \frac{\lambda_h^L}{\lambda_h^L + \lambda_n^L} \text{ for LSMT} \quad (1)$$

$$a_H = \frac{\lambda_h^H}{\lambda_h^H + \lambda_n^H} \text{ for HSMT} \quad (2)$$

The total relative mobility for both HSMT and LSMT is

$$a_{HL} = \frac{\lambda_h^H + \lambda_h^L}{\lambda_h^H + \lambda_n^H + \lambda_n^L + \lambda_h^L} \quad (3)$$

The offered load per cell is,

$$T_{off} = \frac{\lambda_h^L + \lambda_n^L + \lambda_h^H + \lambda_n^H}{\lambda_H} \quad (4)$$

Where $\mu_H = 1/T_H$ and T_H is the channel holding time

4 Analysis

4.1 Analysis of the Model with Prioritized Handoff Procedure

Let n be the number of femtocell in the femtocellular area. The total offered load in the system is $T_{off}^{tot} = n T_{off}$.

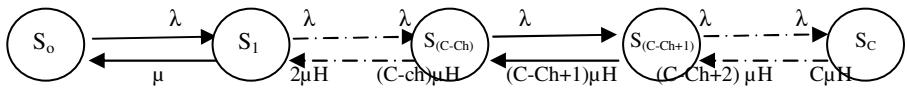
The total number of channels in the system is $C_s = n C$

The steady state probabilities that j channels are busy in every femtocell, can be derived from fig. 1

$$P_j = \begin{cases} \frac{\lambda_n^L + \lambda_n^H + \lambda_h^L + \lambda_h^H)^j}{j! \mu_H^j} P_0 & \text{for } j=1, 2, \dots, C - C_h \\ \frac{\lambda_n^L + \lambda_n^H + \lambda_h^L + \lambda_h^H)^{C-C_h} (\lambda_h^L + \lambda_h^H)^{j-(C-C_h)}}{j! \mu_H^j} P_0 & \text{for } j=C - C_h + 1, \dots, C \end{cases} \quad (5)$$

where

$$P_0 = \sum_{k=0}^{C-C_h} \left[\frac{\lambda_n^L + \lambda_n^H + \lambda_h^L + \lambda_h^H)^k}{k! \mu_H^k} \right]^{-1} + \sum_{k=C-C_h+1}^C \frac{(\lambda_n^L + \lambda_n^H + \lambda_h^L + \lambda_h^H)^{C-C_h} (\lambda_h^L + \lambda_h^H)^{k-(C-C_h)}}{k! \mu_H^k} \quad (6)$$



$$\text{Where } \lambda = \lambda_n^L + \lambda_h^L + \lambda_n^H + \lambda_h^H$$

Fig. 1. State transition diagram for femtocell for the existing one layer architecture

The blocking probability (P_B) for a new call (either HSMT or LSMT) per femtocell is the sum of probabilities that the state number (j) of the femtocell is greater than ($C - C_h$). Hence

$$P_B = \sum_{j=C-C_h}^C P_j \quad (7)$$

The blocking probability for a new call of HSMT per femtocell is

$$P_B^{HSMT} = \frac{\lambda_n^H}{\lambda_n^L + \lambda_n^H} P_B \quad (8)$$

And the blocking probability for a new call of LSMT per femtocell is

$$P_B^{LSMT} = \frac{\lambda_n^L}{\lambda_n^H + \lambda_n^L} P_B \quad (9)$$

The probability of handoff attempt failure P_{fh} is the probability that the state number of the femtocell is equal to C . Thus, $P_{fh} = P_c$

The handoff blocking probability of HSMT is

$$P_{fh}^{HSMT} = \frac{\lambda_h^H}{\lambda_h^L + \lambda_h^H} P_{fh} \quad (10)$$

And the handoff blocking probability of LSMT is

$$P_{fh}^{LSMT} = \frac{\lambda_h^L}{\lambda_h^H + \lambda_h^L} P_{fh} \quad (11)$$

In this model, the ratio of guard channels towards the total available channel in every femtocell is fixed.

4.2 The Proposed Guard Channel Ratio Technique Based on a Two Low Layer of HCN

A guard channel ratio technique is proposed in order to determine the optimized number of guard channels that assigned to picocells and to femtocells. The purpose of this technique is optimized determination to decrease both the call blocking probability of

HSMT (new and handoff) and the blocking of handoff calls of LSMT. This network has two low layers architecture the lower femtocellular layer and the small higher, the picocellular layer. The implementation of picocell is achieved by using multiple base stations to serve the same area. The picocell layer provide services only to new and handoff calls of HSMT and the femtocell provides services only to new and handoff calls of LSMT. Also homogeneous traffic is considered in all femtocell and picocell and the T_h is the same for picocell and the femtocell.

Let n be the number of femtocell that consist the femto cellular layer. Let C_s be the total number of channels in the system. In the femto Cellular layer, priority is given to handoff calls attempts by assigning guard channels (C_{hf}) exclusively for handoff calls of LSMT among the C_f channels in a call. The remaining ($C_f - C_{hf}$) channels are shared by both new and handoff calls of LSMT and let C_p be the channels assigned to picocell. Priority is given to handoff attempts by assigning guard channels (C_{hp}) exclusively for hand off calls of HSMT among the C_p channels in the picocell. The remaining ($C_p - C_{hp}$) channels are shared by both new and handoff calls of HSMT.

$$\text{Hence, } C_s = nC_f + C_p \quad (12)$$

The ratios of guard channels for picocell is $gC_p = C_{hp}/C_p$

And for femtocell is $gC_f = C_{hf}/C_f$

The mean rate of new and handoff calls of HSMT is λ_n^H and λ_h^H respectively per cell, so the mean rate of new and handoff calls in the picocell is $n\lambda_n^H$ and $n\lambda_h^H$ respectively. The proposed channel assignment scheme, assigns the ratios both gC_p , gC_f and C_p/C_f according to a_L , a_H , a_{HL} and T_{off}^{tot} contributing to improvement of the QoS of HSMT and LSMT

The steady state probabilities that j channels are busy in a femtocell can be derived from fig. 2

$$P_j^f = \begin{cases} \frac{\lambda_n^L + \lambda_h^L)^j}{j! \mu_H^j} P_0^f & \text{for } j=1,2,\dots,C_f - C_{hf} \\ \frac{\lambda_n^L + \lambda_h^L)^{C_f - C_{hf}} \lambda_h^{j-(C_f - C_{hf})}}{j! \mu_H^j} p_0^f & \text{for } j=C_f - C_{hf} + 1 \dots C_f \end{cases} \quad (13)$$

where

$$P_0^f = \left[\sum_{k=0}^{C_f - C_{hf}} \frac{(\lambda_n^L + \lambda_h^L)^k}{k! \mu_H^k} + \sum_{k=C_f - C_{hf} + 1}^{C_f} \frac{(\lambda_n^L + \lambda_h^L)^{C_f - C_{hf}} \lambda_h^{k-(C_f - C_{hf})}}{k! \mu_H^k} \right]^{-1} \quad (14)$$

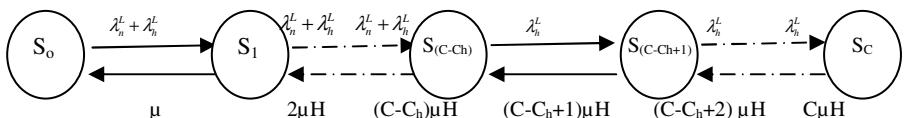


Fig. 2. State transition diagram for every femtocell in proposed channel assignment technique

The blocking probability for a new call of LSMT per femtocell is the sum of probabilities that the state number of the femtocell is greater than $C_{hf} - C_f$. Hence

$$P_B^f = \sum_{j=C_f - C_{hf}}^{C_f} P_j^f \quad (15)$$

The probability of handoff attempt failure P_{Fh}^f is the probability that the state number of the femtocell is equal to C_f . Thus,

$$P_{Fh}^f = P_C^f \quad (16)$$

Now, we measure the cost function of the system which uses system's data as the average new call origination rate and the average handoff attempt rate per cell. This cost function can be expressed as

$$CF = \frac{\lambda_n P_B + \lambda_h P_{Fh}}{\lambda_n + \lambda_h} \quad (17)$$

Therefore, the QoS for HSMT and handoff calls of LSMT assured while allowing high utilization of channels. The proposed model based on the two low layer of HCN is to assure both the required blocking probability of HSMT and the handoff blocking probability of LSMT.

5 Result

In this section, we compare the handoff blocking probability of two low layers (i.e. picocell and femtocell) without and with guard channels ratio technique for different values of C_{hp}/C_p and C_{hf}/C_f . Figure 3(a) and 3(b) present the handoff failure blocking probability of LSMT and HSMT for 40% and 60% of guard channel, $C_s=8$, $C_f=5$, $C_p=5$, against total offered traffic load in LSMT and HSMT respectively.

While figures 3(c) and 3(d) present the cost function (CF) for LSMT and HSMT against T_{off}^{tot} , figure 3(a) and (b), show an improvement in the handoff failure blocking probability of LSMT and HSMT, as a result of using the proposed technique as well as adjusting the ratio's C_p/C_s , C_f/C_s according to T_{off}^{tot} to, a_L , a_H , and a_{HL} . Figure 3(c) and 3(d) show that the cost function of LSMT and HSMT against total traffic loads of LSMT and HSMT. In these figures, the cost function decrease when we added 40% of guard channel against T_{off}^{tot} and also more decreases when we added 60% of guard channel against T_{off}^{tot} as shown in Table1 and Table2. Hence, added 60% of guard channel is much better for getting good result.

Handoff failure blocking probability of LSMT

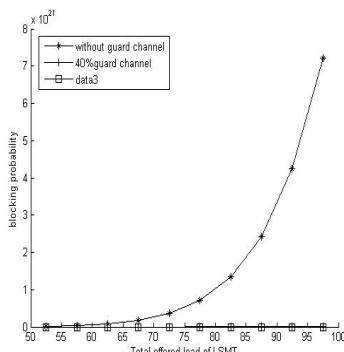


Fig. 3(a)..

Handoff failure blocking probability of HSMT

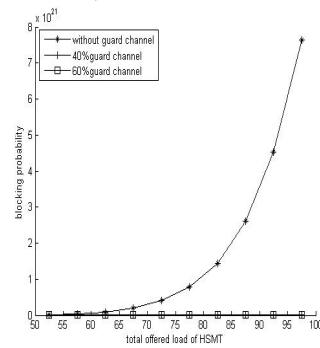


Fig. 3(b).

Cost function of LSMT

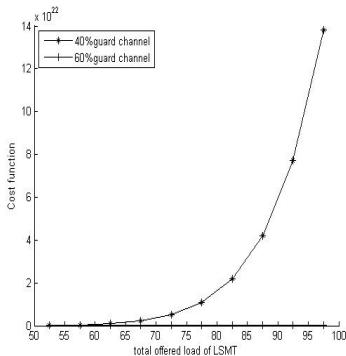


Fig. 3(c).

Cost function of HSMT

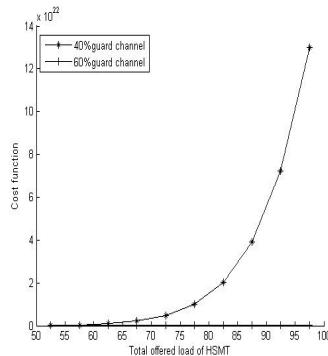


Fig. 3(d).

Table 1. Handoff failure blocking probability of LSMT

T_{off}	P_{BLSMT}	P_{Bf} (40% Guard Channel)	P_{Bf} (60% of Guard Channel)
52.5	1.4318e+019	2.2782e+005	1.8405e+003
57.5	3.5795e+019	3.0154e+005	1.9372e+003
62.5	8.28388e+019	3.8968e+005	2.0340e+003
67.5	1.7962e+020	4.9353e+005	2.1308e+003
72.5	3.6836e+020	6.1439e+005	2.2275e+003
77.5	7.1985e+020	7.5354e+005	2.3243e+003
82.5	1.3487e+021	9.1228e+005	2.4211e+003
87.5	2.4348e+021	1.0919e+006	2.5178e+003
92.5	4.2527e+021	1.2937e+006	2.6146e+003
97.5	7.2123e+021	1.5189 e+006	2.7114e+003

Table 2. Cost Function of LSMT

T_{off}	CF_1 (40% guard channel)	CF_2 (60% guard channel)
52.5	1.4835e+020	8.1654e to 017
57.5	4.0528e+020	1.8596e + 018
62.5	1.0182e+021	3.942 e + 018
67.5	2.3827e+021	7.9333e +018
72.5	5.2467e+021	1.5143e + 019
77.5	1.0106e+022	2.7682e+019
82.5	2.1862e+022	4.8728e+019
87.5	3.9037e+022	8.2958e+019
92.5	7.7331e+022	1.3711e+020
97.5	1.3828e+023	2.2068e+020

6 Conclusion

In this paper, we proposed a dynamic adjustment of guard channel ratio technique to determine the optimized number of guard channels which assigned both to picocells and femtocells. By using this technique, we added to 40% and 60% guard channels in cellular system and achieved much better performance with 60% of guard channel for both to new and handoff calls of HSMT and handoff calls of LSMT. The cost function of HSMT and LSMT has been optimized having a minimum effect on the cost function of LSMT.

References

- [1] Panoutsopoulos, Kotsopoulos, S., Loannou, C., Louvros, S.: Priority Technique for optimizing Handover Procedure in Personal Communication System. *Electronics Letters* 36(7) (March 30, 2000)
- [2] Chandrashekhar, V., Andrews, J.C.: Femtocell Networks, A Survey. *IEEE Communication Magazine* 46(49), 59–67 (2008)
- [3] Wu, X.: Supporting Quality of Services (QoS) in Overlaid Wireless Networks, PhD. Thesis. University of California Davis, USA (2001)
- [4] Kudoh, E., Shibuya, A., Koboto, S.: Pico cell Network for Local Positioning and Information System. In: 26th Annual Conference of the IEEE Industrial Electronic Society, vol. 2, pp. 1165–1170 (October 2000)
- [5] Jain, R.K., Agarwal, N.K.: Hierarchical Cellular Structures in High Capacity Cellular Communication System. *IJACSA International Journal of Advanced Computer Science and Application* 12(9) (2011)
- [6] Li, B., Wu, C., Fujuda, A.: Performance Analysis of Flexible Hierarchical Cellular Systems with a Bandwidth Efficient Handoff Schemes. *IEEE Transactions on Vehicular Technology* 50(4), 971–980 (2001)
- [7] Moon, J.-M., Cho, D.-H.: Efficient Handoff Algorithm for Inbound Mobility in Hierarchical Macro/Femtocell Networks. *IEEE Communication Letters* 13(10), 755–757 (2009)
- [8] Zhang, J., et al.: Femtocell Technology and Deployment. A John Wiley and Sons, Ltd. (2010)

- [9] Parwani, K., Purohit, G.N.: Performance Measures of Mobile Communication in a Hierarchical Cellular System. In: ICDe COM International Conference IEEE Xplore (2011)
- [10] Parwani, K., Purohit, G.N., Sharma, G.: Performance Evaluation of Mobile Communication in an Hierarchical Cellular Systems with Different Schemes. In: Das, V.V., Thankachan, N. (eds.) CIIT 2011. CCIS, vol. 250, pp. 677–683. Springer, Heidelberg (2011)
- [11] Parwani, K., Purohit, G.N.: Performance Measures of Hierarchical Cellular Networks on Queuing Handoff Calls. In: Proc. of Int. Conf. on Advances in Computer Science, pp. 25–29 (2011), doi:02, ACS.2011.02.528
- [12] Lounnou, K., Kotsopoulos, S., Stavroulakis, P.: Optimizing the QoS of High Speed Moving Terminals in Cellular Networks. International Journal of Communications System 16, 851–863, doi:10.1002
- [13] Lim, B.L., Haurence Wang, W.C.: Hierarchical Optimization of Microcellular call handoff. IEEE Transation, Vech. 48(2), 459–466 (1999)
- [14] Ioannou, K., Louvros, S.: Optimizing the Handover Call Blocking Probability in Cellular Networks with High Speed Moving Terminals. IEEE Communications Letters 6(10) (October 2002)
- [15] Louvros, S., Py Larinow, J.: Handoff Multiple Queue Model in Microcellular Network. Computer Communications 30, 396–403 (2007), doi:10.1010 Com.2006. 09.008
- [16] Salih, T., Fidanboiyu, K.M.: Performance Analysis and Modeling of Tier Cellular Networks with Queuing Handoff Calls. In: IEEE International Symposium on Computers and Communication, ISCC 2003 (2003)
- [17] Passas, N., Paskalis, S., Vali, D.: Quality-Of-Service Oriented Medium Access Control for Wireless ATM Networks. IEEE Communication Magazine, 42–50 (November 1997)
- [18] <http://www.femtoforum.org>
- [19] <http://www.thinkfemtocell.com>

Application Development and Cost Analysis for Content Based Publish Subscribe Model in Mobile Environment

Medha A.Shah and P.J. Kulkarni

Walchand College of Engineering, Sangli.(MS) India
Shah.medha@gmail.com, pjk_walchand@rediffmail.com

Abstract. Content-based publish-subscribe is emerging as a communication paradigm able to cope with the needs of scalability, flexibility, and reconfigurability typical of highly dynamic distributed applications. However very few efforts concentrates on application development and cost analysis for content based publish subscribe model in real time mobile environment. In this paper we illustrate the application developed for content based publish subscribe in wired and wireless network. Also we present the result of cost analysis performed using cost model in real time mobile environment. The result of the experiment which is carried out using mobile devices show accordance with cost analysis, which verifies the correctness and usefulness of cost model.

Keywords: Publish subscribe model, content based, Mobile Ad-HOC network, REDS, Cost model.

1 Introduction

Mobile computing is changing the way researchers, developers and users think about distributed computing systems. Today's hardware is experiencing a rapid grow of computing and communication capabilities quite always corresponding to a symmetric reduction in size of the resulting devices. This trend poses the basis for many different computing scenarios where the final user may be equipped with multiple, small computing devices without constraining its ability to move in the surrounding environment. Among those scenarios, one of the most challenging is that of Mobile Ad-Hoc Networks (MANETs). In this case, the network infrastructure is totally absent and nodes agree to relay each other's packets towards their ultimate destinations, thus forming an overall cooperative infrastructure. A number of situations can be considered as an ideal deployment scenario for MANETs. In particular, all those cases where a fixed communication infrastructure does not exist or cannot be relied upon, such as during disaster recovery can gain advantages from the unique features of Mobile Ad-Hoc Networks.

However, this kind of network environments are more difficult to implement than traditional fixed networks because of their inherent dynamicity and the lack of global knowledge on the network topology. The issues one has to face in designing and implementing software systems to be deployed on MANETs impact on all network layers, from the access medium up to the application level. In this context in particular, the developer has typically to deal with frequent disconnections and highly varying

data transfer rates. Here we have developed application that works in real time MANET using REDS as middleware. Cost analysis helps to take major decisions while deploying any kind of application in real time MANET.

In light of the peculiar features of Mobile Ad-Hoc Networks, many well known communication paradigms cannot find in this infrastructure-less network environment, a comfortable deployment scenario. However, the strong decoupling between information producers and consumers the Publish-Subscribe model fosters seems very well fitted to loosely coupled scenarios such as MANETs. In the Publish-Subscribe interaction style, subscribers express interest in some kind of event (or event pattern) and can be possibly notified in case some other node in the system publishes an event matching their interests. Events are routed from publishers to the intended subscribers by a dispatching infrastructure, this last typically implemented with a set of interconnected dispatching servers. All the communications between publishers and subscribers are asynchronous as well as fully decoupled in time and space [11]. However, the Publish-Subscribe communication paradigm could not be deployed on MANETs without taking into account the inherent dynamicity of the environment, mainly due to nodes mobility. For this reason, event-based systems to be deployed on MANETs have to be provided with suitable reconfiguration mechanisms [9] capable of facing issues like frequent link breakages and node joining or leaving the system at unpredictable points in time.

1.1 Wireless Mobile Ad-Hoc Networks

Here we briefly sketch the main features and characteristics of wireless Mobile Ad-Hoc Networks. A Mobile Ad Hoc Network [7] is an autonomous system of mobile hosts connected by wireless links, the union of which forms an arbitrary graph. The hosts are free to move randomly and organize themselves arbitrarily; thus, the wireless topology may change rapidly and unpredictably. Such a network may operate in a standalone fashion, or may be connected to the larger Internet. Such networks are of interest because they do not require any prior investment in fixed infrastructure. Instead, the network nodes agree to relay each other's packets toward their ultimate destinations, and the nodes automatically form an overall cooperative infrastructure.

Sometimes, a fixed infrastructure exists but cannot be relied upon, such as during disaster recovery. Finally, existing services may not provide adequate service, or may be too expensive. Mobile Ad-Hoc Networks are attracting a lot of attention these days due to the little efforts needed to deploy them. Moreover, in emergency services such as disaster recovery these networks are the only viable choice. Though ad hoc networks are attractive, they are more difficult to implement than fixed networks. Fixed networks take advantage of their static nature in two ways. First, they proactively distribute network topology information among nodes, and each node pre-computes routes through that topology using relatively inexpensive algorithms. Second fixed networks embed routing hints in node addresses because the complete topology of a large network is too unwieldy to process or distribute globally. Neither of these techniques works well for networks with mobile nodes because movement invalidates topology information and permanent node addresses cannot include dynamic location information. Moreover, a set of characteristics are peculiar of wireless Mobile Ad-Hoc Networks, these can be summarized as follows:

- New members can join or leave the system at any point in time.
- There are no base stations to provide connectivity to backbone hosts or to other mobile hosts.
- There is no need for handover or location management.
- Each mobile host can act as a router, forwarding packets from one mobile host to another.
- Communication connectivity is fairly "weak", network topology changes are frequent and unpredictable.

1.2 Publish-Subscribe Flavors and Variations

The basic Publish-Subscribe paradigm can be implemented in various ways, each of them being well-suited to some particular application or deployment scenario. Two characteristics are worth a particular attention: the way subscribers can express the events they are interested in (usually called subscription schema) and the particular architecture of the dispatching infrastructure (often termed as interconnection topology). These two orthogonal aspects are now briefly introduced.

Subscription schema

The way subscribers can express their interest in a particular event strongly influences the subscription language and hence the overall efficiency of the system. The exiting literature distinguishes among topic-based, content-based and type-based systems. We explore here content based schema.

- Content-based systems use a different approach with subscriptions expressed in terms of the actual content or properties of the published events. This implies that the way an event is represented should be carefully designed in order to precisely define the semantics of subscriptions with respect to events. The dispatching infrastructure should also be able to efficiently match the content of a given event against a potentially very large set of different subscriptions. Usually, the subscription language defines a set of constraints the events must satisfy in order to be forwarded toward the interested subscribers. These constraints are typically composed of a set of name-value pairs and a set of well-defined comparators that can be logically combined to form a complex subscription pattern. Each value has also a type for ensuring well-defined ordering and equality relations. Examples of a subscription and a matching event for a content-based system are as following

A subscription in a content-based system.

```
string manufacturer = "Maserati"
string model = "A6 GCS" _ "Mexico 3300" integer price < 500.000
```

An event matching the subscription mentioned above

```
string manufacturer = "Maserati" string model = "A6 GCS"
integer price = 499.999
```

2 Cost Model

2.1 System Parameters

In this section, cost analysis model for publish/subscribe systems is discussed.

Following basic system parameters are used to analyze the cost.

- α (publish rate): Publisher's event generation is governed by Poisson process with average inter arrival time of $1/\alpha$.
- β (request rate or process (reference access) rate): This parameter for two meanings: (1) subscriber's access rate of published events, and (2) request rate of client in the client/server models. These rates are also governed by Poisson process
- $c_{ps}(\alpha)$ (publish/subscribe cost per event): Cost required for an event publish. C_{ps} is divided into two parts: (1) c_{pub} : ES(Event Source) publish events to EBS (Event Brokering System), and (2) c_{sub} : EBS(Event Brokering System) relays the events to ED(Event Displayer) which registered for the events.
- $s(n)$ (effect of sharing among n subscribers): For example, server can deliver events with low cost when it broadcasts event to many subscribers. It will be between $1/n$ and 1.
- t_{ps} (time delay for publish/subscribe): Time delay for publishing an event. t_{ps} is divided into two parts: (1) t_{pub} : time delay for publish, ES(Event Source) publish events to EBS(Event Brokering System), and (2) t_{sub} : time delay for subscribe,
- ED(Event Displayer) subscribes events from EBS(Event Brokering System).

2.2 Cost Analysis of Publish/Subscribe Model

The cost of publish subscribe model is analyzed by following way. The following costs are calculated.

- (1) Conceptual total cost (e.g., the number of message, amount of message, or time delay) per unit time for given model.
- (2) Cost for each access by client (or subscriber).
- (3) Time delay for access after subscriber's (or client's) intention.
- (4) Time delay between event occurrence and notification to subscriber (or recognition by client). Cost can be the number of messages, amount of message, or time delay.

Since it is assumed that c_{pub} is cost for that ES(Event Source) publish events to EBS(Event Brokering System), and c_{sub} is cost for that ED(Event Displayer) subscribes events from EBS(Event Brokering System), cost of publish/subscribe model for each event publish and subscribe is $c_{pub} + n s(n)c_{sub}$. Here n is the average number of subscriber and $s(n)$ is sharing effect among n nodes. When publish rate is α , cost per time unit[1] is:

$$\alpha (C_{pub} + nS(n)C_{sub}) \quad (1)$$

The cost in the view point of subscriber (per each event access of subscriber) is calculated as follows. Analyzed three performance metrics, (1) conceptual cost for each access, (2) time delay for subscriber to access event after its intention, (3) and time delay until notification to subscriber after event occurring. The average number of event occurred before each access is cost for each access:

$$\sum_{i=0}^{\infty} \frac{\beta}{\alpha + \beta} \left(\frac{\alpha}{\alpha + \beta} \right)^i = \frac{\alpha}{\beta}$$

where c_{pub} is shared among n subscriber and c_{sub} is required for each subscriber. Thus, average cost for each access is:

$$\frac{\alpha}{\beta} \left(\frac{C_{pub}}{n} + C_{sub} \right) \quad (2)$$

There is no time delay for access after subscriber's intention since event has already been received. Time delay between event occurrence and notification to subscriber is:

$$t_{ps} = t_{pub} + t_{sub} \quad (3)$$

3 REDS Middleware

REDS (REconfiguring Dispatching System)[12] is an open source middleware. It implements arbitrary topological reconfigurations of the message dispatching network independently from the underlying networking scenario.

It includes two approaches i.e.

- a) *Configuration of Middleware Architecture* and addresses the problem of enabling the selection of different mechanisms for different deployment scenarios. Thus, for instance, it is possible to select the routing strategy (or network transport, or matching semantics, and so on) that is most appropriate for the application at hand. Moreover, proper extension mechanisms are in place that enables programmers to define their own variants, if needed.
- b) *Dynamic reconfiguration of topology of message dispatching network* and therefore addresses the needs of highly dynamic systems where the topology of the network undergoes continuous change (e.g., mobile and peer-to-peer systems).

4 Application Development Using REDS for Wired Network

We developed an application that will demonstrate the communication between the brokers and subscribers in **Wired Network**. The publisher publishes the list of songs

associated with each artist. The subscriber is supposed to subscribe the desired songs of the artist of his interest. The songs chosen by the subscriber of his interest reaches to publisher through the content based publish subscribe system. We have used various API provided by REDS[13]. The subscription schema[15] format used for this application is as follows.

- P1: [class: 'HINDI'] [ARTIST:'A'][PAYLOAD:"AB"]

Here P1 schema suggest class of songs as 'HINDI' , the artist whose song is to be subscribed is 'A' and actual content of the song is payload 'AB'

4.1 Experiment Setup and Result Analysis in Wired Network

We performed the experiment using seven desktop PCs connected in wired network. In all the three scenarios we kept five number of brokers along with only one publisher and varying number of subscribers. The following scenario take into account single publisher, single subscriber and many brokers. Following graph shows the delay i.e. the time required for messages to reach the desired subscriber.

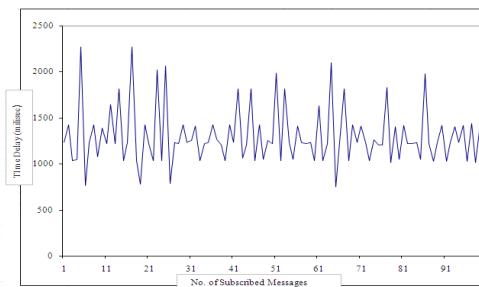


Fig. 1.

Using REDS, We start with basic scenario which includes only single publisher, single subscriber and many brokers. The results of this experiment are formulated in a graph as shown in above figure 1. The graph shows that delay i.e. time required for messages to reach to the desired subscriber. The above experiment was conducted 5 numer of times with variation in number of clients ,number of publishers and number of subscribers. It is observed that as number of brokers increases the time required to reach the message to its destination also increases. We observed that there is minimum packet loss so Message delivery ratio is nearly 100% in wired network.

4.2 Evaluation Through – Real World Experiment in Wireless Network

We integrated our component in REDS(Reconfigurable dispatching System) , a content based publish subscribe middleware. Our module performs real time analysis of content based routing protocol in MANET. We were clearly limited to small scale scenarios. Nevertheless, the experiments we ran provide interesting insights about the trends and values of cost analysis model.

Scenario 1. Following tables describe the simulation parameters.

Parameters	Values
No. of Nodes	6
No. of Publishers	Vary from 3-6
No. of Subscribers	Vary from 3-6
Wireless Device Range	35m
CBR	1, 2, 4, 8, 16, 24 events/sec
Area	60*150 m ²
Mobility model	Random way point

In this scenario each message was generated with a 0.1 probability of matching a subscription. Subscriptions were allowed to change dynamically. In this experiment we ran actual content based application traffic, with messages delivered only to the nodes that expressed an interest in them. The test was taken for 5 minutes, with measures starting after the first minute.

Metrics- We measured the message delivery ratio (MDR) i.e. the ratio between the published messages actually delivered to a client and the overall number of published messages matching at least one subscription at that client.

Results – Figure 2. shows the Message Delivery Ratio (MDR) against the publish frequency. Given the high dynamicity of our scenario, where mobility induces frequent network partitions, the values in the chart are good, especially by considering that no additional measure is taken to recover lost messages. Figure 3 focuses on the message delivery ratio over time. The message delivery ratio drops upon disconnection. By looking to values we determined that MDR reaches the lowest values either when both publisher and subscriber belong to different partitions or when multiple disconnections and reconfigurations occur concurrently, increasing the time required to restore connectivity.

Scenario 2- To perform the cost analysis , we conducted empirical experiment using mobile clients and a message brokering system. The purpose of our experiment was to get actual c_{ps} (t_{ps}) which is publish/subscribe cost (i.e. time delay) per event) for numbers of clients in a practical environment. The experiment environment consists of REDS system which is message brokering system . We have experimented using 6 to 12 laptops and PDAs.

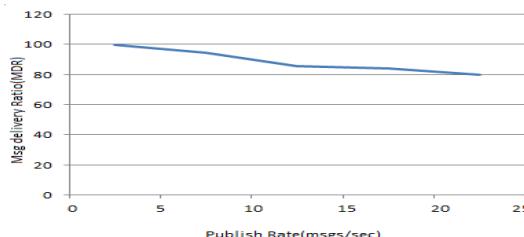


Fig. 2. Publish rate v/s MDR

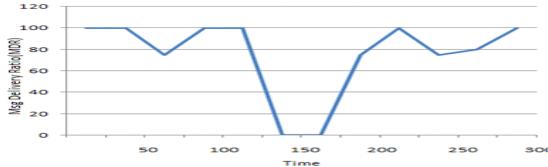


Fig. 3. Message delivery ratio over time.

Metrics.- We performed two types of experiments. First is the experiment to measure the data transition time between an event source (publisher) and an event display (subscriber) by varying the number of clients. The second experiment is to measure the conceptual cost per transaction for varying number of clients. Also we have measured the cost for each access of subscriber.

Result- The experiment result of the data transition time of publish/subscribe message (t_{ps}) is shown in figure d. It is observed that when number of clients increases then time required for data transition also goes on increasing. Data transition time is calculated by using time required for publish plus time required for subscribe. The result of the cost for each access for number of clients is shown in figure e. It is observed that when number of clients increases then cost for each access decreases. This happens because c_{pub} is distributed among all nodes involved. Figure f. shows the relationship between conceptual total cost per time unit and number of clients when number of nodes increases then conceptual total cost per time unit decreases. This happens because of sharing effect among nodes.

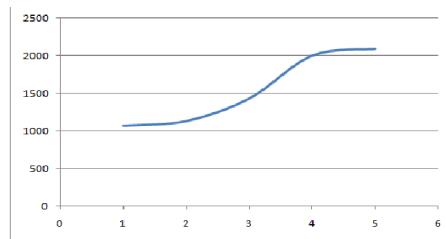


Fig. 4. Data Transition time v/s No. of Clients

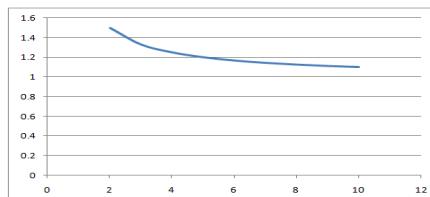


Fig. 5. The cost for each access v/s Number of clients

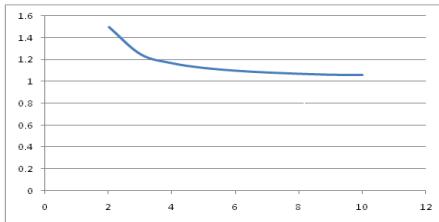


Fig. 6. Conceptual total cost per time unit v/s Number of clients

5 Conclusion

Although content based publish/subscribe system has been popular in distributed real-time system recently, cost analysis has not been done and verified yet. In this paper, we have illustrated the behavior of content based publish subscribe system in wired network and wireless network. We have used cost analysis model [6] for publish/subscribe systems. The empirical result from our test bed verifies cost model. By providing the simulation result and the empirical result which is based on cost analysis model, we give theoretical proof to the known claim, the publish subscribe system is well suited for distributed real-time system.

We have drawn following conclusions like minimum packet loss is observed when content based publish subscribe systems are deployed in wired network. Also Message delivery ratio is near about 100% in wired network. When content based publish subscribe systems evaluated in wireless network message delivery ratio drops upon disconnection .When multiple disconnections and reconfigurations occurs the time required to store the connectivity also increases. It is observed that as number of clients increases the time required for data transition also go on increasing and the cost for each access decreases due to sharing effect of the nodes.

From observation we can conclude that cost analysis plays a major role while deploying any real time application in MANET.

Acknowledgement. This project is funded by AICTE under Research Promotion Scheme (RPS) for the duration from 2009 -2012.

References

1. Banavar, G.: An efficient multicast protocol for content-based publish subscribe System. In: Proc. of the 19th International Conference on Distributed Computing (1999)
2. Camp, T., Boleng, J., Davies, V.: A survey of mobility models for ad hoc network research. Wireless Communications and Mobile Computing. Special Issue on Mobile Ad Hoc Networking 2(5), 483–502 (2002),
<http://cite-seer.ist.psu.edu/camp02survey.html>
3. Oh, S., Kim, J., Fox, G.: Real-time performance analysis for publish/subscribe systems. In: Proceedings of Future Generation Comp. Syst. (2010)
4. Toh, K.C.: Ad hoc Mobile Wireless Networks. Prentice Hall PTR, Upper Saddle River (2002)

5. Baldoni, R., Beraldì, R., Querzoni, L., CugoLa, G., Migliavacca, M.: Content-Based Routing in Highly Dynamic Mobile Ad Hoc Networks (October 24, 2005)
6. Mottola, L., Cugola, G., Picco, G.P.: A Self Repairing Tree Topology Enabling Content-Based Routing in Mobile Ad Hoc Networks. *IEEE Transactions on Mobile Computing* 7(8) (August 2008)
7. Baldoni, R., Beraldì, R., Cugola, G., Migliavacca, M., Querzoni, L.: Structure-Less Content-Based Routing in Mobile Ad Hoc Networks. In: Proc. IEEE Int'l Conf. Pervasive Services, ICPS (2005)
8. Eugster, P., Felber, P., Guerraoui, R., Kermarrec, A.-M.: The many faces of publish/subscribe. *ACM Computing Surveys* 2(35) (June 2003),
<http://citeseek.ist.psu.edu/637776.html>
9. Cugola, G., Picco, G.P.: REDS: A Reconfigurable dispatching system. Dipartimento di Elettronica e Informazione. Politecnico di Milano, Italy
10. REDS Web Page, <http://zeus.elet.polimi.it/reds> (2006)
11. An Adaptive Approach to Content-Based Subscription in Mobile Ad Hoc Networks" Proceedings of the Second IEEE Annual on Per-vasive Computing and Communications Workshops (PERCOMW 2004). IEEE (2004) 0-7695-2106-1/04 \$ 20.00 © 2004
12. Content-Based Routing with On-Demand Multicast" Proceedings of the 24th International Conference on Distributed Computing Systems Workshops (ICDCSW 2004). IEEE (2004) 0-7695-2087-1/04 \$20.00 © 2004

Energy Aware AODV (EA-AODV) Using Variable Range Transmission

Pinki Nayak, Rekha Agarwal, and Seema Verma

¹ Amity School of Engg. and Technology, New Delhi-110061, India,
Also a research scholar at Banasthali University, Rajasthan
pinki_dua@yahoo.com

² Amity School of Engg. and Technology, New Delhi-110061, India
rarun96@yahoo.com

³ AIM and ACT, Banasthali University, Rajasthan-304022, India
seemaverma3@yahoo.com

Abstract. Energy conservation is an important issue in Mobile Ad Hoc Networks (MANETs) as most of the nodes are powered by a battery source which has limited energy reservoir and it also becomes very difficult to recharge or replace the battery of the nodes. Existing MANET routing protocols s. a. DSR and AODV use common transmission range for transfer of data and does not consider energy status of nodes. This paper discusses a new energy aware scheme (EA-AODV) based on AODV using variable transmission range. The protocols are simulated using Network Simulator-2 and comparisons are made to analyze their performance based on energy consumption and network lifetime metrics. The results show that EA-AODV makes effective node energy utilization.

1 Introduction

A Mobile Ad hoc Network (MANET) is a network formed without any central administration. It consists of nodes that use a wireless interface to send and receive packet data. These nodes in the network are mobile and can serve as routers and hosts, thus can forward packets on behalf of other nodes and run user applications. This allows people and devices to seamlessly inter-network in areas with no pre-existing communication infrastructure. Significant examples of MANET include establishing survivable dynamic communications for emergency/rescue operations, disaster relief efforts, military networks, business indoor application, home intelligence devices [13].

Developing routing protocols for MANETs has been a challenging task because of its dynamic topology, bandwidth constrained wireless links and resource (energy) constrained nodes. Many proactive and reactive protocols have been proposed which try to satisfy various properties, like: efficient utilization of bandwidth, battery capacity, fast route convergence, optimization of metrics (like throughput and end-to-end delay), and freedom from loops.

This paper tries to address the problem of energy efficient routing to increase the lifetime of the network. Mobile hosts, today are powered by battery, therefore efficient utilization of battery energy becomes very important. The energy resources of actively participating nodes get depleted faster than other nodes, which in some cases, may lead to partitioning of the network, thus decreasing the lifetime of the network. For this reason, reducing the energy consumption is an important issue in ad hoc wireless

networks [6]. The three major ways of increasing the life of a node are efficient battery management, transmission power management and system power management. In this paper, a scheme has been proposed to minimize the energy consumption at the nodes, thus maximizing the network lifetime. Transmission power control approach is used to adjust the power levels at node. Common power levels are used during Route Discovery. New power levels are calculated between every pair of nodes based on distance.

The rest of the paper is organized as follows. Section 2 presents a discussion on the related work in energy aware routing protocols. Section 3 gives a brief description of the existing routing protocols. In section 4, the detailed working of the proposed EA-AODV is discussed. Section 5 includes the simulation environment setup used in NS-2 simulator. The simulation results are explained in section 6. Finally, section 7 gives conclusions.

2 Related Work

Many research efforts have been devoted for developing energy efficient routing algorithms. Nodes energy is minimized not only during active communication but also when they are in inactive state. Transmission power control and load distribution are two approaches used to minimize the active communication energy of individual nodes and sleep/power-down mode to minimize energy of nodes during inactivity. In transmission power control approach choosing a high transmission power reduces the number of forwarding nodes needed to reach the required destination, but creates excessive interference in a medium that is commonly shared whereas, choosing a lower transmission power reduces the interference seen by potential transmitters but packets require more forwarding nodes to reach their required destination. The specific goal of the load distribution approach [1], [9], [12] is to balance the energy usage of all mobile nodes by selecting a route with underutilized nodes rather than selecting the shortest route [15], [18].

Some research proposals based on transmission power control are discussed in [8], [7]. Flow Augmentation Routing (FAR) [3] finds the optimal routing path in a static network, for a given sourcedestination pair that minimizes the sum of link costs along the path. Online Max-Min Routing (OMM) [14] for wireless ad-hoc networks optimizes the lifetime of the network as well as the lifetime of individual nodes by maximizing the minimal residual power, which helps to prevent the occurrence of overloaded nodes. Power-aware Localized Routing (PLR) [17] is a localized, fully distributed energy-aware routing algorithm, with the assumption that a source node has the location information of its neighbors and the destination. The main goal of Minimum Energy Routing (MER) [5] is not to provide energy efficient paths but to make the given path energy efficient by adjusting the transmission power just enough to reach to the next hop node. The authors in [8] investigates the impact of variable range power control on physical layer and network layer connectivity and shows that variable range increases network lifetime over common range transmission. An optimal transmission range for nodes during flooding the route request messages is given in [7].

3 Routing Protocol

Routing in MANET depends on factors like topology, selection of routers, location of request initiator etc. in finding the path quickly and efficiently. The traditional routing

algorithms are classified as Proactive and Reactive protocols. These algorithms lack energy awareness of the nodes in the network. Reactive routing protocols discover or maintain a route as needed. This reduces overhead that is created by proactive protocols. Reactive routing protocols can be classified as source routing and hop by hop routing. Dynamic Source Routing (DSR) is an example of source routing, whereas Ad-hoc On Demand Distance Vector Routing (AODV) is a hop by hop routing protocol.

DSR (Dynamic Source Routing) [11] is an on-demand, simple and efficient routing protocol for multi-hop wireless ad hoc networks of mobile nodes. DSR uses source routing whereby all the routing information is maintained (continuously updated) at mobile nodes instead of relying on the routing table at each intermediate device. The protocol is composed of two main mechanisms-Route Discovery and Route Maintenance, which works together entirely, on-demand. The protocol allows multiple routes to destination, loop-free routing, support for unidirectional links, use of only soft state in routing and rapid discovery when routes in the network change.

AODV (Ad hoc on-demand distance vector) [2] is a dynamic, self-starting, loop free, multi-hop on-demand routing for mobile wireless ad hoc networks. AODV discovers paths without source routing and maintains route table instead of route cache. It allows mobile nodes to respond to link breakages, changes in network topology in a timely manner through Route Error (RERR), Route Request (RREQ) and Route Reply (RREP) messages. It maintains active routes only while they are in use using sequence numbers and delete unused routes (stale).

4 Proposed Algorithm

The main aim of the proposed algorithm is to make the network energy aware. Energy efficient design of the protocol is generated by varying the transmission range of the nodes. Variable transmission range means controlling the power level for each packet in a distributed manner at each node, thus affecting energy consumption of the network. Choosing a higher transmission range reduces the number of nodes needed to reach the destination but creates large interference, whereas reducing the transmission range demands more number of forwarding nodes leading to less energy utilization. Each node communicates with the neighboring nodes during Route Discovery phase. Once the route is known, each individual node then controls the transmission range as per the distance between source and destination node, so that optimum energy is utilized for packet transmission. The proposed algorithm is explained below:

The steps involved in EA-AODV are:

Step 1 : When any node needs to send data, it generates the Route Request (RREQ) packet and broadcast it to its neighbor with initial common transmission range of 250m.

Step 2 : The route reply messages from the intermediate nodes contain two fields locX and locY that stores the location of the node sending the route reply. Step 3 : In AODV, the path is established for the first RREP received. But in EA-AODV, each node waits for a time (T_{wait}) till it receives all the RREP messages destined for the node.

Step 4 : The node then calculates the distances between the nodes from where the RREP message is received and itself. This is done using own location and the locations of the intermediates nodes.

Step 5 : Now, the node with minimum distance is selected calculated in step 4 and its location is also updated in the routing table as two entries (n_{hopX}) and (n_{hopY}).

Step 6 : The transmission energy is then calculated using the Friis transmission equation in free space (Appendix) based on the distance, which is the distance between the current node and the next hop node in the said algorithm. The received power threshold of the node is kept constant throughout.

Step 7 : The route between source and destination is maintained for data transfer.

Step 8 : If the route is broken, repeat from step 1.

5 Simulation Setup

The simulations are carried out using the event driven simulation tool Network Simulator-2 (ver. 2.34) and the wireless extensions provided by CMU. The used channel is Wireless channel/Wireless Physical, Propagation model is Free Space Propagation Model, Queuing model is Drop Tail/Priority Queue, Mobility model is Random Waypoint model and MAC protocol is 802.11. The simulation setup consists of an area of 800 X 800 m² with different number of nodes for each simulation. To emulate the dynamic environment, all the nodes move around in the entire region. Varying speeds with minimum 2m/sec and maximum of 40 m/sec have been considered. Constant Bit Rate (CBR) traffic source with packet size of 512 bytes is taken. Different source-destination pair (5 25 connections) was used to establish the routes. All the simulations were run for a period of 200 sec. The initial energy of each node was set as 100 Joules with transmission and reception power of 5 W and 1W respectively. Table 1 gives the simulation parameters used.

For tables use

Table 1. Simulation Parameters

Parameter	Value
No of Nodes	20
Grid Area	800m × 800m
No. of Connections	5, 10, 15, 20, 25
Pause Time	0 sec
Speed	5 m/s, 10m/s, 40m/s
Traffic Model	CBR
Data Packet Size	512 bytes
Data Packet Interva	14 packets/sec
Simulation Time	200 sec
Initial Energy of Node	100 J
Transmitted Power	5 W
Received Power	1W
Idle State Power	0.0005 W
Sleep State Power	0.0002 W
Transition from sleep to active state	.03W

6 Results and Discussion

Simulations have been conducted to compare the performances of AODV and EA-AODV protocols. Performance comparisons have been made on two parameters namely Network Lifetime through number of exhausted nodes and Average Residual Energy [4].

6.1 Performance Comparison of Network Lifetime for Two Protocols

Network Lifetime is a vital performance metric while comparing routing protocols. The network lifetime can be defined as [16]

- the time taken for k nodes to die
- the time taken for the first node to die
- the time taken for all the nodes in the network to die

In this paper, the first definition is adopted.

Fig 1a and 1b show the comparison of network lifetime between AODV and EA-AODV with nodes moving at speed of 10 m/sec and 40 m/sec respectively. Number of connections is taken as 10. It is observed that energy consumption of participating nodes in the network starts increasing from 91 sec of the simulation time for AODV and from 96 sec for EA-AODV when the speed of nodes is 10m/sec. As the speed is increased to 40m/sec (Fig. 1b), EA-AODV still shows a better distributed network lifetime as compared to AODV. In AODV, the 20 participating nodes consume all the energy within 101 sec of simulation time, whereas in EA-AODV the number of nodes consuming all energy reduced. This is because, as the transmitted power is adjusted according to the shortest distance in EA-AODV, the available node energy is effectively used decreasing the number of energy exhausted nodes.

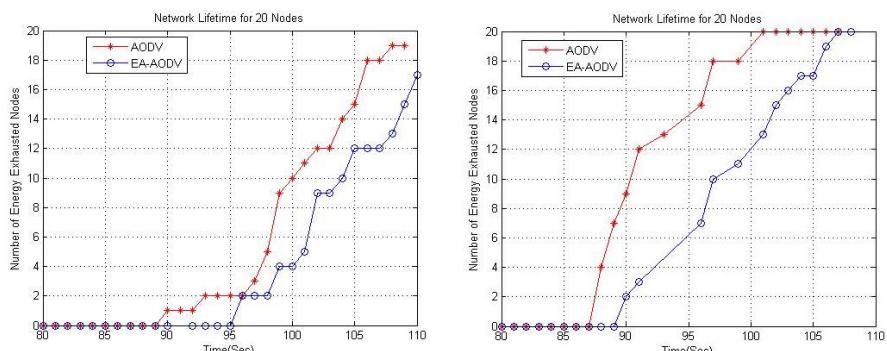


Fig. 1. a) Network Lifetime for Speed 10m/sec

b) Network Lifetime for speed 40m/sec

6.2 Performance Comparison of Average Residual Energy for Two Protocols

Average residual energy gives the remaining energy of the network and in turn reflects the network utilization time. It is seen from Fig. 2 that the average residual energy of both the protocols decreases as the simulation time progresses. EA-AODV shows superior performance than AODV. At simulation time of 130 sec, the average residual energy of AODV is 62J and that of EA-AODV is 69J. Again the variable transmit power keeps the average residual higher in EA-AODV.

Fig. 3 gives the variation of average residual energy with number of nodes in the network for different speeds. It is clearly seen that EA-AODV outperforms AODV under various mobility conditions. At higher speed (40 m/sec, Fig 3b)) the average residual energy is less as compared lower speed (10 m/sec, Fig. 3a)). This is because the nodes are more mobile at higher speeds causing frequent path breaks, which requires new

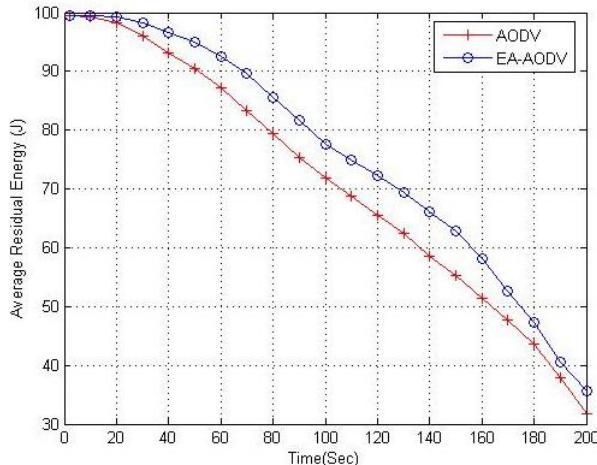


Fig. 2. Average Residual Energy variation with Time for speed of 10 m/sec

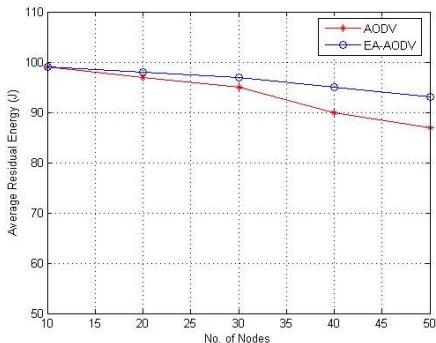
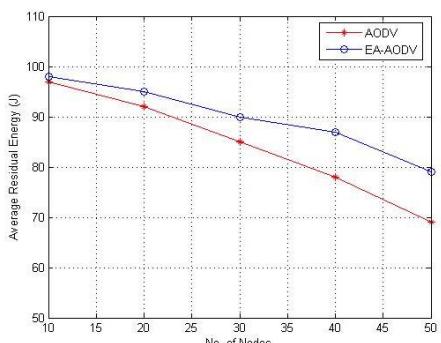


Fig. 3. a) Average Residual Energy variation with number of nodes at 10m/sec



b) Average Residual Energy variation with number of nodes at 40m/sec

route discovery, increasing routing overhead. This causes increase in energy consumption, hence reducing the residual energy. For Fig. 4, simulations are carried for 50 nodes at speed of 10m/sec. Variation of Average Residual Energy with number of connections (half of the number of nodes) is plotted. As seen, Average residual energy decreases with number of connections as number of nodes transmitting data i.e. the network load is now more. EA-AODV shows better variation over AODV.

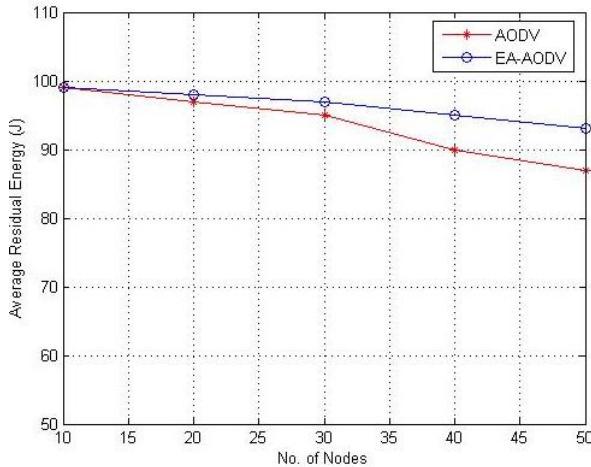


Fig. 4. Variation of Average Residual Energy with No. of Connections for 50 nodes at speed of 10m/sec

7 Conclusion

Energy efficiency has been one of the main problems in MANET routing protocols. The proposed work aims at designing an efficient energy aware routing scheme for MANETs taking variable range transmission into consideration. Simulation results show that EA-AODV shows superior performance as compared to common range AODV in terms of energy consumption and improves network lifetime. Each node's energy state has a big influence on the entire network lifetime. At the time of route selection, EA-AODV takes care of the distance between the nodes choosing nodes with least distance thereby decreasing the energy required to transmit data. Energy consumption at each node is thus improved using variable range transmission in proposed scheme. The average residual energy of the network is increased and also a remarkable improvement in network lifetime is observed even at higher mobility in EA-AODV as compared to AODV. This is due to the transmitter power adjustments done at each node before transmitting the data, which makes utilization of different nodes in the network effective. Future work includes simulation of the proposed scheme for sparse mediums and real life scenarios and also for other metrics like link layer overhead, pause time, path optimality etc.

References

1. Le, A.N., Toh, C.K., Cho, Y.Z.: Load balanced routing protocols for ad hoc mobile wireless networks. *IEEE Communications Magazine* 47(8), 78–84 (2009)
2. Das, S.R., Perkins, C.E., Royer, E.M.: Ad hoc on demand distance vector routing. In: IETF Internet Draft (2001)
3. Chang, J.-H., Tassiulas, L.: Energy conserving routing in wireless ad-hoc networks. In: Proceedings of the Conf. on Computer Communications (IEEE Infocom 2000), pp. 22–31 (2000)
4. Hu, Y., Rongsheng, D., Liu, J.: Ecprpa: An energy efficient routing protocol for ad hoc networks. *Journal of Networks*, 748–755 (2010)
5. Doshi, S., Brown, T.X.: Minimum energy routing schemes for a wireless ad hoc network. In: Proceedings of the Conference on Computer Communications (IEEE Infocom 2002) (2002)
6. Fotino, M., De Rango, F., Marano, S.: Ee-olsr: Energy efficient olsr routing protocol for mobile ad-hoc networks. In: Proceedings of IEEE Military Communications Conference, MILCOM 2008, pp. 1–7 (2008)
7. Simplot-Ryl, D., Ingelrest, F., Stojmenovic, I.: Optimal transmission radius for energy efficient broadcasting protocols in ad hoc networks. *IEEE Transactions on Parallel and Distributed Systems* 17(6), 536–547 (2006)
8. Gomez, J., Campbell, A.T.: Using variable-range transmission power control in wireless ad hoc networks. *IEEE Transactions on Mobile Computing* 6(1), 1–13 (2007)
9. Jung, S., Zelikovsky, S., Hundewale, N.: Energy efficient node caching and load balancing enhancement of reactive ad hoc routing protocols. *Journal of Universal Computer Science* 13, 110–132 (2007)
10. LAN MAN Standards Committee IEEE Computer Society. Wireless lan medium access control and physical layer specifications. IEEE 802.11 Standard (1999)
11. Johnson, D., Broch, J., Maltz, D.: The dynamic source routing protocol for mobile ad hoc networks. In: IETF Internet Draft (1999)
12. Gondal, I., Iqbal, M., Dooley, L.: A novel load balancing technique for proactive energy loss mitigation in ubiquitous networks. In: IEEE International Conference on Consumer Communications and Networks (CCNC 2006), vol. 1, pp. 157–167 (2006)
13. Murthy, C.S.R., Manoj, B.S.: Ad hoc wireless networks: Architectures and protocols. Prentice Hall (2004)
14. Aslam, J., Li, Q., Rus, D.: Online power-aware routing in wireless ad-hoc networks. In: Proceedings of Intl. Conf. on Mobile Computing and Networking (MobiCom 2001) (2001)
15. Hundewale, N., Jung, S., Zelikovsky, A.: Energy efficiency of load balancing in manet routing protocols. In: International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks, pp. 476–483 (2005)
16. Senouci, S.M., Pujolle, G.: Energy efficient routing in wireless ad hoc networks. In: IEEE International Conference on Communication, vol. 5, pp. 1057–1061 (2004)
17. Stojmenovic, I., Lin, X.: Power-aware localized routing in wireless networks. *IEEE Transactions on Parallel and Distributed Systems* 12(11), 1122–1133 (2001)
18. Zeng, C., Jiang, W., Li, Z., Jin H.: Load balancing routing algorithm for ad hoc networks. In: International Conference on Mobile Ad-hoc and Sensor Networks, pp. 334–339 (2009)

Appendix

The transmit power of the node needs to be modified, to vary the transmission range. In the proposed algorithm, free space propagation has been chosen for simulation. Friis transmission equation is used to calculate the transmit power of nodes, given as:

$$P_r = P_t G_r G_t \left(\frac{\lambda}{4\pi R} \right)^2 \quad (1)$$

where, P_r and P_t are received and transmit powers respectively, G_t and G_r are the transmitter and receiver antenna gain, R is the distance between the nodes and λ is the wavelength.

Table 2. Transmit Power for different distances

Distance	P_t
50	-9.2
100	-3.23
150	0.3
200	2.8
250	4.73
400	8.9

For the simulations, G_t and G_r are taken as unity, λ is taken as 0.125 m (at 2.4 GHz operating frequency). The constant received power P_r has been chosen as -84 dBm as IEEE 802.11 [10] MAC protocol specifies a received power range from -81.0 dBm to -110 dBm at transmission bandwidth of 2 Mbps. Thus, based on the measured distance between nodes, P_t can be calculated using equation 1. Table 2 gives some values of P_t with changing R . As seen the transmitter power requirement increases with increase in distance between nodes.

Automatic Speech Recognizer Using Digital Signal Processor

Raghavendra M. Shet¹ and Raghunath S. Holambe²

¹ Department Of Instrumentation technology
B.V.B College of Engineering and Technology
Hubli, Karnataka, India

² Department Of Instrumentation Engineering
S.G.G.S Institute of Engineering and Technology
Nanded, Maharashtra, India

Abstract. The use of speech recognition [2] [3] techniques in many practical applications has demonstrated the need for an Automatic speech Recognizer (ASR), it is a complex machine developed with the purpose to understand human speech. The conventional methods for speech recognition, such as HMM (Hidden Markov Model) and DTW (Dynamic Time Warping), are very complicated and time consuming. To apply Digital Signal Processor TMS320C6713 Digital signal processing Starter Kit (DSK) board is an attempt to implement a laboratory based Automatic speech Recognizer [1]. The proposed approach in this paper simplifies the algorithm using Linear Predictive Cepstral coefficients (LPCC) and Vector Quantization (VQ). The paper includes a performance evaluation of the above techniques on Matlab and application evaluation on DSK board. The database on which the training and testing was carried out is created in-house laboratory under calm and noise free environment.

Keywords: DSK, ΔLPCC and VQ.

1 Introduction

This speech interface is increasingly used in office automation, factory automation and home automation devices. By far, most of these systems are large-scale ones. In this paper an attempt is made to realize an ASR in lab using DSK board to verify its functional capability. The speech recognition system described in this paper is a DSK board containing the speech functions of prompt, playback, training and recognition. The chip integrates a 32 bit TMS320C6713 Digital Signal Processor (DSPs) core chip, on-chip RAM, on-chip ROM, CODEC and other peripheral circuits. The rest of this paper is organized as follows. In Section 2, we describe the heterogeneous architecture of the speech recognition on board, i.e., DSK board. In Section 3, the algorithmic details are described. Evaluation results on the test board are given in Section 4. In the last section, we summarize the performance and application fields of our speech recognition on board.

2 Hardware Architecture

The proposed ASR has a heterogeneous architecture that is composed of an high-performance 225MHz 32-bit Floating-Point Digital Signal Processor (DSPs) [7], AIC23 stereo CODEC, Four Position User DIP Switch and Four User LEDs, On-board Flash and SDRAM. The block diagram is shown in Fig.1. The details of each block are described as below.

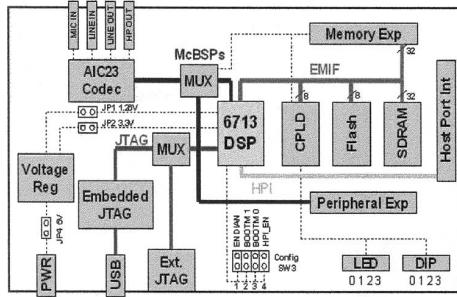


Fig. 1. Block diagram of TMS320C6713 DSK board.

- 32-bit DSP core

It uses a high speed 225MHz floating point processor capable of handling complex DSP algorithm. The 6713 DSP Starter Kit (DSK) is a low-cost platform which lets customers evaluate and develop applications for the Texas Instruments C67X DSP family. It uses a VLIW Core, VLIW is a processor architecture that allows many instructions to be issued (8 on the 6713 DSP) in a single clock while still allowing for very high clock rates [7]. 192Kbytes Internal Memory, 64Kbytes L2 Cache/RAM, 4Kb Program/Data Caches,

- Stereo CODEC

The DSK uses a Texas Instruments AIC23 (part #TLV320AIC23) stereo codec for input and output of audio signals. The codec samples analog signals on the microphone or line inputs and converts them into digital data that can be processed by the DSP with a sampling frequency option of 8 KHz, 16 KHz, 24 KHz, 44.8 KHz and 96 KHz which are sufficient for any speech signal processing implementation. When the DSP is finished with the data it uses the codec to convert the samples back into analog signals on the line and headphone outputs so the user can hear the output.

3 Software Algorithm

3.1 Basic Software Module Using Dsk

The speech recognition using DSK is composed of two modules training module and testing module (speech recognition module). The software algorithm block diagram for both the modules is showed in Fig.2.

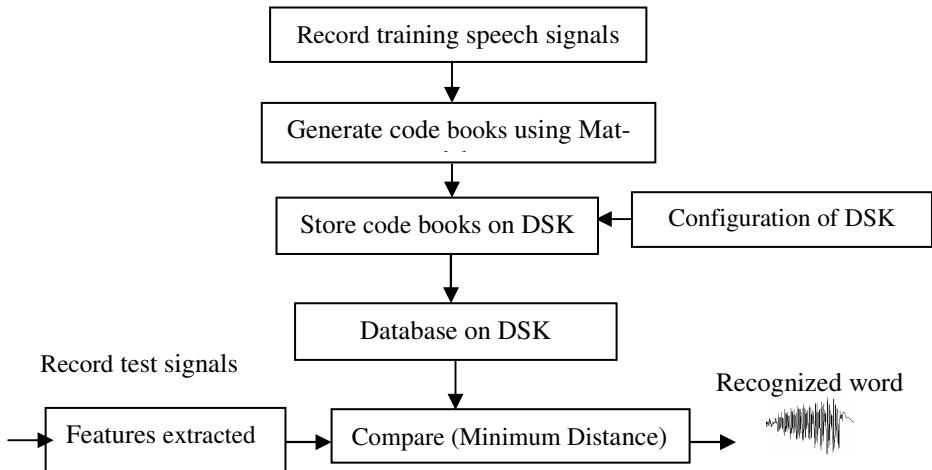


Fig. 2. Speech recognition system Using TMS320C6713 DSK

3.2 Software Algorithm

The conventional front-end process methods of speech recognition are employed. The input signal is sampled at 8k Hz and recorded using Gold Wave v5.12 software to create a data base of our own.

First database “database 1” and “database 2” include three isolated words UP, DOWN and ONE uttered by students with 10 samples of each word for training and testing stage. Second database “new” consists of ONE, TWO.....Ten digits uttered by students with 10 samples of each word for training stage and 20 samples of each word for testing stage.

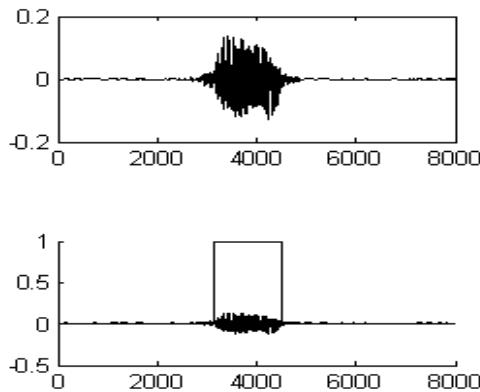
The recorded training speech signals are used in Matlab software to generate codebooks for the above two databases using VQ i.e. 3 codebooks of 16 vectors each for the first database of 3 words and 10 codebooks of 16 vectors each for the second database of 10 words. These generated codebooks are stored in the memory of TMS320C6713 DSK board as files.

3.3 Training Module

3.3.1 Front End Processing

The input signal is pre-emphasized and end point detected to get an active speech signal. Then it is segmented to 15msec blocks windowing using a Hamming window.

During the process, the most crucial step is endpoint detection. It is well known that the performance of a template based word recognition system is very sensitive to the variations of endpoints. Inaccurate endpoint detection will decrease the performance of the speech recognizer. Finding active speech involves finding short time energy estimate E_s , short time power estimate P_s and short time zero crossing rate Z_s . For the speech signal $s_1(n)$ these measures are calculated as follows [4],

**Fig. 3.** End point detection

$$Es_1(m) = \sum_{n=ml}^{(m+1)l} s_1^2(n) \quad (1)$$

$$Ps_1(m) = \frac{1}{L} \sum_{n=ml}^{(m+1)l} s_1^2(n), \quad (2)$$

$$Zs_1(m) = \frac{1}{L} \sum_{n=ml}^{(m+1)l} \frac{|\text{sgn}(s_1(n)) - \text{sgn}(s_1(n-1))|}{2} \quad (3)$$

where

$$\text{sgn}(s_1(n)) = \begin{cases} +1, & s_1(n) \geq 0 \\ -1, & s_1(n) < 0 \end{cases}, \quad (4)$$

Where,

m - is the number of blocks for analysis of speech samples,

l - is number of samples in each block i.e. 125 samples,

n - is the number of samples in the speech signal.

The continuous signal is further blocked/framed into smaller frames(generally 50% overlap is considered).

$$W(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{l-1}\right), 0 \leq n \leq l-1 \quad (5)$$

3.3.2 Feature Extraction

A variety of techniques for this task are available in the literature such as linear predictive coefficients (LPC), Mel cepstrum (MFCC), spectrograph and wavelet packet parameters. The purpose of these modules is to convert the speech waveform to some type of parametric representation. The simplest way of getting this is Linear Predictive Cepstral Coefficients (LPCC) [6].

The main idea behind linear prediction is to extract the vocal tract parameters. A model of the vocal tract can be seen in Fig.4.,

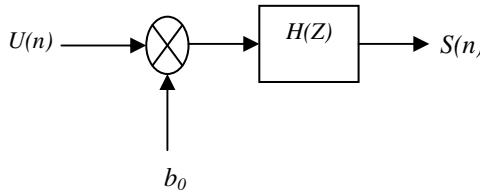


Fig. 4. Model of vocal tract

It uses a autocorrelation method and Levinson-Durbin recursion to extract the tract model. Each frame of windowed signal is auto correlated, where the highest autocorrelation value p , is the order of the linear predictive coefficients (LPC) analysis. Typically value of p is between 8 to 16, and with $p = 8$ being the value used for most systems. The method for converting from autocorrelation coefficients to an LPC parameter set is known as Durbin's method.

The cepstral coefficients are the coefficients of the Fourier transform representation of the log magnitude spectrum, these are more robust, reliable feature set for speech recognition than the LPC parameters.

$$C_0 = \ln \sigma^2 \quad (6)$$

Where, σ^2 is the gain term in the LPC model.

$$C(m') = a_{m'} + \sum_{k=1}^{m'-1} \left(\frac{k}{m'} \right) C_k a_{(m'-k)}, 1 \leq m' \leq p \quad (7)$$

$$\Delta C(m') = \frac{\partial C_{m'}(t)}{\partial t} = \Delta C_{m'}(t) \approx \mu \sum_{k=K}^K K C_{m'}(t+K) \quad (8)$$

Where, $C(m')$ and $\Delta C(m')$ are Cepstral and Temporal Cepstral [6]

3.3.3 Speech Modeling

The feature extracted patterns are further used in testing stage, where they are matched with the features of input test speech signal. In this work VQ [6] is used for generating the reference.

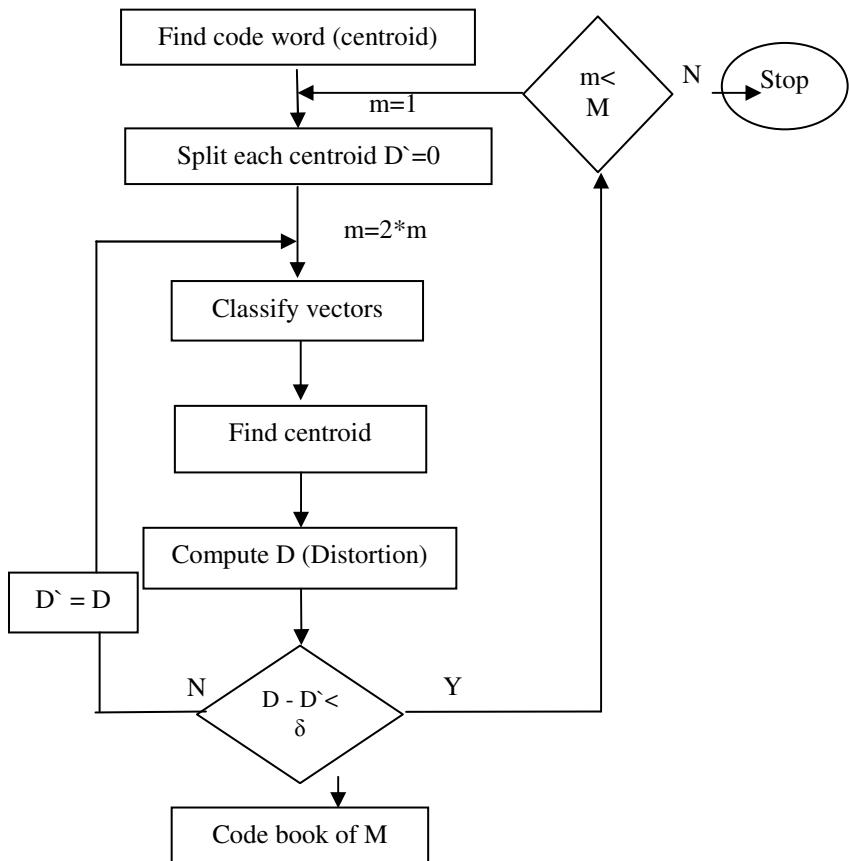


Fig. 5. Flow diagram of binary split code book generation algorithm

3.4 Testing Module

First three blocks are same as training stage, to obtain the test features of speech signal and to form a test feature matrix. The comparison between these features extracted and the reference features is made via a decision rule, the mathematical rule could be in the form of a dissimilarity or distance measure (distortion) that evaluates the

closeness between two vectors i.e. Cepstral distance [6]. This Cepstral distance is averaged minimum distance between the test vector and the reference vectors, and is given by,

$$K_i = \arg \min[d(x, \mu_k)] \quad (9)$$

Where, i is the number of words used in codebook for training and $d(x, \mu_k)$ is the distance between the test feature vector (x) and the reference vector (μ_k).

4 Experimental Results

Here the results are evaluated in two different ways. First, the performance (recognition rate) is tested using Matlab. Testing for first database is performed containing 3 words.

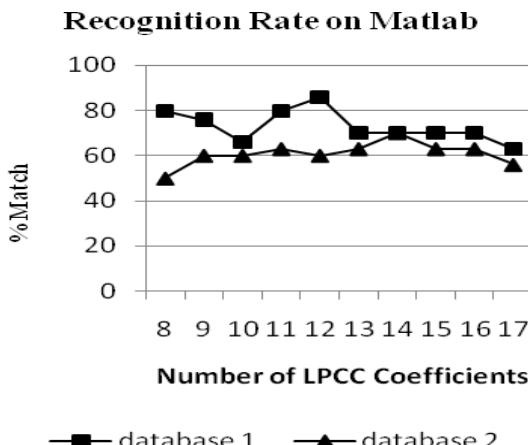


Fig. 6. Evaluation results using Matlab

Second, the functional evaluation of the kit is performed by using test sample T1, T2, T3, D, O and U in “File.h” format individually or through the MICIN connector of the AIC23 codec, the results are verified with the help of corresponding LED on-board and the results are as shown in the Fig.7.

From the above result it is evident that the ASR can be implemented on Laboratory DSK board. The recognition rate depends on the database used and the algorithm used for Feature extraction and modeling.

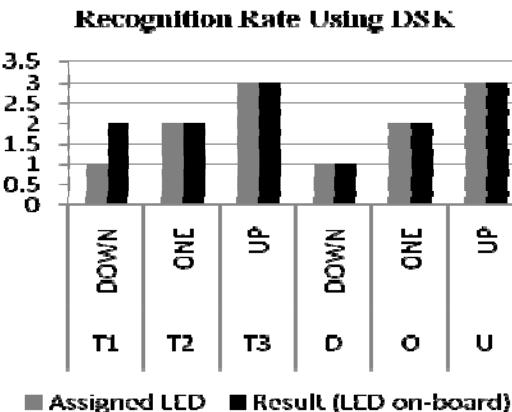


Fig. 7. Evaluation results using DSK

5 Conclusion

This speech recognition approach using DSK can be extended to continuous speech. The performance of the ASR system with microphone variability can be done. Focus on real world robustness i.e. the performance of the system under noisy condition. Speech recognition can be performed on different languages.

An algorithm for performing speech-recognition on a 32-bit DSP core is presented. The main contribution of the work is the use of simple laboratory DSK board for the development of a hardware based speech recognizer. This can be used as an embedded system for control tasks in devices such as a wheel chair for physically challenged persons, voice control commands etc.

References

- [1] Yuanyuan, S., Jia, L., Runsheng, L.: Single-Chip Speech Recognition System Based on 8051 Microcontroller Core. *IEEE Transactions on Consumer Electronics* 47(1) (February 2001)
- [2] Peacocke, R., Graf, D.: An Introduction to Speech and Speaker Recognition. *IEEE Computing Research Laboratory* (August 1990)
- [3] Murveit, H.Y., Brodersen, R.W.: An Integrated-Circuit-Bases Speech Recognition System. *IEEE Transactions on Acoustics, Speech and Signal Processing ASSP- 34(6)* (December 1986)
- [4] Saxena, A., Singh, A.: A Microprocessor Based Speech Recognizer For Isolated Hindi Digits
- [5] Iwata, T., Ishizuka, H., Watari, M., Hoshi, T., Kawakami, Y., Mizuno, M.: Speech Recognition. In: *IEEE International Solid State Circuits Conference* (February 24, 1983)
- [6] Rabiner, L., Jung, B.H.: *Fundamental of Speech Recognition*. Pearson Education (2003)
- [7] Chassaing, R.: *Digital Signal Processing And Applications With The C6713 And C6416 DSK*. Wiley Interscience publication (2005)

- [8] Singh, A., Srinivasan, S.: Digital Signal Processing-implementation Using DSP Microprocessors With Examples From TMS320C54XX. Thomson publication (2006)
- [9] Venkataramani, B., Bhaskar, M.: Digital Signal Processors Architecture, Programming and Applications. Tata McGraw Hill publication (2002)
- [10] TMS320C6000Assembly Language Tools User's Guide, Literature Number: SPRU186N (April 2004)
- [11] TMS320C6000 CPU and Instruction Set Reference Guide, Literature Number: SPRU189F (October 2000)

Pre Decision Based Handoff in Multi Network Environment

Manoj Sharma¹ and R.K. Khola²

¹ R.S., Faculty of Engineering & Technology, M.D.University, Rohtak, Haryana, India
neelmanoj@gmail.com

² Department of Electronics & Communication Engineering
P.D.M College of Engineering, Bahadurgarh, Haryana, India
sharmamanoj_brcm@rediffmail.com

Abstract. Wireless networking is becoming an increasingly important and popular way of providing global information access to users on the move. One of the main challenges for seamless mobility is the availability of simple and robust vertical handoff algorithms, which allow a mobile node to roam among heterogeneous wireless networks. The next generation of mobile networks will support not just simple mobile connectivity but access to evolving smart space environments. In this paper, we propose a novel vertical handoff decision algorithm for overlay wireless networks consisting of heterogeneous wireless environment. This paper presents the proposal of optimal network selection algorithm in wireless heterogeneous environment that is based on TOPSIS method when solving the multi criteria analysis. The target network is selected using TOPSIS based decision algorithm which, in addition to usual parameters, also takes a prediction of the RSS into account

1 Introduction

The next generation of wireless networks is a mixture of heterogeneous wireless technologies. This mix includes networks with different radio access technologies, architectures, protocols, and services [1]. In this the mobile terminals (MTs) are equipped with multiple radio interfaces. Thus, mobile users are capable of switching connections among heterogeneous networks to use the network that better fits their communication and connectivity needs according to the *Always Best Connected* (ABC) concept [2]. Vertical handoff support is responsible for service continuity when a connection needs to migrate across heterogeneous wireless access networks. It generally involves three phases [3], [4]: *system discovery*, *vertical handoff decision*, and *vertical handoff execution*. During the system discovery phase, the mobile terminal (MT) receives advertised information from different access networks. These messages may include their access costs and QoS parameters for different services. In the vertical handoff decision phase, the MT determines whether the current connection should keep using the same network or switch to another one. The decision is based on the information it received during the system discovery phase, and the current state conditions. In the vertical handoff execution phase, the connections are seamlessly migrated from the existing network to another. This process involves authentication,

authorization, and also the transfer of context information. Vertical handoff decision involves a tradeoff among many handoff metrics including quality of service (QoS) requirements (such as network conditions and system performance), mobile terminal conditions, power requirements, application types, user preferences, and a price model. Using these metrics involves the optimization of key parameters (attributes), including signal strength, network coverage area, data rate, reliability, security, battery power, network latency, mobile velocity, and service cost. These parameters may be of different levels of importance to the vertical handoff decision.

2 Related Work

There exist some network selection algorithms in the open literature. Jesus Ruben, Gallardo-Medina et al. [5] proposed the use of the decision scheme named VIKOR for vertical handoff decision in B3G or heterogeneous wireless networks. Jonathan Rodriguez [6] presents a middleware architecture that supports multimedia services across inter-technology radio access networks in a secure and seamless manner. In [7] enhanced media independent handover framework and its mobility management mechanism based on IEEE 802.21 was presented. Eng Hwee Ong and Jamil Y. Khan [8] proposes a novel measurement-based network selection technique to augment handover decision of existing cost function approach, through handover initiation, so that heterogeneity of a dynamic multi-RAT environment can be exploited optimal. In [9] QingHe proposes a fuzzy logic based VHDA in heterogeneous networks. The VHDA first give the vertical handoff decision criteria and then the VHDA is employed to calculate PEVs using fuzzy logic theory. The neuro-fuzzy predictor was used to predict the RSS in cellular network and WLAN in [10]. Based on the fuzzy predictor, it proposed the fuzzy inference mechanism to determine the possibility of handoff according to fuzzy decision algorithm. In [11] multi-criteria vertical handoff decision algorithm for heterogeneous Wireless Wide Area Network (WWAN) and Wireless Local Area Network (WLAN) environment was proposed. In [12] an intelligent approach is used for vertical handover decision is proposed. The intelligence is based on the fuzzy logic. Fuzzy logic is used for network selection and decision making for vertical handover.

3 Predictions of RSS

Here an algorithm is proposed which is used for predetermination of reverse signal strength. The algorithm will help to reduce the call dropping probability in vertical handoff. In this vertical handoff algorithm, the Predictive Reverse Signal Strength (PRss) is used to decide when to start a vertical handoff. If and only if the PRs in the networks fit to the Reverse Signal Strength (Rss) thresholds of the networks, then vertical handoff procedure will be triggered. Thus, it will help in pre determination of reverse signal strength. In this the Mobile Terminal (MT) samples the Rss periodically in the procedure of moving. With a few sampled Rss's stored in the database of MT, PRss can be achieved. This can be achieved as following:

$$r(t+1) = \gamma(t).I(t) + v(t)$$

$$= [\gamma 1(t) \quad \gamma 2(t)][r(t) \quad r(t-1)]^T + v(t) \quad (1)$$

in which, t is the current time; $I(t)$ is the input signal matrix at t , $I(k)=[r(k) \quad r(k-1)]$; $\gamma(t)$ and $v(t)$ describe the predict index matrix at the time t . We adopt the LMS algorithm to reach the optimal predicted index, that is, γ and v are optimal to guarantee the error square minimal between the prediction values of $(t+1)$ and t . The index formulations are present as following:

$$\gamma(t+1) = \gamma(t) + \mu.e(t).I(t)$$

$$v(t+1) = v(t) + \mu.e(t) \quad (2)$$

in which, $e(t)=d(t+1)-r(t+1)$, μ is the fixed step size. An example of Rss prediction is shown in Figure 1 for a particular wireless environment [13]. The algorithm is simple and accurate to predict the Rss trend of future time.

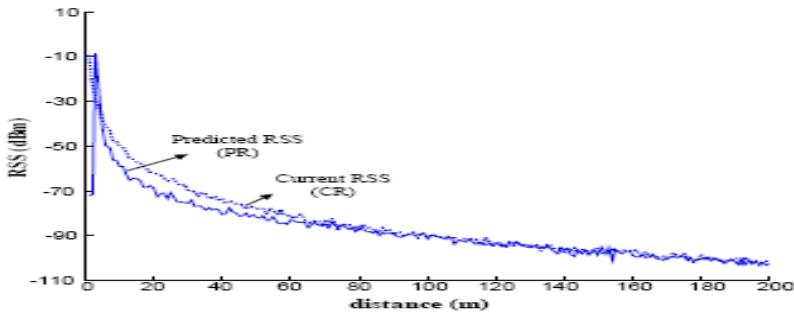


Fig. 1. The curves of Predicted RSS and Current RSS.

4 Proposed Method for Handoff Decision

TOPSIS multi criteria decision method was developed by C. Hwang and K. Yoon in 1981. This concept is based on the theory that the chosen alternative should have the shortest Euclid distance from the ideal solution and the longest from the unideal solution. In this proposed algorithm the following network parameters are chosen for vertical handover decision model: Quality of Service (QoS), Available bandwidth (B), and data transfer cost (C). There may be some other parameters such as Delay, jitter, error rate, battery level etc. which can also be used to decide the network selection.

In order to compare the attributes of different values and different units of measurement it is necessary to use the process of normalization. Normalization is needed

to ensure that the values in different units are meaningful. The normalized function is given by (3)

$$N(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (3)$$

Where $x=Q_oS$, B, C. In (3), RSS, B and C denote the original values; N (Q_oS), N (B), and N (C) are the normalized values.

In this paper we are using user defined subjective weights. Our three weight sets are {0.6, 0.2, 0.2} (weight set 1), {1/3, 1/3, 1/3} (weight set 2), and {0.2, 0.6, 0.2} (weight set 3), respectively.

The flow chart for the proposed algorithm is shown in figure 2. The network that provides the highest rank value is selected as the best network to handoff from the current access network according to the mobile terminal conditions, network conditions, service and application requirements, cost of service, and user preferences.

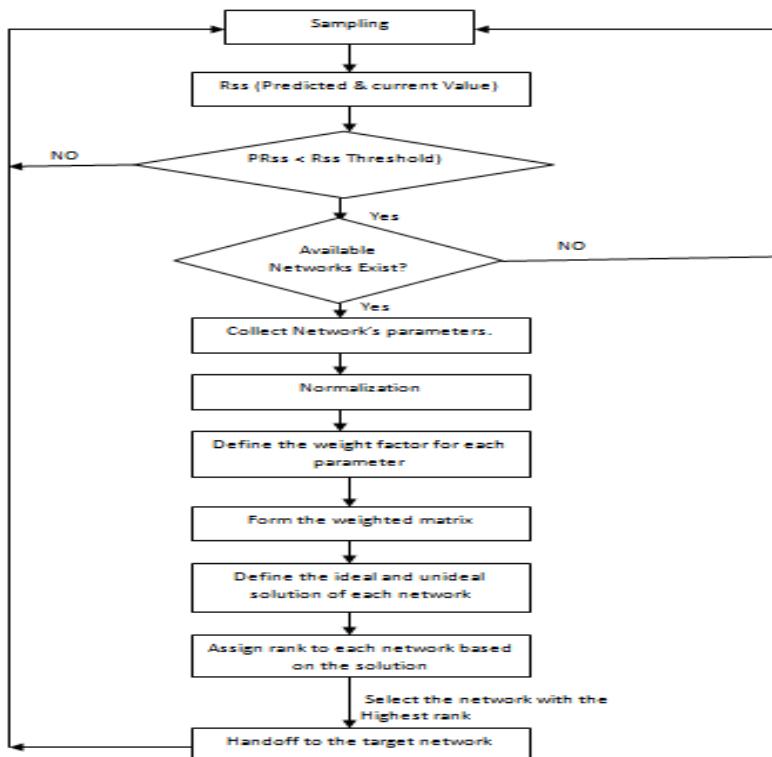


Fig. 2. Flow Chart of Proposed Algorithm.

Weight factors have been defined and now we can continue with TOPSIS method. First we have to create the weighted matrix $\|M_{ij}=W_j.N_{ij}\|$.

Where W_j is the weight factor of j^{th} parameter and N_{ij} is the normalized value of j^{th} parameter for i^{th} network.

The ideal solution is given by the set

$$I^+ = \{(\max M_{ij} | j \in \{1,2,3\})\} \quad (4)$$

and the unideal solution is given by the set

$$I^- = \{(\min M_{ij} | j \in \{1,2,3\})\} \quad (5)$$

With the help of ideal and unideal solution the Euclid alternative distance can be calculated as

$$D_i^+ = \sqrt{\sum_{j=1}^3 (M_{ij} - I_j^+)^2} \quad (6)$$

And

$$D_i^- = \sqrt{\sum_{j=1}^3 (M_{ij} - I_j^-)^2} \quad (7)$$

Now the rank of the networks is calculated by

$$D_i = \frac{D_i^+}{D_i^+ + D_i^-}, D \in (0,1) \quad (8)$$

from (8), the best network is the one with the largest relative closeness to the ideal solution.

5 Experimental Results and Analysis

The results of proposed algorithm are shown through three different cases for the four available access networks. According to the first case (Table 1), the user's preferable parameter is QoS and for this preference the weight factor W is 0.6. After the result of simulation (implementation of the algorithm) network 4 is declared as optimal network. Network 4 is also optimal network in terms of data transfer cost.

Table 1. Case 1

	QoS	Bandwidth	Cost
Normalized Matrix			
Network 1	.30	.67	.12

Table 1. (*continued*)

Network 2	.82	.33	.29
Network 3	.45	.11	.24
Network 4	.89	.43	.35
Weight W	.6	.2	.2
Weighted Matrix			
Network 1	.18	.13	.02
Network 2	.49	.06	.05
Network 3	.27	.02	.04
Network 4	.53	.08	.07

In the second case (Table 2) the user has defined equal weight factor $W=1/3$ to each parameter. After the result of simulation network 3 is found to be the optimal network. The network 3 is better in terms of QoS and Bandwidth. The results of the testing that presents this case is shown in Table 2.

Table 2. Case 2

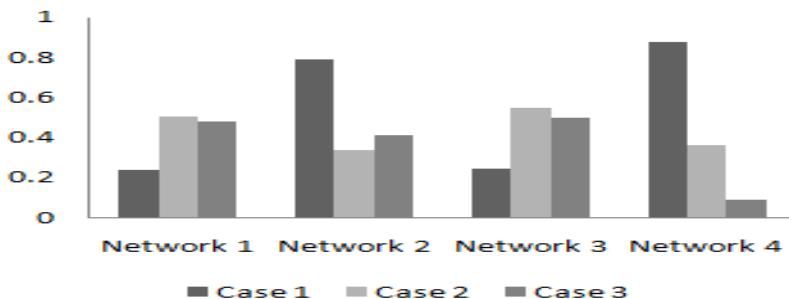
	QoS	Bandwidth	Cost
Normalized Matrix			
Network 1	.13	.43	.90
Network 2	.56	.19	.30
Network 3	.78	.76	.20
Network 4	.19	.27	.65
Weight W	1/3	1/3	1/3
Weighted Matrix			
Network 1	.04	.14	.30
Network 2	.18	.06	.10
Network 3	.26	.25	.06
Network 4	.06	.09	.21

For the third case (Table 3) the user has defined priority to bandwidth. The weight factor for the bandwidth $W=0.6$ and for QoS and Cost the weight factor is equal and its value is 0.2. After the result of simulation the network 3 is declared as the optimal network because it provides much better bandwidth than other networks. Although the network 1 is optimal for QoS and network 2 is optimal for Cost but here these parameters are of less importance.

Table 3. Case 3

	QoS	Bandwidth	Cost
Normalized Matrix			
Network 1	.93	.38	.24
Network 2	.26	.45	.65
Network 3	.33	.62	.19
Network 4	.11	.40	.22
Weight W	.2	.6	.2
Weighted Matrix			
Network 1	.18	.22	.04
Network 2	.05	.27	.13
Network 3	.06	.37	.03
Network 4	.02	.24	.04

Figure 3 shows the rank of each of the networks for the three different cases. From this it is seen that N/W 4 is optimal for the first case, N/W 3 in second case and N/W 3 for the third case.

**Fig. 3.** Graph showing the rank of the networks for each case

6 Conclusion

In this paper a method for network selection in heterogeneous wireless environment is proposed. A method for Rss prediction is proposed. With the help of Rss prediction, handoff procedure can be started in advance & it will also reduce the call-dropping probability. After this an algorithm for network selection is proposed. We have chosen QoS, bandwidth & Cost as the parameters on which the handover decision is taken. The algorithm proposed here for the optimal network selection is based on TOPSIS method. We have also tested some cases with the proposed algorithm and based upon it optimal network was selected.

References

- [1] Akyildiz, I., Xie, J., Mohanty, S.: A Survey of Mobility Management in Next-Generation All-IP-Based Wireless Systems. *IEEE Wireless Communications* 11(4), 16–28 (2004)
- [2] Gustafsson, E., Jonsson, A.: Always Best Connected. *IEEE Wireless Communications* 10(1), 49–55 (2003)
- [3] McNair, J., Zhu, F.: Vertical Handoffs in Fourth-generation Multinetwork Environments. *IEEE Wireless Communications* 11(3), 8–15 (2004)
- [4] Chen, W., Liu, J., Huang, H.: An Adaptive Scheme for Vertical Handoff in Wireless Overlay Networks. In: Proc. of ICPAD 2004, Newport Beach, CA (July 2004)
- [5] Gallardo-Medina, J.R., Pineda-Rico, U., Stevens-Navarro, E.: VIKOR Method for Vertical Handoff Decision in Beyond 3G Wireless Networks. In: Proc. of 6th IEEE International Conference on Electrical Engineering, Computing Science & Automatic Control, January 10-13 (2009)
- [6] Rodriguez, J.: A Middleware Architecture Supporting Seamless and Secure Multimedia Services across an Inter technology Radio Access Network. *IEEE Wireless Communications* 16(5), 24–31 (2009)
- [7] Wang, Y., Zhang, P., Zhou, Y., Yuan, J., Liu, F., Li, G.: Handover Management in Enhanced MIH Framework for Heterogeneous Wireless Networks Environment. *Wireless Personal Communications* 52(3), 615–636 (2010)
- [8] Ong, E.H., Khan, J.Y.: On Optimal Network Selection in a Dynamic Multi-RAT Environment. *IEEE Communication Letters* 14(3), 217–219 (2010)
- [9] He, Q.: A Fuzzy Logic Based Vertical Handoff Decision Algorithm between WWAN and WLAN. In: 2nd IEEE Conference on Networking & Digital Society, May 30-31, pp. 561–564 (2010)
- [10] Lin, C.J., Tsai, I.-T., Lee, C.Y.: An adaptive fuzzy predictor based handoff algorithm for heterogeneous network. *IEEE Annual Meeting of the Fuzzy Information* 12, 944–947 (2004)
- [11] Sharma, M., Khola, R.K.: An Intelligent Handover Decision System for Multi Network Environment. *Journal of Electrical & Electronics Engineering* 9(2), 1067–1072
- [12] Sharma, M., Khola, R.K.: An Intelligent Approach for Handover Decision in Heterogeneous Wireless Environment. *International Journal of Engineering* 4(5), 452–462
- [13] Xia, L., Jiang, L.-G., He, C.: A Novel Fuzzy Logic Vertical Handoff Algorithm with Aid of Differential prediction and Pre-Decision Method. In: IEEE National Conference on Communications, June 24–28, pp. 5665–5670 (2007)

A Swarm Inspired Probabilistic Path Selection with Congestion Control in MANETs

Subhankar Joardar¹, Vandana Bhattacherjee², and Debasis Giri¹

¹ Department of CSE, Haldia Institute of Technology

² Department of CSE, Birla Institute of Technology

subhankarranchi@yahoo.co.in vbhattacharya@bitmesra.ac.in,
debases_giri@hotmail.com

Abstract. Congestion is a major problem in mobile ad-hoc networks. In mobile ad hoc networks congestion creates delay in transmission and also loss of the packet that causes wastage of time and energy on recovery. The Wireless Networks have to play an important role to adopt and execute a large no of innovative application. New challenges have come considering the major limitations of the ad hoc network like node's limited processing power, balance the load of network (to maintain the computation of the node). To overcome the above problem some algorithm is invoked and there may be huge amount of packet loss and this leads to decrease the lifetime of the network. The objective of our proposed algorithm is to identify the congestion areas between source and its neighboring nodes to the destination and thus it will help to avoid the congestion of the network in the intermediate links and also minimize the packet loss in the network. Using a new mathematical model considering the swarm-based ant intelligence concept, we found an efficient congestion control mechanism (Ant's probabilistic transition rule).

Keywords: Ad hoc networks, Conditional probability, Congestion, Distance vector, Swarm intelligence, Transmission queue length.

1 Introduction

Mobile Ad hoc Networks are movable and also able to communicate within themselves using a wireless physical medium where there is no need of pre existing network infrastructure. These networks can form stand-alone groups of wireless terminals, which can be again a part of any other cellular system or a fixed network. The main characteristic which draws the attention is that they are able to configure themselves without a centralized administration.

The Mobile ad hoc network has several drawbacks that are not found in fixed networks. The frequent change in network topology due to the mobility of the nodes causes a great deal of control information onto the network. The small capacity of batteries and the bandwidth limitation of wireless channels are other factors. The scalability is also a major factor because the network performance degrades quickly as the no of nodes increases.

1.1 Routing in Mobile Ad-Hoc Wireless Networks

Specially configured routing protocols are engaged in order to establish routes between nodes which are more than a single hop. The ability to trace routes in spite of a dynamic topology is the unique feature of these protocols. These protocols can be categorized into two main types: Reactive (On-demand) and Proactive (Table-driven). Evaluating the routes continuously within the network is done by proactive protocols, so when a packet needs to be forwarded the route is already known and can be immediately used. Reactive protocols appeal to a route determination procedure on demand only.

1.2 Congestion in Mobile Ad-Hoc Wireless Networks

In mobile ad hoc networks congestion occurs due to the shared wireless channel and dynamic topology, packet transmissions suffer from interference and fading. The network load is burdened through the transmission errors. There is an increasing demand of multimedia communications in MANETs so large amount of real-time traffic involves high bandwidth and it is liable to congestion. Congestion leads to packet losses and bandwidth degradation and also wastes time and energy on congestion recovery.

2 Related work

The continuous research on congestion minimization for MANET still needs more new techniques. This section will illustrate the research related to congestion control.

A distributed multi-path DSR protocol (MPDSR) was developed to enhance the quality of service [1]. This protocol forwards the outgoing packets using multiple no of paths and with point to point reliability. Split Multi-path Routing protocol [2] uses a per-packet allocation technique to distribute the data packets into multiple paths which prevents nodes from being congested in heavily loaded traffic conditions.

A novel technique Congestion Adaptive Routing in Mobile Ad Hoc Networks (CRP) [3] was proposed. In CRP each node on a route in the network sends a warn message to its previous node when it is to be congested. Then the previous node uses a new route to bypass the congestion and get the first non congested node on the route. Traffic will be distributed probabilistically over these routes and there will be less chance of congestion occurrence.

A dynamic load aware based load-balanced routing for ad hoc networks (DLBL) algorithm was developed. The DLBL [4] uses intermediate node routing load metric that helps the protocol to discover a route with less network congestion. When a link fails because of the node movement the algorithm uses the path maintenance to re join the broken links to get a route from the source to the destination.

Another novel algorithm proposed is a Congestion-aware routing metric (CARM) [5] which considers Mac-overhead, data-rate and buffer queue delay to select low congestion with high throughput routes. Another mechanism that CARM applied was

the use of link data-rate categories to prevent the mismatched link data-rate. CARM performed locally.

An ant based DSR [6] will inform the source node about QoS available to destination node such as acceptable delay, jitter and energy in the case of multimedia and real time application. They proposed DSR using ACO and called AntDSR(ADSR).

Another novel algorithm proposed a QoS-aware routing protocol [7] which consider two scheme admission control and feedback for the Qos requirements. Their major works on the approximation of bandwidth estimation to calculate the rest amount of bandwidth available at each node.

A biologically inspired routing algorithm [8] for multi-hop networks considers the concept of stigmergy and evaluates the optimal and sub-optimal routes without the whole system-wide connectivity information. They also incorporate MAC layer information into the routing decision can prevent the forward data traffic into the congested areas under situation like various nodal mobility, network density and data loads.

3 Proposed Work

In this section we will discuss Swarm intelligence, Distance metric and the transmission Queue length.

3.1 Swarm Behavior

This algorithm is inspired by the foraging behavior of biological ant colony [9], when they find paths to food sources. Each ant deposits a chemical called pheromone when they move from source to the destination and the foragers follow such pheromone trails. By that more ants are attracted by these pheromone trails and in turn reinforce them even more. This natural phenomenon (stigmergy) plays a role in developing and manipulating local information.

3.2 Congestion Metric

The major parameters to control congestion areas probabilistically are the two metrics in a node.

3.2.1 Distance Metric

The first metric is the distance between the single hop by using the Time of arrival [10] estimating the time of signal traveling between nodes (TOA) provides the distance between two nodes. We can use ultrasound sensor for measuring range between nodes. For this purpose very accurate rubidium-based oscillators must be used. In order to calculate the TOA parameter the nodes must have a common clock, or exchange timing information by certain protocols such as two-way ranging protocol.

The distance metric d_{ij} where j is the neighbor node of i .

$$d_{ij} = C * (d_{\text{remote}} - d_{\text{local}})/2, \quad C \approx 3 * 10^8 \text{ m/s} \quad (1)$$

The time between starting the transmission of a data packet and receiving the corresponding immediate acknowledgement will refer as remote delay(d_{remote}) and receiving one data packet and sending out the immediate acknowledgement will be the duration as local delay(d_{local}). C is measured as the speed of light.

3.2.2 Transmission Queue Length Metric

The second metric is a heuristic value transmission queue length in the i_{th} node towards the neighboring nodes. The transmission queue length [8] can affect the packet latency and packet drop because of the size of the packet length.

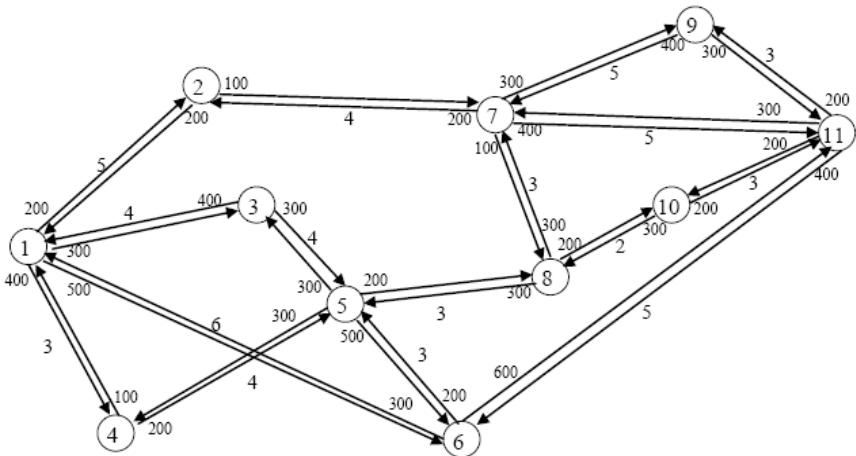


Fig. 1. Mobile nodes with distance and out going transmission queue

$$T_{i,j} = 1 - \frac{q_{i,j}}{\sum_{j \in N_b^{[i]}} q_{i,j}} \quad (2)$$

where q_{ij} is the outgoing queue length in terms of bits waiting to be sent to the link between i and j and N is all the nodes in the network and $N_b^{[i]}$ is the neighboring nodes of the node i . We have considered some random values as transmission queue in terms of bits. Using (2), we find the goodness of heuristic values means small queue data length will have higher heuristic value. From the figure 1, all the nodes have symmetric link between them. We consider some hypothetical value as distance and transmission queue length, using (1), we found the congestion metric $T_{i,j}$ in the table 1 [8].

Table 1. Goodness of congestion metric $T_{i,j}$

$j \Rightarrow$ $i \Downarrow$	1	2	3	4	5	6	7	8	9	10	11
1		.85	.78	.71		.64					
2	.33						.66				
3	.42				.57						
4	.66				.33						
5			.76	.76		.61		.84			
6	.66				.77						.33
7		.80						.90	.70		.66
8					.62		.62			.75	
9							.42				.57
10								.40			.60
11						.63	.72		.81	.81	

3.3 Ant Based Probabilistic Path Selection

When a source node s , needs to route a destination d without knowing the global topology, using the initial pheromone information available the ant will chose its next hop with probability $P_{i,j}$. We have proposed an algorithm to find the distance of neighbour node from the i_{th} node and also we calculate the pheromone deposition, pheromone evaporation and residual pheromone on the link $l_{i,j}$.

$$\Delta\tau_{ij}(t) = mvl_{i,j} * \frac{1}{d_{i,j}}, j \in N_b^{[i]} \quad (3)$$

The distance $d_{i,j}$ from the node i to its neighboring nodes can be calculated by using the equation 1, we assume some random distance metric for our algorithm. $mvl_{i,j}$ is the minimum distance selected from the neighbor nodes of i . The $\Delta\tau_{ij}(t)$ is the reciprocal of the distance $d_{i,j}$ on the link. We found that the less distance have more reciprocal value of rate of pheromone deposition and vice versa.

$$\tau_{ij}^d(t) = \tau_{ij}(t) + \Delta\tau_{ij}(t)\rho \quad (4)$$

$$\tau_{ij}^e(t) = \tau_{ij}^d(t)(1 - \rho) \quad (5)$$

$\tau_{ij}(t) = 0.1$ is the initial pheromone deposition and the ρ [11] specifies the rate at which pheromone evaporates means ant “forget” previous decision for value $\rho = 1$

the pheromone evaporates rapidly and search is random, when $\rho=0$ results in slower evaporation rates, $\rho \in [0,1]$. We take $\rho = 0.6$. The final pheromone deposition is $\tau_{ij}^d(t)$. The final pheromone evaporates from the link is $\tau_{ij}^e(t)$. The residual pheromone on the link is.

$$\tau_{ij}^r(t) = \tau_{ij}^d(t) - \tau_{ij}^e(t) \quad (6)$$

Using (3) : $\Delta\tau_{1,2}(t) = 3*1/5 = .6$, $\Delta\tau_{1,3}(t) = 3*1/4 = .75$, $\Delta\tau_{1,4}(t) = 3*1/3 = 1$,

$$\Delta\tau_{1,6}(t) = 3*1/6 = .5$$

Using (4) : $\tau_{1,2}^d(t) = .1 + (.6 * .6) = .46$, $\tau_{1,3}^d(t) = .1 + (.75 * .6) = .55$, $\tau_{1,4}^d(t) = .1 + (1 * .6) = .7$, $\tau_{1,6}^d(t) = .1 + (.5 * .6) = .5$

Using (5) : $\tau_{1,2}^e(t) = .46 * (1 - .6) = .18$, $\tau_{1,3}^e(t) = .55 * (1 - .6) = .22$,

$$\tau_{1,4}^e(t) = .7 * (1 - .6) = .28$$
, $\tau_{1,6}^e(t) = .4 * (1 - .6) = .16$

Using (6) : $l_{1,2}(t) = .46 - .18 = .28$, $l_{1,3}(t) = .55 - .22 = .33$, $l_{1,4}(t) = .7 - .28 = .42$,

$$l_{1,6}(t) = .4 - .16 = .24$$

The link l_{ij} from node i shows when the distance is less compared to other neighboring nodes the more pheromone deposition occurs. The nodes residual pheromone deposition metric in the table 2.

Table 2. Residual pheromone metric $\tau_{ij}(t)$

$j \Rightarrow$ $i \Downarrow$	1	2	3	4	5	6	7	8	9	10	11
1		.28	.33	.42		.24					
2	.42						.35				
3	.42				.42						
4	.42				.33						
5			.33	.33		.42		.42			
6	.33				.42						.20
7		.33						.42	.28		.28
8					.30		.30			.42	
9						.28					.42
10								.42			.30
11						.20	.28		.42	.42	

When the pheromone information are available the ant will chose its next hop with probability $P_{i,j}$. We adopt this concept by using Baye's probability theorem

$$P_{i,j}(t) = \frac{\left[\tau_{ij}(t) \right]^\alpha \left[T_{ij} \right]^\beta}{\sum_{j \in N_b^{[i]}} \left[\tau_{ij}(t) \right]^\alpha \left[T_{ij} \right]^\beta}, \quad \text{for } i = 1 \text{ to } N, \quad (7)$$

where $\alpha \geq 0$ is the relative importance of the pheromone trail. $\beta \geq 0$ is the relative importance of the congestion. α and β are the two tuneable parameter will control relative weight of pheromone trail and heuristic value T_{ij} .

In the following, we calculate the probability corresponding to the node i using j.

$$P_{1,2} = (.28^3 * .85^1) / (.28^3 * .85^1) + (.33^3 * .78^1) + (.42^3 * .71^1) + (.24^3 * .64^1) = .17$$

$$P_{1,3} = (.33^3 * .78^1) / (.28^3 * .85^1) + (.33^3 * .78^1) + (.42^3 * .71^1) + (.24^3 * .64^1) = .25$$

$$P_{1,4} = (.42^3 * .71^1) / (.28^3 * .85^1) + (.33^3 * .78^1) + (.42^3 * .71^1) + (.24^3 * .64^1) = .48$$

$$P_{1,6} = (.24^3 * .64^1) / (.28^3 * .85^1) + (.33^3 * .78^1) + (.42^3 * .71^1) + (.24^3 * .64^1) = .08$$

Analogously, we can calculate the probability corresponding to other nodes which are shown in table 3.

Table 3. Probability metric on l_{ij}

i, j	$P_{i,j}(t)$	$P_{i,j}(t)$
	$\alpha > \beta$	$\alpha < \beta$
1,2	.17	.31
1,3	.25	.28
1,4	.48	.27
1,6	.08	.11
2,1	.46	.13
2,7	.53	.86
3,1	.42	.28
3,5	.57	.71
4,1	.80	.91
4,5	.19	.08
5,3	.16	.22
5,4	.16	.22
5,6	.27	.15
5,8	.38	.39
6,1	.28	.32
6,5	.68	.65

i, j	$P_{i,j}(t)$	$P_{i,j}(t)$
	$\alpha >$	$\alpha <$
6,11	.03	.02
7,2	.23	.26
7,8	.53	.48
7,9	.12	.15
7,11	.10	.09
8,5	.18	.22
8,7	.18	.22
8,10	.62	.55
9,7	.17	.21
9,11	.82	.78
10,8	.64	.29
10,11	.35	.70
11,6	.03	.08
11,7	.11	.17
11,9	.42	.37
11,10	.42	.37

3.4 Algorithm P²scm : Node i Processing

Algorithm P²scm()

Input:

d_{ij} : Distance between the nodes.
 q_{ij} : Transmission queue length.
 α : Relative importance of pheromone deposition.
 β : Relative importance of congestion.
 ρ : Rate at which pheromone evaporates.

Output:

$P_{i,j}(\tau_{ij}(t))$: Goodness of Congestion
 $P_{i,j}(T_{i,j})$: Residual pheromon

BEGIN

1. Update parameter for Probabilistic path selection
2. $\tau_{ij}(t) \leftarrow$ Pheromone residual on the link.
3. $T_{i,j} \leftarrow$ Probability of congestion it discards.
4. **if** ($\alpha > \beta$) **then**
5. **if** ($\tau_{ij}(t) > \tau_{ij \in N_b}(t)$) **then**
6. Max.Prob_{i,j}($\tau_{ij}(t)$)
7. **else**
8. **if** ($\max(\tau_{ij}(t)) = \tau_{ij \in N_b}(t)$) **then**
9. Max.Prob_{i,j}($T_{i,j}$)
10. **end if**
11. **end if**
12. **else**
13. **if** ($\alpha < \beta$) **then**
14. **if** ($T_{i,j} > T_{i,j \in N_b}$) **then**
15. Max.Prob_{i,j}($T_{i,j}$)
16. **else**
17. **if** ($\max(T_{i,j}) = T_{i,j \in N_b}$) **then**
18. Max.Prob_{i,j}($\tau_{ij}(t)$)
19. **end if**
20. **end if**
21. **end if**
22. **end if**
23. **END**

3.5 Description of the Algorithm

Our proposed p²scm algorithm presents the pseudo code of how a node i process the out going queue and by using swarm intelligence how it will select the path probabilistically to its neighbor node j .

The node i first calculate the distance d_{ij} between all its neighbor nodes. By using the smallest distance $mvl_{i,j}$ the node i evaluate the relative distance between all its neighbor node. P²scm is also evaluate the $\Delta\tau_{ij}(t)$ which is reciprocal to the distance d_{ij} where the relative pheromone deposition occurs on the link $l_{i,j}$. Now p²scm evaluate the residual pheromone on the link $l_{i,j}$ from the line 2-9 of the above algorithm.

Secondly p²scm evaluate the probability of congestion $(1 - \frac{q_{i,j}}{\sum_{j \in N_b^{[i]}} q_{i,j}})$ it discards

in the line 10.

In our proposed algorithm from line no 11-19 when $\alpha > \beta$ it first checks the pheromone ($\tau_{ij}(t)$) deposition in a particular link have the highest value within its neighbor or not. If found yes then the maximum probability $P_{i,j}$ will be on that link $l_{i,j}$. Next it checks a pair of paths with highest deposition and same value, if found yes within that, less congested ($T_{i,j}$) path will have the highest probability.

Similarly from line no 20 -28 when the $\alpha < \beta$, if $T_{i,j}$ have the highest value within its neighbor then the highest probability will be on that link $l_{i,j}$. Next it checks a pair of paths with highest and same value of $T_{i,j}$ or not if found yes within that, more pheromone ($\tau_{ij}(t)$) deposited path will have the highest probability on that link $l_{i,j}$.

4 Analysis

P²scm at first from node i and the two tunable parameter $\alpha > \beta$ we observe the higher pheromone on the link l_{ij} will have higher probability of path selection and higher goodness of congestion means small queue data length will have low probability, if the deposition is same on both the link, we observe then p²scm by using $\alpha > \beta$ and conditional probability evaluates the less congested path means queue data length is small . In our example node 3 to its neighbor node 1 and 2 from table 1, table 2 and table 3.

Also by making $\alpha < \beta$ p²scm consider the less congested path probabilistically means when the queue data length is same on both the link, the higher probability comes with higher pheromone concentration on the link and if the congestion is same within a pair of paths then it evaluates the paths which have more pheromone deposition.

By using two tunable parameters p²scm decides the path probabilistically which have less congested or more pheromone deposition when the nodes have dynamic topology at real time situation.

5 Conclusion

In this paper, we propose a new mathematical model for congestion control, here we have developed a single hop congestion aware probabilistic path selection algorithm which employs the swarm intelligence of biological ant. By probabilistic approach this algorithm can prevent forward data traffic into the congested areas. If more congestion occurs in all the links of a node, the algorithms by using the tunable parameter that can shift the mode of transmission and can easily avoid the congested area. By using the above stated mathematical model and some hypothetical data we tried to control the congestion at local node level.

References

1. Leung, R., Jilei, L., Poon, E., Chan, A.C., Baochun, L.: MPDSR: a QoS-aware multi-path dynamic source routing protocol for wireless ad-hoc networks. In: Proc. of 26th Annual IEEE Conference on Local Computer Networks, pp. 132–141 (2001)
2. Sung, J.L., Gerla, M.: Split Multi-path Routing with Maximally Disjoint Paths in Ad Hoc Networks. In: Proc. Int. IEEE Conf. on Comm., pp. 3201–3205 (2001)
3. Tran, D., Raghavendra, H.: Congestion Adaptive Routing in Mobile Ad Hoc Networks. *IEEE Transactions on Parallel and Distributed Systems* 17, 1294–1305 (2006)
4. Zheng, X., Guo, W., Liu, R., Tian, Y.: A New Dynamic Load-aware Based Load-balanced Routing for Ad Hoc Networks, pp. 407–411. IEEE (2004)
5. Chen, X., Haley, M.J., Jayalath, A.D.S.: Congestion-Aware Routing Protocol for Mobile Ad Hoc Networks. In: Proceedings of IEEE Conference on Vehicular Technology, pp. 21–25 (2007), doi: 10.1109/VETECF.2007.21
6. Asokan, R., Natarajan, A.M., Venkatesh, C.: Ant Based Dynamic Source Routing Protocol to Support Multiple Quality of Service (QoS) Metrics in Mobile Ad Hoc Networks. *International Journal of Computer Science and Security* 2(3), 48–56 (2008)
7. Chen, L., Heinzelman, W.B.: QoS-Aware Routing Based on Bandwidth Estimation for Mobile Ad Hoc Networks. *IEEE on Selected Areas in Communications* 23(3) (2005)
8. Zhenyu, L., Kwiatkowska, M.Z., Constantinou, C.: A Biologically Inspired Congestion Control Routing Algorithm For MANETs. In: IEEE PerCom 2005 (2005)
9. Beckers, R., Deneubourg, J.L., Goss, S.: Trails and u turns in the selection of the shortest path by the ant *lasius niger*. *Journal of Theoretical Biology* 159, 397–415 (1992)
10. Günther, A., Hoene, C.: Measuring round trip times to determine the distance between WLAN nodes. In: Boutaba, R., Almeroth, K.C., Puigjaner, R., Shen, S., Black, J.P. (eds.) *NETWORKING 2005. LNCS*, vol. 3462, pp. 768–779. Springer, Heidelberg (2005)
11. Jawahar, A.C.S.: Ant Colony Optimization for Mobile Ad-Hoc Networks,
ajaychak@eden.rutgers.edu

A Hierarchical CPN Model for Mobility Analysis in Zone Based MANET

Moitreyee Dasgupta, Sankhayan Chaudhury, and Nabendu Chaki

University of Calcutta, India

92 APC Road, Kolkata 700009, India

moitreyeed@yahoo.com, sankhayan@gmail.com, nabendu@ieee.org

Abstract. The crucial decision issues in MANET like load balancing or link discovery after an existing link breakage for routing are dependent on the mobility of the participating nodes. These decisions will be accurate if the changes within the network due to mobility can be apprehended efficiently. In this paper, we have proposed a formal modeling approach to assess the changes in terms of node density in a zone based network. A two layered Colored Petri net is proposed for describing the mobility of a single node and the resulting changes within a network due to this mobility. The model is expected to provide useful information through analysis and that helps a lot for taking these critical decisions.

Keywords: mobility analysis, MANET, Hierarchical CPN, Colored Petri net.

1 Introduction

Mobile Ad hoc network (MANET) is based on wireless multi hop architecture without fixed infrastructure and prior configuration of the nodes. In general, the scalability issue for these networks are being addressed by the zone (cluster) based structure [1] where entire network is divided into number of zones. One distinguished node is responsible for maintaining the activities within a zone i.e. the other participant nodes of the zone must be registered under that distinguished node for communication. The most challenging issue for these networks is the prediction of the mobility of the individual nodes. One node in a specific zone may move to some adjacent zone and needs registration under a new zonal authority. As a result the node-density within the zones is changing with time. These frequent changes within a network affects various crucial issues such as routing, resource allocation etc. Moreover, this dynamic topology due to mobility of the nodes makes MANET more prone to frequent link breakage. The objective of this work is to propose a model through which the changes in zonal scenario due to mobility of the individual nodes can be analyzed. The general approach for analysis of ad hoc network is to use the existing simulation tools but the result of simulating an algorithm may be different, depending on the selected tool, because of important divergence between simulators [8]. Thus here we have used the alternative approach of formal modeling through Colored Petri net. Here the changes in node-density of the clusters within network can be predicted by the proposed one. One can analyze the model for apprehend the rate

of change of node density within zones and these outcome can be effectively used for optimized distribution of bandwidth or routing decision after link breakage.

2 Related Work

Colored Petri Nets (CPN) [2] is a discrete-event modeling language combining Petri nets with the functional programming language Standard ML. Petri nets provide the foundation of the graphical notation and the basic primitives for modeling concurrency, communication, and synchronization. Standard ML provides the primitives for the definition of data types, describing data manipulation, and for creating compact and parameterized models. A CPN model of a system is an executable model representing the states of the system and the events (transitions) that can cause the system to change state. CPN Tool [2] is an industrial-strength computer tool for constructing and analyzing CPN models. Using CPN Tools, it is possible to investigate the behaviour of the modeled system using simulation, to verify properties by means of state space methods and model checking, and to conduct simulation-based performance analysis.

High order Petri nets are the class of Petri nets in which Petri nets themselves are considered as tokens or first-class objects. Here the Petri nets may be the values of parameters and variables as well as the results are computation performed. K. Hoffmann & T. Mossakowski [3] has proposed the concept of Algebraic Higher Order Nets, which allow to have dynamic tokens like graphs or (ordinary low-level) Petri nets. This is an extension of Petri nets including the concept of the higher-order algebraic specification language HasCasl [4]. J. W. Janneck & R. Esser [5] has explored the usefulness of some of the high order Petri net models and their application domain and illustrates them with small and medium sized application examples.

M. A Azagomi & Ali Khalili [6] proposed a CPN model for modeling and performance evaluation of a energy aware MAC protocol for wireless sensor network named S-MAC. Using the hierarchical capabilities of CPN, a wireless environment has been modeled & performance has been evaluated. Here, the proposed hierarchical CPN can grab much more details of the process and increases flexibility & scalability for deploying a model in wireless scenario, but no model checking has been performed.

An Object oriented Petri net has been used by A. Masri et al. [7] for modeling & performance analysis of the Distributed Co-ordinated Function (DCF) of IEEE 802.11b. Along with the different classes of the Petri net, a detailed model for DCF procedure for IEEE 802.11b has been presented which precise back-off procedure and time synchronization. Simulation of the model has been done for dense network with different number of workstations. The results have been compared with the ns2 simulation results to prove the correctness and quality of the proposed OOPN model.

M. Rajaratnam & F. Takawira [9] has modeled hand-off traffic for micro-cellular boundaries where mobile may request for hand-off many times at many different cellular boundary and therefore undergo a series of hand-offs that may alter their traffic profile. Hand-off has been modeled using the mean & variance of its probability distribution to make a descent attempt to capture the change in the traffic

profile that occurs during the hand-off process. Through simulation they have shown that hand-off traffic is smooth traffic process and two moment representation of hand-off traffic is superior to the single moment representation. The proposed model is not complete in the sense that they have only considered the cell-blocking mode.

A channel allocation schemes have been proposed by Rita Jain [10] to improve bandwidth efficiency, whereas handoff management scheme, based on bandwidth reservation, has been proposed to guarantee a low connection dropping rate. In this paper author has taken both the issues together to provide guaranteed QoS and effective utilization of the bandwidth. S. Chinara and S. Rath [1] has proposed a modeling methodology of dynamic topology adaptive clustering algorithm for mobile ad hoc network using Colored Petri net which enhances the network life time by handling the node energy in an efficient manner.

In this paper, an attempt has been made to model the mobility of nodes within a zone based MANET divided in several zones. The model is used to study and analyze the changes in the density of the nodes due to this mobility. The results of analysis can be used efficiently for taking decision about demand-based bandwidth allocation, improving fault tolerance, etc.

3 The Proposed CPN Model

Here, the node density in the different zones of the network is being analyzed by modeling individual node mobility with the help of CPN. We assume that the deployment area is divided into nine different zones called East (E), West (W), North (N), South (S), NE, NW, SE, SW, and Centre (C) as shown in figure-1. Every node can move to its adjacent zones based on the direction of movement. The activities in each zone are governed by a Zone Representative node for that zone.

NW (Violet)	N (Indigo)	NE (Blue)
W (Green)	C (Yellow)	E (Orange)
SW (Red)	S (Pink)	SE (Black)

Fig. 1. Different Zones of the Network Deployment Area

It is assumed that a GPS device is attached to every node so that each node can find out the location of itself. The topology of the network in MANET changes dynamically due to mobility. The nodes may reach out of service area of its current Zone Representative, and thus need to register itself to new Zone Representative(s) of the adjacent zone(s).

In this paper, a CPN based 2-layer hierarchical model has been proposed to represent the above scenario. The proposed model would be useful towards predicting the node density and therefore in making decisions for load-based bandwidth allocation, routing, and building fault tolerance into the network. The lower layer deals with the individual node mobility based on its color, velocity and direction,

while the upper layer models the rate of change of registration process with different zone heads. Thus these two layers collectively will be able to model the change in scenario in terms of node density within a network.

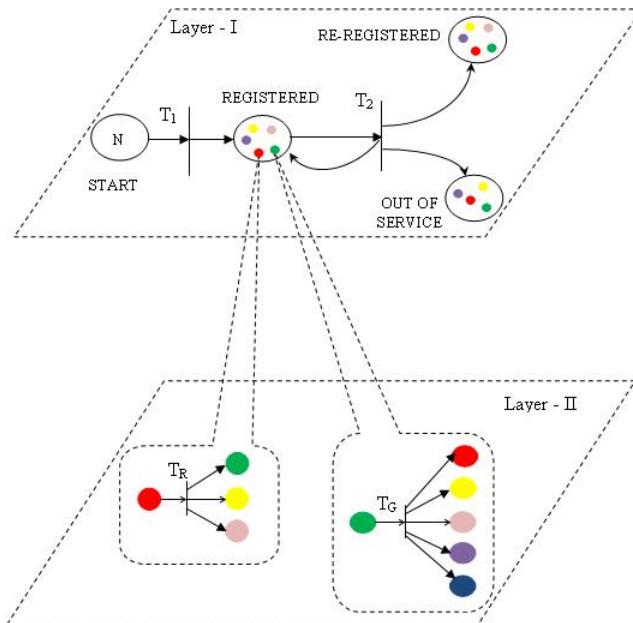


Fig. 2. The 2-layer CPN Model for a Zone based MANET

In layer-2 of the proposed CPN model, there are nine (9) different places corresponding to nine (9) zones. Each node in the network is represented by a token belonging to one and exactly one of these nine (9) places at any given instance. All tokens belonging to a place are of the same color. Thus each zone is associated with a particular color. This association has been illustrated in figure 1. Thus label of a zone as described above and its color both can be used to imply the nodes belonging to that zone. In rest of the paper we have used zone labels and zone colors in an interchangeable manner; e.g. West (W) zone is same as the Green zone, and similarly the SE zone is identical with the Black zone. Whenever a node is moving from one zone to another and thus registering under a different Zone Representative, the node will take the color of the zone it is joining.

In the top layer, there are only 4 places. This is invariant of the total number of nodes in the network and their mobility pattern. The START place does not distinguish between tokens in terms of colors. The place will have exactly N tokens where N is the total number of nodes in the network. This place actually does not represent the dynamics of a MANET and is used only to keep the model bounded. When a node is associated with a particular zone based on its geographic position, the token gets a color value as explained earlier. In each of the rest three places in the top layer, there would be at most nine (9) colored tokens. A token for a particular color

appears in the REGISTERED place, as soon as there is at least one node in the corresponding zone of the network. Each of the tokens in the REGISTERED place in the top layer represents a CPN subnet in layer-2 with a single transition. The color of the input place for the transition in the subnet would be same as the color of the token in the top layer. The subnet thus represents the movement of nodes from this color zone to its neighbouring zones. This has been demonstrated in figure 2.

Each token in the top layer, excepting those in the start place, is labelled like $\langle Color, Count \rangle$, which stand for the node color and the number of tokens of that color respectively. Thus if there are ten nodes in the SW zone then in layer 2, there would be ten tokens in the place corresponding to Red color. The corresponding token in the REGISTERED place in layer-1 would be labelled as $\langle Red, 10 \rangle$.

We assume that the session begins with N nodes in the network that are yet to be registered with any Zone Representative. Thus initially there would be N tokens in the START place and none of the remaining 3 places would have any token. The 9 places in layer 2 are yet to get any token. As mentioned earlier, the START place has been incorporated in the model just to ensure boundedness in the proposed model for a network with a finite number of nodes. The initial marking for the entire model would thus be $(N, 0, 0, 0)$ for layer-1 and $(0, 0, 0, 0, 0, 0, 0, 0, 0)$ for layer-2.

4 Transitions and the Firing Rule

There would be two transitions T_1 and T_2 in the top layer and 9 transitions $T_V, T_I, T_B, T_G, T_Y, T_O, T_R, T_P, T_{BL}$ in layer-2. Each transition in layer 2 has a distinct input place for a particular color. Transition T_1 is fired every time a node in the system registers with some Zone Representative after a new session starts. A token would be consumed from the START place and the *count* of a token in the REGISTERED place would be increased by 1 depending on the zone in which the node is registered. After registration of all the nodes in the network, the total of the count values for all the nine tokens in REGISTERED place would be N. Here, we assume that enough bandwidth and resources are available initially for registration of each node in the network. The transition T_2 fires every time a node migrates to a different zone. The REGISTERED place forms both an input and output place for T_2 . The *count* of the token for the color zone that the node is leaving would be reduced by 1 and the count for the token for the color zone that it is migrating to would be increased by 1. Transition T_2 has two more output places. If the node successfully registers with the Zone Representative of the new zone, then token count in the RE-REGISTERED place would be increased by 1, otherwise count for the OUT-OF-SERVICE would be increased. For layer-2 each zone has a different CPN subnet with single input place and the number of output places depends on the position of the color zone corresponding to the input place. For each subnet there is only one transition, firing which would result in possible change in zone for a mobile node. Transitions are marked by T_V, T_I, T_B and so on where, V, I, B are the corresponding color code of different zone as shown in figure 1. Every time a node moves from its input zone to any other adjacent zone, the transition fires and the token count of the input place is reduced by 1 while incrementing the token count of any of the output places.

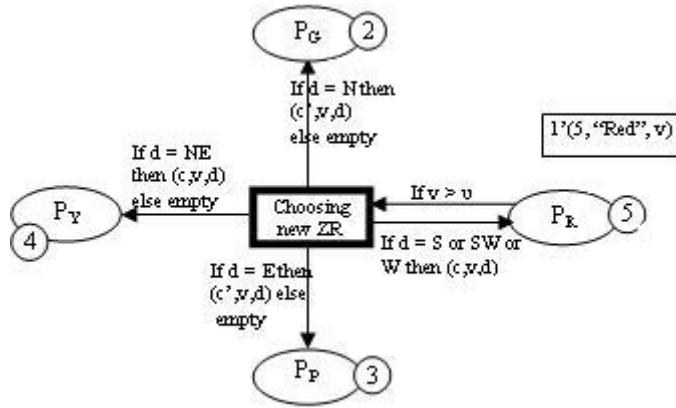


Fig. 3. Layer 2 CPN model of individual node mobility of place P_E

Suppose we are considering the CPN subnet of Red zone and Green zone. The input place for Red zone is P_R and the three output adjacent places according to figure 1 are P_G , P_Y and P_P . The Green zone has five output places against one input place and these are P_R , P_Y , P_P , P_V and P_I as shown in figure 2. At any point of time if the marking for Red zone is $(5, 2, 4, 3)$ and suppose a token moves from Red zone to Yellow zone, then the resulting marking would be $(4, 2, 5, 3)$.

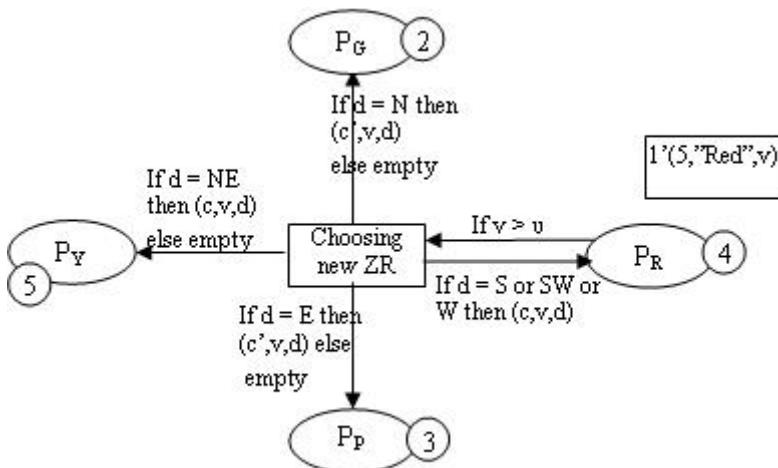


Fig. 4. Marking change of CPN model for P_E

The CPN model for Red zone is shown in figure 3. Figure 4 depicts the marking change after the transition fires. The arc expressions on the input arcs of transitions together with the tokens on the input places determine whether the transition is enabled. So, if the required tokens are present in the input place, tokens will migrate to the output place based on the results of the arc expressions.

The formal CPN model for layer 2 is shown in figure 5. We assume the initial marking is $(3, 0, 0, 0)$. After successive firing of transition T_1 , i.e., when all the nodes gets registered under some zone representative based on their location, the marking of the place changed to $(0, 3, 0, 0)$. Now suppose nodes from P_B and P_p get successfully migrated to orange and yellow zone and the token from P_R failed to shift to P_Y . The resulting marking would be then $(0, 3, 2, 1)$. From here we can conclude that Yellow has already reached to the saturation zone for the bandwidth allocation and max capacity of that zone is 1. The transition T_2 has a guard function “if band < α ” where α is the required bandwidth for the node to migrate while band is the available bandwidth for that zone. The guard function must evaluate to true for the arc expressions to be enabled.

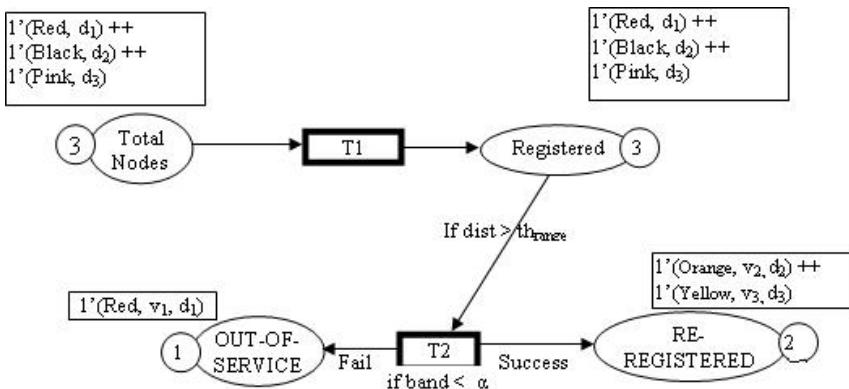


Fig. 5. Formal CPN model for layer 1

Let's now discuss the mobility of nodes within and beyond a zone. Suppose a node X is in the Orange zone. Now, if the node moves in the east direction, then it will continue to be in the current zone until it exceeds the transmission range. If the movement is in north direction, X may enter the Blue zone and the transition will be to PN. After entering the Blue zone, node X is likely to be registered with the Zone Representative of the Blue zone. However, due to some reason like limited bandwidth, etc., the registration may be denied. In the lower layer, there would still be a token for node X in PN after the transition fires. In the upper layer, the count of blue token will increase either in the RE-REGISTERED place or in the place corresponding to OUT OF SERVICE place, depending on the actual status of node X in the blue zone. If the number of tokens in the Orange zone was n before this migration, then the new label for the orange token in REGISTERED place would be $<Orange, n-1>$.

If the node is re-registered in the Blue zone, then the count of Blue token in the RE-REGISTERED place would be increased by 1. On the contrary, if the re-registration for X fails in the Blue zone then the count for Blue token in the OUT OF SERVICE place would be increased by 1 instead. In an ideal case, when all the nodes are re-registered with the Zone Representative of the target zone for each migration, then there would be no token at all in the OUT OF SERVICE place.

In figure 2, there is a Pink token in the RE-REGISTERED place while there is no Pink token in the OUT OF SERVICE place. This is to be interpreted as all the nodes in the Pink (South) zone are registered with the Zone Representative. The other possible places where node X may migrate from the Orange zone are NE, C, S and SE based on the direction of the movement. The mobility pattern of other nodes of rest of the eight places can be modeled in the same way.

5 Performance Analysis

Lemma 1: *At any given instance of time, the count of token for a particular color token in the REGISTERED place is equal to the sum of the count of tokens of the same color in the RE-REGISTERED and OUT OF SERVICE places.*

Proof: Let's prove this by the method of contradiction.

Let's assume that the counts for a particular color token are $cnt1$, $cnt2$, and cnt in the RE-REGISTERED place, OUT OF SERVICE place, and in the REGISTERED place respectively.

Let's first assume that $cnt > cnt1 + cnt2$. This means that the total number of nodes that physically exists in a zone at the given time is greater than number of nodes either registered with the Zone Representative or has failed to do so. Thus there has to be some node in the zone which is neither registered nor has it failed registration.

However, this is in conflict with the rule of firing for the CPN as described earlier. Thus the assumption fails. It is inferred that,

$$cnt \neq cnt1 + cnt2 \dots \quad (1)$$

Let's next assume that $cnt < cnt1 + cnt2$. This means that the total number of nodes that physically exists in a zone at the given time is less than number of nodes either registered with the Zone Representative or has failed to do so. It is trivial to state that this too is in conflict with the rule of firing for the proposed CPN. Thus the assumption fails. It is inferred that,

$$cnt \neq cnt1 + cnt2 \dots \quad (2)$$

Therefore, from equations 1 and 2, we conclude that $cnt = cnt1 + cnt2$.

Lemma 2: *For any finite integer N , if the initial marking for layer-1 CPN model is $\mu = (N, 0, 0, 0)$ and network deployment area is divided to K -zones, then the proposed model is K -bounded if $K > N$, otherwise network is N -bounded.*

Proof: According to our initial marking, the START place is N -bounded. Now every time transition T_1 fires 1 token shifts from START place to REGISTERED place. So, the REGISTER place can not have more than n token, thus also bounded by N . A token in the places named REGISTER, RE-REGISTER and OUT-OF-SERVICE means at least one token of corresponding color zone, the max number can not exceed the total number of zone which is finite. Successive shift of token by transition T_2 will only increase the token count of the token of corresponding zone. Thus REGISTER, RE-REGISTER and OUT-OF-SERVICE places are K -bounded. Therefore, the proposed CPN model is K -bounded if $K > N$, otherwise network is N -bounded.

From the definition of safeness, we can say that the model is also K-safe or N-safe based on the value of these two variables.

6 Conclusion

Here the colored hierarchical Petri net is presented to provide a formal model about the mobility of the participating nodes in a zone based mobile ad-hoc network. The model is able to reflect all the relevant events to mobility those are necessary to feel the changes within a network. Our objective is to model these typical events as these events are the key issues on which most of the important decision depends in a MANET. The model is verified and proved to be bounded. CPN tool can be used for estimating the values after simulation and these analyses will be helpful for taking various critical decisions.

References

1. Chinara, S., Rath, K.S.: Modeling of a Topology Adaptive Clustering Algorithm for Mobile Ad Hoc Networks using Coloured Petri Nets. World Academy of Science, Engineering and Technology 69 (2010)
2. Jensen, K., Kristensen, S.L., Wells, L.: Coloured Petri net and CPN Tool for modelling and validation for concurrent systems. International Journal on Software Tools for Technology Transfer (2007)
3. Hoffmann, K., Mossakowski, T.: Algebraic higher-order nets: Graphs and petri nets as tokens. In: Wirsing, M., Pattinson, D., Hennicker, R. (eds.) WADT 2003. LNCS, vol. 2755, pp. 253–267. Springer, Heidelberg (2003)
4. Schröder, L., Mossakowski, T.: HASCASL: Towards Integrated Specification and Development of Functional Programs. In: Kirchner, H., Ringeissen, C. (eds.) AMAST 2002. LNCS, vol. 2422, pp. 99–116. Springer, Heidelberg (2002)
5. Janneck, W.J., Esser, R.: Higher-order Petri net modelling –techniques and applications. Application and Theory of Petri nets: Formal Methods in Software Engineering and Defence Systems 12 (2002)
6. Azgomí, A.M., Khalili, A.: of Sensor Medium Access Control Protocol Using Coloured Petri Nets. Electronic Notes in Theoretical Computer Science 242, 31–42 (2009)
7. Masri, A., Bourdeaud'huy, T., Toguyeni, A.: Performance Analysis of IEEE 802.11b Wireless Networks with Object Oriented Petri Nets. Electronic Notes in Theoretical Computer Science 242, 73–85 (2009)
8. Cavin, D., Sasson, Y., Schiper, A.: On the Accuracy of MANET Simulators. In: Proc. of POMC, pp. 38–43. ACM Press (2002)
9. Rajaratnam, M.G., Takawira, F.: Hand-off Traffic Modelling in Cellular Networks. In: Global Telecommunications Conference, GLOBECOM (1997)
10. Jain, R.: Integrating Distributed Channel Allocation and Handoff Management for Cellular Networks. In: IET-UK International Conference on Information and Communication Technology in Electrical Sciences, ICTES 2007 (2007)

Sensor Deployment for Mobile Object Tracking in Wireless Sensor Networks

Yingchi Mao and Ting Yin

College of Computer and Information, Hohai University, Nanjing, 210098, China
maoyingchi@gmail.com

Abstract. Mobile object tracking is a key application of wireless sensor network based surveillance systems. Sensor deployment is an important factor in tracking performance. In this paper, we consider the problem of optimal sensor deployment for mobile object tracking using the node sequences model without considering the sensing quality and the specific movement trace. We determine that there are three patterns that can reduce tracking error rate: regular tiling pattern, random pattern, and irregular pattern. We analyze the properties of these patterns and provide an upper bound of the tracking error rate for each pattern in the worst case scenario. We also provide theoretical analysis and simulation evaluations to demonstrate that the irregular pattern outperforms the other two patterns.

Keywords: sensor networks, deployment pattern, node sequences.

1 Introduction

The emergence of the wireless sensor network (WSN) has made it possible to detect and track the presence of a malicious intruder in a region of interest [1]. WSNs for mission-critical applications (such as target detection, object tracking, and security surveillance) often face the fundamental challenge of meeting stringent performance requirements imposed by users. For instance, a surveillance application may require any intruder to be detected and traced with a high accuracy. Therefore, the positions at which sensors are placed play an important part in affecting detection performance of a WSN. However, finding the optimal sensor deployment for target tracking is challenging due to the uncertainty in physical environments, sensing irregularity and unknown mobile target movement trace.

Unlike previous work, this paper focuses on an optimal deployment pattern to improve the tracking performance. A tracking scheme with node sequences is the basis of our work. Tian He et al. proposed a robust framework for tracking mobile targets using unreliable node sequences [2]. Without assuming the movement pattern or noise model, and without accurate range-based localization, target tracking is accomplished by processing node sequences, which can be easily obtained by ordering related sensor nodes according to their sensing

results of the mobile target. In fact, the detected node sequences reflect the relative distance relationships among the target and the sensor nodes within the known position. In their paper, the authors only discuss the design of the tracking scheme, but do not mention whether different deployment patterns affect the tracking performance. They also do not mention how to optimally place the sensors to enhance the tracking accuracy. Therefore, this paper discusses the sensor deployment problem for target tracking using the node sequences model.

2 Related Work

Sensor node deployment has been studied intensively in the literature [3]- [5]. The optimizing objectives of deployment include coverage, connectivity, power-saving, target detection and tracking, and security.

For example, different deployment patterns achieving full coverage and k -connectivity under different ratios of the sensor communication range to the sensing range for homogeneous wireless sensor networks were studied [5]. In [6], two deployment patterns, random and deterministic, in the situation of node failure and placement errors, are analyzed. Also, three patterns of node deployment in terms of coverage, energy consumption, and message transfer delay were investigated [7]: uniform random, square grid, and tri-hexagon tiling. However, those patterns of node deployment for mobile target tracking were not addressed.

The most related work is [3], [4]. Their optimizing objective of deployment is object detection. Two recent works about optimal sensor placement based on data fusion for object detection schemes have been proposed. In [3], the authors address the following issues: What is the best way to deploy sensors in order to meet the detection requirements in a mean squared sense, while maintaining a specified false alarm probability? They presented an optimal control theory and model the system as a linear quadratic regulator with the locations serving as control parameters to solve this problem. Fast sensor placement algorithms based on a probabilistic data fusion model based on sensor data fusion are proposed in [4]. The authors formulate the sensor placement problem for fusion-based target detection as a constrained optimization problem.

3 Preliminaries

In this section, we briefly introduce the target tracking scheme with node sequence, and describe the network model and problem formulation.

3.1 Main Idea of Node Sequences

The proposed tracking of mobile targets framework by Tian He et al. [2] includes three parts: map diving and neighborhood graph building, detection node sequences, and tracking with unreliable node sequences processing.

First, after the deployment of sensor nodes, the monitored area can be divided into lots of small regions, called faces, according to the positions of the sensor

nodes. Given two sensor nodes with known positions, the whole area can be divided into two parts by the perpendicular bisector. By the rules of geometry, every position point in the gray area under bisector $Div(1, 2)$ is closer to node 1 than to node 2. The division of the map is based on the fact that, ideally, the geographic distance between a sensor node and the target has a monotonic impact on the sensing readings. So, the signature sequence S_f of a face f reflects the geographic distance between a sensor node and the target. For example, in Fig. ??(b), face f_1 has a signature node sequence of $S_{f_1} : (S_1, S_2, S_3)$, we have $\forall p \in f_1, dis(p, S_1) > dis(p, S_2) > dis(p, S_3)$, where $dis(p, S_i), i = 1, 2, 3$ denotes the distance between point p and sensor S_i .

3.2 Problem Formulation

Given N sensor nodes that can be deployed in a finite, two-dimensional planar region A , A can be divided into lots of small faces by the perpendicular bisector $Div(i, j)$ between any two node pairs (S_i, S_j) . In order to reduce the tracking error rate with node sequences, the goal of the optimal sensor deployment pattern is to minimize the largest area of a face, denoted MAX_{face} , and minimize the area variance of all the faces, denoted VAR_{face} . The average area of all the faces is denoted $MEAN_{face}$. Thus, the optimal sensor deployment problem is transformed to optimal line arrangement problem. In the next section, we analyze the properties of map division under the regular, random and irregular patterns, and give the corresponding upper bound of MAX_{face} .

4 Three Deployment Patterns Categories

4.1 Regular Pattern

Following Grunbaum and Shephard [9], a tiling is said to be *regular* if the symmetry group of the tiling acts transitively on the *flags* of the tiling, where a flag is a triple consisting of a mutual incident vertex, edge, and tile of the tiling. This means that for every pair of flags, there is a symmetry operation mapping the first flag to the second. This is equivalent to the tiling being an edge-to-edge tiling by congruent regular polygons. There must be six equilateral triangles, four squares and three hexagons at a vertex, yielding the three regular tessellations: triangular tiling, square tiling and hexagonal tiling, respectively. In this section, we discuss the results of map division in three regular patterns.

Triangular Tiling. For a single triangular tiling, the area can be divided into 6 identical faces by three vertices. The area of each face is $Area(face) = \sqrt{3}/24d^2$; d denotes the distance between two neighboring sensor nodes. In the following part, d is set to 10 meters, as the unit distance between two neighboring nodes. For a single triangle tiling, we have $MAX_{face} = 7.2169$, $MEAN_{face} = 7.2169$, and $VAR_{face} = 0.0$.

For multiple triangular tiling, we discuss the results of map division with 16, 25, and 36 nodes.

- 16nodes: $MAX_{face} = 5.1095$, $MEAN_{face} = 0.9259$, and $VAR_{face} = 1.0081$.
- 25nodes: $MAX_{face} = 3.9690$, $MEAN_{face} = 0.2566$, and $VAR_{face} = 0.3608$.
- 36nodes: $MAX_{face} = 2.9211$, $MEAN_{face} = 0.0906$, and $VAR_{face} = 0.1501$.

Square Tiling. For multiple square tiling, we discuss the results of map division with 16, 25 and 36 nodes.

- 16 nodes: $MAX_{face} = 7.5000$, $MEAN_{face} = 1.2931$, and $VAR_{face} = 1.3590$.
- 25 nodes: $MAX_{face} = 4.1667$, $MEAN_{face} = 0.3722$, and $VAR_{face} = 0.5241$.
- 36 nodes: $MAX_{face} = 3.1250$, $MEAN_{face} = 0.1312$, and $VAR_{face} = 0.2110$.

Hexagonal Tiling. For multiple square tiling, we discuss the results of map division with 16, 25 and 36 nodes.

- 16 nodes: $MAX_{face}=10.8253$, $MEAN_{face} = 1.2778$, and $VAR_{face} = 1.5298$.
- 25 nodes: $MAX_{face} = 8.6045$, $MEAN_{face} = 0.4764$, and $VAR_{face} = 0.7399$.
- 36 nodes: $MAX_{face} = 6.4952$, $MEAN_{face} = 0.1641$, and $VAR_{face} = 0.3013$.

4.2 Random Pattern

Given one monitored region A , N sensor nodes are randomly deployed. Using the random pattern is the simplest way to deploy large-scale sensor systems. Similarly, we discuss the results of map division with 16, 25, and 36 nodes.

- 16 nodes: $MAX_{face} = 4.9527$, $MEAN_{face}=0.1214$, and $VAR_{face} = 0.4405$.
- 25 nodes: $MAX_{face} = 2.6284$, $MEAN_{face} = 0.0303$, and $VAR_{face} = 0.0653$.
- 36 nodes: $MAX_{face} = 1.2517$, $MEAN_{face} = 0.0212$, and $VAR_{face} = 0.0410$.

From the numerical results shown in Fig. 2, it can be seen that the maximum area and area variance of faces using the random pattern are both smaller than those using triangular tiling. Therefore, the tracking scheme using the random deployment pattern can have higher tracking accuracy. In addition, for the purpose of deployment convenience, it is easier to place sensor nodes using the random pattern rather than using other regular patterns.

Now, the question is whether the random pattern is the optimal pattern for target tracking using node sequences. Due to the uneven distribution of sensor nodes, some faces are larger and some are smaller. To solve the uneven distribution problem in the random pattern, we propose two kinds of irregular patterns combining the properties of the random pattern and the regular pattern.

4.3 Irregular Pattern

To reduce the tracking error rate using the node sequence tracking scheme, the deployment pattern should minimize the area of the largest face as well as area variance of all faces. Based on Arrangement of Lines theorem, the deployment pattern should ensure that the arrangement of bisectors among any two node pairs in the monitored region is simple.

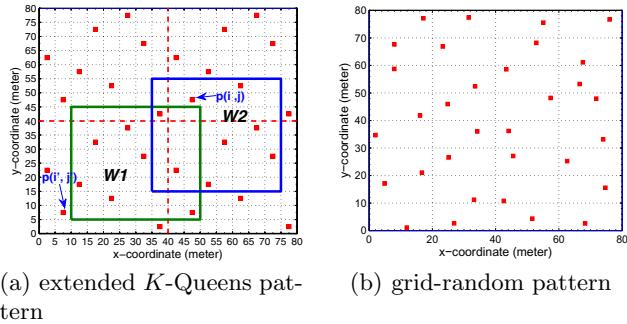


Fig. 1. Examples of irregular pattern.

Definition 1. (*Independent Set*): Given one finite, two-dimensional planar region A and N sensor nodes, if the arrangement of bisectors between any two sensor node pairs is simple, we call the set of nodes as an *IndependentSet*. In other words, there are no three bisectors intersecting at a common point and no two bisectors are parallel.

On the other hand, the independent set is not well-distributed, which generates the uneven map division. Target tracking using node sequences can result in a higher estimation error rate of real traces. Therefore, the optimal sensor deployment pattern should minimize the area of the largest face as well as maintain the uneven distribution of faces.

In this paper, two irregular patterns are proposed: extended N-Queens pattern and grid-random pattern.

N-Queens Pattern In this subsection, we discuss the map division performance of the N -Queens pattern.

Definition 2. (*N-Queens Pattern*): Given that one finite, two-dimensional planar region A is divided into a $N \times N$ chessboard, N sensor nodes are placed in such a way that two nodes share the same row, column, or diagonal.

Similarly, we give the results of map division with 16, 25, and 36 nodes using the N -Queens pattern.

- 16 nodes: $MAX_{face} = 3.2115$, $MEAN_{face} = 0.1035$, and $VAR_{face} = 0.2209$.
- 25 nodes: $MAX_{face} = 1.8635$, $MEAN_{face} = 0.0283$, and $VAR_{face} = 0.0692$.
- 36 nodes: $MAX_{face} = 0.9992$, $MEAN_{face} = 0.0156$, and $VAR_{face} = 0.0378$.

For the large-scale sensor networks, it is difficult to compute and deploy N sensor nodes in the monitored field using the N -Queens Pattern. In real applications, map division can be done locally with sensing range instead of including all the sensors. So, the extended K -Queens pattern is proposed.

Definition 3. (*Extended K-Queens Pattern*): Given one finite, two-dimensional planar region A and N sensor nodes, A is divided into $m = \lceil N/K \rceil$ sub-regions. Each sub-region is further divided into a $K \times K$ chessboard, and K sensor nodes are placed with the same K -Queens pattern in each sub-region.

Given region A and one window W with the size of the $K \times K$ chessboard, N sensor nodes are deployed by the extended K -Queens pattern. When moving window W to any direction in A , the positions of sensor nodes in W are still in the K -Queens pattern. For example, 32 sensor nodes are deployed using the extended 8-Queens pattern in the 80×80 meter² unit grid. A is divided into four sub-regions, and every 8 sensor nodes are placed in each sub-region with the same 8-Queens pattern. The size of window W is 40×40 meter². As shown in Fig. 1(a), no two nodes share the same row, column, or diagonal in W_1 and W_2 , respectively.

Theorem 1. Given N sensor nodes deployed by the extended K -Queens pattern in the monitored region A , moving one window W with the size of $K \times K$ to any direction in A , the positions of sensor nodes in W are still in the K -Queens pattern.

Proof: Assume $\forall p(i, j) \in A$, $p(i, j) = 1$ if there is one sensor node in the position $p(i, j)$; otherwise, $p(i, j) = 0$. Assume the positions of sensor nodes in W are not in the K -Queens pattern. Without loss of generality, $\exists p(i, j) \in A$, $p(i + l, j) \in A$, ($0 < l < K$), we have $p(i, j) = 1$ and $p(i + l, j) = 1$. According to Definition 3, $p(i', j') = 1$ and $p(i' + l, j') = 1$, where $i' = i - (\lfloor i/K \rfloor) * K$, $j' = j - (\lfloor j/K \rfloor) * K$. $p(i', j')$ and $p(i' + l, j')$ are the position points in the original sub-region, shown in Fig. 1(a). Two nodes in positions $p(i', j')$ and $p(i' + l, j')$ share the same row, however, the positions of sensor nodes within the original sub-region are in the K -Queens pattern. Thus, the positions of sensor nodes in W are in the K -Queens pattern. \square

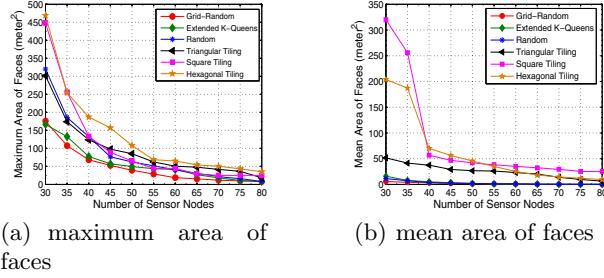
Grid-Random Pattern In this subsection, we discuss the map division performance using the grid-random pattern.

Definition 4. (*Grid-Random Pattern*): Given one finite, two-dimensional planar region A and N sensor nodes, A is divided into $N = m_1 \times m_2$ regular grids. Each sensor node is randomly placed in each grid.

For example, 32 sensor nodes are deployed using the grid-random pattern in the 80×80 meter² unit grid. A is divided into 4×8 regular grids. As shown in Fig. 1(b), sensor nodes are well distributed in A due to one sensor being in one grid. In addition, the random position of a sensor node in one grid can ensure that a larger number of faces is obtained, and further reduce the area of the largest face.

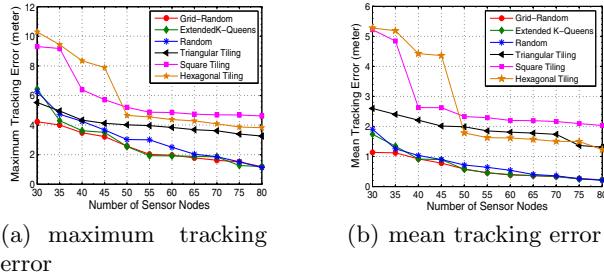
Similarly, we give the results of map division with 16, 25 and 36 nodes.

- 16 nodes: $MAX_{face} = 3.2258$, $MEAN_{face} = 0.1001$, and $VAR_{face} = 0.2202$.
- 25 nodes: $MAX_{face} = 1.7061$, $MEAN_{face} = 0.0279$, and $VAR_{face} = 0.0631$.
- 36 nodes: $MAX_{face} = 0.7893$, $MEAN_{face} = 0.0152$, and $VAR_{face} = 0.0342$.



(a) maximum area of faces

(b) mean area of faces

Fig. 2. Results of map division using different patterns

(a) maximum tracking error

(b) mean tracking error

Fig. 3. Tracking error rate using different patterns

4.4 Discussion

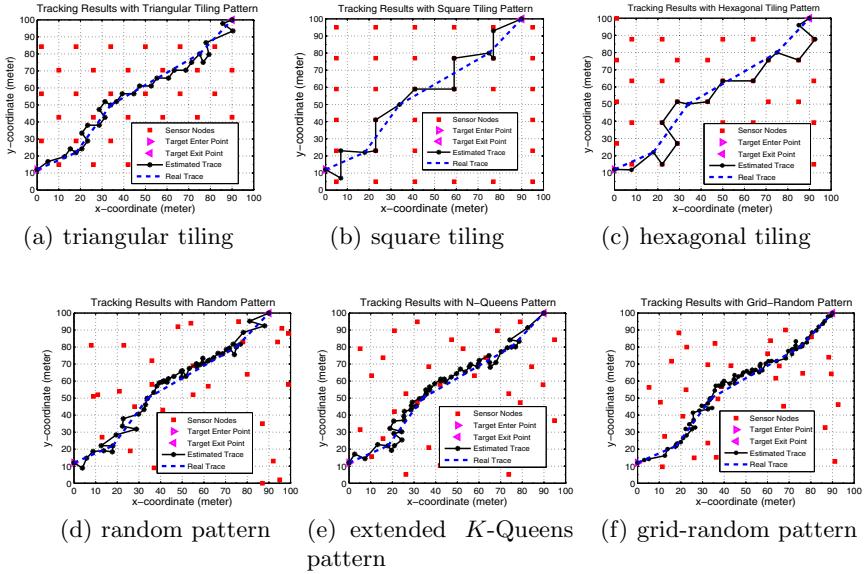
In this section, we analyze the properties of map division using different deployment patterns. Due to the limitation of the length, we omit the proof of the following theorem.

Lemma 1. *Given one segment from 0 to 1, randomly choose N points from that segment. The bound of maximum and minimal distance of any two neighboring numbers is $O(\log N/N)$ and $O(1/N^2)$, respectively.*

Theorem 2. *Given two-dimensional planar region A , N sensor nodes are randomly deployed in A . A finite set of bisectors L between any two node pairs can subdivide A into lots of cell complexes. The bound of the diameter and area of the maximum cell is $O(\log N/N^2)$ and $O((\log N/N^2)^2)$, respectively.*

Corollary 1. *Given two-dimensional planar region A , N sensor nodes are deployed with the square tiling in A . A finite set of bisectors L between any two node pairs can subdivide A into lots of cell complexes. The bound of the diameter and area of the maximum cell is $O(\sqrt{N}/N^2)$ and $O(1/N^3)$, respectively.*

Corollary 2. *Given two-dimensional planar region A , N sensor nodes are deployed using the grid-random pattern in A . A finite set of bisectors L between any two node pairs can subdivide A into lots of cell complexes. The bound of the diameter and area of the maximum cell is $O(c_1/N^2)$ and $O(c_2/N^4)$, respectively, where c_1 and c_2 are constant.*

**Fig. 4.** Tracking simulation example - curve path

5 Simulation Evaluation

In this section, we evaluate the tracking performance with the irregular pattern proposed in this paper, and compare it with three regular patterns and the random pattern.

Table 1. Simulation Settings

Parameters	Values
Monitored Region	$100 \times 100 \text{ meter}^2$
Number of Sensor Nodes	[30, 80]
Sensing Range	10 meters
Sensing Sampling Rate	10 Hz
Target Velocity & Random between Deployment Pattern	1 – 5(meters/s) triangular tiling, square tiling, hexagonal tiling, random pattern, extended K-Queens pattern, grid-random pattern

5.1 Simulation Settings and Methodology

In the simulations, we model the monitored area as a 100×100 unit map. A movement trace of the mobile target is generated with the random waypoint model [8]. Due to the locality of sensor nodes for target detection at any time instance, the effective range R is set to twice the sensing range ($2R_s$) in the

simulation experiments. The tracking scheme using node sequences proposed by [2] is used. To compare the tracking performance of different deployment patterns, we evaluate these patterns in terms of maximum area and mean area of faces as well as the maximum and mean tracking error rate. First, the map division results are illustrated. Second, we evaluate the tracking accuracy. The mean tracking error rate is defined as the averaged error rate of all points in the trace.

5.2 Simulation Results

Map Division. We compare the results of map division using three categories of patterns under a different number of sensor nodes, ranging from 30 to 80, in steps of 5. Fig. 2 illustrates the area of the largest face and the mean area of all faces, after map division, using three categories of patterns. As Fig. 2 shows: (i) the irregular pattern outperforms the regular and random patterns, especially in a sparse sensor network; (ii) with the increasing number of deployed sensor nodes, the maximum and mean area of faces is reduced; (iii) three categories of patterns present similar results as map division when the density of sensor nodes is high enough.

Tracking Accuracy. In order to evaluate different deployment patterns, we also compare the tracking performance in terms of maximum and mean tracking error rate under a different number of sensor nodes. Simulation results shown in Fig. 3 indicate that (i) the irregular pattern greatly outperforms the regular pattern with the increasing density of sensor nodes; (ii) the irregular pattern gets a small tracking performance gain, especially in a high density network; (iii) for any deployment pattern, we can enhance the tracking accuracy considerably when the number of sensor nodes increases from 30 to 60.

Visualized Simulation Example. We further give an intuitive comparison among the three categories of patterns. Fig. 4 illustrates the examples of differences between an estimated trace and a real movement trace with 36 nodes deployed using three categories of patterns. Fig. 4(a)-(f) show all the position points estimated using three categories of patterns, respectively. From Fig. 4(a)-(f), compared to the square and hexagonal tiling, it is obvious that position points, given by triangular tiling, are much more closely distributed around the true trace. Meanwhile, the estimated trace using the irregular pattern is closer to the real trace than those using the random pattern and triangular tiling pattern. Moreover, the grid-random pattern has the highest tracking accuracy among the six patterns.

6 Conclusion

To deploy sensor nodes for target tracking using node sequences, this paper presents and studies three categories of deployment patterns: regular tiling, random patterns, and irregular pattern. We propose two kinds of irregular patterns:

extended K -Queens and grid-random. To compare the tracking performance of the three categories of patterns, we analyze the properties of each pattern and the upper bound of the tracking error rate in the worst case scenario. Theoretical analysis and simulation experiments show that the proposed irregular pattern outperforms the regular tiling pattern and random pattern.

References

1. Mainwaring, A., Polastre, J., Szewczyk, R., Culler, D., Anderson, J.: Wireless Sensor Network for Habit Monitoring. In: Proc. of ACM International Conference of Wireless Sensor Networks, and Application (WSNA). ACM Press (2002)
2. Zhong, Z., Zhu, T., Wang, D., He, T.: Tracking with Unreliable Node Sequences. In: Proc. of 28th IEEE International Conference on Computer Communications (INFOCOM). IEEE Press, New York (2009)
3. Ababnah, A., Natarajan, B.: Optimal Sensor Deployment for Value-Fusion Based Detection. In: Proc. of IEEE GLOBAL Communications Conference (GLOBECOM). IEEE Press, New York (2009)
4. Yuan, Z., Tan, R., Xing, G., Lu, C., Chen, Y., Wang, J.: Fast Sensor Placement Algorithms for Fusion-based Target Detection. In: Proc. of 29th IEEE Real-Time Systems Symposium (RTTS). IEEE Press, New York (2008)
5. Bai, X., Yun, Z., Xuan, D., Jia, W., Zhao, W.: Pattern Mutation in Wireless Sensor Deploymnet. In: Proc. of 29th IEEE International Conference on Computer Communications (INFOCOM). IEEE Press, New York (2010)
6. Balister, P., Kumar, S.: Random *vs.* Deterministic Deployment of Sensors in the Presence of Failure and Placement Errors. In: Proc. of 28th IEEE International Conference on Computer Communications (INFOCOM). IEEE Press, New York (2009)
7. Poe, W.Y., Schmitt, J.B.: Node Deployment in Large Wireless Sensor Networks: Coverage, Energy Consumption, and Worst-Case Delay. In: Proc. of ACM Asian Internet Engineering Conference (AINTEC). ACM Press (2009)
8. Johnson, D.B., Maltz, D.A.: Dynamic Source Routing in Ad Hoc Wireless Networks. *J. Mobile Computing* (1996)
9. Grunbaum, B., Shephard, G.C.: Tilings and Patterns. W. H. Freeman and Company (1987)

ELRM: A Generic Framework for Location Privacy in LBS

Muhamed Ilyas¹ and R. Vijayakumar²

¹ Research Scholar, School of Computer Science, Mahatma Gandhi University
Kottayam, Kerala, India

Muhamed.ilyas@gmail.com

² School of Computer Science, Mahatma Gandhi University
Kottayam, Kerala, India
Kiran2k@bsnl.in

Abstract. Recent advances in mobile communication and development of sophisticated equipments lead to the wide spread use of Location Based Services (LBS). A major concern for large-scale deployment of LBSs is the potential abuse of their client location data, which may imply sensitive personal information. Protecting location information of the mobile user is challenging because a location itself may reveal user identity. Several schemes have been proposed for location cloaking. In our paper, we propose a generic Enhanced LBS Reference Model (ELRM), which describes the concept, the architecture and the functionalities for location privacy in LBS. As per the architecture, the system ensures location privacy, without trusting anybody including the peers or LBS servers. The system is fully distributed and evaluation shows its efficiency and high level of privacy with QoS.

Keywords: Location privacy, Location Based Services, Location Cloaking, Distributed Query Processing.

1 Introduction

The last decade showed an accelerated development of mobile and Internet technologies. Internet technology with globally connected mobile networks introduces new business models and the development of service architecture. Location-Based Services (LBS) are such an example. Location based services (LBS) are Internet services that provide information or enable communication based on the location of users and/or resources at specific times. Service providers envision offering many new services based on a user's location as well as augmenting many existing services with location information [3]. At the same time, LBSs poses a new threat, i.e., privacy preservation. For example, someone wants to have dinner and is searching for a restaurant using the Internet. In order to get more accurate and useful research results, more terms such as the mobile user's location, the type of food, etc. should be included in his search criteria. Unfortunately, if the queries are not securely managed, it could be possible for a third party to retrieve the mobile user's personal sensitive information such as his location information, his habit, etc. In this case, even if an

individual does not directly release personal information to the service provider, this provider may become aware of the sensitive information if it has to provide a service to such an individual [4].

Research in the field of privacy preservation in pervasive computing has mainly concentrated on techniques for anonymous communication [1], access control and obfuscation [6, 7], dummy requests [5], or on a combination of such techniques. Many of these techniques are based on a central server called Location anonymizer (LA). In this case, the mobile user has to submit his/her location identifier to the LA, and LA cloaks the location using different models developed, like K-anonymity, before submitting the query to the LBS. Location Cloaking with a centralized architecture must trust the central third party server with their identities, locations and queries. However, there are a number of disadvantages for centralized approaches, such as a single point of failure, bottlenecks due to communication overhead, and privacy threats as these systems store all information in a single place. To overcome these problems, several decentralized approaches have been proposed. [22].

We propose a distributed approach to protect user privacy in LBS that does not need a centralized server for location cloaking and does not trust any one including participating peers. Our approach is similar to the work proposed by [5], but with cluster based peer selection algorithm and an enhanced distributed peer cloaking method. Our approach uses the capabilities of current mobile systems to form ad-hoc Wireless Personal Area Networks (WPANs) using technologies like Bluetooth. As per the system, a user who needs a location services, called query initiator, initially forms a peer group of n individuals based on a cluster algorithm. Then it randomly selects a peer, called query requestor, to forward the query, on behalf of the query initiator, to the LBS server. A major challenge of this approach is the selection of the query requestor with uniform probability. It ensures that even if the LSP has access to the information that currently n devices form an ad-hoc network, the LSP is only able to identify the query initiator with a probability of $1/n$ [5].

In our approach, the user and the peers do not reveal their exact location to each other. Instead the actual positions are obfuscated with an imprecise location like circle. For maximal privacy protection this approach combines obfuscation with K-anonymity [5]. If a user requires a location service, our algorithm computes a minimum bounding circle (called Global Cloaking Area, GCA), that encloses obfuscated locations of all his peers. The GCA contains the obfuscated location of the user and the obfuscated locations of all other K-1 peers.

In summary, our contributions in this paper are as follows:

We propose a heuristic algorithm to compute the obfuscated location, called Self Cloaked Area (SCA), of the user and all its participating peers. Self Cloaking is done individually by the query initiator and all participating peers.

We develop a Greedy algorithm for generating a user's K-anonymous obfuscated location from available n SCAs. In each iteration the algorithm checks the K-anonymity and continues until K-anonymity level is met. Unlike in [5], where the selection of peers to meet the K-anonymity is done by the query initiator, our work distributes this process among peers. Each peer calculates, whether it's Self Cloaked Area is within the GCA and is eligible to participate in the obfuscation process to meet the K-anonymity level.

We present a near-uniform random selection algorithm to select a query requestor without revealing their identities. In paper [5] they presented a decentralized approach to protect user privacy during the access of LBSs using wireless ad-hoc networks. Users do not need to trust any involved party, including their peers, the LSP or the infrastructure. We extend this work by further decentralizing location obfuscation among peers with less computational overheads, and also introducing simple greedy algorithm for a near-uniform random selection for any type of peer distribution.

2 Related Works

Location anonymity and privacy awareness in location-based services has been extensively studied as a solution to protect user privacy in recent literatures. The objective is to allow the mobile user to request services without compromising his/her privacy, especially location privacy [8]. Several privacy protection techniques have already been proposed. Based on the underlying methodologies, these techniques can be divided into three categories: Cloaking, Transformation, pseudonym.

Spatial Cloaking is the most widely used privacy preserving technique for users accessing LBS. The main idea behind cloaking approaches is to blur a user's exact location in a larger cloaked region and to make him/her indistinguishable among the set of other (real or dummy) users located in the cloaked region. Spatial cloaking can be grouped in to Centralized and Decentralized, depending on where the cloaking is taking place.

Many existing approaches in spatial cloaking is based on a centralized architecture. These approaches rely on the existence of a trusted Intermediary server called location anonymizer which protects a user's private location and identity information from an untrusted location server (e.g., Mokbel et al., 2006; Gruteser and Grunwald, 2003; Gedik and Liu, 2005a, 2005b; Du et al., 2007) [9 -14]. The main idea in centralized cloaking is to put an anonymiser between the users and the location server to prevent the server from learning users' precise location information and identities. Every location-based query is first sent to the anonymizer, which transforms the user's exact location to a cloaked area (i.e., rectangle or circle) and forwards the query to the LBS server for that cloaked area. While different cloaking algorithms are proposed for cloaking a user's location, the common objective is to blur a user's location in an area of size at least A_{min} and/or among a set of at least $k - 1$ other users. Depending on the approach, these parameters can be specified by each user independently, or are chosen as system parameters. During the second phase, the privacy-aware location server, which is modified to process a cloaked region query, generates a candidate list which is guaranteed to include the nearest neighbor of any point inside the cloaked region. This list is then transferred to the client side for further refinement to obtain the final result set [15]. The blurred spatial area can be based either on the k-anonymity concept [Samarati 2001; Sweeney 2002a, 2002b] [16, 17, 18] (i.e., the area should contain at least k users) or on a graph model that represents a road network [Duckham and Kulik 2005]. [19]

However the centralized approach has several disadvantages. This approach requires an anonymiser, as sophisticated as the location server itself, to act as a proxy between users and the server per query. There are chances for single point of failure/attack and bottleneck due to communication overhead. Another important drawback is that, in many scenarios cloaking users' location information in a larger region or among $k - 1$ other users does not protect user's location information. This is due to the fact that based on user distributions in the space and the value of k (or similarly size of the cloaked region), precise user location can be derived using several techniques. To overcome these limitations, decentralized approaches have been proposed that construct cloaked region. The approaches proposed by Chow et al. (2006)[20] and Ghinita et al. (2007b, 2007c)[20,21] assume users communicate with each other to collaboratively form a cloaked region. Ghinita et al. (2007b) propose a hierarchical overlay network resembling a distributed B+ tree for constructing the cloaked region that overcomes the above drawback. However, it suffers from very slow response time. Ghinita et al. (2007c)[21] propose methods which provide stronger privacy than Chow et al. (2006) for various distributions and do not suffer from slow response time of Ghinita et al. (2007b). The authors propose a distributed method to find a random set of k adjacent users based on their 1-D Hilbert ordering. Finally, Duckham and Kulik (2005)[19] propose a graph model to represent possible user's locations and denote the cloaked region by a set of vertices in the graph. The client progressively gives more information about her precise location until the query result set reaches her desired accuracy.

Tanzima Hashem, Lars Kulik [5] have developed a decentralized approach to protect user privacy during the access of LBSs using wireless ad-hoc networks. Users do not need to trust any involved party, including their peers, the LSP or the infrastructure provider. It exploits the wireless advantage that all users in communication range can overhear a message to anonymize the communication among users.

Our work is similar to [5] but with less communication overhead and with a different cloaking method. Simulation results show that it has less communication overhead and high quality of service.

3 System Architecture and Location Cloaking

We present a decentralized system that employs the power of ad-hoc networking for obfuscating the user location from third party LBS servers. User and each participating peer cloaks their location as surrounding circles, where the user location cannot be identified by an adversary. The user, who wants to access the LBS service, first determines the size of the circle (Called Global Cloaked Area) in terms of its radius. Also it determines $K - 1$ number of participating users for cloaking the location. The cloaking process proceeds in two stages. Initially, the query initiator calculates its Self Cloaked Area (SCA). Then it sends a broadcast message to all its peers, which are normally one hop away from the user. The message contains three parameters; the pseudo IP address of the query initiator, the parameters of its SC, and

the parameters of the initial GCA. On receiving the message, each peer calculates its SCA. Then it performs a spatial ‘within’ operation to identify that GCA contains its SCA. The peer returns its SCA along with Boolean result of the spatial operation. The query initiator then checks for the K-anonymity and continue the process with hope distances > 1 until the K-anonymity is met. The GCA will be a minimum radii circle which encloses all SCAs. Then our Random Selection Algorithm selects a query requestor to forward the query along with GCA to *LBS Server*.

3.1 Generating SCA

In our approach, we generate a circular cloaked area which contains the peer’s exact location anywhere in the circle. We have developed a heuristic algorithm to generate the cloaked circle. Let (x, y) be the exact location of the mobile user. (The location might be received from GPS or any other means). In order to obtain maximum anonymity, we cloak the point (x, y) with a surrounding circle. But if we generate such a circle, adversary can easily identify the location of the user, as it may be centre of the circle. So we generate a pseudo circle with centre (x_0, y_0) that must include the point (x, y) anywhere in the circle. Let R be the radius of the Self cloaked circle. In order to find a random point (x_0, y_0) , we randomly choose a distance value r , where $r \leq R$, and an angle θ , where $0 \leq \theta \leq 2\pi$.

$$\text{Then, } \begin{aligned} x_0 &= x + r \cos \theta \\ y_0 &= y + r \sin \theta, \text{ where } r \leq R \end{aligned}$$

The cloaked area, with radius R , then can be calculated. This ensures that the original location of the mobile user is within or on the boundary of the newly created obfuscated circle with centre (x_0, y_0) .

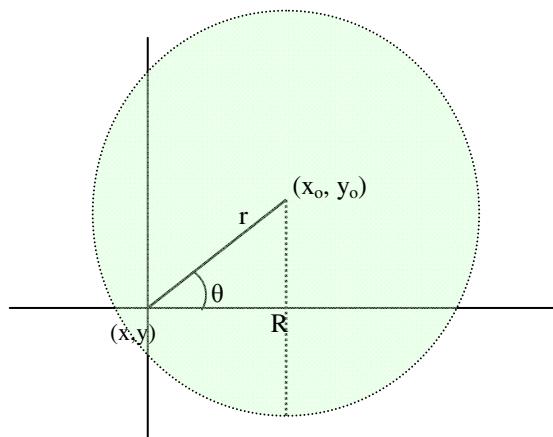


Fig. 1. Self Cloaking

Algorithm 1: ComputeSCA

Input:

x_0, y_0 : Centre of the GCA circle
 RGCA : Radius of GCA
 Ref : x_i, y_i, R

Output:

x_i : X-coordinate of the obfuscated user position
 y_i : Y-coordinate of the obfuscated user position
 R : Radius of SCA (Obfuscated circle)

1. $x \leftarrow$ X-coordinate of the user position from GPS
2. y (Y-coordinate of the user position from GPS)
3. Let r (Random value where $0 < r \leq R$)
4. θ (Random value where $0 \leq \theta \leq 2\pi$)
5. R (Required radius of the obfuscated circle (SCA))
6. $x_i = x + r \cos \theta$
7. $y_i = y + r \sin \theta$
8. Compute Spatial Within operation for SCA with RGCA as radius
9. return x_i, y_i, R as reference
10. Return true if spatial within operation is true

3.2 Generating GCA

Let K_l and K_h be lowest and highest anonymity level of the query initiator j. Also let R_l and R_h be the lowest and highest radius of the circular area which the query initiator wants to obfuscate, which we call the Globally Cloaked Area (GCA). R is selected in such a way that it balances the K-anonymity and an optimal area which includes all other $K-1$ self cloaked peers. At the first step, the Query initiator sends a message to all its 1-hop peers requesting its pseudonym, the obfuscated origin (x_{pi}, y_{pi}) of the Self Cloaked Area (SCA) and the radius of SCA; if SCA of the peer Pi is within the Globally Cloaked Area (GCA).

Initially the message is sent down to all 1-hop peers with a value R. If the query initiator fails to find sufficient number of SCAs within the limits of GCA, query initiator either decrement the value of K or it increments the value of R. This process continues until the value of $K \geq K_l$ and the value of R reaches R_h .

3.3 The Greedy Algorithm

We present an algorithm to compute the Globally Cloaked Area of the query initiator j. This is Greedy algorithm which executes until the desired level of anonymity is obtained. Assume that K is the desired anonymity level of j and R_l and R_h be the lowest and highest radius of the Globally Cloaked Area. i.e., the area of Globally Cloaked Area is $\pi R_l^2 \leq GCA \leq \pi R_h^2$. To compute the GCA, we have to find the smallest circle r that encloses a K-subset (including j's SCA) from the n SCAs. The SCA of a user i is described by its obfuscated centre (x_i, y_i) and the radius r_i . If the

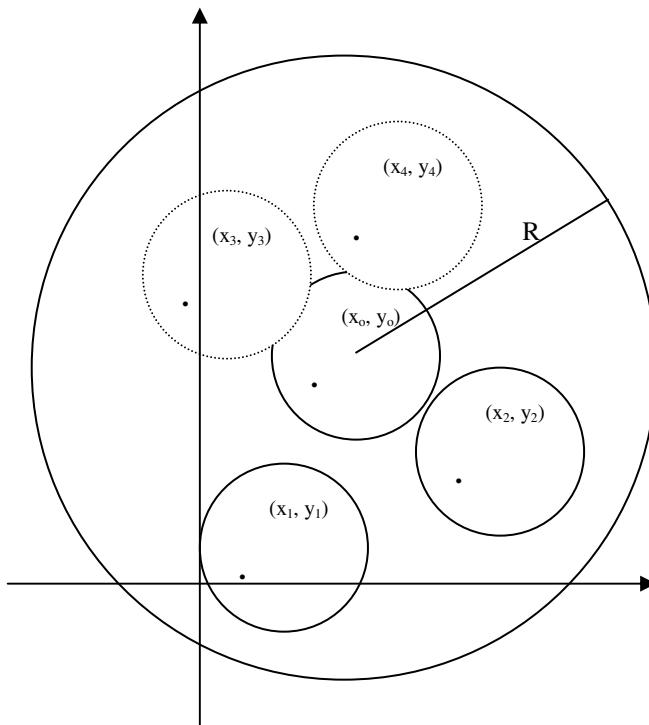


Fig. 2. Globally Cloaked Area

initial GCA with radius R_l and query initiators obfuscated centre (x_0, y_0) , is not able to contain all $K-1$ SCAs then the value of R (Radius of GCA) is incremented in each step until R_l reaches the value R_h . We have developed a greedy-based algorithm (GBA) with a time complexity of $O(h)$, where h is the hop count. Initially the message is broadcasted to all 1-hop peers. All peers receiving this message, computes its own SCA, perform a spatial ‘within’ operation with GCA (The ‘within’ predicate in spatial geometry returns t (TRUE) if the first geometry is completely inside the second geometry). If the result of the ‘Within’ operation is true, then function returns the obfuscated centre of the SCA, and the radius of the SCA. The numbers of SCAs are counted after each hop iteration, if the number of SCAs are below the K -anonymity level, then we increment the hop and continue the process until the system desired parameter h_{max} is reached.

Algorithm 2: Compute GCA

Input:

- h_{max} : The maximum hop count
- K : The anonymity level
- R_l : The minimum radius of the GCA

R_h : The maximum radius of the GCA

X_0, Y_0 : Obfuscated coordinates of Query initiator

Output: A, The Globally Cloaked Area of the Query initiator that covers K-SCAs

1. $A \leftarrow \alpha$
2. Let S be the set of SCA
3. $S \leftarrow \alpha$
4. Compute SCA of Query initiator
5. $x_0 \leftarrow X\text{-coordinate of the centre of SCA of Query initiator}$
6. $y_0 \leftarrow Y\text{ coordinate of the centre of SCA of Query initiator}$
7. $AGCA \leftarrow \text{Initial Area of GCA with circle } R_l$
8. for $R = R_l$ to R_h step 1 do
 - i. if Compute SCA = true then add SCA to S
 - ii. Count \leftarrow Total count of Peers after Computing their SCAs and within the geometrical boundary of AGCA
9. for $h = 1$ to h_{\max} step 1 do
 - b. if Count < K then continue else stop
10. if count < K the continue to step 8 else stop
11. Stop

Once the Optimal GCA has been computed, the query initiator randomly finds a query requestor from the set of available peers within the GCA. The peers are addressed with their pseudonyms that are sent back from peers, during the process of computing SCAs. The query is then forwarded to the query requestor, and it is submitted to the LBS server with GCA. Since the GCA is an obfuscated area, adversary may find it difficult to locate the query initiator. Moreover all IPs are encrypted with their pseudonyms. The LBS server returns a list of results applicable for this GCA. From this set of result the query initiator filter the values that he is interested in.

4 Experimental Evaluation

We evaluate our system with Greedy based GCA computation in [5]. The time complexity of GGC in [5] is $O(n \log n)$ where as our system the time complexity is $O(h)$ where h is the hop of the peers that are included in the GCA. We set our simulation for experiments in this article on a Pentium 2.8 GHz and 1 GB RAM with varying K and R for GCA. For all K the system has shown remarkable performance compared to GGC in [5], because all SCA computation and the selection of optimal GCA are done at peers simultaneously, instead of computing at the system of query initiator. Figure 3 shows the average response time for GGC and for our system. We assume that the distributions of objects are normal. Figure 4 shows the computational cost for generating GCA for a set of 50 users. Since, in our system, the SCAs are calculated at the peers, the value of the anonymity level K or the numbers of peer does not affect the total computational cost. All the computations are distributed in our system simultaneously. Where as in the case GGC the GCA is calculated by the query initiator and the computational cost of GCA is directly proportional to the value of K and the number of peers n .

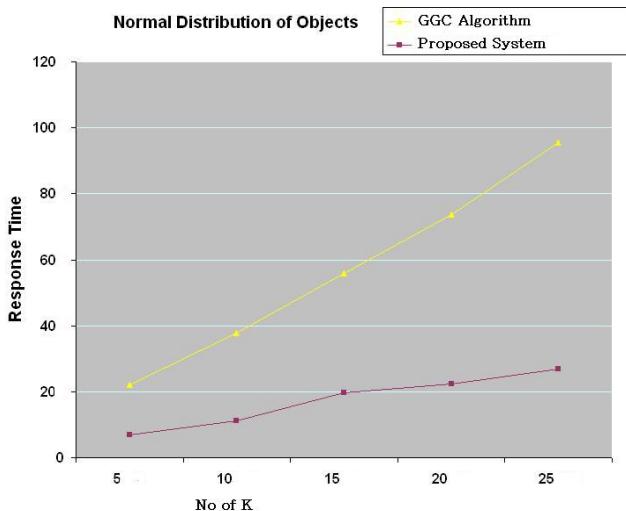


Fig. 3. Response time for K subset of 25 SCAs

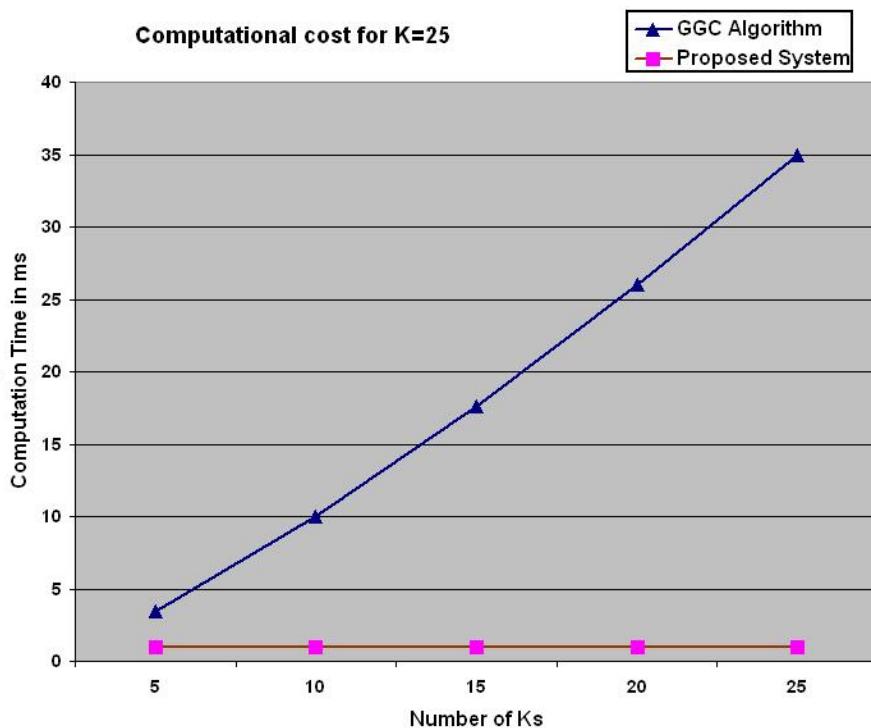


Fig. 4. Computational cost for a subset of 25 SCAs

5 Conclusion

In this paper, we have presented a distributed location obfuscation method, to protect location privacy for Location Based services (LBS). In this case, the query initiator does not want to trust anyone including the peers and any third party service providers. Even peers do not reveal their exact locations, instead presents only their Self cloaked Areas. The system we presented fully distributes all computations, and even GCA calculation is done at peers. We have also presented algorithms which needs less computation time at the query initiator, compared to our previous works, due to the fully distributed approach. We also evaluated our system with different size of K-anonymity level and shows good accuracy and optimal results.

We plan to extend our work, with a network assisted approach, where the numbers of participating peers are less than the anonymity level K. We are also investigating the possibility of moving peers and dynamic Self Cloaked Areas (SCA).

References

- [1] Al-Muhtadi, J., Campbell, R.H., Kapadia, A., Mickunas, M.D., Yi, S.: Routing Through the Mist: Privacy Preserving Communication in Ubiquitous Computing Environments. In: Proc. of ICDCS 2002, pp. 74–83. IEEE Computer Society (2002)
- [2] Bettini, C., Mascetti, S., Wang, X.S.: Privacy protection through anonymity in location-based services. To appear in Handbook of Database Security: Applications and Trends. Springer (2007)
- [3] Sun, Y., La Porta, T.F., Kermani, P.: A Flexible Privacy-Enhanced Location-Based Services System Framework and Practice. *IEEE Transactions on Mobile Computing* 8(3), 304–321 (2009)
- [4] Yao, L., Lin, C., Kong, X., Xia, F., Wu, G.: A Clustering-based Location Privacy Protection Scheme for Pervasive Computing. In: IEEE/ACM International Conference on Green Computing and Communications & IEEE/ACM International Conference on Cyber, Physical and Social Computing (2010)
- [5] Hashem, T., Kulik, L.: Don't trust anyone": Privacy protection for location-based services. *Pervasive and Mobile Computing* 7, 44–59 (2011)
- [6] Gandon, F.L., Sadeh, N.M.: A Semantic E-Wallet to Reconcile Privacy and Context Awareness. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 385–401. Springer, Heidelberg (2003)
- [7] Wishart, R., Henricksen, K., Indulska, J.: Context Obfuscation for Privacy via Ontological Descriptions. In: Strang, T., Linnhoff-Popien, C. (eds.) LoCA 2005. LNCS, vol. 3479, pp. 276–288. Springer, Heidelberg (2005)
- [8] Hu, H., Xu, J., Senior Member, IEEE: 2PASS: Bandwidth-Optimized Location Cloaking for Anonymous Location-Based Services
- [9] Gruteser, M., Grunwald, D.: Anonymous usage of location-based services through spatial and temporal cloaking. In: MobiSys 2003: Proc. of the 1st Int. Conf. on Mobile Systems, Applications and Services, pp. 31–42 (2003)
- [10] Mokbel, M.F., Chow, C.-Y., Aref, W.G.: The new casper: query processing for location services without compromising privacy. In: VLDB 2006: Proc. of the 32nd Int. Conf. on Very Large Data Bases, pp. 763–774 (2006)

- [11] Gedik, B., Liu, L.: Protecting location privacy with personalized k-anonymity: architecture and algorithms. *IEEE Transactions on Mobile Computing* 7(1), 1–18 (2008)
- [12] Kalnis, P., Ghinita, G., Mouratidis, K., Papadias, D.: Preventing location-based identity inference in anonymous spatial queries. *IEEE Transactions on Knowledge and Data Engineering* 19(12), 1719–1733 (2007)
- [13] Gedik, B., Liu, L.: Location privacy in mobile systems: a personalized anonymization model. In: ICDCS 2005: Proc. of the 25th IEEE Int. Conf. on Distributed Computing Systems, pp. 620–629 (2005)
- [14] Bettini, C., Wang, X., Jajodia, S.: Protecting Privacy Against Location-Based Personal Identification. In: Jonker, W., Petković, M. (eds.) SDM 2005. LNCS, vol. 3674, pp. 185–199. Springer, Heidelberg (2005)
- [15] Khoshgozaran, A., Shahabi, C.: A taxonomy of approaches to preserve location privacy in location-based services
- [16] Samarati, P.: Protecting respondents' identities in microdata release. *IEEE Trans. Knowl Data Engin.* 13(6), 1010–1027 (2001)
- [17] Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. *Inter. J. Uncert. Fuzz. Knowl.-Based Syst.* 10(5), 571–588 (2002a)
- [18] Sweeney, L.: k-anonymity: A model for protecting privacy. *Inter. J. Uncert. Fuzz. Knowl.-Based Syst.* 10(5), 557–570 (2002b)
- [19] Duckham, M., Kulik, L.: A formal model of obfuscation and negotiation for location privacy. In: Proceedings of the International Conference on Pervasive Computing (2005)
- [20] Chow, C., Mokbel, M., Liu, X.: A peer-to-peer spatial cloaking algorithm for anonymous location-based service. In: GIS, pp. 171–178 (2006)
- [21] Ghinita, G., Kalnis, P., Skiadopoulos, S.: PRIVE: anonymous location-based queries in distributed mobile systems. In: WWW, pp. 371–380 (2007b)
- [22] Ghinita, G., Kalnis, P., Skiadopoulos, S.: MobiHide: a mobile peer-to-peer system for anonymous location-based queries. In: SSTD, pp. 221–238 (2007c)

Energy Efficient Administrator Based Secure Routing in MANET

Himadri Nath Saha¹, Debika Bhattacharyya¹, and P.K. Banerjee²

¹ Department of Computer Science and Engineering, Institute of Engineering & Management, West Bengal, India

² Department of Electronics and Tele Communication Engineering, Jadavpur university, West Bengal, India

him_shree_2004@yahoo.com, bdebika@yahoo.com

Abstract. The lack of static infrastructure causes several issues in mobile Ad Hoc network, such as energy utilization, node authentication and secure routing. In this paper we propose a new scheme for energy efficient secure routing of data packets in MANET. This approach will reduce the computational overhead to make it more energy efficient than existing schemes. As there is no stationary infrastructure, each node in MANET acts a router that forwards data packets to other nodes. Therefore selection of effective, suitable, adaptive and robust routing scheme is of utmost importance. We have reduced the amount of network activity for each node required to route a data packet. This leads to lesser wastage of energy and increases security. Our simulation results will show how this is energy efficient and secure. Finally we have discussed how this scheme prevents various attacks which may jeopardize any wireless network.

Keywords: Administrator, Associative node, Traversed Administrator field, Watch nodes, Backtracking bit, Routing, Manet.

1 Introduction

A major role to globally reduce energy consumption will be played by Ad hoc routing technologies. The communication in case of mobile Ad hoc network, MANET, is mainly based on the radio signals transmitted by the node. Again MANET, being a wireless network, is quite different from the common mobile communication. In mobile[1] communication bridge networks within its own range are used by the nodes to communicate with other nodes. The bridge networks act mainly as base stations which the source node needs to contact while sending a data packet to its destination. We need to remember that the nodes are constantly moving and thus when a node goes out of the range of a base station it must contact its new base station which it finds in its range. But in MANET[2] there is no base station or any other infrastructure, helping to setup or perform the network activity required. Thus in this case the nodes are the routers transferring the data packets themselves. Hence a robust and good routing protocol that will perform all the functions but with an optimized network activity to decrease the network traffic as well as make the transmission fast is very essential. Thus while building our routing protocol we kept in mind these three factors – making the transmission fast, decreasing the network traffic and properly utilizing the energy consumed by each router.

This paper deals with how the entire network needs to be setup from the start, algorithms required to implement the protocol and finally implementation of the entire network using snapshots of a network showing how the algorithm works when a data packet is sent from one node to another. Section 1 gives a brief introduction to our protocol; in section 2 we mention some previous work related to our scheme. In the next section we elaborate the entire scheme and explain the algorithm of our protocol. Section 5 illustrates the simulation results of our protocol. The following sections explains the security aspects of our protocol and compares its performance with existing protocols. We name our scheme as energy efficient administrator based secured routing , abbreviated as EEABSR, in MANET.

2 Related Work

S. Matri[3] proposed to trace malicious nodes by using watchdog. According to this system, whenever a node forwards a data packet, the watch dog of the node checks whether the next node which receives the packet also sends the packet by listening to the broadcast signal of the next node. If the next node does not forward the packet within a predefined threshold time, the watchdog detects malicious behavior and accuses the node for aberration. This proposal has two shortcomings:

1. To monitor the behavior of nodes two or more hops away, the watch node has to trust the information from other nodes, which introduces the vulnerability of malicious activity.
2. The *watchdog* cannot differentiate between misbehavior and ambiguous collisions, receiver collisions, controlled transmission power and other such false alarms that might be generated during the data sending through the network.

We have used this concept in the form of a watch node in this protocol and have tried to eliminate the difficulties plaguing the watchdog by associating two watch nodes to each admin node.

Gonzalez [4] presents a methodology, for detecting packet forwarding misbehavior, which is based on the principle of flow conservation in a network. That states that if all neighbors of a node v_j are queried for

- i. The amount of packets sent to v_j to forward and
- ii. The amount of packets forwarded by v_j to them,

The total amount of packets sent to and received from v_j must be equal. They assume a threshold value for non malicious packet drop. A node v_i maintains a table with two metrics T_{ij} and R_{ij} , which contains an entry for each node v_j to which v_i has respectively transmitted packets to or received packets from. Node v_i increments T_{ij} on successful transmission of a packet to v_j for v_j to forward to another node, and increments R_{ij} on successful receipt of a packet forwarded by v_j that did not originate at v_j . All nodes in the network continuously monitor their neighbors and update the list of those they have heard recently. The algorithm requires fewer nodes to overhear each others' received and transmitted packets since it uses statistics accumulated by each node as it transmits to and receives data from its neighbors. Since there is no collaborative consensus mechanism, such an algorithm may lead to false accusations against correctly behaving nodes.

3 The Scheme

Every node in a MANET has a range of itself i.e. no node is capable of transmitting a data packet to an infinite distance. The nodes which fall in the range of a particular node are called its Neighboring nodes. In our algorithm we have alternatively used friend nodes for neighbor nodes, both of them being the same. In the network, three types of nodes have been used:

1. Common nodes
2. Associative nodes
3. Administrator nodes
4. Watch Nodes

The classification is based on the range and the position of the nodes in the network. But to understand classification we firstly need to understand how the entire network is set up. After a stipulated time period each node checks for its neighboring nodes, i.e. which nodes are present within its range. From this friend list, a list is prepared which contains all the neighboring nodes for all the nodes in the network.

Next, the node compares its previous and present list to check for network change and reports any difference to its administrator. The administrator node always lies in the range of the node in question.

We will describe how an Administrator node is chosen later. If there is no change in the topology of the network, there is no need to choose an administrator node; however for any change in the network, the previous Administrator nodes will choose a new Admin node which leads us to the discussion on what an Admin node is and how it is selected.

ADMINISTRATOR NODES

Now the topic of selecting a new Admin arises. If there is a need to elect a new Admin, all the nodes send their neighbor's list to their Admins(old admin) which all the Admins exchange among themselves thereby giving each Admin the knowledge of the neighbors of each and every node in the entire network. A list with the names of all the nodes and their corresponding neighbors written beside them is prepared and sorted according to the highest number of neighbors each node has and all the possible nodes are selected in a top-down sequence. If the list of names of all the nodes is a subset of the neighbor list of that node we will designate it as the Admin node else we have to take any two nodes from the comprehensive sorted list in a top-down sequence. Next the union of the neighbors of the two selected nodes is considered, if the entire list results as a subset of the union those two nodes are considered as the Admin nodes. However for a negative result we take any three nodes from the sorted comprehensive list in a top-down sequence and continue the process. We continue this process increasing the number of nodes considered by unity until the subset criteria is satisfied after which the Admins are selected as those nodes whose friend list covers the complete network.

ASSOCIATIVE AND COMMON NODES

The Associative nodes are nodes lying in the region common to multiple Admin nodes. If any Admin node does not have an Admin node or an associative node attached to it, then an associative node pair is selected. It is a pair of nodes through

which Admin can communicate with the next Admin node. All the nodes in the network excepting the Admin nodes and the associative nodes are the common nodes.

Let us consider the following network:

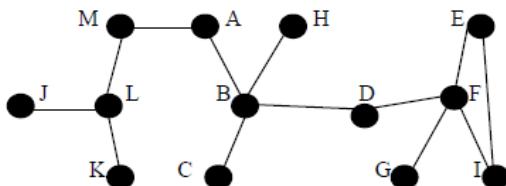


Fig. 1. A snapshot of an Ad-Hoc network

For the above network in fig. 1 the neighbor list has to be prepared first and only then can we assign the admin nodes. The neighbor list is as follows:

Nodes	Friends
A	A,B,M
B	A,B,C,D,H
C	B,C
D	B,D,F
E	E,F,I
F	D,E,F,G,I
G	F,G
H	B,H
I	E,F,I
J	J,L
K	K,L
L	J,K,L,M
M	A,L,M

The list in top down sequence is as follows:

Nodes	Friends
B	A,B,C,D,H
F	D,E,F,G,I
L	J,K,L,M
A	A,B,M
D	B,D,F
E	E,F,I
I	E,F,I
M	A,L,M
C	B,C
G	F,G
H	B,H
J	J,L
K	K,L

As we choose the admin nodes, it is clearly visible that node B alone does not cover the whole network, its neighbor nodes A, B, C, D, H are not all the nodes of the network. Hence we choose a pair of nodes and their union is considered. Hence F along with B is chosen and the resultant union gives us the nodes A, B, C, D, E, F, G, H and I. However still some nodes are missing, so we need to take a third node for union. We take L and get all the nodes. Now we can see node D is neighbor of both admin nodes B and F. Hence D acts as the associative node for admin nodes B and F. It is clear that there is no associative node attached to admin node L. So we choose A, M pair by which admin L can communicate with admin node B. So A, M pair is called the special associative node pair.

Hence in our example,

Admin nodes: B, F, L

Associative node: D

Common nodes: A, C, E, G, H, I, J, K, M

Associative node pair: A, M

WATCH NODES

Watch nodes have been used basically to promote security in the network. As is the case with our protocol, if we are able to implement security in the admin and the associative nodes, we can guarantee that the total network is secured from any attacks since the common nodes do not have any role to play in transmission of data apart from sending or receiving of data packets. Hence we have added two watch nodes to each admin and Associative node which checks after every time interval if any data packet entering into an Admin or Associative node goes out of the node within a stipulated time period, failing which it issues a warning to its previous admin that the node maybe a malicious node and the node is not used for transmission of data through the network. We have used the two neighboring nodes of each admin as the watch nodes in our protocol.

In the network in fig.1 if a data packet is sent from node A to node I then the path followed will be

A-B-D-F-I

When the data packet is at admin B, any two nodes from A, C, D and H will be selected as watch nodes and will keep an eye on the admin node B for any aberrant behavior.

This protocol also has certain aspects such as battery life, admin reselection and back tracking.

BATTERY LIFE

We realize that the nodes in an Ad-hoc network are constantly in motion and hence will run out of battery power sooner or later. However for our protocol, the battery life is particularly important for the Admin nodes since they perform the maximum amount of work. Hence when the power of a certain admin decreases below a certain level we need to get that battery to re-charge before it can take part in any retransmission. We have developed our protocol in such a way that for every admin, there is a special field for the battery life and if the threshold value of the battery of an

admin is reached, it immediately withdraws itself from the network and recharges and admin reselection takes place. After recharging it can be again reconsidered in the network.

ADMIN RESELECTION

There may be a few cases when admin reselection is required. If an admin node is found to be malicious it is blocked from the network immediately and the network chooses a new admin again in the same way as described above. Again if the battery power of an admin gets drained off completely then it is suspended temporarily from the network for recharging and admin reselection takes place.

BACK TRACKING

In many cases it may so happen that when a data packet reaches an admin, it has multiple paths to move to. The data packet may choose one path, but the destination may very well be on the other path. In such cases a back tracking is required. We have added two bits along with the data packet which records the last traversed admin and then forwards the data packet, if the data packet does not find its destination on the traversed path then it backtracks to the admin where it finds a multiple path and then moves to the other alternative path. In this way we re-transmit the data from an admin with multiple paths to reach the destination.

4 Algorithm

- **Select_admin**

Step 1: Every node which enters the network broadcasts hello packets //new node insertion

Step2: If there is no response then

There is no need to flood the neighbor list

The node itself is the admin

Step 3: If there is any response then

Update the friend list

The nodes flood their neighbor list across the network

Send a special request for presence of admin in the network //Admin_present = 1 or Admin_present = 0

Step 4: If there is no admin // Admin_present = 0

The node with the least ID number calls Compute()

The result is flooded across the network

Step 5: Else if there are previous admins

//Admin_present = 1

Then the previous admin with the least ID number will call Compute()

The result is flooded across the network

Step 6: After a certain time period every node in the network broadcasts hello packets //check for deletion or relocation of nodes

Step 7: Continue from Step 2

- **Compute()**

Step 1: Sort the friend list in descending order of number of friends

Step 2: union_result = 0

Step 3: While (union_result is a subset of entire_list)

Check neighbor list

Take next highest entry of list in descending order

Union_result = Union_result + neighbor_of_node[i]

if (union_result = entire_friend_list)

Set the nodes as admins

End If

End While

Step 4: Nodes common to multiple admins are associative nodes

Step 5: Other nodes are common nodes

- **Packet_sending**

Step 1: If sender = receiver

Sender is same as receiver, so packet will always be successfully sent

Else Step 2

Step 2: Packet is sent to the admin of the node

Set sender=admin

Traversed_admin_field =sender

If multiple path possible from current admin

Then

 Set Backtracking_bit = 1 for current admin

 /* Backtracking_bit is a bit field for every admin node whose value indicates whether backtracking is possible or not from that node onwards */

End If

/* Traversed_admin_field is a one dimensional array which stores the id of the admin nodes which have already been traversed by the data packet. This is used to prevent loopback.

When a packet reaches a node, it checks the Traversed_admin_field for the next admin's entry. If it finds the admin in the array, then it checks for the value of the Backtracking_bit. If Backtracking_bit=1, only then it allows the packet to move to the next admin */

Step 3: While (packet is not sent to the receiver)

 If sender = receiver

 Packet is sent successfully

 Generate and Send Ack

 Else

 If receiver is neighbor of admin

 Packet sent

 Generate and Send Ack

 Else if receiver is not neighbor of admin

 Packet sent to the next admin /* if current admin is within the range of the next admin */

```

        Sender=next admin
Traversed_admin_field =sender
    If multiple path possible from current admin Then
        Set Backtracking_bit = 1 for current admin
        /* Bactracking_bit is a bit field for every admin node whose value
           indicates whether backtracking is possible or not from that node
           onwards */
    End If
Else if current admin is not within next admin's range
    Packet sent to associative node to send it to next admin
    Traversed_admin_field =sender
    If multiple path possible from current admin Then
        Set Backtracking_bit = 1
    End If
Else
    Packet sent to special associative
    pair nodes to send it to next admin
    Traversed_admin_field =sender
    If multiple path possible from current admin Then
        Set Bactracking_bit = 1
    End If
If (next admin is not in traversed_admin field)OR (admin is in
traversed_admin_field and Backtracking_bit=1)
    Send packet to the next admin
    Sender=admin
    Traversed_admin_field =sender
    If multiple path possible from current admin Then
        Set Bactracking_bit = 1 for current admin
    End If
    ElseIf Bactracking_bit = 1 for current node
        Try alternative path
        Set Bactracking_bit = 0 for current node /* if all alternative paths
           have been exhausted */
    End if
    End if
End if
If all admins have been traversed atleast once but receiver not found /* receiver left
the network or failed or no such receiver id exists*/
    Then
    Drop packet
    Break while
    End If
    /* In this case, Sender does not get Ack and it assumes that either the packet was
lost in transit and did not reach the receiver OR the receiver is not present in the
network, hence retransmits the packet once more */
End While
End if

```

- **Watch_node**

Step 1: While(data_packet is at an admin[j])

Set friend_node[i] and friend_node[i+1] of admin[j] as watch nodes
End While

Step 2: Set count = 0

Step 3: If packet_sending(admin[j]) == false

```
/* Watch_node monitors the admin network traffic pattern */
Watch_node detects malicious activity
Report activity
count = count +1;
If count == 3 /* if the watch_nodes detect malicious activity
more than 3times */
    remove admin from network
    Select_admin();
End If
End If
```

- **Battery_Life**

Step 1: Set threshold_value for battery life of each admin

Step 2: While battery_life > threshold_value

```
perform network activity
battery_life = battery_life - 1;
```

End While

Step 3: If Admin(battery_life < threshold_value) then

Remove admin from network

Recharge admin

Select_admin(if allowed by network topology)

Else

Remove admin from network

Recharge admin

Insert the recharged admin into the network

End If

- **Admin_Failure**

If current admin crashes

admin reselection takes place

/*Only if network does not gets disconnected due to admin failure*/

Else

disconnected part of the network stalls until network topology changes or admin recovers, whichever is earlier

End If

5. Simulation Results

We have done simulation on java platform and based on simulation results we have plotted different graphs shown in section 8. Mainly we have created MANET dynamically based on user input and then simulate different attacks .Lastly we have

observed different performance issues like energy consumption ,delay and no of packet dropped whenever routing is going on.

6 Security Aspect

For any protocol, maintaining security[5] is absolutely necessary. If it does not secure data transmission among, then there is no point in using the protocol simply because it does not guarantee the proper delivery of correct data to correct receiver. There are many security attacks possible on a network. The ones this protocol guards against are:

1. Hello Flooding[6]: In this attack a dishonest node sends a repeating message through the network causing network congestion. Our protocol deals with this attack by associating a timestamp with every data packet; if a data packet, having the same timestamp[7] value and same source node number containing the same data, is repeated then the receiving node will simply discard the data packet.
2. Co-operative Black Hole attack [8]: In this attack a group of dishonest nodes act as a black hole, i.e., when a node receives a data packet it circulates the packet among them without sending it out of the loop, hence the data packet never reaches the destination. This protocol deals with re-transmission in such a way so as to stop this attack. If there is a requirement of a data packet to be sent back along the path it had come, then it is noted that the previous path does not have the receiver and the data packet can no longer go in that direction. In other words, our protocol avoids data being sent through loops in the network. Hence this attack is avoided.
3. Black Hole attack [9]: Black holes are those nodes in a network where incoming traffic is silently discarded but the source has no information that the data did not reach the intended recipient. This is one of the biggest security attacks that occur in a MANET.

We have used the concept of watch nodes in our network. Watch nodes act as guardians and check if an Admin node is correctly forwarding a data packet or not. If it finds out that on multiple occasions a data packet is not being forwarded by an Admin node, it assumes that the Admin may be malicious and will simply discard[10] the Admin node from the network. Since this protocol is very lightweight and the computations depend solely on the admin nodes, hence the security of the admin nodes ensures that the network is secured.

4. Gray Hole Attack:[11] A gray hole is similar to a black hole but it starts its action after it has been the part of a network for some time. It will behave as normally when the network starts its functioning but after a certain amount of time it will either consume all or some of the data packets that come its way.

Like the previous attack, the watch nodes once again detect if any Admin node does not successfully forward a data packet and if a node tries to act as a gray hole; it will suspend the node for aberration.

5. Sleep Deprivation[12]: The attacker attempts to consume batteries of other nodes by requesting routes, or by forwarding unnecessary packets.

This protocol has no route request mechanism, the entire route is based on the dynamic nature and it is decided during the packet sending. Hence this stops sleep deprivation. However if unnecessary packets are sent, hello flooding attack is stopped by checking the time stamp value. If packets having different time stamps are sent, then it is very difficult to distinguish a real data packet from an unnecessary one, in such a case, the battery life is drained. This protocol is designed in such a way that if the battery life of an Admin node is below a threshold level, it simply disconnects itself from the network until it can recharge itself. Hence although the battery power is drained off, the network activity does not stop.

7 Comparative Study with Existing Schemes

AODV has a lot of network activity associated with it since there are routing packets transmitted all over the network to know the desired route. In this protocol however, the admin nodes take the duty of network transmission and so the overall load on the network decreases many folds. The network traffic depends on the dynamic nature of the network, lower the amount of changes in the network, lower will be the network traffic. For DSDV each node connects to all other nodes in order to maintain their routing table[16]. However in this case, the admin nodes do this work and so there is a very low routing maintenance required for the networks. DSR has higher delay than our scheme.

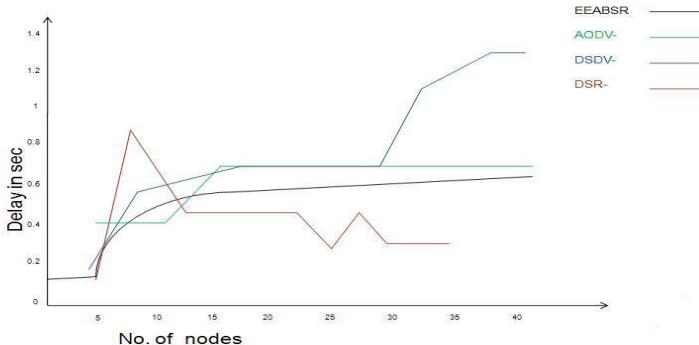


Fig. 2. Graphical comparison of EEABSR with existing protocols

The algorithmic complexity of $O(n^2)$ gives the parabolic nature[17] of the graph. As the number of nodes increases, the network congestion increases, gradually increasing the delay. But the delay is much stable compared to the other protocols.

In our protocol packets will be dropped only when the admin battery gets exhausted below a threshold[18]. In the meantime a new admin will be selected & the previous admin will be recharged. It performs consistently well.

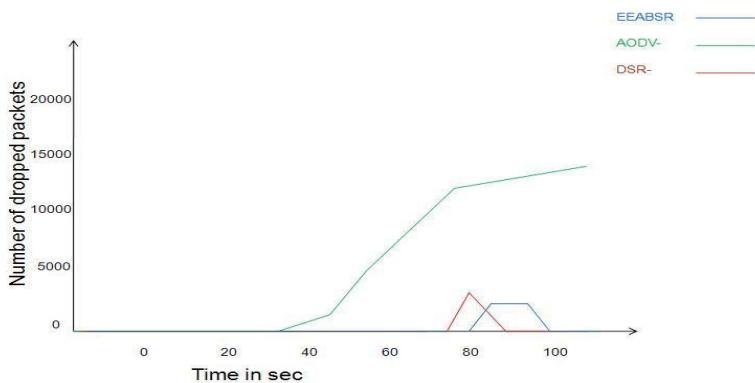


Fig. 3. Packets Dropped by EEABSR, AODV & DSR

We investigated the amount of energy consumed by participating[19] nodes according to the network card activities. We observed that overhearing consumed most of the energy. Idle power and overhearing effects dominate the energy consumption in the simulation of a dense network.

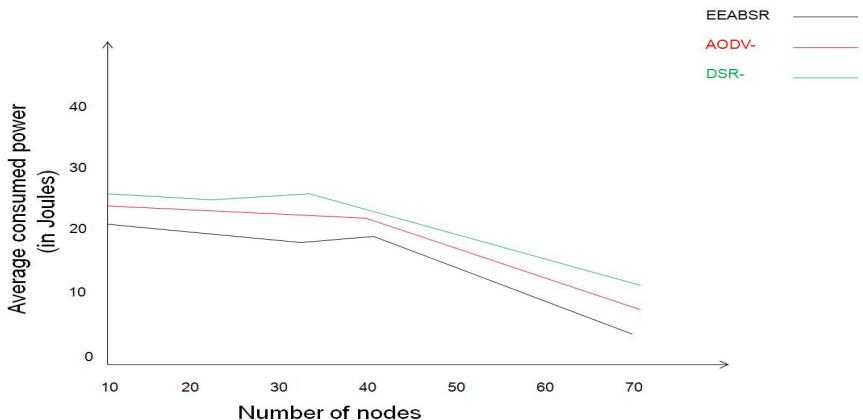


Fig. 4. Average consumed power by EEABSR, AODV & DSR

In our protocol, the network traffic is mostly restricted in between the admins and hence the overall overhead is reduced, thus reducing the overall energy consumption.

8 Conclusion

The algorithm actually is a basic design built to reduce network overhead[20] and computations and also to ensure absolute security[21] of the data packet transmitting through the network and it performs very efficiently in that respect. The rate of

topology [22] change must remain less or medium. If the topology changes constantly the protocol may be vulnerable. This protocol will be much more optimal[23] compared to the existing protocols such as DSDV[24] and also AODV[25] unless a small network is considered. There are various protocols which can send data very fast but then they have a lot of overhead attached to them. This protocol leads to a decent solution as it sends data at an optimal speed while taking care of the computational overhead.

References

1. Rubin, I., Behzad, A., Zhang, R., Luo, H., Caballero, E.: TBONE: A Mobile-Backbone Protocol for Ad Hoc Wireless Networks. In: Proceedings of IEEE Aerospace Conference, vol. 6, pp. 2727–2740 (2002)
2. Liu, K., Deng, J., Member, IEEE, Varshney, P.K., Fellow, IEEE, Balakrishnan, K., Member IEEE: An Acknowledgment-Based Approach for the Detection of Routing Misbehavior in MANETs. *IEEE Transaction on Mobile Computing* 6(5) (May 2007)
3. Matri, S., Giuli, T.J., Lai, K., Baker, M.: Mitigating Routing Misbehaviour in Mobile Ad Hoc Networks. In: Proceedings of Mobicom 2000, pp. 255–265 (2000)
4. Gonzalez- Computers & Security 25(18), 736–744 (2000)
5. Papadimitratos, P., Haas, Z.J.: Secure message transmission in mobile ad hoc networks, *Ad Hoc Networks*, pp. 193–209. IEEE (2003)
6. Hu, Y.-C., Perrig, A., Johnson, D.B.: Ariadne: a secure on-demand routing protocol for ad hoc networks. In: ACM MobiCom 2002, pp. 12–23 (2002)
7. Hu, Y.-C., Johnson, D.B., Perrig, A.: SEAD: Secure efficient distance vector routing for mobile wireless ad hoc networks. In: Proceedings of the 4th IEEE Workshop on Mobile Computing Systems and Applications (WMCSA 2002), pp. 3–13 (2002)
8. Komala, C.R., Shetty, S., Padmashree, S., Elevarasi, E.: Wireless Ad hoc Mobile Networks. In: National Conference on Computing, Communication and Technology, pp. 168–174 (2010)
9. Zhou, L., Hass, Z.: Securing Ad Hoc Networks. *IEEE Network Magazine* 13, 24–30 (1999)
10. Hu, Y.-C., Perrig, A., Johnson, D.B.: Wormhole Attacks in Wireless Networks. *IEEE Journal on Selected Areas in Communication*, 370–380 (2006)
11. Johnson, D.B., Maltz, D.A.: The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks, IETF Draft, 49 pages (October 1999)
12. Sanzgiri, K., Dahill, B., Levine, B.N., Shields, C., Royer, E.M.: A secure routing protocol for ad hoc networks. In: Proceedings of the 10th IEEE International Conference on Network Protocols, pp. 78–87 (2002)
13. Perkins, C.E., Royer, E.M., Das, S.R.: Ad Hoc On-demand Distance Vector Routing, IETF Draft (October 1999)
14. Perkins, C.E., Bhagwat, P.: Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers. *Comp. Comm. Rev.*, 234–244 (October 1994)
15. Das, S.R., Perkins, C.E., Royer, E.M.: Performance Comparison of Two On-demand Routing Protocols for Ad Hoc Networks
16. Perkins, C., Belding-Royer, E., Das, S.: Ad hoc On-Demand Distance Vector (AODV) Routing (2003)

17. Avoiding Black Hole and Cooperative Black Hole Attacks in Wireless Ad hoc Networks, <http://www.scribd.com/doc/26788447/Avoiding-Black-Hole-and-Cooperative-Black-Hole-Attacks-in-Wireless-Ad-hoc-Networks>
18. Razak, S.A., Furnell, S., Clarke, N., Brooke, P.: A Two-Tier Intrusion Detection System for Mobile Ad Hoc Networks- A Friend Approach. Springer, Heidelberg (2006)
19. Lee, J.-S., Chang, C.-C.: Secure communications for cluster-based ad hoc networks using node identities. *Journal of Network and Computer Applications, International Journal of Computer Science and Security* 1(1), 67 (2006)
20. Johnson, D.B., Maltz, D.A., Broch, J.: DSR: The Dynamic Source Routing Protocol for Multi-Hop Wireless Ad Hoc Networks
21. Zhang, Y., Lee, W.: Intrusion Detection in Wireless Ad Hoc Networks. In: Proceedings of Mobicom 2000, pp. 275–283 (2000)
22. Khokhar, R.H., Ngadi, M.A., Mandala, S.: A Review of Current Routing Attacks in Mobile Ad Hoc Networks. *International Journal of Computer Science and Security* 2(3), 18–29; International Conference on Wireless Networks (ICWN 2003), Las Vegas, Nevada, USA, pp. 570– 575 (2003)
23. Blum, J.J., Member, IEEE, kandarian, A.E., Member, IEEE: A Reliable Link-Layer Protocol for Robust and Scalable Intervehicle Communications. *IEEE Transactions on Intelligent Transportation Systems* 8(1) (March 2007)
24. Shanthi, N., Lganesan, Ramar, K.: Study of Different Attacks on Multicast Mobile Ad-hoc Network. *Journal of Theoretical and Applied Information Technology*, 45
25. Klein-Berndt, L.: A Quick Guide to AODV Routing

Effect of Mobility on Trust in Mobile Ad-Hoc Network

Amit Kumar Raikwar

Department of Computer Science and Engineering
Motilal Nehru National Institute of Technology Allahabad,
Uttar Pradesh -211004, India
amitfrm.mnnit@gmail.com

Abstract. In MANET environment intermediate nodes on a communication path are expected to forward packets of source node towards destination node. This enhances the communication range of mobile nodes beyond their wireless transmission range. From the security point of view a proper mechanism should be there, so that nodes can trust each other for safe delivery of their data. Therefore trust establishment is considered as an important approach to defend safe communication in MANET. In the previous years several trust models have been proposed to enhance the security of mobile ad hoc network. However an important factor, mobility of node, before establishing the trust relationship among nodes is yet not discussed. In this paper trust model is been proposed. In this model the Trustworthiness of each node is calculated with the mobility factor. Minimal configuration, quick deployment and absence of central governing authority make ad hoc networks suitable for emergency situations. Natural disasters, military conflicts are some real life scenarios where trust among nodes is very important. Thus this model will be very effective in these situations. Speed of node is used to decide on the trust value for the trustee node.

Keywords: Ad hoc network, Trust, Mobility, Trustworthiness.

1 Introduction

1.1 Trust

Trust is defined as the subjective belief of one node in the willingness and ability of another node to follow the protocol of the ad hoc network in order to process a certain transaction. The definition has the following implications.

1. *Trust is unidirectional* A node that trusts another one might not be trusted from the other side in turn.
2. *Trust is subjective* each node computes the trust values from its own criteria and thresholds. Even when all nodes have the same information, one node could trust a particular other one, while others do not.
3. Trust is the estimated probability that a transaction succeed. The notation of trust cannot give guarantees and contains the risk of frauds. Thus, Security-Related applications must not solely rely on trust management system.

1.2 Mobile Ad hoc Network

A Mobile Ad hoc NETwork (MANET) is a collection of wireless nodes communicating with each other in the absence of any infrastructure. Due to the availability of small and inexpensive wireless communicating devices, the MANET research field has attracted a lot of attention from academia and industry in the recent years.

In the near future, MANETs could be used in various applications such as mobile classrooms, battlefield communication and disaster relief applications.

Misbehaving Node Most of the approaches for ad hoc network assume that the node readily and honestly follow the protocol. Unfortunately, this assumption does not always hold in reality. Misbehaving in ad hoc networks can take place at different levels of the system architecture.

(a) On the communication level, nodes can refuse to forward messages of others in order to save bandwidth and energy. Many misbehaving nodes would lead to a low reliability of the system. But such nodes can be detected, because every node can eavesdrop the network traffic of adjacent nodes. Most of the research on reputation and trust models in ad hoc networks focus on this issue.

Why should a node participate in the system, but deny to follow the protocol properly?

There are basically four reasons for this kind of behavior.

(a) Cost

In most application areas, the economic dominant strategy in a de-centralized system with many participants is to limit the quality of the offered services.

(b) Attacks

Nodes in ad hoc networks can be subject to attacks. A node might try to expel peers it competes with, block particular services or data, or tamper with transactions of other nodes. This can be done with Denial-of-Service (DoS) attacks.

(c) Social Issues

Collaboration in ad hoc networks has many social aspects. For instance, nodes could take revenge for failed transactions or try to gain attraction by disturbing the network.

(d) Technical Problems

Misbehavior can be the result of technical problems. It is hard to distinguish between nodes with technical difficulties and nodes which intentionally do not follow the protocol. However, from the perspective of other nodes it is not important whether a node is faulty by intention or not.

1.3 Mobility Model

The mobility model is designed to describe the movement pattern of mobile users, and how their location, velocity and time change over time. There are various mobility models available, different mobility model can be differentiated according to their spatial and temporal dependencies.

(a) Spatial dependency

It is a measure of how two nodes are dependent in their motions.

(b) Temporal dependency

It is a measure of how current velocity is related to previous velocity of a node.

The most commonly used mobility models are

- (a) Random Way point Mobility Model
- (b) Random Point Group Mobility(RPGM)
- (c) Freeway Mobility Model
- (d) Manhatten Mobility Model

Since the Random Way point Mobility Model is used in the model. Thus Only Random Way point mobility is discussed in detail here.

Random Mobility Model. In random-based mobility models, the mobile nodes move randomly and freely without restrictions. To be more specific, the destination, speed and direction are all chosen randomly and independently of other nodes.

Random Way Point Mobility Model. The Random Way point Model was first proposed by Johnson and Maltz. Because of its simplicity and availability it became the most widely used mobility model. The Random Way point mobility model work as follows.

As the simulation starts, each mobile node randomly selects one location in the simulation field as the destination. It then travels towards this destination with constant velocity chosen uniformly and randomly from $[0, V_{max}]$, where the parameter V_{max} is the maximum allowable velocity for every mobile node. The velocity and direction of a node are chosen independently of other nodes. Upon reaching the destination, the node stops for a duration defined by the pause time parameter T_{pause} . If $T_{\text{pause}} = 0$, this leads to continuous mobility. After this duration, it again chooses another random destination in the simulation field and moves towards it. The whole process is repeated again and again until the simulation ends. As an example, the movement trace of a node is shown in the figure.

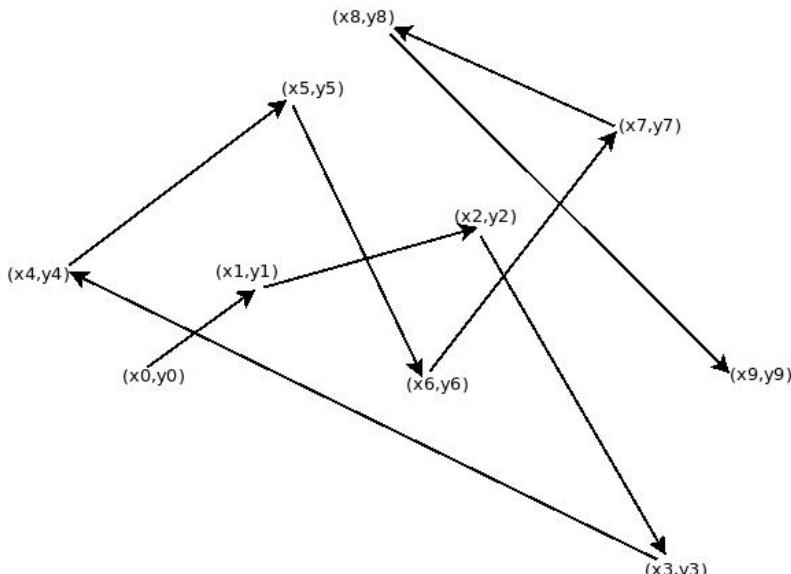


Fig. 1. Random Way Point Mobility Model

1.4 Calculation of Trustworthiness for Two Nodes

Lets the two nodes be i and h for which Trustworthiness will be calculated. Furthermore Trustworthiness at each time instant k of a node is given $T_{h,i}(t, c)$ is defined as a combination of trust and confidence values that node h has regarding i . For calculating trust and confidence a beta distribution in combination with a modified Bayesian approach is used (cf [10]). The mean of the beta calculation is used as trust value and the variance is used as the confidence value. The trust t^k as used in [10] is dened as the mean value $\mu(a_k, b_k)$, of the beta (a_k, b_k) distribution corresponding to the pdf $f_k(r)$ as follows:

$$t^k = \mu(a_k, b_k) = \frac{a_k}{a_k + b_k}$$

The confidence value, c^k as defined in [10], associated with the trust value t^k in terms of standard deviation $\sigma(a_k, b_k)$ corresponding to pdf $f_k(r)$ as follows:

$$c^k = 1 - \sqrt{12}\sigma(a_k, b_k)$$

In [11] aging is applied to the trust and confidence values based on an observation window mechanism. In this concept a weighted average of N trust values from the past are used in order to age older values according to following equation.

$$t_{(h,i)}(k) = \frac{2}{N(N+1)} \sum_{l=1}^N l \cdot t_{h,i}(k - N + l - 1) \quad (1)$$

Definition: The *Trustworthiness* $T_{(t,c)}$ for specific trust and confidence values $t, c \in R$ is dened by the following equation (as proposed in [10]):

$$T(t, c) = 1 - \frac{\sqrt{\frac{(t-1)^2}{x^2} + \frac{(t-1)^2}{y^2}}}{\sqrt{\frac{1}{x^2} + \frac{1}{y^2}}} \quad (2)$$

In this equation the parameters x and y are referred as Trustworthiness parameters. The impact of the choice of parameters x and y on the mapping of (t, c) to T . The parameters x and y in the equation above can be used to adjust the influence of trust and confidence on the Trustworthiness.

- a) When $x > y$, implies that the confidence value has greater weight than the trust value for the majority of points on ellipse.
- b) When $x = y$, the ellipse becomes a circle. Here the value of t and c have equal weight in determining T .
- c) When $x < y$, implies that the trust value has greater weight than the confidence value for the majority of the ellipse points.

2 Context for the Model

In our model the nodes are assumed to perform fixed range transmission using omnidirectional antenna. The random way point model is used to calculate the mobility of

a node. In random way point model at every instant, a node randomly chooses a destination and moves towards it. The velocity is chosen randomly from a uniform distribution $[V_{min}, V_{max}]$, where V_{max} is the maximum allowable velocity for every mobile node. After reaching the destination the node stops for duration. This is called the pause time parameter. After this duration, it again chooses a random destination and repeats the whole process until whole simulation ends. Average speed of the node can be calculated by,

$$V_{avg} = \frac{V_{max}-V_{min}}{2} \quad (3)$$

Our approach in treating mobility as a trust correction factor rely on the observation that the number of packets dropped increases with the increase in the speed of node. Suppose there's an trustee node who wishes to forward the packet to the destination with the help of intermediate nodes. The previous assumptions assumes node to be in rest and stable for most of the packet transfer time and thus trust on a node was calculated assuming node to be in stable state. Now taking a more practical scenario let us assume the trustee node to be moving while the packet is being under transmission. Thus the packet dropped will increase, decreasing the trust value and so the Trustworthiness. The packet dropped is proportional to the speed of node. The more is the speed of node, the more will be the packet dropped. Thus based on the above observation a correction factor called "Mobility factor" is been proposed. The mobility factor is inversely proportional to the speed of node. Thus with the increase in speed of node the mobility factor will decrease and so the Trustworthiness.

Thus the Truster node will decide the Trustworthiness value for trustee nodes on the basis of the Trustworthiness and the mobility factor of the trustee node. If the speed of node is between the minimum speed of a node and the calculated average speed. The mobility factor μ_t is set to 1. When this 1 is multiplied as a correction factor multiplied with the trust. The trust value will not be affected so is the value of Trustworthiness. Thus it can be inferred that the Trustworthiness value will not differ if the speed of node is between V_{min} and V_{avg} . If the speed of node exceeds the average speed but is less than the maximum speed. The value for mobility factor is reduced linearly with the linear increase in speed of node. Thus the value for Trustworthiness will decrease with the increase in speed of node. If the speed of node exceeds the maximum allowable speed of the node. The mobility factor t of trustee node is reduced to 0, exempting the trustee node from taking part in further transmission. Thus the truster node will follow a different route exempting the malicious node.

Thus mathematically,

$$Trustworthiness(T) \propto trust \quad (4)$$

$$trust \propto Mobilityfactor (\mu_t) \quad (5)$$

$$Mobility factor (\mu_t) \propto \frac{1}{Speed of the Node (V^t)} \quad (6)$$

$$Trustworthiness \propto Mobilityfactor (\mu_t) \quad (7)$$

$$Trustworthiness \propto \frac{1}{Speed of the Node (V^t)} \quad (8)$$

As the speed of the node increases, the value of its Trustworthiness decreases.

3 Background and Related Work

The Our main focus is revolving around the calculation of trust based on the mobility of a node. Therefore firstly the fundamental concept of trust and trust management system is defined below. The different mobility models are then seen thus random way point mobility model is discussed.

The motion of a node for a small period of time is defined by a mobility vector. The mobility vector is chosen by the node independently for a random period of time [1]. The mobility vector gives the speed and direction of the node. A mobility model that allow for varying speeds during a movement has also been proposed [2]. The effect of mobility on the capacity of ad hoc network was analyzed in [3]. The random way point model is the most widely and popularly used model for simulation. Several studies of this model have revealed various flaws or unexpected properties. Mobility models such as in [4] [5] have been shown to suffer from speed decay Phenomenon. In speed decay phenomenon the average speed of mobile nodes decays with time. In [6] it has been shown from simulation that the random way point model does not lead to a uniform distribution of nodes location. Stochastic properties of certain characteristics of the random waypoint model have been analyzed in [7] [8]. Studies of ad hoc networks have measured mobility in terms of the changes in the underlying transmission graph. C. Zouridaki et.al in [10] presented Hermes, a quantitative trust establishment framework for MANETs, which is designed to improve reliability of packet forwarding over multi-hop routes in the presence of potentially malicious nodes. The framework defined two metrics, trust and confidence, which are computed using a Bayesian approach based on empirical first-hand observations of packet forwarding behavior by neighbor nodes. The trust and confidence metrics are mapped into a single trustworthiness metric. Peter Ebinger et.al in [11] proposed a system for trust Evaluation and Reputation Exchange for cooperative intrusion detection in MANETs(TEREC). They propose to split the reputation information into two values trust and confidence. This allows each node to successively determine the reliability of other nodes without the need or reliance on a static, pre-established trust infrastructure which requires significant overhead and cannot be recovered once compromised. TEREC is evaluated via simulation and its performance measured in the presence of an increasing amount of malicious nodes. Evaluation results show that a benign majority of nodes prevail over malicious attacking nodes based on reputation estimations.

4 The Proposed Approach

4.1 Calculation of Trustworthiness for Two Nodes

In random way point model, let us assume that each nodes movement is characterized by non overlapping time period $X_1, X_2 \dots X_n$. Let the node cover some random distance D_i with random speed V_i . Let the time required be X_i .

Let, V^t be the speed of the node at time instant $t > 0$.

Let F_V^t and f_V^t be its corresponding cumulative distribution function(cdf) and probability distribution function(pdf).

As explained in [12], the model need to analyze these functions for calculating speed of node in respect of $t \rightarrow \infty$.

V^t has the following limit probability function i.e. cdf and pdf :

$$\lim F_V^t(z) = \frac{E[D]}{E[X]} \int_0^z \frac{1}{x} dF_V(x) \quad (9)$$

$$\lim f_V^t(z) = \frac{E[D]}{E[X]} \cdot \frac{1}{z} f_V(Z) \quad (10)$$

where, $E[D]$ and $E[X]$ are the mean of distance D and renewal period X.

4.2 Calculation for Average Speed of Node

Considering the random way point mobility model. A node chooses a random co-ordinates (x, y) within a grid then move to another coordinate choosing its speed randomly between $[V_{min}, V_{max}]$ for a certain period of time. Then at every T seconds this process is repeated. The pause time is assumed to be zero in this model and therefore the node is in continuous movement. The mean speed for nodes in motion will be $/2$, but the average speed over all nodes will be less. Average speed of the node is calculated from the speed, uniformly chosen from $[V_{min}, V_{max}]$.

Where $V_{max} > V_{min} > 0$.

Average speed of the node is:

$$V_{avg} = \frac{V_{max} - V_{min}}{2} \quad (11)$$

4.3 Mobility Factor (μ_t)

Mobility factor is a dimensionless quantity whose value ranges from 0 and 1. The value of mobility factor depends totally on the speed of node. Mobility factor is inversely proportional to the speed of node. The mobility factor remains constant till the speed of node is between the minimum and the average speed of node. Thereafter with the increase in the speed of node the mobility factor will decrease linearly.

In this section the mobility factor is calculated μ_t for nodes in a network. As discussed in section 1, the approach revolve around the assumption that the number of packets dropped will increase with the increase in the speed of node. With the increase in number of dropped packets the trust value for node will decrease and so the Trustworthiness. Thus based on this fact the Trustworthiness of a node is calculated. The mobility factor as a correction factor is multiplied with the trust. The Truster node will decide the best forwarder among the neighboring trustee nodes on the basis of the value of Trustworthiness with mobility factor.

let the mobility factor be denoted by μ_t .

On the basis of the speed of node the value of t can be calculated by the following algorithm,

Theoretically it can be inferred that,

If the speed of node is between the minimum speed of a node and the calculated average speed. The mobility factor μ_t is set to 1. When this 1 is multiplied as a correction factor multiplied with the trust. The trust value will not be affected so the value of Trustworthiness. Thus it can be inferred that the Trustworthiness value will

not differ if the speed of node is between V_{min} and V_{avg} . If the speed of node exceeds the average speed but is less than the maximum speed. The value for mobility factor is reduced linearly with the linear increase in speed of node. Thus the value for Trustworthiness will decrease with the increase in speed of node. If the speed of node exceeds the maximum allowable speed of the node. The mobility factor μ_t of trustee node is reduced to 0, exempting the trustee node from taking part in further transmission. Thus the truster node will follow a different route exempting the malicious node.

Algorithm 1. Mobility Factor

```

1. if ( $V_{Min} \leq Speed\ of\ Node \leq V_{avg}$ ) do
2.   Mobility factor ( $\mu_t$ ) = 1
3. else
4.   if ( $V_{avg} < Speed\ of\ Node \leq (0.75\ V_{avg} + 0.25\ V_{Max})$ ) do
5.     Mobility factor ( $\mu_t$ ) = 0.8
6.   else
7.     if ( $(0.75\ V_{avg} + 0.25\ V_{Max}) < Speed\ of\ Node \leq (0.5\ (V_{avg} + V_{Max}))$ ) do
8.       Mobility factor ( $\mu_t$ ) = 0.6
9.     else
10.      if ( $0.5\ (V_{avg} + V_{Max}) < Speed\ of\ Node \leq (0.75\ V_{Max} + 0.25V_{avg})$ ) do
11.        Mobility factor ( $\mu_t$ ) = 0.4
12.      else
13.        if ( $(0.75\ V_{Max} + 0.25\ V_{avg}) < Speed\ of\ Node \leq V_{Max}$ ) do
14.          Mobility factor ( $\mu_t$ ) = 0.2
15.        else
16.          if ( $Speed\ of\ Node > V_{Max}$ ) do
17.            Mobility factor ( $\mu_t$ ) = 0
18.          end if
19.        end if
20.      end if
21.    end if
22.  end if
23. end if
```

5 Calculation of Final Trustworthiness (T_F) on a Node

Thus Trustworthiness is calculated by merely multiplying the mobility factor with the trust as a correction factor.

The Updated Trustworthiness is denoted by T.

Thus,

$$T = t \times \mu_t \quad (12)$$

Substituting the value of Trustworthiness in above equation,

$$T_F = \frac{1 - \sqrt{\frac{(t-1)^2 + (t-1)^2}{x^2 + y^2}}}{\sqrt{\frac{1}{x^2} + \frac{1}{y^2}}} \times \mu_t \quad (13)$$

The Trustworthiness calculated is for a particular instant of time and is named as Current Trustworthiness. To calculate Trustworthiness based on both the present and past behaviour of the node, node maintains a trust table consisting of 3 fields. The combination of both the present and past behaviour of a node is given by the Updated Trustworthiness. The trust table with each node consists of three fields namely.

Table 1. Trust table with each node

Node Number	Present_Trustworthiness	Previous_Trustworthiness	Updated_Trustworthiness
1	0.40	0.30	0.35
2	0.60	0.50	0.25
3	0.40	0.20	0.30
4	0.70	0.60	0.65

- a) **Present Trustworthiness** Trustworthiness calculated at current time.
- b) **Previous Trustworthiness** Trustworthiness calculated previously.
- c) **Updated Trustworthiness** The node will calculate the updated Trustworthiness according to the following algorithm.

Algorithm 2. Trustworthiness

```

1. if (Present_Trustworthiness ≥ Previous_Trustworthiness) do
2.   Updated_Trustworthiness =  $\frac{\text{Previous_Trustworthiness} + \text{present_Trustworthiness}}{2}$ 
3. else
4.   Updated_Trustworthiness =  $\frac{\text{Previous_Trustworthiness}}{2}$ 
5. end if
6. Previous_Trustworthiness = Updated_Trustworthiness

```

6 Simulation

The Trustworthiness The simulation is done in matlab and the results are analyzed with the plotted graphs. This section compares the final result for Trustworthiness of a node with and without the mobility factor in it through a graph. For simplicity of understanding the Updated Trustworthiness with mobility factor is represented by TRUST. In the below shown graph all the other factors other than mobility factor are kept constant. The value for Trustworthiness and TRUST are then analyzed. As seen in the graph it can be inferred that after multiplying mobility factor, if the mobility factor is 1, the TRUST value is same as Trustworthiness. Moreover with the increase in speed of node the mobility factor decreases and so is the TRUST. It is also observed through

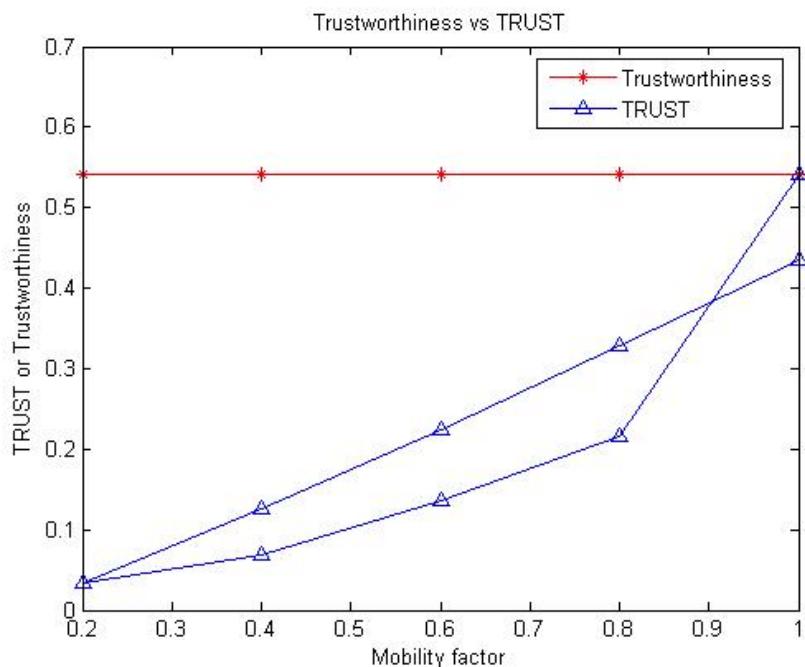


Fig. 2. Trustworthiness vs Mobility

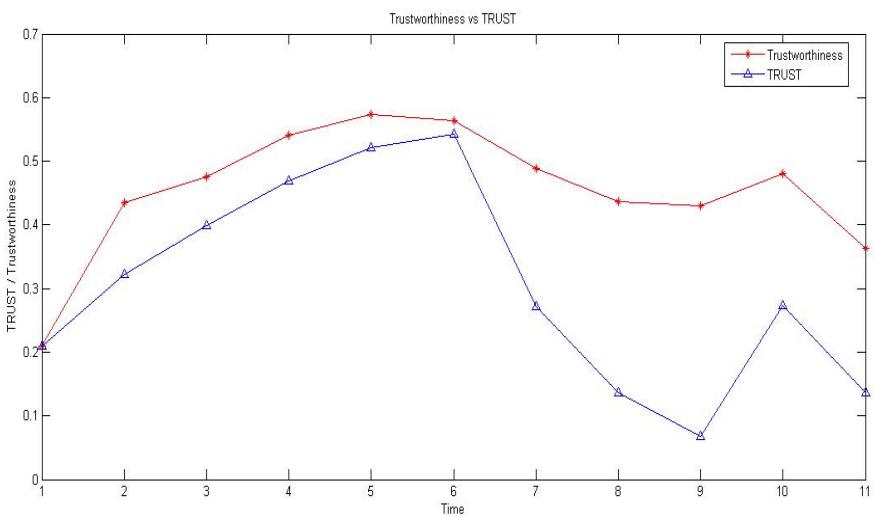


Fig. 3. Trustworthiness vs Time

graph that the TRUST of node falls sharply if there seems a decrease in trust but rises slowly and linearly with the increase in trust. Thus node have to be more disciplined, so that it can come into the completely trusted region. Thus our mobility factor make the trust model much secure as now node will behave in a more disciplined way with the fear or losing their previously gained Trustworthiness.

This graph shows how the value for Trustworthiness and TRUST, rise and falls with the time. It can be inferred that the TRUST value rises slowly but falls drastically. The Sharp fall in TRUST value force the nodes to behave as a small misbehaving can throw them in a region of untrusted node. Thus with the fear of loosing its previously gained trust. Thus this make the model more secure.

7 Conclusion and Future Work

In this paper a new Trust model has been proposed for mobile nodes through which it explicitly represented and managed to present mobility as an important factor in their trust relationship with others nodes. This model managed to calculate the Trustworthiness of a node based on both his past record and the current movement of a node. Simulation results demonstrated the effectiveness in distinguishing the malicious node from the non malicious node with greater discipline. Thus our mobility factor when multiplied as a correction factor to the Trustworthiness make the Trust model more secure. The studies and observations done in this model are based on the Random Way Point mobility model. As a future work the study can be done with other mobility models and simulating it in NS-2 and then comparing the effectiveness of all models.

References

1. Liang, B., Haas, Z.H.: Virtual Backbone generation and maintenance in ad-hoc network mobility management. In: Proceedings of the IEEE INFOCOM (March 2000)
2. Schindelhauer, C., Lukovszki, T., Riihrup, S., Volbert, K.: Worst case mobility in ad hoc networks. In: Proceedings of the 15th Annual ACM Symposium on Parallelism in Algorithm and Architecture, pp. 230–239 (June 2003)
3. Grossglauser, M., Tse, D.: Mobility increases the capacity of ad-hoc wireless networks. In: Proceedings of the Twentieth Annual Joint Conference of the IEEE Computer and Communication Societies (INFOCOM 2001), April 22-26, pp. 1360–1369. IEEE Computer Society, Los Alamitos (2001)
4. Yoon, J., Liu, M., Noble, B.: Random waypoint considered harmful. In: Infocom 2003 (April 2003)
5. Royer, E.M., Melliar-Smith, P.M., Moster, L.E.: An Analysis of the optimum node density for ad hoc networks. In: IEEE International Conference on Communication (ICC) (June 2001)
6. Bettstetter, C., Krause, O.: On border effects in modelling and simulation of wireless ad hoc networks. In: 3rd IEEE International Conference on Mobile and Wireless Communication Networks (MWCN) (August 2001)

7. Bettstetter, C., Hartenstein, H., Perez-Costa, X.: Stochastic properties of the random waypoint mobility model: Epoch length, direction distribution, and cell change rate. In: ACM MSWiM (September 2002)
8. Bettstetter, C.: Topology properties of ad hoc networks with random waypoint mobility. In: ACM MobiHoc (June 2003)
9. Sztompka, P.: Trust: A Sociological Theory, Seidman, S., Alexander, J.C.(eds.) Cambridge University Press (2000)
10. Zouridakis, C., Mark, B.L., Hejmo, M., Thomas, R.K.: A quantitative trust establishment framework for reliable data packet delivery in manets. In: SASN 2005: Proceedings of 3rd ACM Workshop on Security of Ad hoc and Sensor Network (2005)
11. Ebinger, P., Bibmeyer, N.: TEREC: Trust Evaluation and Reputation Exchange for Cooperative Intrusion Detection in MANETs. In: Seventh Annual Communication Networks and Services Research Conference (2009)
12. Lin, G., Noubir, G., Rajaram, R.: Mobility Models for Ad hoc Network Simulation. In: IEEE Infocom 2004 (2004)

Securing Systems after Deployment

David(DJ) Nea¹ and Syed (Shawon) Rahman²

¹ Capella University
Minneapolis, MN 55402, USA
`DJ@neal.ws`

² Dept. of Computer Science & Engg.
University of Hawaii-Hilo
200 W. Kawili St, Hilo, HI 96720, USA
`srahman@hawaii.edu`

Abstract. Applications are generally designed and developed with little regards to security consideration. Fortunately, there is abundant of processes and technologies today that is available that can be used to easily secure an application while it is in the maintenance phase. In this paper, we have discussed how we can use symmetric and asymmetric cryptography methods and security architecture can be created to protect a system from various cipher attacks after deployment.

Keywords: System Security, Cryptography, Security Architecture, Database Security, Application Security, Cipher Attacks, deployment, software maintenance

1 Introduction

Securing an application system after deployment can be accomplished by mapping out a security architecture that addresses all the end points that can be vulnerable to that application. Encryption technologies are available using methods provided by secret key cryptography and public key cryptography that can be used to create a security plan that evaluate the role encryption will play in protecting the rewritten application's data. By analyzing the architecture and design issues within the database management system, the application and the operating system the various cryptography methods can be used to defend the newly rewritten application against an abundance of cipher attacks.

2 Security Architecture

The security architecture needs to determine if the application is going to be based on a centralized or de-centralized structure. The role-based access control system is going to be centralized if the application exists in a Windows Active Directory domain using Kerberos, which will also be used to control access to the SQL server database, FTP Server, Exchange Server, and File server for the various assets that is available by the application system. The Security Diagram is an outline of all resources potentially involved with the interaction of an application. Individuals will be granted access to certain resources depending on their role and job functions within an organization is outlined in Figure 1.

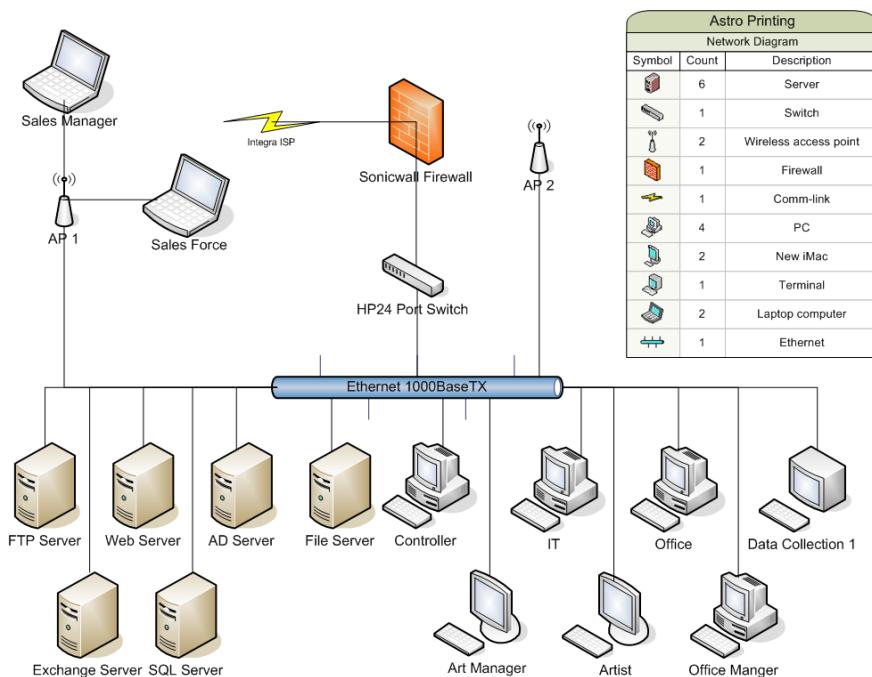


Fig. 1. Security Resources available within an organization

Based on the security architecture setup both symmetric and asymmetric methods will be applied to the application project. Symmetric will be applied to secure data sets, databases and application code; while asymmetric will be applied to secure transmission between servers and clients via IPSec and SSL/TLS protocols.

Symmetric encryption will be used to store passwords in a centralized database provided by Windows Active Directory services using the Kerberos protocol. Asymmetric encryption will be used in a decentralized environment with each digital certificate stored on each server which is to be used to provide secured communication via SSL/TSL protocols. Based on the chosen architecture the cryptography methods chosen for the application will include both symmetric and asymmetric encryption.

Symmetric methods include Advanced Encryption Standard (AES) for long term storage and Secure Hash Algorithm (SHA) for one way has functions [3]. Asymmetric method includes the RSA encryption to be used with public and private keys for digital certificates provided by a Certificate Authority. Current list of technologies that can be used with an application project:

- Centrify Express for single sign-on for Mac users, www.centrify.com
- Fortify 360 for application vulnerability testing, www.fortify.com
- Veracode for application vulnerability testing, www.veracode.com
- GoDaddy.com for SSL digital certificate, www.godaddy.com
- VMware Workstation for secure development platform, www.vmware.com

The development servers and production servers should be harden Windows servers which will only be used as a web server and SQL server. Microsoft Baseline Security

Application will be used to harden the web server and SQL Server for best practices to keep only essential services running for best protection. Communication between the web server and database server will be setup to use IPSec on a different network segment. Availability to the server will only be provided using IPSec with all of the internal traffic behind the firewall. External availability will only be permitted to the public via the web server that is to be used to host the application. By providing IPSec services between servers, this prevents eavesdropping and replay cipher attacks on the local network. To thwart against client cipher attacks, users are required to meet the organizations password policy when setting up their own password [3].

3 Cipher Considerations

The Active Directory system will be setup on a role-base access control system so that users, developers, managers, and project leaders access the correct content of the project based on roles and functions. Asymmetric encryption will be needed to provide integrity and confidentiality to the project's methods of accessing the application over the internet using digital certificates from a Certification Authority. Based on the chosen architecture the cryptography methods chosen for the application will include both symmetric and asymmetric encryption. Symmetric methods include Advanced Encryption Standard (AES) for long term storage and Secure Hash Algorithm (SHA) for one way has functions [3]. Asymmetric method includes the RSA encryption to be used with public and private keys for digital certificates provided by a Certificate Authority. Figure 2 outlines the use of encryption used over the footprint of the whole project.

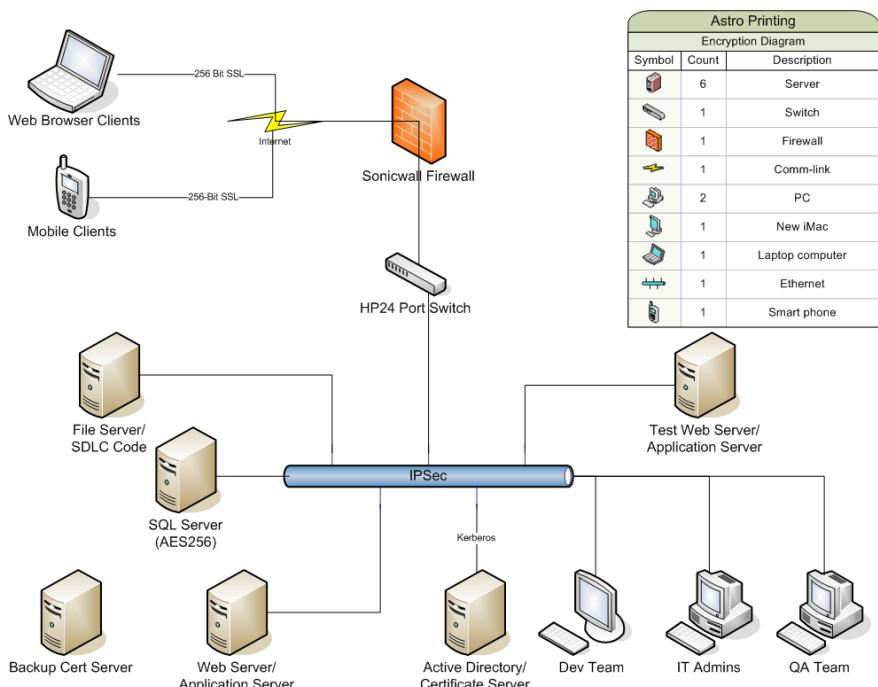


Fig. 2. Encryption Diagram

4 Database Security Considerations

Database security considerations consists of the following, platform and network security, principles and database object security, application security, and SQL Server security tools, utilities, views and functions [4].

Platform and network security considerations include the physical security, operating system security, and the SQL Server operating system files security. Physical security considerations can include locked doors to restrict access to the physical database server and equipment that is used to access that server, such as routers, switches, cabling, and backup tapes [4]. Based on Microsoft's recommendation it is better protection to backup databases first to files using NTFS encryption file system (EFS) than to backup to tape, rather than using the weaker protection of media set password and backup set password [4]. Considerations need to be taken and followed to keep unauthorized users off the SQL Server network. This can be done by assigning users to specific tabular data streams (TDS) endpoints within the database, and to restrict network access by disabling null sessions [4]. Operating system (OS) security considerations can include applying OS service packs and upgrades to ensure the latest security enhancements are implemented. Plus, the OS firewall is to be configured properly to protect the SQL Server services, integration services, analysis services, reporting services, and to ensure the surface area of the SQL Server is limited with "least privileges" for the required services [4]. Since this project is including the Internet information server (IIS), then security considerations need to be made to lockdown IIS security with SQL Server as well [4]. SQL Server operating system files security considerations can include limiting permissions to the SQL Server program files, database files, and log files [4]. To accomplish this, options would be to utilize NTFS access control lists to restrict the whole database structure and platform.

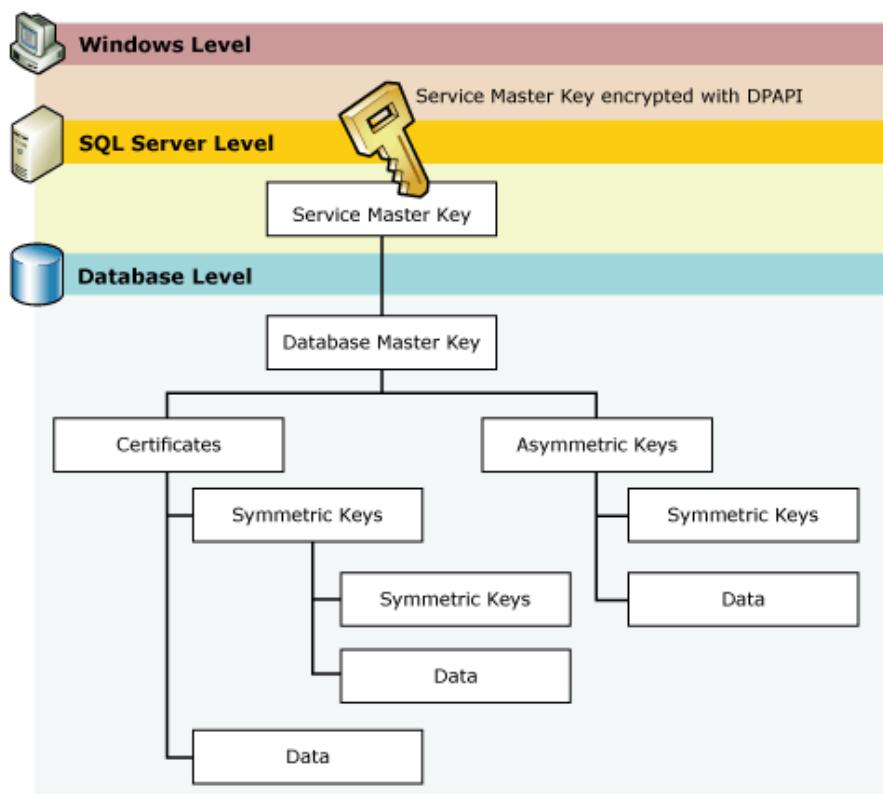
Principles and database object security is the next section to consider for database security. Principles are SQL Server entities that request resources and have a security identifier (SID) and securable are the resources that can be supplied by SQL Server database engine [4]. When creating the application, developers need to document the process which securable within the SQL Server needs to be assigned to the least restrictive principles based on the assigned role-based access control. The securable are in Table 1 [4].

To enhance security by limiting data loss that can occur from the problems of passing access control, encryption can be used for that rare occurrence when access control has been compromised. SQL Server provides three mechanisms for encryption, certificates, asymmetric keys and symmetric keys [4].

Based on the developer's needs for encryption Figure 3 displays the different levels of encryption that SQL Server will be able to encrypt data based on different hierarchical key management infrastructures [4]. Any parameters already created within the application that uses SQL Server will have to be converted to use "parameterized SQL" calls so that the database doesn't get exposed to SQL injection attacks that can compromise the database information or structure [1]. It is safer to use "parameterized SQL" calls within MS SQL Server so that the inputs coming from the web application is treated as text and not special SQL code [1]. The last session to consider for database security is application security in the entire SQL Server security tool, utilities, views and

Table 1. Securable within the SQL Server: Source [4]

Server	Database	Schema
Endpoint	User	Type
Login	Role	XML Schema Collection
Database	Application Role	Aggregate
	Assembly	Constraint
	Message Type	Function
	Route	Procedure
	Service	Queue
	Remote Service Binding	Statistic
	Full text Catalog	Synonym
	Certificate	Table
	Asymmetric Key	View
	Symmetric Key	
	Contract	
	Schema	

**Fig. 3.** SQL Server 2005 Encryption Hierarchy: Source [4]

functions [4]. Database management fundamentals require tools and utilities which are used by various users for supporting, maintaining, and developing. Encrypting connections to SQL server ensures data stays confidential but at the cost or performance [4]. It is important to approve and make sure that all tools are configured correctly and consistently to reduce exposure of risk all developing application.

5 Application Security Considerations

Inherently, a lot of applications are designed to run as a non-privilege user with access rights designated to processes and services within the web server to run only what is needed. All servers within the project are going to go through Windows Operating System hardening to limit the attack surface to a minimum [6]. The goal to hardening the OS is to create security in depth that supplies confidentiality, integrity, and availability to the project. Encryption is designed into the project to provide confidentiality and integrity, which is used with the secure socket layer (SSL) for transmission between the clients and the Web server and Kerberos version 5 protocol with IPSec to communicate between all IPSec enabled computers on the network [5]. The project systems will all have the standard company's anti-virus and host firewall protection to protect the web server from malware that could compromise the projects availability and integrity. The company's network firewall will be configured to only allow port 80 and port 443 to communicate directly from the clients from outside the firewall. All patches to the server and client will be planned out and scheduled regularly to make sure all security patches within the intranet are updated to ensure phishing scams exploits in Internet Explorer (IE) on the client side is minimized [1]. Secure design considerations that are directly within the application that creates confidentiality, integrity and availability is provided by making sure input validation, cross site scripting, buffer overflows, SQL injection, and error handling is handled within the application. Input validation is going to process on the client side for simple communication server and server side validation is going to be included so control within the application can canonicalize inputs before processing. To help minimize cross site scripting (XSS) attacks third party software such as Acunetix [1] and HP's WebInspector [8] will be used to scan source code during development and all user input is to be encoded to HTML entity encoding to protect against JavaScript-based XSS attacks [1]. To protect the application from SQL injection attacks the database users assigned to be use with SQL Server for the application will have limited access rights to only have rights for the specific database and all database queries are to be processed by SQL user-defined stored procedures [1]. By this method all encryption within the database is processed by the database user inside the application database to ensure confidentiality and integrity for the projects data sets. Finally error handling will include standards that are to specify how each type of error is to be handled, how each error is logged and what information is to be returned to the user with the exception that no ODBC errors or to return users globally within the whole project [1].

There are a few common insecure application design issues that could have adverse effects. First insecure application design is on the notion that the application is one big module. To break up the application into several smaller modules limits the attack surface area as a section of the application is compromised. The second insecure application design is surrounded on the concept of logging. Due to limited resources logging is

only enabled in certain sections of the application and other sections is dependent on Windows Server Event Logging, which by company standards gets overwritten every 30 days. And lastly if the company's Active Directory (AD) server is taken down then this can prohibit the certificates to be available for encrypting, decrypting and authentication between the web server, database server and various clients within the firewall. This has a direct effect on the applications availability and integrity but does not affect the confidentiality considering all encrypt data stays encrypted.

6 Types of Cipher Attacks

When determining application security and secure designs it is important to recognize the main types of cipher attacks that can be a threat to the application. When using symmetric encryption schemes there are basically two types of fronts to protect from, the first is cryptanalysis and the second is brute-force attacks [3]. Cryptanalysis attacks exploit that characteristics of algorithm to use the key which could be catastrophic considering all future encryption cannot be protected [3]. According to MoRUN.net resources there are various cipher attacks that can be broken down by the following list [9] such as Algebraic attack, Algorithmic attack, Birthday attack, Brute Force Attack, Chosen ciphertext attack, Chosen plaintext attack, Ciphertext-only attack, Dictionary attack, Differential cryptanalysis, Known plaintext attack, Meet-in-the-middle attack, Middleperson attack, Precomputation attack.

7 Cryptography Types of Cipher Attacks

To help defend the organization's application against cipher attacks it is going to use industry standard Advanced Encryption Standard 256 (AES 256) when possible for symmetric encryption and to use 256-Bit SSL encryption for asymmetric encryption. It will also require that all generated keys be stored on a physically separated network from the production network which will be using the encryption keys. The development servers and production servers will be harden Windows servers which will only be used as a web server and SQL server. Microsoft Baseline Security Application will be used to harden the web server and SQL Server for best practices to keep only essential services running for best protection. Communication between the web server and database server will be setup to use IPSec on a different network segment. Availability to the server will only be provided using IPSec with all of the internal traffic behind the firewall. External availability will only be permitted to the public via the web server that is to be used to host the application. By providing IPSec services between servers, this prevents eavesdropping and replay cipher attacks on the local network [5]. To thwart against client cipher attacks, users are required to meet the organizations password policy when setting up their own password [3].

8 Conclusion

In conclusion, securing an application system after deployment can be accomplished by identifying the security architecture needs and evaluating the role encryption plays

in protecting the application's end points to defend against cipher attacks by using various cryptography technologies. By analyzing the architecture and design issues within the database management system, the application and the operating system cryptography is used to address the need to disguise critical information to secure an insecure application.

References

- [1] Raval, V., Fichadia, A.: Risks, Controls and Security Concepts and Applications. Wiley & Sons, Inc., Hoboken (2007)
- [2] Schumacher, R.: How to easily prevent SQL injection attacks [Web log message] (January 25, 2011), Retrieved from
<http://blogs.enterprisedb.com/2011/01/25/how-to-easily-prevent-sql-injection-attacks/>
- [3] Stallings, W., Brown, L.: Computer Security Principles and Practice. Pearson Prentice Hall, Upper Saddle River (2008)
- [4] Microsoft, MSDN Library (2011a), Retrieved from
<http://msdn.microsoft.com/en-us/library/>
- [5] Microsoft, IPSec (2011b), Retrieved from
<http://technet.microsoft.com/en-us/network/bb531150>
- [6] The University of Texas at Austin, Windows 2003 Server Hardening Checklist (2009), Retrieved, from <http://security.utexas.edu/admin/win2003.html>
- [7] Acunetix, Audit your website security with Acunetix Web Vulnerability Scanner (2011), Retrieved from <http://www.acunetix.com/vulnerability-scanner/>
- [8] Hewlett-Packard Development Company, L.P., HP WebInspect software (2011), http://www.hp.com/cda/hpms/display/main/hpms_content.jsp?zn=bto&cp=1-11-201-200^9570_4000_100__ (retrieved February 19, 2011)
- [9] MoRUN.net, Some Types of Attacks on Cryptosystems (2011), Retrieved from http://www.encryptionanddecryption.com/encryption/types_of_attacks.html
- [10] National Institute of Standards and Technology, Security Considerations in the System Development Life Cycle (2008), Retrieved from
<http://csrc.nist.gov/publications/nistpubs/800-64.../SP800-64-Revision2.pdf>
- [11] Halton, M., Rahman, S.: The Top 10 Best Cloud-Security Practices in Next-Generation Networking. International Journal of Communication Networks and Distributed Systems (IJCNDs); Special Issue on: Recent Advances in Next-Generation and Resource-Constrained Converged Networks 8(½), 70–84 (2012)
- [12] Mohr, S., Rahman, S.: IT Security Issues within the Video Game Industry. International Journal of Computer Science & Information Technology (IJCSIT) 3(5) (October 2011)
- [13] Dees, K., Rahman, S.: Enhancing Infrastructure Security in Real Estate. International Journal of Computer Networks & Communications (IJCNC) ISSN: 0974-9322 [Online]; 0975- 2293 [Print]
- [14] Hood, D., Rahman, S.: IT Security Plan for Flight Simulation Program. International Journal of Computer Science, Engineering and Applications (IJCSEA) 1(5) (October 2011) ISSN: 2230-9616 [Online]; 2231-0088 [Print]

- [15] Schuett, M., Rahman, S.: Information Security Synthesis in Online Universities. International Journal of Network Security & Its Applications (IJNSA) 3(5) (September 2011) ISSN: 0974-9330(online); 0975-2307 (Print)
- [16] Slaughter, J., Rahman, S.: Information Security Plan for Flight Simulator Applications. International Journal of Computer Science & Information Technology (IJCSIT) 3(3) (June 2011) ISSN: 0975-3826(online); 0975-4660 (Print)
- [17] Benson, K., Rahman, S.: Security Risks in Mechanical Engineering Industries. International Journal of Computer Science and Engineering Survey (IJCSES) ISSN: 0976-2760 (Online); 0976-3252 (Print)

On the Security of Two Certificateless Signature Schemes

Young-Ran Lee

Division of Fusion and Convergence of Mathematical Sciences, National Institute
for Mathematical Sciences, Daejeon, 305-811, Korea
yrlee@nims.re.kr

Abstract. Recently, Xiao *et al.* proposed a strong designated verifier certificateless signature scheme. Zhang *et al.* claimed that Xiao *et al.*'s scheme is vulnerable to key replacement attacks. In this paper, we show that Zhang *et al.*'s cryptanalysis on Xiao *et al.*'s scheme is incorrect and Xiao *et al.*'s scheme is insecure against key replacement attacks. On the other hand, Li *et al.* proposed a certificateless signature scheme without MapToPoint. It is shown that an adversary who replaces the public key of a signer can forge valid signatures for that signer without knowledge of the signer's private key.

1 Introduction

In traditional public key cryptography (PKC), a user's public key is just a random string that is unrelated to his identity, and hence needs to be authenticated. The authenticity of public key is provided in the form of a certificate, which is a digital signature generated by the Certificate Authority (CA) and provides a trusted link between the public key and the identity of the user. However, the Public Key Infrastructure (PKI) required to support such certificated-based system has many problems including storage, revocation and distribution. To reduce the certificate management overhead in traditional PKC, Shamir [13] proposed the identity based PKC (ID-PKC), in which each user's public key is predetermined by the information that uniquely identifies him/her (such as his/her email address). The corresponding private key is generated by a trusted third party called Private Key Generator (PKG). Since the user's public information can be used their public keys, there is no need to manage and control certificates. Nevertheless, the ID-PKC has the inherent key escrow problem : the PKC knows all users' private keys and all users must unconditionally trust the PKG.

Certificateless PKC (CL-PKC) was introduced by Al-Riyami and Paterson [1] to overcome the drawbacks of both traditional PKC and ID-PKC. In the system, the private key of each user is a combination of a user-chosen secret value and a partial private key obtained from a Key Generation Center (KGC), in such a way that the key escrow problem can be solved. Due to the nature of CL-PKC, there are two types of adversaries. A type I adversary captures attacks launched by outsiders. Because of the lack of authentication information (such as a X. 509 certificate in PKI) for public keys, so an adversary can replace them with public keys of its choice. In contrast, a type II adversary is stated with respect to malicious-but-passive KGC who controls the generation of a master

public/secret key pair. Ever since its introduction in [1], there have been quite a number of certificateless signature (CLS) schemes proposed [19,7,20,3,9,6,5,11,2,16,15]. However, only few schemes [20,3,9] are known to be secure against these adversaries.

In 1996, Jakobsson *et al.* [10] proposed a designated verifier signature (DVS) scheme in which the signature produced can only be verified by a specific user, but no one else. However, Saeednia *et al.* [14] pointed out that universal verifiability of DVS could reveal the identity of the signer, and further proposed a strong designated verifier signature (SDVS) scheme which enables protection of the signer's anonymity from a third party. Recently, several certificateless SDVS (CL-SDVS) schemes are proposed [8,4,18,17,21]. Very recently, Xiao *et al.* [17] presented a CL-SDVS scheme and showed that their scheme is secure against type I and type II attacks. Unfortunately, we show that their scheme is insecure against type I attacks in this paper. Although Zhang *et al.* [21] have already pointed out the weakness of the scheme, their cryptanalysis is incorrect. We also give some comments about that.

On the other hand, Li *et al.* proposed a CLS scheme without MapToPoint. The authors claimed that their scheme is relatively efficient, because it does not need the special MapToPoint hash function. Generally, it is known that the efficiency of the MapToPoint function is much lower than the general hash functions. Nevertheless, the proposed scheme is insecure against type I attacks.

The rest of this paper is organized as follows. In Section 2, we review Xiao *et al.*'s CL-SDVS scheme, and show that the scheme does not resist against public key replacement attacks. We also show that Zhang *et al.*'s cryptanalysis on Xiao *et al.*'s scheme is incorrect in Section 2. In Section 3, we recall Li *et al.*'s scheme and then show its insecurity against public key replacement attacks. The conclusion is given in Section 4.

2 Cryptanalysis of Xiao *et al.*'s CL-SDVS Scheme

Recently, Xiao *et al.* proposed a CL-SDVS scheme, and claimed that their scheme is provably secure in the random oracle model under the Computational Bilinear Diffie-Hellman(CBDH) assumption [17]. In this section, we show that their scheme is vulnerable to key replacement attacks. On the other hand, Zhang *et al.* pointed out the weakness of Xiao *et al.*'s scheme against public key replacement attacks, and also proposed a modification of the scheme [21]. We show that Zhang *et al.*'s cryptanalysis on Xiao *et al.*'s scheme is incorrect, and also give a comment on their improvement. To analyze the security of Xiao *et al.*'s CL-SDVS-scheme, we first review their scheme as follows.

2.1 Review of Xiao *et al.*'s CL-SDVS Scheme

Let \mathbb{G}_1 denote an additive group of prime order q , \mathbb{G}_2 be a multiplicative group of the same order, and $e : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$ be a bilinear pairing.

Setup. Given a security parameter k , the KGC chooses an arbitrary generator P and a random $s \in \mathbb{Z}_q^*$ as a master secret key, and computes a system public key $P_{pub} = sP$.

Let $H_1 : \{0, 1\}^* \rightarrow \mathbb{G}_1$ and $H_1 : \{0, 1\}^* \times \mathbb{G}_1^4 \rightarrow \mathbb{Z}_q^*$ be two cryptographic hash functions. Finally, the KGC publishes the system parameters $Params = \langle k, \mathbb{G}_1, \mathbb{G}_2, e, q, P, P_{pub}, H_1, H_2 \rangle$.

Partial-Private-Key-Extract. Given $Params$, a master secret key s , and a user identity $ID \in \{0, 1\}^*$, the KGC computes $Q_{ID} = H_1(ID) \in \mathbb{G}_1$ and outputs a partial private key $D_{ID} = sQ_{ID} \in \mathbb{G}_1$.

User-Key-Generation. Given $Params$, the user with identity ID chooses a random $x_{ID} \in \mathbb{Z}_q^*$ as his secret value and computes his public key $PK_{ID} = x_{ID}P$.

CL-SDVS-Sign. To generate a designated verifier signature on a message m under the signer A 's identity ID_A and public key PK_A , and the designated verifier B 's identity ID_B and public key PK_B , the signer A performs the following steps:

1. Chooses a random $r \in \mathbb{Z}_q^*$, and computes $U = rQ_B$.
2. Computes $h = H_2(m, U, Q_B, PK_A, x_A PK_B) \in \mathbb{Z}_q^*$.
3. Computes $V = (h/r)D_A$.
4. Outputs $\sigma = (U, V)$ as a signer A 's signature on m for the designated verifier B .

CL-SDVS-Verify. Given a signature $\sigma = (U, V)$, to verify the signature σ on m for an identity ID_A and public key PK_A , the designated verifier B with public key PK_B performs the following steps:

1. Computes $h = H_2(m, U, Q_B, PK_A, x_B PK_A) \in \mathbb{Z}_q^*$.
2. Accepts the signature if and only if the following equation holds:

$$e(V, U) \stackrel{?}{=} e(Q_A, D_B)^h.$$

CL-SDVS-Simulation. The designated verifier B can compute an indistinguishable signature $\hat{\sigma} = (\hat{U}, \hat{V})$ on the message m by executing the following steps:

1. Chooses a random $\hat{r} \in \mathbb{Z}_q^*$, and computes $\hat{U} = \hat{r}Q_A$.
2. Sets $\hat{h} = H_1(m, \hat{U}, Q_B, PK_Q, x_B PK_A)$.
3. Computes $\hat{V} = (\hat{h}/\hat{r})D_B$.

2.2 Weakness of Xiao et al.'s Scheme

Now, we will show that Xiao et al.'s CL-SDVS scheme is vulnerable to key replacement attacks. Specifically, their scheme is insecure against strong type I adversaries \mathcal{A}_I . In the first formal security model for CLS proposed by Huang et al. [7], the **Sign** oracle is defined such that it should provide a valid signature σ no matter whether the public key of the identity has been replaced or not. In particular, this definition requires **Sign** oracle to return a valid signature under the public key chosen by the adversary even if the queried public key is replaced by himself while the corresponding user secret key is not given. As first termed by Huang et al. [6], if the **Sign** oracle is defined in this way, the corresponding types of security are called **Super-Sign** type I and **Super-Sign** type II security. This is a very strong notion of security. On the other hand, in the **Strong-Sign** oracle, named by the same authors, the adversary is required to provide the user secret key sk_{ID} which is used to generate the replaced public key pk'_{ID} .

In [17], the authors claimed that their scheme is provably secure against the type I adversary in the random oracle model under the hardness of CBDHP (Computational Bilinear Diffie-Hellman Problem) assumption. Unfortunately, there exists a strong type I adversary who can always forge a valid CL-SDVS on any message as below:

Phase 1. In this phase, the adversary replaces a target user's public key PK_A with other public key of its own choices, and makes a sign query to the Strong-Sign oracle.

- \mathcal{A}_I chooses $x' \in_R \mathbb{Z}_p^*$ and replaces the target user's public key PK_A with $PK' = x'P$.
- In the phase of oracle queries, \mathcal{A}_I issues a sign query by submitting (m, ID_A, x') . Then a signature $\sigma = \text{CL-SDVS-Sign}(m, D_A, x', PK', ID_B, PK_B)$ is given to \mathcal{A}_I . Note that according to the algorithm CL-SDVS-Sign, the signature is of the form $\sigma = (rQ_B, (h/r)D_A)$, where $h = H_2(m, U, Q_B, x'P, x'PK_B)$.
- Given $\sigma = (U, V)$, the value of h can be easily derived as $h = H_2(m, U, Q_B, x'P, x'PK_B)$, where $(PK', SK') = (x'P, x')$ is the replaced key pair of the target user A. And then, \mathcal{A}_I can compute $V' = \frac{1}{h}V = \frac{1}{r}D_A$ by using h .

Phase 2. In this phase, the adversary can forge a signature on any message M . The following steps are performed by \mathcal{A}_I :

- Chooses a random $r' \in \mathbb{Z}_q^*$.
- Computes $U_f = r'U$, $h_f = H_2(M, U_f, Q_B, x'P, x'PK_B)$ and $V_f = (h_f/r')V'$, where $V' = \frac{1}{h}V = \frac{1}{r}D_A$.
- Outputs a forgery $\sigma_f = (U_f, V_f)$ on the message M under (ID_A, PK', ID_B, PK_B) .

The forged message-signature pair (M, σ_f) for $\{ID_A, PK', ID_B, PK_B\}$ will always be accepted since

$$\begin{aligned} e(V_f, U_f) &= e\left(\left(\frac{h_f}{r'}\right)V', r'U\right) \\ &= e\left(\frac{h_f}{r'} \cdot \frac{1}{r}D_A, r'rQ_B\right) \\ &= e(h_f D_A, Q_B) \\ &= e(Q_A, D_B)^{h_f}. \end{aligned}$$

Comments on Zhang et al.'s Improvement and Cryptanalysis. Zhang et al. showed that the Xiao et al.'s CL-SDVS scheme is insecure against public key replacement attacks and proposed its modification [21]. Generally speaking, an SDVS scheme should satisfy three security requirements: UNFORGEABILITY, NON-TRANSFERABILITY, and SIGNER'S AMBIGUITY. In particular, the non-transferability means that any designated verifier cannot transfer the conviction to any third party, i.e., the designated verifier cannot prove to a third party that the signature was produced by the signer or by himself. This property is accomplished by a TRANSCRIPT SIMULATION algorithm, which is run by the designated verifier to produce identically distributed transcripts which are indistinguishable from the signature produced by the signer. In Zhang et al.'s modification, the designated verifier B cannot simulate the signer A 's signature. This is due to the last term $S = r_2Q_B + (r_1 + H(U_2, V))D_A$ and verifying equation of their improved scheme, i.e., the verifier B cannot produce $\hat{\sigma} = (\hat{U}, \hat{V}, \hat{U}_1, \hat{U}_2, \hat{S})$ satisfying the CL-SDVS verification equation $e(\hat{S}, P) = e(Q_A, \hat{U}_1 + H(\hat{U}_2, \hat{V})P_{pub})e(Q_B, \hat{U}_2 + H(\hat{U}_1, \hat{V})P_{pub})$. Thus, their scheme does not satisfy all the necessary properties of SDVS schemes.

Moreover, their cryptanalysis is incorrect. According to [21], the adversary \mathcal{A}_I gets a signature $\sigma = (U, V) = (rQ_B, (h/r)D_A)$ on a message m from the target user, before it attempts to attack. And then, \mathcal{A}_I mounts key replacement attacks as follows:

- (1) Chooses $x' \in \mathbb{Z}_q^*$, and computes $PK' = x'P$ as the signer A 's public key.
- (2) Computes $h' = H_2(m', U, Q_B, PK', x'PK_B)$ and $h = H_2(m, U, Q_B, PK_A, x_BPK_A)$, where U comes from the given signature σ on m .
- (3) Computes $V' = \frac{h'}{h}V$ and sets $U' = U$, where V comes from the given signature σ on m .
- (4) Outputs its forgery on the message m' as $\sigma' = (U', V')$

Since the verification will pass, their cryptanalysis seems to be successful. But, the adversary knows neither of the users' secret values corresponding to the original public keys, i.e., x_A or x_B . Therefore, \mathcal{A}_I cannot compute the value of h in the step (2). Thus, the authors' attack on Xiao *et al.*'s scheme is incorrect.

3 Cryptanalysis of Li *et al.*'s CLS Scheme

Recently, Li *et al.* [12] has proposed a CLS scheme. In this paper, we show that the proposed scheme is insecure against public key replacement attacks, that is, an adversary can forge a valid signature for any message of any user under public key replacement attacks.

3.1 Review of Li *et al.*'s CLS Scheme

Let \mathbb{G}_1 denote an additive group of prime order q , \mathbb{G}_2 be a multiplicative group of the same order, and $e : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$ be a bilinear pairing.

Setup. Given a security parameter k , the KGC chooses an arbitrary generator $P \in \mathbb{G}_1$ and a random $s \in \mathbb{Z}_q^*$, and sets $P_{pub} = sP$ and a master key as $msk = s$. Let $H_1 : \{0, 1\}^* \rightarrow \mathbb{G}_1$ and $H_2 : \{0, 1\}^* \times \mathbb{G}_1 \rightarrow \mathbb{Z}_q^*$ be two hash functions. Then the KGC publishes the system parameters

$$\textit{Params} = (\mathbb{G}_1, \mathbb{G}_2, e, q, P_{pub}, H_1, H_2).$$

Partial-Private-Key-Extract. Given \textit{Params} , msk and a user identity ID_A , the KGC computes $Q_A = H_1(ID_A) \in \mathbb{G}_1$ and outputs a partial private key $D_A = sQ_A \in \mathbb{G}_1$.

Set-Secret-Value. Given \textit{Params} , the user with ID_A chooses $x_A \in \mathbb{Z}_q^*$ and sets x_A as his secret value.

Set-Private-Key. Given \textit{Params} and the partial private key D_A , the user with ID generates his private key $SK_A = x_A D_A \in \mathbb{G}_1$.

Set-Public-Key. Given \textit{Params} and the secret value x_A , the user with ID_A generates his public key $PK_A = x_A P_{pub}$.

Sign. To sign a message m , the signer with identity ID_A , private key SK_A and public key PK_A performs the following:

1. Chooses $r \in_R \mathbb{Z}_q^*$ and sets $U = rQ_A$, where $Q_A = H_1(ID_A)$,
2. Computes $h = H_2(m, U + PK_A)$ and $V = (r + h)SK_A$,
3. Returns $\sigma = (U, V)$ as the signature on the message m .

Verify. Given a signature $\sigma = (U, V)$ for the identity ID_A and public key PK_A on a message m , a verifier computes $h = H_2(m, U + PK_A)$, and then checks whether or not the equation $e(P, V) = e(PK_A, U + hQ_A)$ holds. If the equality holds, outputs Valid, otherwise Invalid.

3.2 Key Replacement Attack on Li *et al.*'s Scheme

Given a public key $PK_A = x_A P_{pub}$, an identity ID_A , and a message m to be signed, \mathcal{A}_I executes the following:

Phase 1. Chooses a random value $k \in \mathbb{Z}_q^*$ and replaces the signer's public key with $PK' = kP$.

Phase 2. In this stage, a signature is generated under the replaced public key.

- Chooses $r \in \mathbb{Z}_q^*$,
- Computes $U = rQ_A$,
- Sets $h = H_2(m, U + PK')$,
- Computes $V = k(U + hQ_A)$, and
- Returns the pair $\sigma = (U, V)$ as the forged signature.

The forged message-signature pair (m, σ) for $\{ID, PK'\}$ can always be accepted, since

$$\begin{aligned} e(P, V) &= e(P, k(U + hQ_A)) \\ &= e(P, U + hQ_A)^k \\ &= e(kP, U + hQ_A) \\ &= e(PK', U + hQ_A). \end{aligned}$$

The attack shows that Li *et al.*'s scheme, while relatively efficient, is not a secure CLS scheme.

4 Conclusion

Due to the lack of public key authentication, there exists an adversary who can replace public keys of arbitrary identities with other public keys of its own choices in a certificateless cryptosystem. However, most of CLS schemes are known to be insecure against these attacks. In this paper, we cryptanalyzed two CLS schemes: Xiao *et al.*'s CL-SDVS scheme and Li *et al.*'s CLS scheme. We exploited the **Strong-Sign** oracle to show that the scheme from [17] is insecure against key replacement attacks. In case of Li *et al.*'s scheme, we were able to show that the adversary who replaces a signer's public key can forge a valid signature for that signer without knowledge of the signer's private key.

Acknowledgement. This research was supported by the National Institute for Mathematical Sciences (NIMS) grant funded by the Korea government (No. A21103).

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. Castro, R., Dahab, R.: Two Notes on the Security of Certificateless Signatures. In: Susilo, W., Liu, J.K., Mu, Y. (eds.) ProvSec 2007. LNCS, vol. 4784, pp. 85–102. Springer, Heidelberg (2007)

3. Choi, K.Y., Park, J.H., Hwang, J.Y., Lee, D.H.: Efficient Certificateless Signature Schemes. In: Katz, J., Yung, M. (eds.) ACNS 2007. LNCS, vol. 4521, pp. 443–458. Springer, Heidelberg (2007)
4. Chen, H., Song, R., Zhang, F., Song, F.: An efficient certificateless short designated verifier signature scheme. In: 4th IEEE International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1–6. IEEE Press, New York (2008)
5. Du, H., Wen, Q.: Efficient and provably-secure certificateless short signature scheme from bilinear pairings. Computer Standards and Interfaces 31(2), 390–394 (2009)
6. Huang, X., Mu, Y., Susilo, W., Wong, D.S., Wu, W.: Certificateless Signature Revisited. In: Pieprzyk, J., Ghodosi, H., Dawson, E. (eds.) ACISP 2007. LNCS, vol. 4586, pp. 308–322. Springer, Heidelberg (2007)
7. Huang, X., Susilo, W., Mu, Y., Zhang, F.: On the Security of Certificateless Signature Schemes from Asiacrypt 2003. In: Desmedt, Y.G., Wang, H., Mu, Y., Li, Y. (eds.) CANS 2005. LNCS, vol. 3810, pp. 13–25. Springer, Heidelberg (2005)
8. Huang, X., Susilo, W., Mu, Y., Zhang, F.: Certificateless designated verifier signature schemes. In: Proceedings of 20th International Conference on Advanced Information Networking and Applications, pp. 15–19. IEEE Press, New York (2006)
9. Hu, B.C., Wong, D.S., Zhang, Z., Deng, X.: Certificateless signature: a new security model and an improved generic construction. Des. Codes Crypt. 42, 109–126 (2007)
10. Jakobsson, M., Sako, K., Impagliazzo, R.: Designated Verifier Proofs and Their Applications. In: Maurer, U.M. (ed.) EUROCRYPT 1996. LNCS, vol. 1070, pp. 143–154. Springer, Heidelberg (1996)
11. Liu, J.K., Au, M.H., Susilo, W.: Self-generated-certificate public key cryptography and certificateless signature/encryption scheme in the standard model. In: Proceedings of ACM Symposium on Information, Computer and Communications Security, pp. 273–283. ACM, New York (2007)
12. Li, F., Liu, P.: An efficient certificateless signature scheme from bilinear pairing. In: Proceedings of IEEE International Conference on Network Computing and Information Security, pp. 35–37. IEEE Press, New York (2011)
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. Saeednia, S., Kremer, S., Markowitch, O.: An Efficient Strong Designated Verifier Signature Scheme. In: Lim, J.-I., Lee, D.-H. (eds.) ICISC 2003. LNCS, vol. 2971, pp. 40–54. Springer, Heidelberg (2004)
15. Tso, R., Yi, X., Huang, X.: Efficient and Short Certificateless Signature. In: Franklin, M.K., Hui, L.C.K., Wong, D.S. (eds.) CANS 2008. LNCS, vol. 5339, pp. 64–79. Springer, Heidelberg (2008)
16. Xiong, H., Qin, Z., Li, F.: An improved certificateless signature scheme secure in the standard model. Fundamenta Informaticae 88, 193–206 (2008)
17. Xiao, Z., Yang, B., Li, S.: Certificateless strong designated verifier signature scheme. In: Proceedings of 2nd International Conference on e-Business and Information System Security, pp. 1–5. IEEE Press, New York (2010)
18. Yang, B., Hu, Z., Xiao, Z.: Efficient certificateless strong designated verifier signature scheme. In: International Conference on Computational Intelligence and Security, pp. 432–436. IEEE Press, New York (2009)
19. Yum, D.H., Lee, P.J.: Generic Construction of Certificateless Signature. In: Wang, H., Pieprzyk, J., Varadharajan, V. (eds.) ACISP 2004. LNCS, vol. 3108, pp. 200–211. Springer, Heidelberg (2004)

20. Zhang, Z., Wong, D.S., Xu, J., Feng, D.: Certificateless Public-Key Signature: Security Model and Efficient Construction. In: Zhou, J., Yung, M., Bao, F. (eds.) ACNS 2006. LNCS, vol. 3989, pp. 293–308. Springer, Heidelberg (2006)
21. Zhang, J., Xie, J.: Breaking a certificateless strong designated verifier signature scheme. In: Proceedings of International Conference on Consumer Electronics, Communications and Networks, pp. 130–133. IEEE Press, New York (2011)

Information Security Using Chains Matrix Multiplication

Ch. Rupa¹ and P.S. Avadhani²

¹ Associate Professor, Dept of CSE, VVIT, Nambur, Andhra Pradesh, India
rupamtch@gmail.com

² Professor, Dept of CS&SE, Andhra University, Visakhapatnam,
Andhra Pradesh, India
psavadhani@yahoo.com

Abstract. Information technology is facing lots of problems, while transmitting sensitive data and confidential data due to rapid growth of technology. The sensitive data is intended to share between only authorized persons, not for all. Information security concentrates on sensitive knowledge pattern that can be exposed when extracting the data. Therefore, researchers, for a long time period, have been investigating paths to improve privacy for sensitive data in information security analysis task process. So, many techniques have been introduced on privacy preserving issues in information security by using cryptography, steganography and their combination. In this paper we proposed an effective and efficient approach based on cryptography and Chains Matrix Multiplication (CMM) steganography uses optimal substructure property for generating random key to hide the data. It has to be improved privacy of sensitive data in information security analysis.

Keywords: Information security, CMM, optimal substructure property,random key.

1 Introduction

As the modern world is gradually becoming paperless with huge amount of information stored and exchanged over the internet, it is essential to have robust security measurements to safeguard the privacy and security of the underlying data. Cryptography techniques [1] have been widely used to encrypt the plain text data, transfer the cipher text over the internet. At present hackers or intruders are conducting cryptanalysis attacks on the ciphertext and retrieving the information due to lack of proper privacy preserving approaches [2].

Information hiding techniques are using to improve the privacy of the data i.e steganography. In this approach secure data is hiding in the cover media such as text files, image files, audio/video files by hiding techniques like LSB, Block Parity, etc [5, 15]. Steganalysis is the science of detecting the presence of hidden data in the cover media files and is emerging in parallel with steganography. Various steganalysis algorithms are using to retrieve the information which is stored at cover media. Hence, the combination of steganography and cryptography is using for improving the privacy and other security services to the information. In this approach, first secret information is encrypted by suitable cryptographic algorithm then the resultant cipher

text is hidden in the cover media. Even though, unauthorized parties are accessing the information from the cover media due to the limitations in the existing systems.

In this paper, proposed a method by Chains Matrix Multiplication (CMM) [9, 19]. It uses random approach instead of sequential for hiding the data in the cover media. Hence it will be reduced the limitations in the existing system. In this proposed method, cryptography and steganography both will be used to enhance privacy preservation in the information security. First, encrypt the secret message using suitable cryptographic mechanism, and hide the ciphertext into the cover media (Image) with the help of randomly generated key ' k' by Chains Matrix Multiplication. CMM consists of two modules for generating ' k' i.e Optimal substructure property and filling the table property [9, 19]. This proposed method improves the quality of stego image along with reduction of limitations in the existing system.

In this paper we proposed a new privacy preserving technique based on chains matrix multiplication to share sensitive information in secure manner. The paper is structure in the following way: Section 2 describes the problems in existing techniques. Section 3 consists of the proposed technique by using chains matrix multiplication. Solutions to the problems in existing technique are to be explained in Section 4. Different case studies related to CMM will be described in section 5. Section 6 describes the results and test analysis.

2 Privacy Preserving Techniques in Information Security

Privacy preserving is natural tradeoff between privacy quantification and information utility. Identification of problems related to all aspects of privacy of the information is significantly growing in real time environment applications. Two current main categories are Cryptography and Steganography to provide security to the information without compromising privacy. Yet both have some disadvantages, in the first one reduced privacy and Security and increased overhead for the second. Some methods published in privacy and security in the information security based on the orientation of the above two categories are as follows.

- Symmetric Key Cryptography [3, 4, 13]
- Asymmetric Key/ Public Cryptography [4, 13]
- Text Steganography [5, 15]
- Spatial Domain Based Steganography [5, 6, 12]
- Transform Domain Based Steganography [5, 6, 8]
- Masking and Filtering [7]

Now days, often companies are establishing and providing their services through out the globe. As a part of their services need to distribute data to their employees who are working at onsite, etc by secure. Again those companies may give data or distributing data to other companies for another purpose. It is possible, number of peoples to share their information through internet. Most traditional secure techniques [1] analyze the data statistically in terms of aggregation and secure mechanisms are also using in this distribution process. Even though, unauthorized parties are accessing the information due to the limitations which are existed in the existing systems in information security. The main limitations [11, 16] are as follows.

2.1 Limitations

- Lack of Privacy and Security in cryptographic techniques.
- In that perspective any one is possible to get view information. There may be a chance to misuse.
- Trusted third parties are not concentrating on both privacy and security of the information.
- Transmitting the secure data by using steganography mechanism. In this, secured key is fixed. More chances for vulnerability.
- Compressing an image by lossy compression technique will removes data that are hidden using Spatial Domain .
- Embed the information in significant area only. If, that area is identified data will be easily extracted.

3 Proposed Technique by Using CMM

In privacy preserving of data, proposing novel technique that allows security services while trying to protect the information. Privacy preserving in information security has become increasingly popular because it allows sharing of privacy sensitive data analysis purposes. The collection, analysis and sharing of specific data for publication raise serious concerns about the privacy of individuals. There are many challenges that require further investigation both from a theoretical and practical point of view. New privacy preserving techniques in information security are protected (individual privacy) and information is to be shared in secret (corporate privacy) manner. The proposed privacy preserving technique uses “Optimal Substructure Property” [8] to improve optimality and also using suitable Encryption technique and Chain Matrix Multiplication (CMM) [9, 19] for achieving these objectives. The operation of substructure property is show in Figure1.

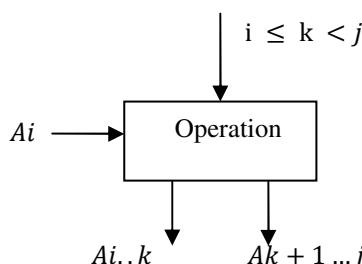


Fig. 1. Operation of substructure property

3.1 Objectives

The main goals of privacy preserving technique are as follows

1. Privacy preserving in information security can offer guarantees about the level of achieved Stego Image quality and quality of original Information for demanding applications.

2. It describes knowledge discovery of data with the integration of cryptographic techniques in terms of privacy preserving manner.
3. Sharing various forms of data without disclosing sensitive information while preserving.
4. Less chances to detect steganography due to less variation between random values and real image values with a statistical analysis.

3.2 Chains Matrix Multiplication (CMM)

A Chain Matrix Multiplication consists of two parts. Those are Optimal Sub Structure Property [8] and Filling in the table Property [9, 19]. Let, corresponding dimensions to the matrix sequence $A1..n$ are $P0, P1, \dots, Pn$. Determine optimal sequence of multiplications for $A1..n$ where $1 \leq i \leq j \leq n$. Search all possible values of 'k' to split the chain of multiplications $Ai..j$ by considering the constraint $i \leq k < j$ such as optimal substructure property. It gives optimal solution to the problem. Hence, needs to apply same procedure recursively. For $1 \leq i \leq j \leq n$, let $M[i, j]$ denotes the minimum number of multiplications needed to compute $Ai..j$. The optimum solution can be described by the following recursive definition.

$$M[i, j] = 0 \text{ where } i \geq j$$

$$\min_{i \leq k < j} (m[i, k] + m[k + 1, j] + P_{i-1} \cdot P_k \cdot P_j) \text{ where } i < j$$

Compute the value of an optimal solution in a bottom up fashion means $M[i, j]$ i.e $(M[1..n, 1..n])$ only defined for $i \leq j$. The important point is that when we use the recursive equation to calculate $M[i, j]$ must have already evaluated $M[i, k]$ and $M[k + 1, j]$. For both cases, the corresponding lengths of the matrix chain are both less than $j - k + 1$. Hence the algorithm should fill by increasing order of the length of the matrix chain. The total functioning of the CMM is shown in Table 1. Let us consider $P0 = 5, P1 = 4, P2 = 6, P3 = 2$ then the corresponding 'k' values are shown in the table 1.

Table 1. Filling the table by CMM

$j \rightarrow 1$	2	3	$\dots n$	
0	120 $k = 1$	88 $k = 1$	158 $k = 3$	$1 \leftarrow i$
	0 $k = 2$	48 $k = 3$	104 $k = 3$	2
		0 $k = 3$	84 $k = 3$	3
			0	:
				n

3.3 Privacy Preserving of Information Using CMM

The efficiency of privacy preserving information methods is based on the two factors i.e level of security and the amount of computation. The proposed method will improves the efficiency of the system by using Optimal Sub Structure and Filling the table properties. This Security system consists of two modules. Module 1 describes the suitable cryptographic system to encrypt the information. Module 2 describes the secure information transmission using proposed steganographic technique CMM. In module 1, encrypts the original data $Ei(I)$ symmetric or asymmetric algorithms. In Module 2, encrypted information $Ei(I)$ is hidden in an image using CMM. Hiding the information into an image and extraction of the information from an image are described in the following algorithms.

- Embedded Algorithm by Chain Matrix Multiplication (CMM)

Input: Original Image

Output: Stego Image

Step 1: Convert an Image into Matrix, ($M[1..n, 1..n]$)

Step 2: Apply Chain Matrix Multiplication to find hiding position (K)

. Step 2.1: Generate Random Numbers ‘ Pi ’ where $i = 0$ to n .

Step 2.2: If $(i == j) or (i > j)$ then

$M[i, j] = 0$ otherwise goto step 3

Step 3: If ($Sizeof(Ei) == No . of Pixels(M[i, j])$) then goto step 4

else goto step 11

Step 4 : Repeat

Step 5: Find ‘min’ value by diagonal wise i.e

$(M[1,2], M[2,3], M[3,4] ... M[1, n])$

Step 6: $min = min(m[i, k] + m[k + 1, j] + Pi - 1, Pk, Pj)$ where $i < k < j$

Step 7: Find ‘ k' .

Step 8: if $(k > 3)$ then $k = k \bmod 3$ and $kvalue[1..n] = k$

Step 9: Hide Encrypted data (Ei) at k th bit of the $M[i, j]$,

Step 10: $Ei++$

Step 11: Until $M[1, n]$ or $(Ei == Null)$

Step 12: Stop.

- Enhanced Original Data Algorithm

Input: Stego Image, $kvalue[]$

Output: Plain Text

Step 1: Convert an Stego Image into Matrix, ($M[1..n, 1..n]$)

Step 2: Apply Chain Matrix Multiplication to find hiding position (K).

Step 2.1: Generate Random Numbers ‘ Pi ’

where $i = 0$ to n

Step 2.2: If $(i == j) or (i > j)$ then

$M[i, j] = 0$ otherwise goto step 4

Step 3: Repeat

Step 4: Retrieve Ei by diagonal wise i.e

$(M[1,2], M[2,3], M[3,4] \dots M[1, n])$ using $kvalue$

Step 5: Decrypt the encrypted data (Ei).

Step 6: Stop.

4 Solutions to the Limitations of the Existing System

Chains matrix multiplication (CMM) based privacy preserving information uses cryptography and optimal sub structure and filling in the table authentication approaches provide all security services to share the sensitive information. Here there is no chance of masquerade the information of senders by intruders because it may changes information based on particular random numbers (Pi) and $kvalue$. Hence it helps to reduce the limitations of the existing privacy preserving approaches by the following way.

- To protect the privacy of individual sharing personal information in secret transmission using Chains matrix multiplication.
- No interference with the third party.
- Dynamically generating random number helps to improve the security to provide services to various end users.
- Reducing the chance of vulnerability through cryptography and steganography by CMM. Random generate number enhance the confusion complexity to attacker and provide more security to the end user.
- During extraction quality of data should be recovered.
- Information is not embedded in a significant area. Hence Information can't be identified easily by intruders.

5 Different Case Studies of CMM Based Approach

All the previous authentication methods are based on either cryptography or steganography by hiding the data in a significant area or both. In Chains Matrix Multiplication (CMM) based privacy preserving in Information Security, cryptography and Optimal Sub Structure and Filling in the table authentication approaches provide all security services to share the sensitive information. Here there is no chance to masquerade the information of end users by intruders why because it may hide the information based on particular instance of $kvalue$

A given problem has Optimal Substructure Property means if optimal solution of the given problem can be obtained efficiently by using optimal solutions of its sub problems. In the application of dynamic programming Richard Bellman's Principle [10] of Optimality is based on the idea that in order to solve a dynamic optimization problem from some starting period ' i ' to some ending period ' j ', one implicitly has to solve sub problems starting from later dates ' s ', where $i < s < j$. This is an example of optimal substructure. The Principle of Optimality is used to derive the Bellman equation [10], which shows how the value of the problem starting from ' i ' is related to the value of the problem starting from s .

For example the shortest path problem [10] has following optimal substructure property: If a node ‘x’ lies in the shortest path from a source node ‘u’ to destination node ‘v’ then the shortest path from ‘u’ to ‘v’ is combination of shortest path from ‘u’ to ‘x’ and shortest path from ‘x’ to ‘v’.

The principle of optimality is whatever the initial state and initial decision are the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

As with the optimal binary search tree, we can observe that if we divide a chain of matrices to be multiplied into two optimal sub-chains:

$(A_1 A_2 A_3 \dots A_k)(A_{k+1} \dots A_n)$ then the optimal parenthesisations of the sub-chains must be composed of optimal chains. If they were not, then we could replace them with cheaper parenthesisations. This property is called as optimal sub-structure is a hallmark of dynamic algorithms: it enables us to solve the small problems (the sub-structure) and use those solutions to generate solutions to larger problems. The cost of multiplying an ‘ $n \times m$ ’ by an ‘ $m \times p$ ’ one is $O(nmp)$ (or $O(n^3)$) for two $n \times n$ ones).

If any vulnerable person attack on the information and suppose he has already get the identity of encrypted method even though it is not sufficient to access the secret information at an image. He has to know about all the identities of hiding positions and these depends on dynamically changed values ‘*kvalue*’ and ‘*Pi*’ values. Hence Steganography using CMM is efficient method to improve privacy preserving in information security. The function is made to be so complex that reversing it is impossible, like trying to unmixed different colored paints in pot.

6 Results and Discussion

This method has been developing to overcome sequence mapping problems and other limitations what are existed in LSB [12], Bock Parity, etc. An original JPG image is shown in Figure 2 and image with a message of 1 KB as shown in Figure 3.



Fig. 2. Original Image



Fig. 3. Stego Image using CMM

Figure 4 shows the embedded message of sequentially embedded on top of the cover image since each bit from the message is sequentially ordered on the cover-image, then it will be easy for the third party to recover the message by retrieving the pixels sequentially starting from the pixel of the image [17]. The embedded message of CMM embedding method is shown in Figure 5. In this approach, each bit is hidden by random key ‘k’ instead of sequence order.

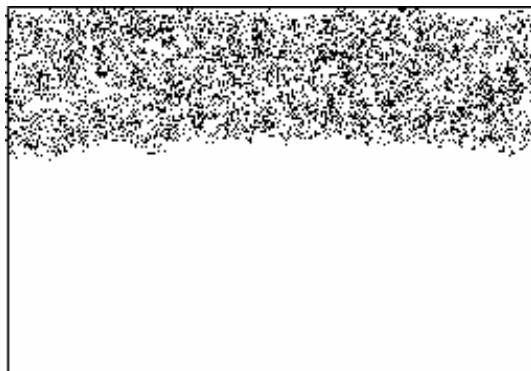


Fig. 4. Sequence Mapping of the Pixels using Existing Methods

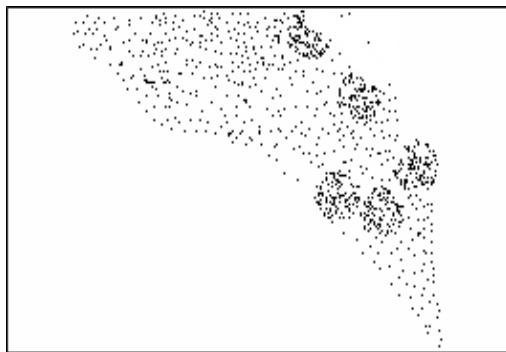
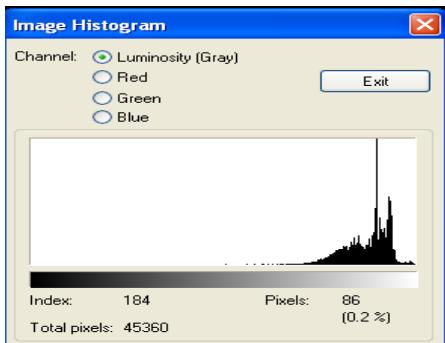
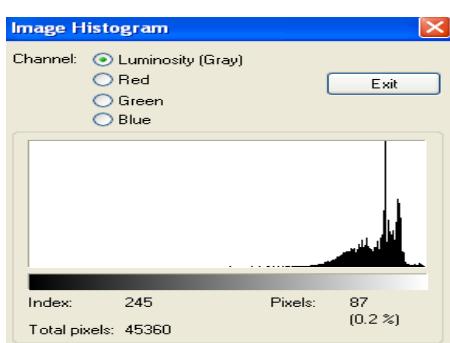


Fig. 5. Random mapping of the pixel by CMM

Figure 6 and Figure 7 are showing the testing results in the form of histograms of the original and stego Images. Very negligible variations are existed in the factor of luminosity [16]. Hence, unauthorized person unable to retrieve the information.

**Fig. 6.** Original Image Histogram**Fig. 7.** Histogram of Stego Image

7 Conclusion

Steganography that uses a key has a better security than non-key steganography. It is difficult to malicious people to recover the embedded message from the proposed approach. In the proposed method, the time taken to generate the random key is depends on the optimal solution by optimal substructure property. It improves privacy preservation of the information. Any Process that modifies the values of the pixels, either directly or indirectly may result in degrading of the quality of the original object.

References

1. Wang, Y., Desmedt, Y.: Perfectly Secure Message Transmission Revisited. *IEEE Transactions on Information Theory* 54(6), 2582–2596 (2008)
2. Bellovin, S.M., et al.: Risking Communications Security: Potential Hazards of the Protect America Act 6(1), 24–33 (2008)
3. O’Melia, S., Elbirt, A.J.: Enhancing the Performance of Symmetric-Key Cryptography via Instruction Set Extensions. *IEEE Trans. VLSI Syst.* 18(11), 1505–1518 (2010)
4. Cachin, C., Camenisch, J.: Encrypting Keys Securely. *IEEE Security & Privacy*, 66–69 (2010)
5. Lin, C.-C., Tsai, W.-H.: Secret image sharing with steganography and authentication. *The Journal of Systems and Software*, 405–414 (2004)
6. Westfeld, A., Pfitzmann, A.: Attacks on Steganographic Systems. In: Proceedings of the 3rd International Conference on Information Systems Security, pp. 16–20. ACM (2007)
7. Dumitrescu, S., Wu, X., Memon, N.: On steganalysis of random lsb embedding in continuous-tone images. In: *IEEE International Conference on Image Processing*, pp. 641–644 (2002)
8. Chikalov, I., Hussain, S., Moshkov, M.: Sequential Optimization of Matrix Chain Multiplication Relative to Different Cost Functions. In: Černá, I., Gyimóthy, T., Hromkovič, J., Jefferey, K., Králović, R., Vukolić, M., Wolf, S. (eds.) *SOFSEM 2011. LNCS*, vol. 6543, pp. 157–165. Springer, Heidelberg (2011)
9. Hu, T.C., Shing, M.T.: Computation of matrix chain products. Part II. *SIAM Journal on Computing* 13(2), 228–251 (1984)

10. Bellman, R.: Dynamic Programming Treatment of the Travelling Salesman Problem. *J. ACM* 9(1), 61–63 (1962)
11. Dinur, L., Nissm, K.: Revealing Information while Preserving Privacy. In: Proceedings of 22nd ACM symposium on Principles of Database Systems, pp. 202–210 (2003)
12. Fridrich, J.J., Goljan, M.: On estimation of secret message length in LSB steganography in spatial domain, pp. 23–34 (2004)
13. Sharma, A., Ojha, V.: Implementation of Cryptography for Privacy Preserving Data mining. *Proceedings of ITDMS* 2(3) (2010)
14. Lindell, Y., Pintas, B.: Preserving Data mining. *Jounal of Cryptology*, 177–206 (2002)
15. Rupa, C., Avadhani, P.S.: Message Encryption Scheme Using Cheating Text. In: *ITNG: Sixth International Conference on Information Technology: New Generations International Journal of Computer Science and Mathematical Applications* (indexed by IEEE, dblp), pp. 470–475 (2009) ISBN: 978-0-7695-3596-8/09
16. Du, R., Guthrie, L.E., Buchy, D.: Steganalysis with JPEG and GIF images, pp. 98–104 (2004)
17. Agaian, S.S., Rodriguez, B.M., Dietrich, G.B.: Steganalysis using modified pixel comparison and complexity measure, pp. 46–57 (2004)
18. Yang, B., Schmucker, M., Funk, W., Busch, C., Sun, S.-H.: Integer DCT-based reversible watermarking for images using companding technique, pp. 404–415 (2004)
19. Bhowmik, B.: Simplified Optimal Parenthesization for matrix chain multi-plication problem using Bottom up practice in 2 – tree structure. *J. of Applied Comp. Sc. & Math* 11, 9–14 (2011)

Formal Security Verification of Secured ECC Based Signcryption Scheme

Atanu Basu¹, Indranil Sengupta¹, and Jamuna Kanta Sing¹

¹ Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur 721302, India
{atanu,isg}@cse.iitkgp.ernet.in

² Department of Computer Science and Engineering
Jadavpur University, Kolkata 700032, India
jksing@ieee.org

Abstract. The signcryption scheme is a primitive in public key cryptography and it is useful where privacy and authenticity are required simultaneously. Our proposed Elliptic Curve Cryptography (ECC) based signcryption scheme preserves all the basic security features with lower overheads like authentication, confidentiality, non-repudiation, unforgeability, forward secrecy as well as public verifiability feature. We have showed through formal security analysis that our proposed scheme is secured against any adversary. The proposed scheme has been implemented in AVISPA, a well-known formal verification tool and the simulation results show that the proposed scheme is secured against any intruder attack.

Keywords: Signcryption, unsigncryption, elliptic curve cryptography, ECDLP, adversary, AVISPA.

1 Introduction

The signcryption scheme [1,2,3,4,5,6] which is a paradigm in public key cryptography implements the function of digital signature and encryption in a single logical step. Though the *signature-then-encryption* [1] scheme offers same functionality (privacy and authentication) as that of signcryption scheme, but in the *signature-then-encryption* scheme digital signature and encryption are implemented sequentially. The signcryption scheme incurs lower computational cost and communication overhead compared to *signature-then-encryption* scheme. A signcryption scheme consists of a pair of algorithm, signcryption algorithm (SA) and unsigncryption algorithm (UA). When SA is applied to a message m of arbitrary length it produces a signcrypted text CT and when UA is applied on CT, it recovers the original message m unambiguously.

The motivation for designing the proposed scheme is that it should incur low computational as well as communication overheads so that it may be applicable to resource constrained wireless mobile devices and the proposed scheme should also support the basic security features to prevent attacks from adversaries.

Our signcryption scheme has been proposed based on Elliptic Curve Cryptography (ECC) [9,10]. This scheme protects all the basic security features like authentication, confidentiality, unforgeability, non-repudiation and forward secrecy. It also supports public verifiability feature which helps to solve disputes. We have extended as well as revised our basic ECC based signcryption scheme in this paper which has been used in secured hierarchical secret sharing scheme [7,8]. In this paper, all the security features have been validated through the formal security proof. We have also used the AVISPA tool [15,16], a formal verification tool as well as model checker to show that our proposed signcryption scheme prevents attacks from active and passive adversary or intruder. Our scheme incurs less computational cost and communication overheads and it is suitable for resource constrained wireless mobile devices.

In this paper, we have presented the related or existing work in the Section 2. The proposed scheme has been discussed in the Section 3, the Section 4 discusses about the performance analysis and the Section 5 discusses about the comparison between our scheme with the other schemes. The Section 6 describes the security analysis of the proposed scheme. The Section 7 describes simulation through AVISPA tool and the Section 8 concludes the paper.

2 Related Work

In this section, we have mainly discussed the existing works of ECC based signcryption schemes. The signcryption scheme was proposed first by Y. Zheng [1]. It is based on ElGamal signature and encryption. After that, Y. Zheng et al. [2] proposed ECC based signcryption scheme where it has been shown that when it is compared with *signature-then-encryption* scheme on elliptic curves, the proposed signcryption scheme can save 58% in computational cost and a 40% in communication overhead. Though this scheme is computationally efficient, it does not support the features like forward secrecy and public verifiability. Peng et al. [3] proposed their ECC based signcryption scheme while proposing a threshold signcryption scheme. But, this scheme does not support features like forward secrecy and public verifiability. Hwang et al. [4] proposed their ECC based signcryption scheme efficiently and it supports all the relevant security features. This scheme has shown the robustness depending on ECDLP (Elliptic Curve Discrete Logarithm Problem) but this scheme does not show whether the scheme can survive from different types of attacks (impersonation attack). Baek et al. [5] showed the robustness of security of a signcryption scheme through formal proof efficiently. Zhou [6] proposed a ECC based signcryption scheme with public verifiability feature. But this scheme is computationally less efficient than Hwang et al. [4] scheme and this scheme does not support the security feature like forward secrecy.

3 Signcryption (SA) and Unsigncryption (UA) Algorithms

We consider the system, network and adversary models which have been considered in our signcryption scheme.

3.1 System, Network and Adversary Model

We consider that Alice sends a message m to Bob through the signcryption scheme. Alice and Bob may use resource constrained wireless mobile devices.

Alice may send message through the public channel to the Bob. The transmission medium may be wired or wireless.

In the adversary model there may exist passive or active adversary. Any passive adversary may eavesdrop any transmitted signcrypted message and tries to open the message. Any active adversary may capture any transmitted message and modifies as well as sends the modified message to Bob through the replay attack. Any intruder may mount impersonation attack in which the entity poses himself as Alice. There may be plaintext or ciphertext attack in which any adversary through analysis of the transmitted messages tries to retrieve the private key of the sender as well as the messages.

3.2 Generation and Distribution of Keys by the Trusted Dealer (TD)

The TD is responsible for the management of private and public keys of Alice and Bob.

A secure elliptic curve $E_p(a,b)$ over finite field $GF(p)$ [9] is chosen where p is a prime number and its base point is G of order q ($q \geq 160$ bits). Alice's private key d_A is chosen randomly from $[1, q - 1]$ and its corresponding public key is Q_A where $Q_A = d_A.G$ which is a point on $E_p(a,b)$. Similarly, Bob's private key d_B is chosen randomly from $[1, q - 1]$ and its corresponding public key is Q_B where $Q_B = d_B.G$ which is also point on $E_p(a,b)$. The private keys of Alice and Bob are sent through secure channel to them. It has been assumed that the private keys of Alice and Bob will remain secured through the operation of the scheme.

The signcryption and unsigncryption algorithms are described below.

3.3 Signcryption Algorithm (SA)

Step 1: Alice chooses a unique random integer, $k \in [1, q - 1]$ from $GF(p)$ for each signcryption session and computes ECC points $T_1^A = k.G = (x_1, y_1)$ and $T_2^A = k.Q_B = (x_2, y_2)$ where x_1 is the x -coordinate, y_1 is the y -coordinate of the point T_1^A and x_2 is the x -coordinate, y_2 is the y -coordinate of the point T_2^A on the elliptic curve $E_p(a,b)$.

Step 2: Alice computes the cyphertext message, $c = m.x_2 \bmod q$.

Step 3: Alice computes a one-way hash value h by $h = H(m)$ where H is a one-way collision resistant hash function [11].

Step 4: Alice computes the message s which includes the digital signature of the message m , $s = d_A - k.h \text{ mod } q$

Alice sends the signcrypted message (c, s, T_1^A) to Bob through public channel.

3.4 Unsignedryption Algorithm (UA)

After receiving the message (c, s, T_1^A) , Bob recovers the message m and verifies validity of the received message whether to accept it or not.

Step 1: Bob computes the ECC point T_1^B , $T_1^B = d_B \cdot T_1^A = (x_3, y_3)$ where x_3 is the x -coordinate and y_3 is the y -coordinate of the point T_1^B on the elliptic curve $E_p(a,b)$.

Step 2: Bob computes $m = c.(x_3)^{-1} \bmod q$. ($\because T_2^A = T_1^B$, shown below)

Step 3: Bob computes a one-way hash value h by $h = H(m)$ where H is a one-way collision resistant hash function [11].

Step 4: Bob computes the ECC point, $T_2^B = s.G + h.T_1^A$

A schematic diagram of the signcryption scheme has been shown in Figure 1 below.

If $T_2^B = Q_A$ where Q_A is the public key of the sender Alice, then Bob accepts the message m otherwise rejects the message. This is termed as **Validity test**.

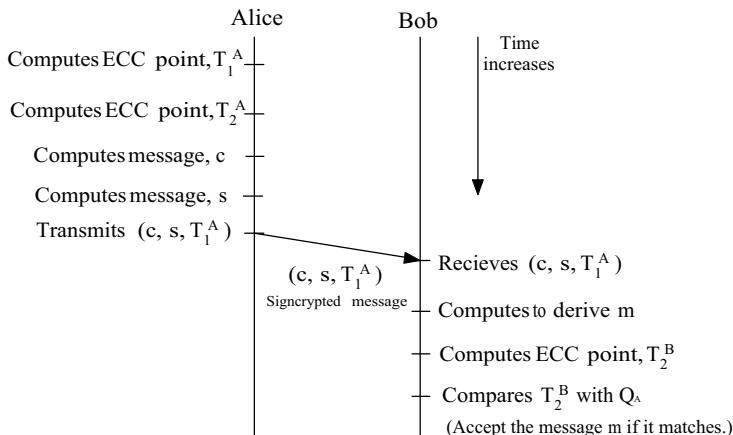


Fig. 1. Pictorial diagram of the signcryption scheme

Theorem 1: If unsigncryption of any signcrypted message passes the validity test, that message is accepted by the receiver.

Proof:

It is proved that

$$T_1^B = d_B \cdot T_1^A = (x_3, y_3) = d_B \cdot k \cdot G = k \cdot (d_B \cdot G) = k \cdot Q_B = T_2^A.$$

It is also proved that

$$\begin{aligned} T_2^B &= s \cdot G + h \cdot T_1^A = (d_A - k \cdot h) \cdot G + h \cdot T_1^A = d_A \cdot G - k \cdot h \cdot G + h \cdot T_1^A \\ &= Q_A - h \cdot (k \cdot G) + h \cdot T_1^A = Q_A - h \cdot T_1^A + h \cdot T_1^A = Q_A. \end{aligned}$$

4 Performance Analysis

The computational overhead and the communication overhead of the scheme have been shown below.

4.1 Computational Overhead

The ECPM (Elliptic Curve Point Multiplication) operation is the most computational intensive operation among other mathematical operations that have been used in this signcryption scheme. The signcryption algorithm (SA) computes two ECPM, two MUL (multiplication), one ADD (addition), one HASH (one-way hash function) operations while the unsigncryption algorithm (UA) computes three ECPM, one ECCADD (ECC addition), one DIV (division), one MUL, one ADD, one HASH operations.

Hasegawa et al. [12] implemented an Elliptic Curve Cryptosystem on a 16-bit microcomputer M16C (10MHz) designed by Mitsubishi Electric Corporation suitable for using in embedded and telecommunication systems. The processing times for scalar multiplication of a randomly given point, a modulo inversion of a given 160-bit data and SHA-1 had been evaluated as 480 msec, 130 msec and 2 msec respectively. Then for signcryption algorithm considering only the ECPM operation, our proposed scheme will take around 2×480 msec = 960 msec while the unsigncryption algorithm will take around 3×480 msec = 1440 msec.

4.2 Communication Overhead

The communication overhead of this signcryption scheme is due to transfer of the message (c, s, T_1^A) to the sender. Then the communication overhead of the scheme is $(|c| + |s| + |T_1^A|)$ bits.

Then,

$$\begin{aligned} &|c| + |s| + |T_1^A| \\ &= |q| + |q| + 2 \cdot |q| \\ &= 4 \cdot |q| \text{ bits.} \end{aligned}$$

For this scheme, if the value of q is 160 bit, then the communication overhead of the transmitted signcrypted message will be 4×160 bits = 640 bits.

Then, it is concluded that the proposed signcryption scheme may be used in resource constrained mobile devices for its lower computational and communication overheads.

5 Comparison with Other Schemes

A comparative study of our scheme with other ECC based signcryption schemes is shown in Table 1 and Table 2 below. The Table 1 shows the comparison based on the properties of confidentiality, integrity, unforgeability, non-repudiation, forward secrecy and public verifiability. The Table 2 shows the comparison based on the computational operations like ECPM, ECPA, DIV, MUL, ADD and HASH.

Table 1. Comparison of different signcryption schemes based on attributes

Schemes	CON	INT	UNF	NON	FOR	VER
Zheng [2]	Yes	Yes	Yes	another scheme	No	No
Hwang [4]	Yes	Yes	Yes	Directly	Yes	Yes
Zhou[6]	Yes	Yes	Yes	Directly	No	Yes
Our scheme	Yes	Yes	Yes	Directly	Yes	Yes

The abbreviated form of the parameters used in the above Table 1 are CON : Confidentiality, INT : Integrity, UNF : Unforgeability, NON : Non-repudiation, FOR : Forward secrecy, VER : Public verifiability.

Table 2. Comparison of different signcryption schemes based on operations

Schemes	Participant	ECPM	ECPA	DIV	MUL	ADD	HASH
Zheng [2]	Sender	1	-	1	1	1	2
	Receiver	2	1	-	2	-	2
Hwang [4]	Sender	2	-	-	1	1	1
	Receiver	3	1	-	-	-	1
Zhou [6]	Sender	2	2	1	2	1	3
	Receiver	4	4	-	1	1	3
Our scheme	Sender	2	-	-	2	1	1
	Receiver	3	1	1	1	1	1

The Table 2 has not included computational overhead for encryption and decryption of the message. Our scheme is an improvement of Hwang et al. [4] scheme and that scheme uses standard symmetric encryption algorithm (DES or AES). But, our scheme does not use any standard symmetric encryption algorithm (DES or AES) and it uses the multiplication operation between randomly chosen key and the message in every signcryption session. This further reduces the computational overhead of our scheme.

6 Security Analysis

We have used the method of proof by contradiction as proposed by Chuang et al. [13]. The Elliptic Curve Discrete Logarithm Problem (ECDLP) has been defined formally similar to the Discrete Logarithm Problem (DLP) [14].

Definition 1: A secure elliptic curve $E_p(a, b)$ where $E_p(a, b) \equiv y^2 = x^3 + a.x + b$, where a & b are two constants of the curve defined over the finite field $GF(p)$ where p is a prime number and q (≥ 160 bits) is the order of the base point G with a point \mathcal{O} (point at infinity or zero point). Let the two points P and Q ($Q = k.P$) $\in E_p(a, b)$ where $k \leftarrow {}_R GF(p)$ which signifies k is chosen randomly from $GF(p)$.

Instance : (P, Q, v) for some $k, v \leftarrow {}_R GF(p)$.

Output : **Yes**, if $Q = v.P$, i.e. $k = v$ and output **No**, otherwise.

Two distributions have been defined below:

$$D_{real} = \{k \leftarrow {}_R GF(p), X = P, Y = Q (= k.P), W = k : (X, Y, W)\},$$

$$D_{rand} = \{v, k \leftarrow {}_R GF(p), X = P, Y = Q (= k.P), W = v : (X, Y, W)\}.$$

The advantage of any probabilistic, polynomial-time, 0/1-valued distinguisher D in solving ECDLP on $E_p(a, b)$ is defined as

$$\begin{aligned} Adv_{D,E_p(a,b)}^{ECDLP} &= |Pr[(X, Y, W) \leftarrow D_{real} : D(X, Y, W) = 1] - Pr[(X, Y, W) \\ &\quad \leftarrow D_{rand} : D(X, Y, W) = 1]| \end{aligned}$$

where the probability $Pr(.)$ is taken over the random choices of k and v . The parameter D is termed to be a (t, ϵ) -ECDLP distinguisher for the $E_p(a, b)$ if D runs at most in time t such that $Adv_{D,E_p(a,b)}^{ECDLP}(t) \geq \epsilon$.

ECDLP assumption: For every probabilistic, polynomial-time 0/1-valued distinguisher D , we must have $Adv_{D,E_p(a,b)}^{ECDLP}(t) \leq \epsilon$, for any sufficiently small $\epsilon > 0$. Therefore, there exists no (t, ϵ) -ECDLP distinguisher for the $E_p(a, b)$.

Now, the Theorem 2 is defined below:

Theorem 2: Under the ECDLP assumption, the proposed signcryption scheme is provably secure against an adversary.

Proof:

It has been assumed that an adversary can solve the ECDLP to find the value k from the points P and Q ($Q = k.P$) $\in E_p(a, b)$. Now, the following oracle has been defined -

Reveal: This outputs the value k through the solution of ECDLP by using the points P, Q ($Q = k.P$) and other elliptic curve public parameters.

The adversary A executes two algorithms, say $Trial1_{SC,A}^{ECDLP}$ (Algorithm 1) and $Trial2_{SC,A}^{ECDLP}$ (Algorithm 2) for the proposed signcryption scheme SC have been shown below. We define $Succ1_{SC,A}^{ECDLP} = Pr[Trial1_{SC,A}^{ECDLP} = 1] - 1$ as defined by Baek et al. [5]. Then the advantage function for $Trial1_{SC,A}^{ECDLP}$ is defined as

$$Adv1_{SC,A}^{ECDLP}(t, q_R) = max_A\{Succ1_{SC,A}^{ECDLP}\},$$

where the maximum is taken over all A with execution time t and q_R is the number of queries to the *Reveal* oracle. We say that the proposed signcryption scheme provides confidentiality, if $Adv1_{SC,A}^{ECDLP}(t, q_R) \leq \epsilon$, for any sufficiently small $\epsilon > 0$.

Now, we define $Succ2_{SC,A}^{ECDLP} = Pr[Trial2_{SC,A}^{ECDLP} = 1] - 1$ as defined by Baek et al. [5]. Then the advantage function for $Trial2_{SC,A}^{ECDLP}$ is defined as

$$Adv2_{SC,A}^{ECDLP}(t, q_R) = \max_A \{Succ2_{SC,A}^{ECDLP}\},$$

where the maximum is taken over all A with execution time t and q_R is the number of queries to the *Reveal* oracle. We say that the proposed signcryption scheme preserves security features like authentication, integrity (replay or man-in-the-middle attack), unforgeability, non-repudiation as well as forward secrecy, if $Adv2_{SC,A}^{ECDLP}(t, q_R) \leq \epsilon$, for any sufficiently small $\epsilon > 0$.

Algorithm 1. $Trial1_{SC,A}^{ECDLP}$

Capture the signcrypted message (c, s, T_1^A) .

Call *Reveal* oracle. Outputs $k \leftarrow \text{Reveal}(E_p(a,b), G, T_1^A)$.

Using the value k , compute $(x_2, y_2) = k.Q_B$.

Retrieve the original message, $m = c.x_2^{-1} \bmod q$.

Algorithm 2. $Trial2_{SC,A}^{ECDLP}$

Capture the signcrypted message (c, s, T_1^A) .

Call *Reveal* oracle. Outputs, $k \leftarrow \text{Reveal}(E_p(a,b), G, T_1^A)$.

Using the value k , compute $(x_2, y_2) = k.Q_B$ and retrieve the original message, $m = x_2^{-1} \bmod q$.

Change m to m' and compute $h' = H(m')$.

Compute $c' = m'.x_2 \bmod q$.

Call *Reveal* oracle. Outputs, $d_A \leftarrow \text{Reveal}(E_p(a, b), G, Q_A)$.

Choose a random integer $k' \in [1, q - 1]$.

Compute $s' = d_A - k'.h' \bmod q$ and $T_1^{A'} = k'.G$.

Send $(c', s', T_1^{A'})$ to the verifier.

Verifier checks if $T_2^B = s'.G + h'.T_1^{A'} = Q_A$, where $h' = H(m')$

If the verification satisfies **then**

return 1

else

return 0

end if

Now, we discuss the security features of the signcryption scheme and the public verifiability feature below based on above discussion.

6.1 Confidentiality

According to $Trial1_{SC,A}^{ECDLP}$, the adversary is able to compute k from T_1^A and thus computes the original message. However, it is a contradiction due to the computational difficulty of the ECDLP. Thus, $Adv1_{SC,A}^{ECDLP}(t, q_R) \leq \epsilon$, for any sufficiently small $\epsilon > 0$. Hence, if any attacker captures the signcrypted message (c, s, T_1^A) , he cannot compute the parameter k from $T_1^A = k.G$ (Step 1 of SA) due to the computational difficulty of the ECDLP. Therefore, the proposed signcryption scheme provides confidentiality feature.

6.2 Authentication

According to $Trial2_{SC,A}^{ECDLP}$, the adversary is able to compute k and d_A . So, the adversary may change the original message m as well as the value s and the verification method ($T_2^B = Q_A$) of the verifier is performed. However, it is again a contradiction due to the computational difficulty of the ECDLP. Thus, $Adv2_{SC,A}^{ECDLP}(t, q_R) \leq \epsilon$, for any sufficiently small $\epsilon > 0$. Since the attacker does not have any ability to change the original message m , the values s and d_A , the adversary is not able to perform replay or man-in-the-middle attack. So, the authentication feature is preserved in this scheme.

6.3 Unforgeability

After eavesdropping of the signcrypted message (c, s, T_1^A) , if any attacker wants to forge the message (c, s, T_1^A) to (c', s', T_1^A) , he needs to get the private key d_A of the sender Alice and the randomly chosen value k . Again, these are not possible due to difficulty of the ECDLP ($Adv2_{SC,A}^{ECDLP}(t, q_R) \leq \epsilon$, for any sufficiently small $\epsilon > 0$). As a result, the proposed scheme provides unforgeability feature.

6.4 Non-repudiation

If the sender Alice denies that she has not sent the signcrypted message (c, s, T_1^A) to receiver Bob, then any trusted third party (TTP) can compute the verification condition $T_2^B = Q_A$ using the public key Q_A of the sender Alice. However, if the condition $T_2^B = Q_A$ satisfies, that ensures that the message has indeed come from the sender Alice and later she cannot deny that she has not sent the message. Thus, the proposed signcryption scheme provides the non-repudiation feature.

6.5 Forward Secrecy

Even if the adversary possesses the private key d_A of Alice at later stage, he cannot recover the previously sent signcrypted messages because he has to get the value k and retrieving the value k is difficult due to the ECDLP. As a result, the adversary is not able to recover the previous original messages and the forward secrecy feature of the proposed scheme is preserved.

6.6 Public Verifiability

If any dispute arises between Alice and Bob, e.g. Alice denies that she has not sent the particular message to Bob, then the dispute can be solved through a TTP without disclosing the secret message and private key of any party. In this case, the Alice will have to send the TTP the value s , the point T_1^A from the received signcrypted message (c, s, T_1^A) and the computed hash value h . The TTP will calculate the value $[s.G + h.T_1^A]$ and if this equals to public key Q_A of the sender Alice, then the TTP declares that the message indeed has come from Alice. Thus, our proposed scheme supports public verifiability feature.

Therefore, it is proved that the proposed signcryption scheme preserves the basic security features as well as the public verifiability feature.

7 Formal Security Verification Using AVISPA Tool

A well-known automated formal verification tool AVISPA (Automated Validation of Internet Security Protocol and Applications) [15,16] is used for security verification of our proposed signcryption scheme. The main advantage of the AVISPA is the ability to use different verification techniques in different backends on the same protocol specification. The verification process helps to check whether the proposed signcryption scheme is protected from any intruder attack. The AVISPA uses Dolev-Yao intruder model [17] where any intruder (active or passive adversary) can eavesdrop any transmitted message, mount masquerading (impersonation attacks) and replay attacks (modify or inject any message) but follows perfect cryptography, i.e. the intruder cannot break the cryptography. The AVISPA framework has been shown in Figure 2 below.

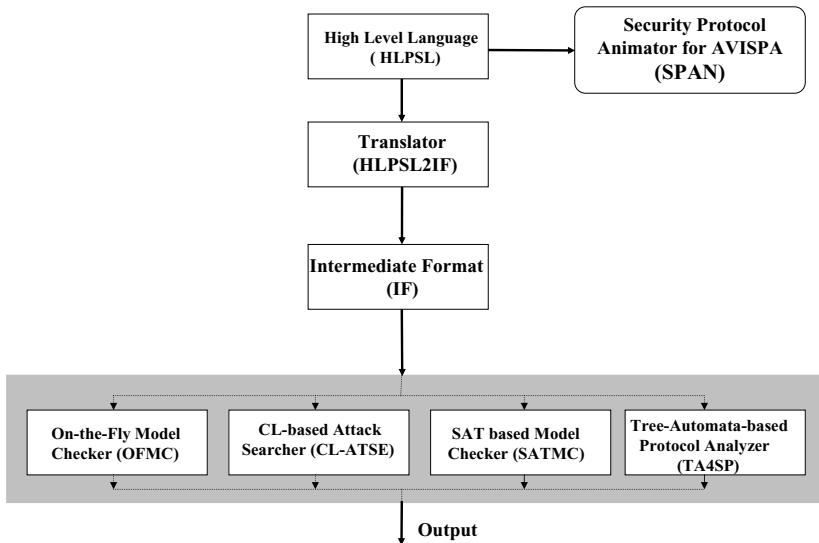


Fig. 2. Pictorial diagram of the AVISPA

A high-level language HLPSL (High Level Protocol Specification Language) is a flexible, expressive, modular and role based formal language and is used in AVISPA for specifying protocols as well as their security properties. The AVISPA translates any HLPSL specification code into IF (Intermediate Format) code. The IF code serves input to the four backends OFMC (On-the-Fly Model Checker), SATMC (SAT-based Model-Checker), CL-ATSE (Constraint-Logic based Attack Searcher) and TA4SP (Tree Automata based on Automatic Approximations for the Analysis of Security Protocols) are integrated into the AVISPA framework. The graphical interface SPAN (Security Protocol ANimator) where the AVISPA has been integrated is used for formal specification code verification and security analysis of the proposed scheme. Any HLPSL code in AVISPA generally consists of role, session, environment and goal sections. The role sections of Alice and Bob, goal section of the HLPSL code have been shown below:

Alice's role-

```
% HLPSL code (ECC-Signcryption.hlpsl)
role alice(A, B : agent,
Qa, Qb : public_key, % Qa - Public key of Alice, Qb - public key of Bob
% dA = inv(Qa) % da - Private key of Alice
K, Q, G : text,
M : message,
H, F1, F2, F3, F4 : function,
SND, RCV : channel(dy))
played_by A def =
local State : nat, IDa, IDb,
N : text
const alice_bob, bob_alice, sk : protocol_id
init State := 0
transition
1. State = 0 /\ RCV(start) =|>
State' := 1 /\ N' := new()
  /\ SND(IDa.IDb.{IDa.N'}_inv(Qa))
  /\ secret(K, sk, A) % K is the randomly selected element
  /\ SND(F1(M.F2(K.Qb)).F3(inv(Qa).K.H(M).Q).F4(K.G))
% H(M) represents hash value of the message
end role
```

Fig. 3. Alice's role

The HLPSL code runs properly in OFMC, CL-ATSE and SAT backend model checkers and the simulation results indicate that no intruder attack has been found for the proposed scheme. Though the proposed signcryption scheme does not use any standard encryption algorithm (DES or AES), but it prevents any type of chosen plaintext or ciphertext attack as in each signcrypted message transfer session new random value k is chosen (Section 3.3). So, it is concluded that even if the security of the scheme depends on ECDLP but it is also immune

from intruder attacks like man-in-the middle attack, replay attacks, chosen plain-text or ciphertext attack.

Bob's role:

```

role bob(A, B : agent,
Qa, Qb : public_key,
% dA = inv(Qa)
K, Q, G : text,
M : message,
H, F1, F2, F3, F4 : function,
SND, RCV : channel (dy)) played_by B def =
local State : nat, IDa, IDb,
N: text
const alice_bob, bob_alice, sk : protocol_id
init State := 1 transition
  1. State = 1 /\ RCV(IDa.IDb.{IDa.N'}_inv(Qa)) =>
    State' := 2 /\ secret (K, sk, A)
      /\ RCV(F1(M.F2(K.Qb)).F3(inv(Qa).K.H(M).Q).F4(K.G))
end role

```

Fig. 4. Bob's role

The goal section:

```

The goal section :
goal
secrecy_of sk
  authentication_on alice_bob
  authentication_on bob_alice
end goal

```

Fig. 5. Goal section of the proposed scheme

8 Conclusion

The signcryption scheme may be used in secured lightweight transaction protocol, e.g. electronic cash payment protocol. This signcryption scheme preserves all the basic security features like authentication, confidentiality, integrity, non-repudiation and forward secrecy efficiently. The scheme also supports public verifiability feature which is useful in case any dispute arises. The scheme is computationally efficient and can be used in resource constrained wireless mobile devices. The formal verification of the scheme as well as simulation results in the AVISPA verification tool shows that the scheme is secured against any adversary.

References

1. Zheng, Y.: Digital Signcryption or How to Achieve Cost (Signature & Encryption) << Cost(Signature) + Cost(Encryption). In: Kaliski Jr., B.S. (ed.) CRYPTO 1997. LNCS, vol. 1294, pp. 165–179. Springer, Heidelberg (1997)
2. Zheng, Y., Imai, H.: How to construct efficient signcryption schemes on elliptic curves. Information Processing Letters 68(5), 227–233 (1998)
3. Peng, C., Xiang, L.: Threshold Signcryption Scheme Based on Elliptic Curve Cryptosystem and Verifiable Secret Sharing. In: International Conference on Wireless Communications, Networking and Mobile Computing, vol. 2, pp. 1182–1185 (September 2005)
4. Hwang, R.-J., Lai, C.-H., Su, F.-F.: An Efficient Signcryption Scheme With Forward Secrecy Based on Elliptic Curve. Applied Mathematics and Computation 167(2), 870–881 (2005)
5. Baek, J., Steinfeld, R., Zheng, Y.: Formal Proofs for the Security of Signcryption. Journal of Cryptology 20(2), 203–235 (2007)
6. Zhou, X.: Improved Signcryption Scheme with Public Verifiability. In: Pacific-Asia Conference on Knowledge Engineering and Software Engineering, pp. 178–181. IEEE (2009)
7. Basu, A., Sengupta, I., Sing, J.K.: Secured Hierarchical secret Sharing using ECC based Signcryption. In: Security and Communication Networks. Wiley (to appear)
8. Basu, A., Sengupta, I., Sing, J.K.: Cryptosystem for secret sharing scheme with hierarchical groups. International Journal of Network Security (to appear)
9. Hankerson, D., Menezes, A., Vanstone, S.: Guide to Elliptic Curve Cryptography. Springer (2004)
10. Vanstone, S.A.: Elliptic curve cryptosystem - The Answer to Strong, Fast Publickey Cryptography for Securing Constrained Environments. Information Security Technical Report 12(2), 78–87 (1997)
11. Stallings, W.: Cryptography and Network Security: Principles and Practices, 4th edn. Pearson Prentice Hall (2006)
12. Hasegawa, T., Nakajima, J., Matsui, M.: A Practical Implementation of Elliptic Curve Cryptosystems over GF(p) on a 16-bit Microcomputer. In: Imai, H., Zheng, Y. (eds.) PKC 1998. LNCS, vol. 1431, pp. 182–194. Springer, Heidelberg (1998)
13. Chuang, Y.-H., Tseng, Y.-M.: An efficient dynamic group key agreement protocol for imbalanced wireless networks. International Journal of Network Management 20(4), 167–180 (2010)
14. Dutta, R., Barua, R.: Provably Secure Constant Round Contributory Group Key Agreement. IEEE Transactions on Information Theory 54(5), 2007–2025 (2008)
15. AVISPA Project. AVISPA protocol library, <http://www.avispaproject.org/>
16. Vigano, L.: Automated Security Protocol Analysis With the AVISPA Tool. Electronic Notes in Theoretical Computer Science, vol. 155, pp. 61–86. Elsevier (2006)
17. Dolev, D., Yao, A.C.-C.: On the security of public key protocols. In: FOCS, pp. 350–357. IEEE (1981)

Universal Steganalysis Using Contourlet Transform*

V. Natarajan and R. Anitha

Department of Mathematics and Computer Applications, PSG College of Technology,
Coimbatore, India
kvn.psg@gmail.com, anitha_nadarajan@mail.psgtech.ac.in

Abstract. This paper proposes a new universal steganalysis method based on contourlet transform with high detection rate. An important aspect of this paper is that it uses the minimum number of features in the transform domain and gives a better accuracy than many of the existing steganalysis methods. Only five features have been extracted from the contourlet transformed image and a back propagation neural network classifier has been used to classify whether the given image is stego image or cover. The efficiency of the proposed method is demonstrated through experimental results. Also its performance is compared with the state of the art wavelet based steganalyzer (WBS), Feature based steganalyzer (FBS) and Contourlet based steganalyzer (CBS). The results show significantly high performance of our method.

Keywords: Steganography, Steganalysis, Contourlet transform, Structural similarity measure.

1 Introduction

Steganography is the art of hiding secret messages by embedding them into digital media while steganalysis is the art of detecting the hidden messages. The purpose of steganalysis is to collect sufficient evidence about the presence of embedded message and to break the security of the carrier. Steganalysis is broadly classified into two categories. One is specific steganalysis, meant for breaking a specific steganography. The other one is universal steganalysis, which can detect the existence of hidden message without knowing the details of steganography algorithms used. Universal steganalysis is also known as blind steganalysis and it is more applicable and practicable [1,2] than the specific steganalysis. Based on the methods used, steganalysis techniques are classified into two classes; signature based steganalysis and statistical based steganalysis. During the initial stage of steganalysis, signature based steganalyzers were used to expose the possibility of hidden information. Specific signature based steganalysis is simple, gives promising results when message is embedded sequentially, but hard to automatize and its reliability is highly questionable[3,4]. In universal steganalysis, using statistical methods and identifying the difference of some statistical characteristics between the cover and stego image becomes a major issue. Due to the tremendous increase in steganography, there is a need for powerful universal steganalyzers which are capable of identifying stego images.

* This work is a part of the Collaborative Directed Basic Research on Smart and Secure Environment project funded by NTRO, Govt.of India.

This paper proposes a new approach to universal steganalysis that is quiet general and does not need any knowledge of the embedding mechanism. This approach utilizes contourlet transform to represent the images. A Gaussian distribution is used to model the contourlet subband coefficients and since skewness and kurtosis of a distribution could be analyzed using the first four moments, the first four normalized statistical moments are considered as the features along with the similarity measure among the frequency bands. The experimental results show the efficiency of our approach when analyzed with various steganography methods.

The rest of the paper is organized as below. Section 2 gives a brief description of related work and section 3 discusses the proposed method. Experimental evaluation of the proposed steganalyzer is given in section 4 and section 5 concludes this paper.

2 Related Work

With the inception of data hiding techniques, the research on steganalysis started in the late 90's. Arooj Nissar et al[7] have given a detailed survey of steganalysis. The first universal steganalyzer was proposed by Avcibas et al.[5] and the same authors improved their previous method in [6]. Jiang N et al. proposed a blind steganalyzer using support vector machine to classify the stego image and cover image. For the first time, Farid et al [8,9] modeled an universal steganalyzer using supervised learning and indicated that the supervised learning is effective for detecting stego images without knowing the statistical property of images and teganoigraphy methods. Xuan et al. presented a universal steganalysis method, which was based on statistical moments of wavelet histogram characteristic functions[10]. This method utilized the Bayes classifier to distinguish stego images form original images. Experimental results indicated that this method work better for LSB, spread spectrum like steganography, F5 and Outguss steganography methods. Lie et al.[11] indicated that in general no single feature is capable of differentiating stego and plain images effectively and a combination of features extracted in different domain will be generally more promising. Based on the best wavelet packet decomposition of images, a universal steganalysis method with high correct detection ratio was proposed by Luo Xiang Yang et al[12]. However, the methods based on wavelet high order statistics cannot perform very well on spatial domain steganography such as LSB steganography[12].

Hedieh Sajedi et al. [12] presented an universal approach to steganalysis called Contourlet Based Steganalysis(CBS), which used statistical moments as well as the log errors between the actual coefficients and predicted coefficients of the contourlet transform as feature for analysis. After feature extraction, a non linear SVM classifier was applied to classify cover and stego images. This method converted the image into grayscale and then processed it. CBS detection rate is very low when message is embedded in medium frequency subbands and this idea is used in [13] to develop a new contourlet based steganography algorithm. So if the algorithm in [13] is used to embed the message, then CBS [12] cannot detect successfully. This fact motivated us to develop an efficient steganalyzer in Contourlet domain.

3 Proposed Scheme

The objective of the proposed scheme is to select the most relevant features using statistical characteristics of the subband coefficients, thus reduce the dimensionality of

feature set and increase the accuracy of detection. In this paper, the first four normalized moments of high frequency, low frequency subband coefficients and structural similarity measure of medium frequency subband coefficients are taken as the feature set. With these five features, a three layer back propagation neural network is trained for further classification. The block diagram of the proposed model is given in figure.1. The following sub sections briefly explain contourlet transformation , how the feature set is extracted from images and how the classification is done.

3.1 Contourlet Transform

The Contourlet transform is a two-dimensional extension of the wavelet transform proposed by Do and Vetterli[14,15] using multiscale and directional filter banks. Contourlet employs an iterated combination of Laplacian Pyramid (LP)[16], for capturing the point discontinuities, and the Directional Filter Bank(DFB)[17] , to gather nearby basis functions and link point discontinuities into linear structures. Contourlet transform is more powerful than the wavelet transform in characterizing images rich of directional details and smooth contours [14, 15].

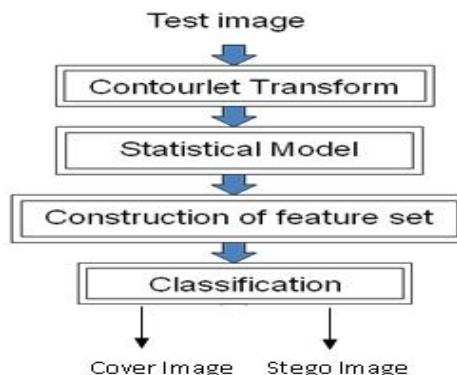


Fig. 1. Block diagram for proposed model

Let the image be a real-valued function $I(t)$ defined on the integer valued Cartesian grid $[2^l, 2^l]$. The Discrete Contourlet Transform with scale j , direction k and level n of $I(t)$ is defined as follows [15,18]:

$$\lambda_{j,k,n}(t) = \sum_{i=0}^3 \sum_{m \in \mathbb{Z}^2} d_k(m) \psi_{j,n}^{(i)}(t)$$

where $d_k(m)$ is the directional coefficient and

$$\psi_{j,n}^{(i)}(t) = \sum_{m \in \mathbb{Z}^2} f_i(m) \phi_{j,n+m}(t)$$

where $\phi(t)$ is the scaling function and $f(t)$ is the spatial domain function.

3.1.1 Subband Coefficient Modeling

The coefficients in the produced sub bands of contourlet transformed image are very appropriate to obtain the texture feature due to coarse to fine directional details of the image in these subbands. Besides, the distribution of the sub bands coefficients is symmetric and unimodal with mean skewness approximately near to zero, though they have not exactly Gaussian distribution [19]. These special characteristics of subband coefficients make them suitable for modeling by Gaussian distribution with density function.

$$f(x, \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2} \quad -\infty < x < \infty$$

where μ and σ are the mean and standard deviation of all the coefficients of sub bands.

3.2 Feature Extraction

There are various methods in the literature to extract the relevant features of digital images based on different transforms or filtering techniques. Even though the accuracy of classifiers is based on the number of suitable features, higher the number of features slower will be the classification. So identifying a minimum number of features which can produce efficient classification is a challenge. In this paper, only 5 features have been used which is very less compared to the number of features used in the existing steganalysis methods.

Contourlet transform is more sparser than wavelet as the majority of the coefficients have amplitudes close to zero. Also the moments of contourlet coefficients are more sensitive to the process of information hiding. The first four normalized moments of the high frequency and low frequency subband coefficients are more sensitive to the process of steganography. Since these moments could be a good measure for skewness and kurtosis due to information hiding, the first four normalized moments have been extracted as features. Moments are computed as below:

$$m_k = \frac{E(X - \mu)^k}{\sigma^{2k}} \quad k=1,2,3 \text{ and } 4.$$

where X represents the coefficients of contourlet subbands. These moments alone are not sufficient to detect the changes in the medium frequency subbands and hence another feature namely structural similarity measure (SSIM) is also included. For estimating SSIM, medium frequency band is split into two equal number of subband groups X and Y respectively. SSIM includes three parts: Luminance Comparison(LC), Contrast Comparison(CC) and Structural Comparison(SC) and they are defined as below[19,20,21]:

$$CC(x, y) = \frac{2\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}, \quad LC(x, y) = \frac{2\mu_x \mu_y}{\mu_x^2 + \mu_y^2}, \quad SC(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$SSIM(X, Y) = [LC(X, Y)][CC(X, Y)][SC(X, Y)]$$

The similarity of the whole image (I) is

$$SSIM(I) = \frac{\sum_{j=1}^n SSIM_j}{n}$$

where n is the number of middle frequency sub bands in the image. The first four normalized moments $m_k(k=1,2,3,4)$ and the similarity measure $SSIM(I)$ form the feature set.

3.3 Classification

A three layer Back Propagation (BP) neural network has been used as a classifier for identifying stego images as well as cover images. All the five extracted features are given as input and based on the samples, a training model is created. The power of back propagation is that it enables us to compute an effective error for each hidden unit, and thus derive a learning rule for the input to hidden weights. The input layer will have 5 neurons and the output layer will have 1 neuron. The number of nodes in the hidden layer is empirically selected such that the mean square error for feed forward networks is minimized. Since less number of features are used BP computations are fast.

4 Experimental Results

The proposed steganalysis have been implemented using MATLAB 7.6.0 with Matlab scripts. In our experiment, for training 12,200 images from Computer Visionimage dataset and INRIA image dataset are used. It contains 5,500 cover images and 6,700 stego images which are generated by different embedding algorithms like LSB, Jsteg, F5, Contsteg etc., Washington image dataset [22] is used for testing the proposed steganalysis method. 100 images are used to test the proposed scheme, with 60 cover images and 40 stego images.

In order to analyze the proposed method, ten typical steganography methods are used. Table 1 gives the detection results of our steganalyzer with respect to payload size and image size. From this table we can see that the average correct detection rate is about 92% which is really a high detection rate.

Table 1. Average correct detection rates for natural images and stego images

Steganography Methods	Average correct detection rates						
	Embedding rates				Different image size		
	100%	50%	25%	512X512	256X256	128X128	64X64
LSB	0.922	0.912	0.894	0.975	0.973	0.895	0.870
PMK	0.916	0.901	0.886	0.970	0.911	0.876	0.883
LTSB	0.939	0.901	0.874	0.893	0.913	0.878	0.846
Jsteg	0.906	0.970	0.854	0.902	0.899	0.873	0.870
F5	0.910	0.969	0.898	0.985	0.942	0.793	0.763
Jphide	0.971	0.899	0.828	0.889	0.884	0.801	0.744
Model based	0.926	0.875	0.781	0.911	0.905	0.813	0.771
Peturb	0.940	0.870	0.825	0.865	0.856	0.828	0.701
Quantization							
Congsteg	0.921	0.897	0.878	0.870	0.856	0.831	0.723
YASS	0.956	0.906	0.844	0.901	0.892	0.879	0.733

The relevancy of the extracted features used in this steganalysis is evaluated using error estimation. Table 2 and figure 2 display the average Median Absolute Error (MAE) of the features .The data shows a higher error than bias for all the embedding algorithms. So it is clear that, with this minimum feature set, the proposed method can able to detect the stego image.

Table 2. Median absolute error and bias for the proposed method

Algorithm	MAE	Bias
LSB	5.91×10^{-3}	-1.70×10^{-4}
PMK	8.38×10^{-3}	-5.29×10^{-4}
LTSB	9.07×10^{-3}	1.51×10^{-4}
Jsteg	3.23×10^{-3}	-2.89×10^{-4}
F5	6.63×10^{-3}	-3.78×10^{-4}
JPHide	7.19×10^{-3}	-1.31×10^{-4}
Model based	4.82×10^{-3}	-2.51×10^{-4}
Perturb	2.33×10^{-3}	1.28×10^{-4}
Quantization		
YASS	4.19×10^{-3}	1.87×10^{-4}
ContSteg	3.25×10^{-3}	0.58×10^{-4}

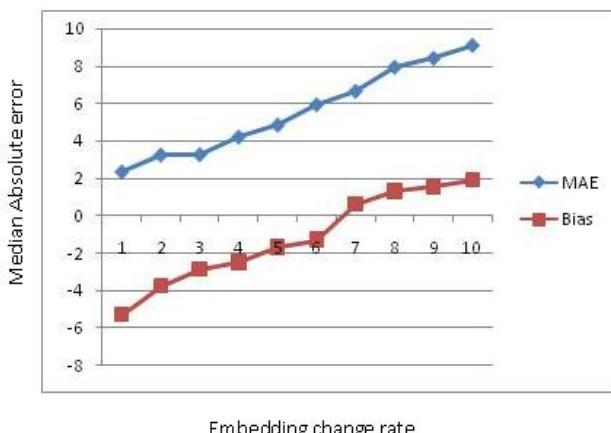


Fig. 2. Median Absolute Error(MAE) and Bias of proposed steganalyzer, with respect to embedding rates

4.1 Comparison with Prior Art

The proposed work is compared with the state of the art Wavelet-based Steganalysis(WBS)[9], Feature based Steganalysis(FBS)[22] and Contourlet-Based Steganalysis (CBS) [12] methods and the results show significant improvement and they are

tabulated in Table 3. Figure 3 shows the comparison of the performance of the proposed method with other steganalyzers against some steganography methods. The Data set used in the proposed scheme for comparison is the Washington dataset which is used in ContSteg [13] and CBS[12].

Table 3. Comparison of average detection accuracy of the proposed scheme with WBS, FBS, CBS on stego image produced by Contsteg

Secret Data Size(bits)	Steganalysis Method	Average detection Accuracy (%)
5000	WBS	51
	FBS	53
	CBS	59
	Proposed method	72
10,000	WBS	53
	FBS	54
	CBS	63
	Proposed method	87
15,000	WBS	58
	FBS	61
	CBS	68
	Proposed method	89

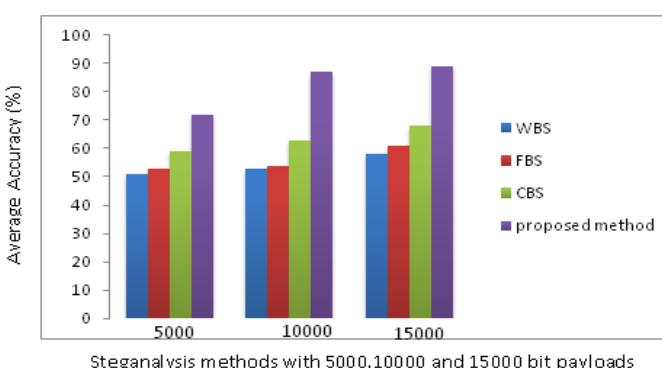


Fig. 3. Comparison of Proposed method with WBS, FBS and CBS

Table 4 shows the comparison of dimension of the feature set and time taken by proposed steganalyzer with that of the existing Fourier, Wavelet and Contourlet based steganalyzers when the algorithms are run on a system with 1GB RAM and P-IV processor. From the results one can infer that proposed scheme is fast with minimum dimensional feature set.

Table 4. The Time evaluation of LSB,FBS,CBS, and proposed method

Steganalysis Method	Dimension of feature set	Average Time (Seconds)
LSB (2006)	-	40.42
FBS(2008)	25	35.32
WBS(2009)	18	25.00
CBS(2010)	8	17.82
Proposed method	5	15.22

Especially proposed scheme gives better results for contourlet based steganography methods [12] and also proposed method is independent of file formats and different image types. The new method based on statistical steganalysis utilizes fewer features and hence the time and computational cost of the new method in extracting the features and detecting the stego image are much less than that of the methods based on feature extraction.

5 Conclusions

A new universal steganalysis method using Contourlet transform is proposed in this paper. Statistical moments and structural similarity of the contourlet coefficients form the feature set. A comparison of the performance of the proposed scheme with other universal steganalyzers like LSB, FBS, WBS and CBS is done using various testing metrics. The average correct detection rate is high, at the same time the dimension of the feature set and the average run time are reduced in this proposed scheme. Furthermore, the method proposed here is an universal blind scheme, which is independent of image type and file format.

References

1. Fridrich, J., Goljan, M.: Practical: Steganalysis of digital images-state of the art. In: Proceedings of SPIE, Security and Watermarking Multimedia Content IV, vol. 4675, pp. 1–13 (2002)
2. McBride, B.T., Peterson, G.L., Gustafson, S.C.: A new blind method for detecting novel steganography. Digit Invest 2, 50–70 (2005)
3. Johnson, N.F., Jajodia, S.: Steganalysis: the investigation of hidden information. In: Proc. IEEE Information Technology Conference, Syracuse, NY (1998)
4. Fridrich, J., Goljan, M.: Practical steganalysis of digital images state of the art. In: Proc. SPIE Photonics West, Electronic Imaging (2002), Security and Watermarking of Multimedia Contents, San Jose, CA, vol. 4675, pp. 1–13 (January 2002)
5. Avcibas, I., Memon, N.D., Sankur, B.: Steganalysis of watermarking techniques using image quality metrics. In: Proceedings of SPIE, Security and Watermarking of Multimedia Content III, vol. 4314, pp. 523–531. SPIE, New York (2001)
6. Avcibas, I., Memon, N., Sankur, B.: Steganalysis using image quality metrics. IEEE Trans. Image Process. 12, 221–229 (2003)

7. Nissar, A., Mir, A.H.: Classification of steganalysis techniques: A study. *Digital Signal Processing*, 1758–1770 (2010)
8. Farid, H.: Detecting hidden messages using higher-order statistical models. In: Proceedings of IEEE International Conference on Image Processing, New York, USA, vol. 2, pp. 905–908 (2002)
9. Lyu, S.W., Farid, H.: Steganalysis using higher-order image statistics. *IEEE Trans. Inf. Forensic Security* 1, 111–119 (2006)
10. Xuan, G., Shi, Y.Q., Gao, J., Zou, D., Yang, C., Zhang, Z., Chai, P., Chen, C.-H., Chen, W.: Steganalysis Based on Multiple Features Formed by Statistical Moments of Wavelet Characteristic Functions. In: Barni, M., Herrera-Joancomartí, J., Katzenbeisser, S., Pérez-González, F. (eds.) IH 2005. LNCS, vol. 3727, pp. 262–277. Springer, Heidelberg (2005)
11. Lie, W.N., Lin, G.S.: A feature-based classification technique for blind image steganalysis. *IEEE Trans. Multimedia* 7, 1007–1083 (2005)
12. Sajedi, H., Jamzad, M.: A Steganalysis method based on contourlet transform coefficients. In: International Conference of Intelligent Information Hiding and Multimedia Signal Processing (2008)
13. Sajedi, H., Jamzad, M.: ContSteg: Contourlet-Based Steganography Method, Wireless Sensor Network, vol. 1(3), pp. 163–170. Scientific Research Publishing (SRP), California (2009)
14. Do, M.N., Vetterli, M.: Contourlets: a directional multiresolution image representation. In: Proc. of IEEE Int. Conf. on Image Process., Piscataway, NJ, vol. 2002(9), pp. 357–360 (2002)
15. Do, M.N., Vetterli, M.: The Contourlet Transform: An Efficient Directional Multiresolution Image Representation. *IEEE Transaction on Image Processing* 14(12), 2091–2106 (2006)
16. Burt, P.J., Adelson, E.H.: The Laplacian pyramid as a compact image code. *IEEE Trans. Commun.* 31(4), 532–540 (1983)
17. Bamberger, R.H., Smith, M.J.T.: A filter bank for the directional decomposition of images: theory and design. *IEEE Trans. Signal Process.* 4, 882–893 (1992)
18. Yazdi, M., Mahyari, A.G.: A new 2D fractal dimension estimation based on Contourlet transform for texture segmentation. *The Arabian Journal for Science and Engineering* 35(13), 293–317 (2010)
19. Mosleh, A., Zargari, F., Azizi, R.: Texture Image Retrieval Using Contourlet Transfrom. In: International Symposium on Signal, Circuits and Systems (2009)
20. Do, M.N., Vetterli, M.: Directional multiscale modeling of images using Contourlet transform. *IEEE Transactions on Image Processing* 15(6), 1610–1620 (2006)
21. Yang, C.L., Wang, F., Xiao, D.: Contourlet Transform based Structural Similarity for image quality assessment. *Intelligent Computing and Intelligent Systems* (2009)
22. Fridrich, J.: Feature-Based Steganalysis for JPEG Images and Its Implications for Future Design of Steganographic Schemes. In: Fridrich, J. (ed.) IH 2004. LNCS, vol. 3200, pp. 67–81. Springer, Heidelberg (2004)

Algorithm for Clustering with Intrusion Detection Using Modified and Hashed K – Means Algorithms

M. Varaprasad Rao¹, A. Damodaram², and N.Ch. Bhatra Charyulu³

¹ Department of Computer Science, MIPGS, Hyderabad – 59

vpr_m@yahoo.com

² Department of Computer Science, JNTU, Hyderabad – 87

damodarama@jntu.ac.in

³ Department of Statistics, Osmania University, Hyderabad – 7

dwarakbhat@osmania.ac.in

Abstract. The k-Means clustering algorithm partition a dataset into meaningful patterns. Intrusion Detection System detects malicious attacks which generally include theft information. It can be found from the studies that clustering based intrusion detection methods may be helpful in detecting unknown attack patterns compared to traditional intrusion detection systems. This paper presents modified k-Means by applying preprocessing and normalization steps. As a result the effectiveness is improved and it overcomes the shortcomings of k-Means. This approach is proposed to work on network intrusion data and the algorithm is experimented with KDD99 dataset and found satisfactory results.

Keywords: Intrusion Detection System, K-Means clustering Algorithm, AIM.

1 Introduction

The collection of a set of similar data objects with respect to a particular (set of) characteristic(s) is said to be a cluster. Cluster Analysis is the name given to a diverse collection of techniques that can be used to classify objects. The classification has the effect of reducing the dimensionality of a data. It is a statistical technique given by US Psychologist Robert Choate Tyron (1935). Cluster analysis identifies and classifies individual objects or variables on the basis of the similarity of the characteristics they possess. It seeks to minimize within-group variance and maximize between-group variance. The result of cluster analysis is a number of heterogeneous groups with homogeneous contents: There are substantial differences between the groups, but the individuals within a single group are as similar as possible. Various clustering algorithms have been designed for various data mining problems.

The K-means algorithm was proposed by Mac Queen in 1967. Modified k-means is given by Alsabti (1998). In k-means, partition the items into k initial clusters, randomly take a set of k- vectors each consisting of p- components as initial centroids, then compute the distance between each object and the k-vectors determined and assign them to the cluster corresponding to minimum distance. Proceed through the list of items assigning an item to the cluster whose centroid is nearest. Recompute the cluster centroids until the cluster centroids stabilize upto a desired level of approximation.

Several authors [3], [4], [5], [6], [7] made attempts on k-means algorithm to increase its effectiveness in producing clusters for many practical applications. Computer network provides usefulness in many ways to mankind. It becomes an integral part of daily life. However, ease of access, outward anonymity, and wide prevalence of saving sensitive information on computers has attracted a large number of criminals and hacker hobbyists. So it is impossible for any computer system to be claimed resistant to network intrusion. Since there is no perfect solution to prevent intrusions from happening, it is very important to be able to detect them at the first moment of occurrences and take actions to minimize the possible damage.

Intrusion Detection [23] is the process of detecting the malicious attacks or threats to any computer or network. The basic task of Intrusion Detection is to audit the log data of a computer which includes network as well as host based data. Intrusion detection process helps to make the network more secure for data transmission, helps to determine the attackers and facilitates to stabilize and increase the lifetime of the network. Clustering algorithms may play a vital role in developing efficient intrusion detection algorithms. These algorithms are able to detect important patterns in huge amount of data. Again, they are having the potential to detect new types of network attacks without any prior knowledge of their existence. Thus clustering techniques are best suited for designing anomaly detection models, means they can deal with new type of attacks.

2 Clustering Algorithms for Intrusion Detection

Y-Means [25] is based on the k-Means Algorithm [5]. In this algorithm partition of the data takes place automatically. It classifies the normal and the abnormal or intrusive clusters. The iteration will continue until there is no empty cluster. Outliers of clusters are removed to form new cluster. If instances are more similar to each other, then the clusters will overlap each other. At last, population ratio of one cluster is above the given threshold all the instances in the cluster will be classified as normal labeled as intrusive.

Intrusion Detection Using Unlabelled Data Technique [26] groups similar data instances into clusters together and then uses distance metrics on clusters to trace out the anomalies. The assumptions for this technique to work are: data instances having same type should be close to each other in feature space under some reasonable metric, while instances with different types must be far apart. The data instances having normal type must be in majority in the training set. It creates clusters from its input data, then labels these clusters as either normal or anomaly and then uses these clusters to classify unseen network data instances as either normal or anomalous.

Hybrid Intrusion Detection Systems [27] is built using two basic models, anomaly detection and misuse detection. Anomaly detection consists of a rule-based model of normal behavior, against which the detected behavior is compared. It has a high detection rate, but the false positive rate is also high. The misuse detection model is built to classify the attack type by comparison with the known types of attack behavior. It has high accuracy, but the detection rate is lower. By combining features of both the models, the Hybrid system gains advantages of both the models.

This paper discuss result analysis of proposed modified k-Means clustering algorithm on various datasets from UCI data repositories and a new approach hashed k-means algorithm is proposed to work on network intrusion data. The algorithm is experimented with KDD99 dataset and found satisfactory results.

3 Modified K-Means Clustering

In this section, an attempt is made to overcome the shortcomings of k-Means to provide the seed values in advance. First preprocess the dataset with one of the minimum, constant, maximum, average, standard deviation options of cleaning method then normalize the data set using min-max, z-score and decimal scaling techniques. Flow chart for the proposed model is shown in the Fig 1.

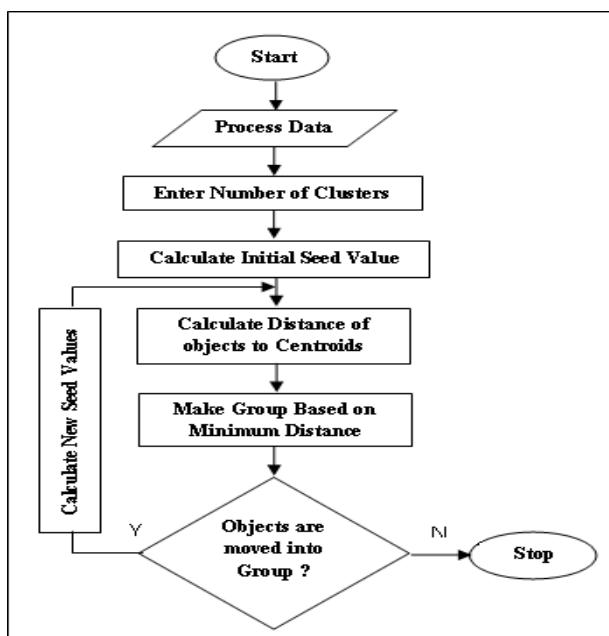


Fig. 1. Flow-Chart of Modified k-Means Algorithm

Let X be the dataset be representing N data objects x_1, x_2, \dots, x_N . Each object has M different attribute values corresponding to the M different attributes. The value of i^{th} object can be given by $X_i = \{x_{i1}, x_{i2}, \dots, x_{iN}\}$. In the first phase the AIM algorithm is applied to find out the value automatically in prior to partition the dataset into ‘ k ’ disjoint subsets. The initial mean set will be generated by the AIM algorithm based on the following distance measures:

$$\text{Distance-Threshold } (D_x) = \mu \pm \sigma$$

Where μ is the grand mean and σ is the standard deviation of the dataset X. Let \bar{x}_i be the mean of the attribute values of i^{th} object belongs to the dataset X and \bar{x} be the mean of the dataset X, then

$$\mu = \frac{1}{N} \sum_{i=1}^N \bar{x}_i \quad \text{and} \quad \sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{x}_i - \bar{x})^2$$

$$\text{Average-Distance } (A_{dx}) = \frac{1}{m} \sum_{i=1}^m d(\text{Setmean}_i, m_i))$$

Where Setmean_i is the initial mean, m_i and is the new cluster mean.

The hash value can be calculated based on the training dataset, which is different for normal and is different for different attack types. A hash table is maintained for all types of known attack types.

Mean and standard deviations are calculated on the training dataset. The last field in the dataset is replaced by the product of the standard deviation and the number of instances of each type, providing each type with a new value. Then, based on the Euclidean distance of the points in the modified training dataset, clusters are formed.

These clusters are used as base for classification of the testing dataset.

In this phase, the KDD99 dataset for intrusion detection is used to train the system. The number of instances for each type of connection is calculated along with the mean and the standard deviation. The hash value is generated as mentioned above. Then the last field is replaced with its hash value and process of cluster formation is carried out. After the cluster formation is over check whether the system is able to detect the intrusion on the system or not. It is done by taking a testing dataset and matching it with the cluster formed already in the training phase.

Algorithm HASHED-K-MEANS ()

Input :

//Input Dataset K
//Number of desired clusters obtained from the method.

Output :

// Set of clusters

Begin

1. Calculate number of instances of each type, the mean and standard deviation of the input.
2. Replace the last field of the labels in the training data with the product of the standard deviation and the number of instances.
3. Assign initial values for means $m_1, m_2 \dots m_k$;

Repeat

- a. Assign each item t_i to the cluster which has the closest mean;
- b. Calculate new mean for each cluster;

Until convergence is met;

End

This proposed model is implemented using PHP and MySQL which prepare the data using different preprocessing techniques. First, Dataset is prepared by applying cleaning method which fill missing values with various options like average, maximum, minimum, constant and standard deviation and remove noise from the dataset. Once the noise is removed from the dataset, volume of data is reduced by applying different normalization approach. After data preparation, modified k-Means algorithm is applied computes sum of squared error of each cluster and mean of SSE.

This section represents the performance evaluation of the proposed model on the River, Iris and datasets from UCI machine learning repository. In every phase, the SSE of each cluster is computed. Finally, Mean of SSE (MSE) is computed for generated clusters which measure the accuracy of the generated clusters. Computed MSE for two clusters with different cleaning methods and normalization approaches; min-max, z-score and decimal scaling for above data sets is shown in Table 1.

Table 1. Computed Error in “Modified k-Means Clustering Algorithm”

Sl No	Data	Cleaning Technique	Error in Modified k-Mean	Error in Modified k-means with Min-Max Method	Error in Modified k-means with z-score Method	Error in Modified k-means with Decimal Scaling Method
1	River	Avg	3.878	0.795	0	0.89
		Const	8.828	0.488	0	1.829
		Min	7.777	0.795	0	0.89
		Max	3.878	0.795	0	0.961
		SD	3.878	0.48	0	1.7
2	Iris	Avg	11.964	6.103	8.631	1.212
		SD	11.989	6.5	0.954	1.128

4 Results and Analysis

The Fig.2 shows one of the initial clusters formed after the training of the system with the Kdd99 dataset. It is found that the normal instances making their own clusters while the other type of connections making their own clusters based on the hash values and the Euclidean distances.

File	Edit	Format	View	Help		
00	0, 00	0.00	0.00	0, 00	smurf,	
00	0.00	0.00	0.00	0.00	smurf,	
00	0.00	0.00	0.00	0.00	smurf,	
00	0.00	0.00	0.00	0.00	smurf,	
00	0.00	0.00	0.00	0.00	smurf,	
00	0.00	0.00	0.00	0.00	smurf,	
00	0.00	0.00	0.00	0.00	smurf,	

Fig. 2. Initial Cluster

The Fig 3 shows the different threats detected which were present in the testing dataset. The proposed algorithm is able to detect unknown type of connections present in the testing dataset.

		File	Edit	Format	View	Help
0		normal				
1		threat				
2		threat				
3		normal				

Fig. 3. Detection Phase

The inter cluster distance of the clusters formed is calculated by finding the Euclidean distance between the cluster means and the clusters members of all the clusters formed. A graph is plotted to study the variations of the clusters formed. A larger inter-cluster distance would mean clusters are more distinct and tight. The Fig 4 shows the graph of inter-cluster distances of the clusters formed by the K-Means and Hashed K-Means algorithms. It can be seen that the clusters formed by the Hashed-K-Means show better and distinct results as compared to normal K-Means algorithm.

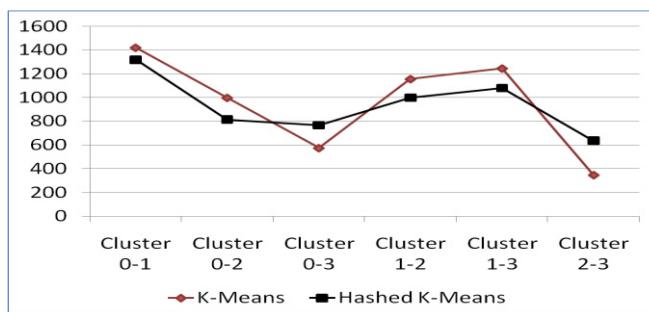


Fig. 4. Inter cluster distances for K means and Hashed K means Algorithm

5 Conclusions

Clustering is an unsupervised learning technique used in data mining. Clustering algorithm never is universal to solve all problems as they are designed on certain assumptions. k-Means is popular for its favorable execution time having some shortcoming like pass number of seed values in advance, does not handle noise and outliers. The proposed k-means is based on K-means and AIM method with some modifications. The efficiency of the algorithm with which intrusions are detected is around 90%-95%. The accuracy of this algorithm depends on the training data taken. Hashed-K-Means Intrusion Detection Algorithm, a chance that; a particular instance might be considered as an abnormal / attack type rather than a normal instance. The algorithm can detect new types of attacks but its classification would be a difficult

task because it's not present in the training data. The comparative analysis of clustering algorithms and result analysis are done for various datasets from UCI data repositories. The proposed model suffers from shortcoming like more time complexity and number of clusters passes in advance.

References

1. Kaufman, L., Rousseeuw, P.J.: *Finding Groups in Data: An introduction to Cluster analysis*. John Wiley, New York (1990) ISBN 0-471-85233-3
2. Velmurugan, T., Santhanam, T.: Computational Complexity between K-Means and KMedoids Clustering Algorithms for Normal and Uniform Distributions of Data Points. *Journal of Computer Science* 6(3), 363–368 (2010)
3. Han, J., Kamber, M.: *Data Mining Concepts and Techniques*, 2nd edn. Morgan Kaufmann Publishers. An Imprint of Elsevier (2006) ISBN 81-312-0535-5
4. Dunham, M.H.: *Data Mining- Introductory and Advanced Concepts*. In: Pearson Education 2006. Proceedings of the World Congress on Engineering, vol. 1 (2009)
5. McQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297. Univ. of California Press, Berkeley (1967)
6. Merz, C., Murphy, P.: UCI Repository of Machine Learning Databases,
<ftp://ftp.ics.uci.edu/pub/machine-learning-databases>
7. Tan, P.-N., Steinback, M., Kumar, V.: *Introduction to Data Mining*. Pearson Education (2007)
8. Patel, V.R., Mehta, R.G.: Clustering Algorithms: A Comprehensive Survey. In: International Conference on Electronics, Information and Communication Systems Engineering, Jodhpur (2011)
9. Oyelade, O.J., Oladipupo, O.O., Obagbuwa, I.C.: Application of kMeans Clustering algorithm for prediction of Students' Academic Performance. *International Journal of Computer Science and Information Security* 7 (2010)
10. Sumitra Devi, K.A., Vijayalakshmi, M.N., Vasantha, R., Abraham, A.: Accomplishment of Circuit Partitioning using VHDL and Clustering Pertaining to VLSI design
11. Tilton, J.C., Marchisio, G., Koperski, K.: NASA's Intelligent Systems Program, NASA Headquarter Code R
12. Ng, R.T., Han, J.: CLARANS:A Method for Clustering Objects for Spatial Data Mining. *IEEE Transaction on Knowledge and Data Engineering* 14(5), 1003–1016 (2002)
13. Seidman, C.: *Data Mining with Microsoft SQL Server 2000*, Technical Reference, ISBN:0-7356-1271-4, <http://amazon.com/Mining-Microsoft-Server-Technical-Reference/dp/0735612714>
14. Noh, S.-K., Kim, Y.-M., Kim, D.K., Noh, B.-N.: Network Anomaly Detection Based on Clustering of Sequence Patterns. In: Gavrilova, M.L., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganá, A., Mun, Y., Choo, H. (eds.) *ICCSA 2006. LNCS*, vol. 3981, pp. 349–358. Springer, Heidelberg (2006)
15. Sahay, S.: Study and Implementation of CHEMELEON algorithm for gene clustering
16. Erman, J., Arlitt, M., Mahanti, A.: Traffic Classification Using Clustering Algorithms. In: *SIGCOMM 2006 Workshops* Pisa, Italy, September 11-15 (2006)
17. Santhisree, K., Damodaram, A.: OPTICS on Sequential Data: Experiments and Test Results. *International Journal of Computer Applications* 5, 1–4 (2010)

18. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. Department of Computer Science, University of Wisconsin, Madison, WI 53706
19. Maheshwari, P., Srivastava, N.: WaveCluster for Remote Sensing Image Retrieval. International Journal on Computer Science and Engineering 3(2) (2011)
20. Scanlan, J., Hartnett, J., Williams, R.: DynamicWEB: Profile Correlation Using COBWEB. In: Sattar, A., Kang, B.-h. (eds.) AI 2006. LNCS (LNAI), vol. 4304, pp. 1059–1063. Springer, Heidelberg (2006)
21. Patel, V.R., Mehta, R.G.: Modified k-Means Clustering Algorithm. In: Das, V.V., Thankachan, N. (eds.) CIIT 2011. CCIS, vol. 250, pp. 209–213. Springer, Heidelberg (2011)
22. Borah, S., Chetry, S.P.K., Singh, P.K.: Hashed-K-Means: A Proposed Intrusion Detection Algorithm. In: Das, V.V. (ed.) CIIT 2011. CCIS, vol. 250, pp. 855–860. Springer, Heidelberg (2011)
23. Sabahi, F., Movaghar, A.: Intrusion Detection: A Survey. In: The Proceedings of 3rd International Conference on Systems and Networks Communications, ICSNC 2008, vol. 1. IEEE (2008) ISBN: 978-0-7695-3371-1
24. Borah, S., Ghose, M.K.: Automatic Initialization of Means (AIM): A Proposed Extension to the K-means Algorithm. International Journal of Information Technology & Knowledge Management 3(2), 247–250 (2010) ISSN: 0973-4414
25. Guan, Y., Ghorbani, A., Belacel, N.: Y-means: A Clustering Method for Intrusion Detection. In: Proceedings of Canadian Conference on Electrical and Computer Engineering, Montreal, Quebec, Canada, May 4-7, pp. 1083–1086 (2003)
26. Portnoy, L., Eskin, E., Stolfo, S.: Intrusion Detection with Unlabeled Data Using Clustering. In: Proceedings of the ACM CSS Workshop on Data Mining Applied to Security (DMSA 2001), Philadelphia, PA, November 5-8 (2001)
27. Yan, K.Q., Wang, S.C., Liu, C.W.: A Hybrid Intrusion Detection System of Cluster-based Wireless Sensor Networks. In: Proceedings of the International Multi-Conference of Engineers and Computer Scientists 2009, IMECS 2009, Hong Kong, March 18-20, vol. I (2009)
28. Dataset is available online at
<http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>

Z Transform Based Digital Image Authentication Using Quantization Index Modulation (Z-DIAQIM)

Nabin Ghoshal¹, Soumit Chowdhury², and Jyotsna Kumar Mandal³

¹ Dept. of Engineering and Technological Studies, University of Kalyani, Kalyani,
Nadia-741235, West Bengal, India

² Dept. of Computer Science and Engineering, Govt. College of Engineering & Ceramic
Technology, 73, A. C. Banerjee Lane, Beliaghata, Kolkata-700010

³ Dept. of Computer Science and Engineering, University of Kalyani, Kalyani,
Nadia-741235, West Bengal, India

{nabin_ghoshal, joy_pinu}@yahoo.co.in, jkm.cse@gmail.com

Abstract. This paper presents a novel Steganographic technique of color image authentication technique based on the Discrete two dimensional Z-Transform using Quantization Index Modulation (QIM). The Z-Transform is applied on sub-image block of size 2 x 2 in row major order of the carrier image for frequency components of the corresponding spatial components. Image authentication is done by hiding secret message/image into the real part of the frequency component of the carrier image. A single bit of secret message/image is embedded in each carrier image byte based on Quantization Index Modulation where a tolerance factor Δ has been used for invisible embedding. After embedding, a delicate re-adjusting phase is applied in all the frequency components of each mask, to keep the pixel values positive and non-fractional in the spatial domain. It is also applicable for secrete data transmission through carrier color image by hiding secrete data. Experimental results proof the robustness and performance of the proposed technique.

1 Introduction

Digital images are transmitted over popular communication channels such as the Internet. For secured communication, image authentication [1, 2, 3] techniques have gained more attention due to their importance for a large number of multimedia applications. Military, medical and quality control images must be protected from alteration so as to maintain their authenticity, a number of approaches have been proposed for this including: conventional cryptography, fragile and semi-fragile watermarking [9, 10] and digital signatures. Digital watermarking is the process of hiding the watermark imperceptibly in the content. This technique was initially used in paper and currency as a measure of authenticity. Copyright [4, 5] abuse is the motivating factor in developing new encryption technologies. One of the driving forces behind the increased use of copyright marking is the growth of the Internet which has allowed images, audio, video, etc to become available in digital [6, 7] form. Though this provides an additional way to distribute material to consumers it has also made it far easier for copies of copyrighted material to be made and distributed. Using the Internet a copy stored on a computer can be shared easily with anybody regardless

of distance often via a peer-to-peer network which doesn't require the material to be stored on a server and therefore makes it harder for the copyright owner to locate and prosecute offending parties. Copyright marking is seen as a partial solution to these problems. The mark can be embedded in any legal versions and will therefore be present in any copies made. This helps the copyright owner [8, 11] to identify who has an illegal copy.

In general, there is a tradeoff between the watermarks embedding [14] strength (the watermark robustness), quality (the watermark invisibility) and security. Increased robustness requires a stronger embedding, which in turn increases the visual degradation of the images. The proposed watermarking scheme adopts a color image as the watermark so human eyes can easily verify the extraction of this visually meaningful watermark. In general, a color image can provide more perceptual information i.e. sufficient evidence against any illegal copyright invasion.

This paper introduces a strong watermarking technique with the help of Z-Transform using the concept of Quantization Index Modulation which satisfies all the necessary criteria that watermarking should satisfy and also provides a good security from all possible attacks. The result of the proposed technique Z-DIAQIM is compared with the existing DFTMCIAWC, DCT-based watermarking method, QFT-based and Spatiochromatic DFT-based watermarking method in terms of visual interpretation, Mean Square Error (MSE), Pick Signal-to-Noise Ratio (PSNR) in dB and Image Fidelity (IF).

This paper is organized as follows: Two Dimensional Discrete Z-Transform and inverse Z-Transform have been presented with expression and simplification.

1.1 Two Dimensional Z Transform

A spatial function $f(n_1, n_2)$ where (n_1, n_2) is a spatial coordinate can be represented in Z-Transform as

$$f(z_1, z_2) = \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} f(n_1, n_2) z_1^{-n_1} z_2^{-n_2} \quad (1)$$

where z_1 and z_2 are both complex numbers consisting of real and an imaginary parts. Since z_1 and z_2 are complex numbers, Let $z_1 = e^{j\omega_1\pi}$ and $z_2 = e^{j\omega_2\pi}$, Where $e^{j\theta} = \cos\theta + j\sin\theta$. Substituting the values of z_1 and z_2 in equation (1), the equation becomes the discrete form of Two Dimensional Z Transformation equation.

$$f(e^{j\omega_1\pi}, e^{j\omega_2\pi}) = \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} f(n_1, n_2) e^{j\omega_1\pi - n_1} e^{j\omega_2\pi - n_2}$$

$$\text{Or } f(\omega_1, \omega_2) = \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} f(n_1, n_2) e^{-j\pi(n_1\omega_1 + n_2\omega_2)} \quad (2)$$

Where ω_1, ω_2 are two frequency variables, varies from $-\infty$ to $+\infty$.

1.2 Two Dimensional Inverse Z Transform

The continuous Inverse Z-Transform of a function $f(n_1, n_2)$ is represented as

$$f(n_1, n_2) = \left(\frac{1}{2\pi j}\right)^2 \iint f(z_1, z_2) z_1^{n_1-1} z_2^{n_2-1} dz_1 dz_2 \quad (3)$$

Where $f(n_1, n_2)$ be a function and $f(z_1, z_2)$ be the Z-Transform of the function $f(n_1, n_2)$. Control integration is for irregular spaces in z-domain.

1.3 Derivation of Inverse Z Transform from Continuous to Discrete Form

Since z_1 and z_2 are complex numbers, Let $z_1 = e^{j\omega_1\pi}$ and $z_2 = e^{j\omega_2\pi}$, where $e^{j\omega\theta} = \cos\omega\theta + j\sin\omega\theta$. Substituting the values of z_1 and z_2 in equation (3), we have a discrete form of inverse Z Transform for two dimensions. Now $z_1 = e^{j\omega_1\pi}$, differentiating this with respect to ω_1 we get $\frac{dz_1}{d\omega_1} = e^{j\omega_1\pi} j\pi$, therefore $dz_1 = e^{j\omega_1\pi} j\pi d\omega_1$ and $z_2 = e^{j\omega_2\pi}$, differentiating this with respect to ω_2 we get $\frac{dz_2}{d\omega_2} = e^{j\omega_2\pi} j\pi$, therefore $dz_2 = e^{j\omega_2\pi} j\pi d\omega_2$. The equation (3) becomes from the above derivation is

$$f(n_1, n_2) = \left(\frac{1}{2\pi j}\right)^2 \iint f(e^{j\omega_1\pi}, e^{j\omega_2\pi}) e^{j\omega_1\pi n_1-1} e^{j\omega_2\pi n_2-1} e^{j\omega_1\pi} j\pi d\omega_1 e^{j\omega_2\pi} j\pi d\omega_2$$

The discrete form of this control integration equation is as follows

$$f(n_1, n_2) = \frac{1}{4} \sum_{\omega_1=-1}^1 \sum_{\omega_2=-1}^1 f(\omega_1, \omega_2) e^{j\pi(n_1\omega_1 + n_2\omega_2)} \quad (4)$$

The equation (4) is the discrete form of Two Dimensional Inverse Z Transform.

2 The Technique

The Insertion of the secrete image is performed in the Z-Domain i.e. the frequency component obtained after performing the Z-Transform on each 2×2 sub-image matrix of the original image matrix. Hence, in order to perform the insertion operation of the secrete image into converted original image byte. A classical data-hiding scheme named Quantization Index Modulation scheme is used with a tolerance factor Δ where the value of tolerance factor Δ is chosen as minimum value such as to be able to detect the embedded watermark from the watermarked color image.

In the Embedding process each bit of secrete binary value is embedded in each byte of the carrier color image. Let $M \times N$ be the size of the original digital color image. Evidently, $M \times N$ bits are to be embedded in the original digital color image.

Let C_w be the image byte in Watermarked color image and C be the original transformed image byte. Let $b[i]$ be the bit to be embedded in image byte C . The embedding or coding step and the detection scheme or the decoding step is as follows:

Coding Step

$$\begin{array}{ll} \text{If } b[i] = 1: & C_w = 2 \Delta \text{Round}(C/(2\Delta)) + \Delta/2 \\ \text{If } b[i] = 0: & C_w = 2 \Delta \text{Round}(C/(2\Delta)) - \Delta/2 \end{array}$$

Decoding Step

$$\text{If } C - 2 \Delta \text{Round}(C/(2\Delta)) > 0, \text{ then } b[i] = 1, \text{ else } b[i] = 0$$

The value of Δ is chosen very carefully. For high value of tolerance factor Δ , Mean Square Error will be increases and subsequently Peak Signal to Noise Ratio decreases. A suitable value is needed for Δ for successful embedding and extraction with less noise integration.

2.1 Insertion Algorithm

Input: A carrier image and secrete message/image.

Output: An authenticated image.

Method: Embedding has been performed on the integer values only while the floating point part has been made intact and has been added after embedding the watermarking bits in the integer part of the pixels values of the source image.

1. Read Image type, dimensions and maximum intensity from source image and write in the output image.
2. Repeat until all pixels have been read from the source image file.
 - 2.1 Repeatedly Take 2×2 blocks of pixels from the matrix at the left and perform Z-Transform of the block of pixels until all pixels in the matrix have been taken.
 - 2.2 Read the secrete image i.e. watermark.
 - 2.3 Embed the watermark bits in the source image using the Coding steps as mentioned above and sustain the watermarked values of each pixel in the 2×2 mask.
 - 2.4 Compute the Inverse Z-Transform of the 2×2 block of pixels.
 - 2.5 If any pixel is found to be of negative value, the maximum negative number is stored and added in the watermarked pixel values such that there is no effect on the bit position where the watermark bit is embedded.
 - 2.6 Compute the Inverse Z-Transform of the block of pixels and the numbers obtained is guaranteed to be of positive values.
 - 2.7 Repeat the steps from 2.1 to 2.6 until all pixels have been transformed.
3. Stop.

2.2 Extraction Algorithm

The Extraction of the secrete image is performed in the Z-Domain i.e. the domain obtained after performing the Z-Transform of the embedded image. Hence, in order to perform the extraction operation of the secrete image from the embedded image.

Input: Authenticated image.

Output: The original image, secrete message/image.

Method: Extraction has been performed on the integer values only while the floating point part has been made intact and has been added after extracting the security bits from the integer part of the pixels values of the source image.

1. Read Image type, dimensions and maximum intensity from embedded image and skip writing in the output image.
2. Repeat until all pixels have been read from the embedded image.
 - 2.1. Repeatedly Take 2×2 blocks of pixels from the matrix at the left and perform Z-Transform of the block of pixels until all pixels in the matrix have been taken.
 - 2.2. Calculate the embedded bits from the embedded image using Decoding steps as mentioned above.
 - 2.3. Convert each 8 bits of 0's and 1's into decimal value and write the value in the output image.
 - 2.4. Repeat the steps from 2.1 to 2.3 until all pixels have been transformed and embedded bits have been calculated.
3. Stop.

3 Result Comparison and Analysis

This section represent the results, discussion and a comparative study of the proposed technique Z-DIAQIM with the DFTMCIAWC technique, DCT-based, QFT-based and Spatiochromatic DFT-based watermarking method in terms of visual interpretation, image fidelity (IF[12]), and peak signal-to-noise ratio (PSNR[12]) analysis and mean square error (MSE[12]). In order to test the robustness of Z-DIAQIM, the technique is applied on more than 100 PPM colour images from which it may be revealed that the algorithm may overcome any type of attack. Distinguishing of carrier and embedded image from human visual system is quite difficult. The original carrier benchmark [13] images ‘Airplane’, ‘Baboon’, ‘Lenna’ and ‘Oakland’ are shown in Fig 1a, 1b, 1c and 1d each with a dimension of 512×512 and 786432 bits of secret information are embedded. Each bit of secret data is embedded in each byte of the carrier image. The embedded images obtained after the insertion of the secret image ‘Earth’ (Fig. 1m) are shown in Fig. 1e, 1f, 1g and 1h using Z-DIAQIM with the value of the tolerance factor Δ is 4. The magnified versions of different embedded images have been shown in fig 2i, 2j, 2k and 2l to study the effect of insertion of the secret data in the carrier images.

Peak signal-to-noise ratio (PSNR) is used to evaluate qualities of the embedded images. Table 1 show that 786432 bits or 98304 bytes of secrete data embedding is done with higher PSNR values for different carrier images. Table 2 shows the PSNR values for Lenna image in existing methods [12] like SCDF, QFT and DCT. In DCT based watermarking scheme do not embed watermarks in every single block of image. Here selectively pick the regions that do not generate visible distortion for embedding, thus decreasing the authenticating data size. In QFT based watermarking compensation mark allows the watermark to be undetected even if the strength of it is high. For low compression factor it can not completely recover the embedded message. In all the existing technique the PSNRs are low, means bit-error rate are high but in Z-DIAQIM

more bytes of secret data can be embedded and the PSNR values are significantly high, means bit-error rate is low. Using Z-DIAQIM technique the average PSNR enhancements are 13.4176, 12.5917 and 13.1154 dB than SCDFT, QFT and DCT respectively with 98304 bytes of embedding capacity. Compare to DFTMCIAWC the average PSNR is slightly increased in Z-DIAQIM but amount of secrete data is increased by 24574 byte.

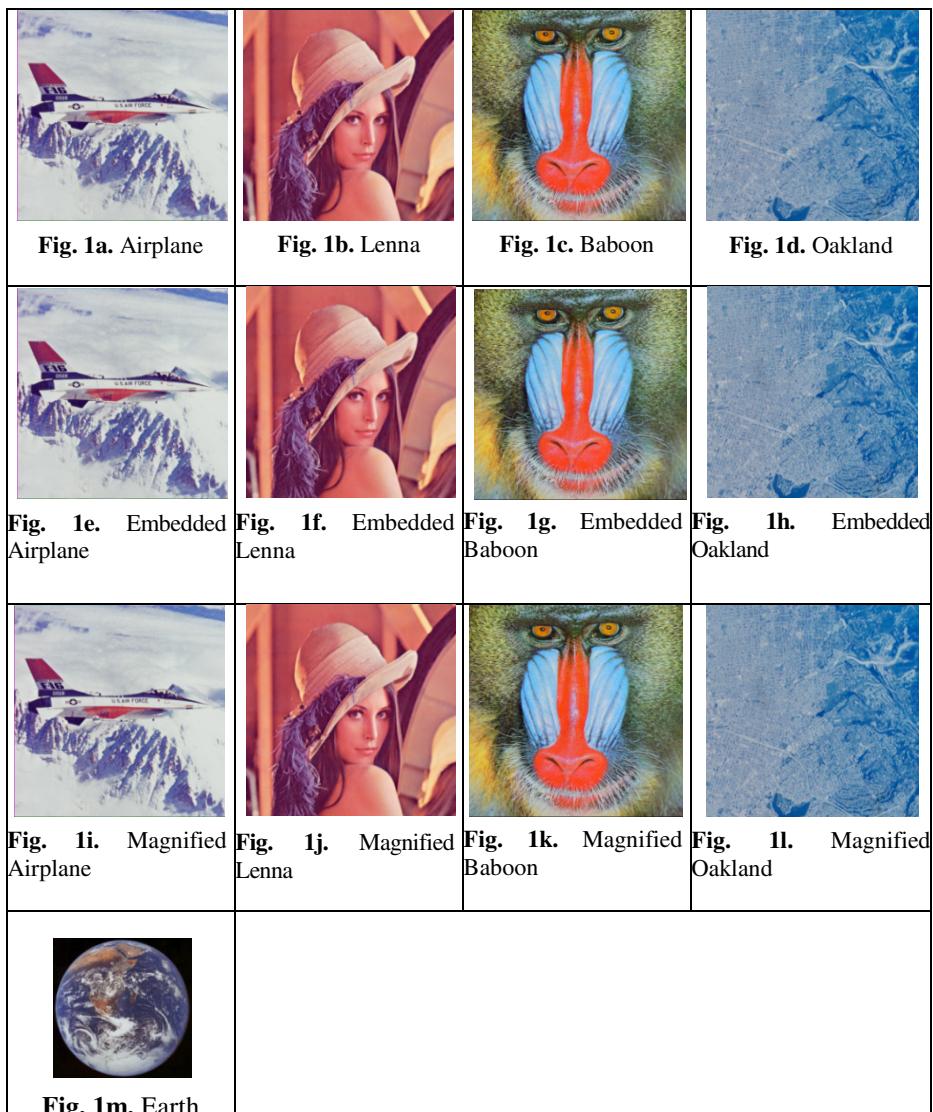


Fig. 1. Visual interpretation of embedded image using Z-DIAQIM and corresponding magnified images after embedding

Table 1. Results of capacities and MSE, PSNR and IF in Z-DIAQIM

Carrier images	Capacity (byte)	MSE	PSNR in dB	IF
Airplane	98302	2.776405	43.695976	0.999723
Baboon	98302	2.848125	43.585213	0.999801
Lenna	98302	3.132698	43.171619	0.999685
Oakland	98302	2.973476	43.398159	0.999863
Peppers	98302	3.108784	43.204899	0.999755
Sailboat	98302	3.028671	43.318283	0.999370
San Diego	98302	3.093292	43.226593	0.999868
Splash	98302	2.701584	43.814621	0.999829
Tiffany	98302	2.501717	44.148422	0.999837
Woodlad	98302	2.827515	43.616756	0.999295
Average	98302	2.899227	43.518054	0.999703

Table 2. Comparison in capacities and PSNR of Z-DIAQIM with some existing techniques [11]

Technique	Capacity (bytes)	PSNR in dB
SCDFT	3840	30.1024
QFT	3840	30.9283
DCT	3840	30.4046
DFTMCIAWC	73728	43.26
Z-DIAQIM	98302	43.52

4 Conclusion

Z-DIAQIM technique is applicable for copyright protection or ownership verification of digital image. In compare to DCT, QFT, SCDFT and DFTMCIAWC the Z-DIAQIM algorithm is more robust authentication process in Z-domain for any types of colour image. The PSNR is high and more embedding capacity of secret data into the carrier images. It is a very hard to detect watermark due to variability of tolerance factor Δ in the insertion process of the secrete bits in the carrier image. So, the proposed technique Z-DIAQIM also provides security from all possible attacks.

Acknowledgement. The author expresses the deep sense of gratitude to the Dept. of Engineering and Technological Studies, University of Kalyani, West Bengal, India, where the work has been carried out.

References

- [1] Ghoshal, N., Mandal, J.K.: A Novel Technique for Image Authentication in Frequency Domain using Discrete Fourier Transformation Technique (IAFDDFTT). Malaysian Journal of Computer Science 21(1), 24–32 (2008) ISSN 0127-9094

- [2] Ghoshal, N., Mandal, J.K.: A Bit Level Image Authentication / Secret Message Transmission Technique (BLIA/SMTT), Association for the Advancement of Modelling & Simulation Technique in Enterprises (AMSE). AMSE Journal of Signal Processing and Pattern Recognition 51(4), 1–13 (2008)
- [3] Ghoshal, N., Mandal, J.K., et al.: Masking based Data Hiding and Image Authentication Technique (MDHIAT). In: Proceedings of 16th International Conference of IEEE on Advanced Computing and Communications ADCOM 2008, Anna University, Chennai, India, December 14–17, pp. 119–122 (2008) ISBN: 978-1-4244-2962-2
- [4] EL-Emam, N.N.: Hiding a large Amount of data with High Security Using Steganography Algorithm. Journal of Computer Science 3(4), 223–232 (2007) ISSN 1549-3636
- [5] Ker, A.: Steganalysis of Embedding in Two Least-Significant Bits. IEEE Transaction on Information Forensics and Security 2(1), 46–54 (2008) ISSN 1556-6013
- [6] Yang, C., Liu, F., Luo, X., Liu, B.: Steganalysis Frameworks of Embedding in Multiple Least Significant Bits. IEEE Transaction on Information Forensics and Security 3(4), 662–672 (2008) ISSN 1556-6013
- [7] Wu, H.C., Wu, N.I., Tsai, C.S., Hwang, M.S.: Image steganographic scheme based on pisel-value differencing and LSB replacement methods. Proc. Inst. Elect. Eng., Vis. Images Signal Processing 152(5), 611–615 (2005)
- [8] Yang, C.H., Weng, C.Y., Wang, S.J., Sun, H.M.: Adaptive Data Hiding in edge areas of Images With Spatial LSB Domain Systems. IEEE Transaction on Information Forensics and Security 3(3), 488–497 (2008) ISSN 1556-6013
- [9] Ahmadi, N., Safabkhsh, R.: A novel DCT-based approach for secure color image watermarking. In: Proc. Int. Conf. Information technology: Coding and Computing, vol. 2, pp. 709–713 (April 2004)
- [10] Bas, P., Biham, N.L., Chassery, J.: Color watermarking using Quaternion Fourier Transformation. In: Proc. ICASSP, Hong Kong, China, pp. 521–524 (June 2003)
- [11] Tsui, T.T., Zhang, X.-P., Androutsos, D.: Color Image Watermarking Using Multidimensional Fourier Transfomation. IEEE Trans. on Info. Forensics and Security 3(1), 16–28 (2008)
- [12] Kutter, M., Petitcolas, F.A.P.: A fair benchmark for image watermarking systems. In: Electronic Imaging 1999, Security and Watermarking for Multimedia Content, San Jose, CA, vol. 3657, pp. 226–239 (1999)
- [13] Weber, A.G.: The usc-sipi image database:
<http://sipi.usc.edu/services/database/Database.html>, Signal and Image Processing Institute at the University of Southern California (October 1997)
- [14] Ghoshal, N., Mandal, J.K.: Discrete Fourier transform based Multimedia Color Image Authentication for Wireless Communication (DFTMCIAWC). In: 2nd International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, Le Royal Meridian Chennai, India, February 28–March 3 (2011) ISBN: 978-1-4577-0787-2/11

Secret Data Hiding within Tolerance Level of Embedding in Quality Songs (DHTL)

Uttam Kr. Mondal¹ and J.K. Mandal²

¹ Dept. of CSE & IT, College of Engg. & Management, Kolaghat, W.B, India
uttam_ku_82@yahoo.co.in

² Dept. of CSE, University of Kalyani, Nadia (W.B.), India
jk.m.cse@gmail.com

Abstract. Embedding message into audio signal especially within songs without compromising its audible quality is one of the growing research areas. The techniques are used for different purposes in practical life. Various methods are evolved for hiding information into audio signals and quality is maintained only by calculating the acceptance ratio of embedding data using human perception. Therefore, an efficient way of finding the limit of embedding information in song signal is one of the challenging issues. In this paper, using modulation of channel capacity in modified form is used to determine the embedded data and song signal trade-off ratio for getting future guideline embedding information in song signal with correlation among embedded data is fabricated. A comparative study has been made with similar existing techniques for performance analysis and experimental results are also supported with mathematical formula based on Microsoft WAVE ("*.wav") stereo sound file.

Keywords: Embedding secret message in quality song, linear coding, song authentication, tolerance level of embedding message, encoding lower frequencies.

1 Introduction

Today's creative organizations facing competitive market for spreading business as well as retaining their goodwill. Creating a quality product involved a lot of investments. People are finding easier way to invest less money and producing product for existence in this contemporary market. Some of them are applying technology for making piracy of original versions and producing lower price products. This intension is a frequent phenomenon for digital audio/video industry with improvement of digital editing technology [5]. Even, it is quite harder to listeners to find the original from pirated versions. Therefore, it is a big challenge for business persons, computer professionals or other concern people to enhance the security criteria regarding originality of songs [1, 2] and this protect the product from the release of duplicate versions.

In this paper, a framework for identifying a particular song with the help of unique secret code fabricated through hiding some secret information with help of linear coding technique over the song signal without affecting its quality has been presented. Proposed technique is fabricated by decomposing frequency components of the signal. Secret information is embedded in lower frequency region with a balance of its

quality of constituted components. Embedded signals with secrete code can easily distinguish the original from similar available songs. Finally, incorporating channel coding theorem in modified form for calculating acceptance ration of hiding message into song signal is done without affecting its quality. It is experimentally observed that proposed technique will not affect the song quality but provide a level of security to protect the piracy of song signal without violating the acceptance level.

Organization of the paper is as follows. Encoding and embedding secret message are presented in section 2.1 and 2.2 respectively. The authentication procedure has been depicted in section 2.3, estimation of embedded message is given in section 2.4 that of extraction in section 2.5. Experimental results are given in section 3. Conclusions are drawn in section 4. References are given at end.

2 The Technique

The scheme fabricates the secret key with help of linear coding technique in lower frequency region (1-200 Hz) [which is not used by audio systems] as well as encoding the lower frequencies (1-200 Hz) followed by embedding secret key in the encoded frequency region. The operation is done with an approximate estimation of limit of impurities impacted with song signal to reain the audible quality of signal. Algorithms termed as DHTL-ELF and DHTL-ESM are proposed as double security measure, the details of which are given in section 2.1 and section 2.2 respectively.

2.1 Encoding Lower Frequencies (DHTL - ELF)

Encoding lower frequencies in the specific positions of signal (1-200 Hz) is performed with help of linear coding technique over GF (8) ["GF" stand for "Galois field"]. The procedure of Encoding lower frequencies is depicted in the following algorithm.

Step 1: Take last 4 digits of 5 consecutive rows of magnitude values of frequency set, i.e., from 1st to 4th frequency positions and put into a two dimensional array of size 4 by 4. If any of these digits greater than 7, put 7 in this position and extra value (original value - 7) with all other extra value or 0 save in somewhere for generating secure code in high frequency region[above audible range]. if the value is .0910, then take 0710 for 1st row of the taken array and save 0.0200 with row number in some place for generating secure code[which will use in the process of section 2.2].

Step 2: Take another array of similar size and populate values with values of 2nd to 5th rows of frequency set, i.e., one position ahead of previous window (of step 1) and convert the data set as step 1 as well as save extra values for similar purpose as described in step1.

Step 3: Add elements of above two arrays and put the result in another array [of GF(2³) , Primitive polynomial = D³+D+1 (11 decimal)] say, C and replace the value of C in the place of B of the specified frequency positions of original song signal.

Step 4: Continue step 1 to 3 until B window (of step 2) reaches to the 200 Hz position.

The value of 1st C matrix needs to put in specified positions of higher frequency region for getting original frequency value set in the time of decoding process [which is described in section 2.5]. If the song is stereo type, then the above method may be repeated for the second channel also. Therefore, if any value changes in processing, the above relationship in lower frequency region will break and can easily detect the error.

2.2 Embedding Secret Message (DHTL - ESM)

Embedding the storage extra values [in step 1 and 2 of section 2.1] and values of 1st C window in the higher frequency region will create another security criteria over original song signal without hampering its audible quality. The embedding process as follows.

- i. Make equal the magnitude values of two channel of stereo type song as it will do not affect over audible quality of song signal [2, 3].
- ii. Separate each digit position of extra value and convert into equivalent lower magnitude value. Let, if the extra value is .0200, then, separated lower magnitude values are 0, 0.0002, 0 and 0 respectively.
- iii. Add the magnitude values in the higher frequency region of song signal [above 20,000 Hz] as follows.

Let, the magnitude value is V, V value will add to i^{th} position, then the same will add with alternate channel of $(i+1)^{\text{th}}$ position, i.e.,

$$\begin{aligned} x(20000+i, s) &= x(20000+i, s) + V \\ x(20001+i, s1) &= x(20001+i, s1) + V, \end{aligned}$$

where $s = 1$ or 2 . and $s1 = [(s+1) \bmod 2 + 1]$.

- iv. Continue the step i to iii until all extra values are added in the higher frequency region of consecutive locations. The storage 1st C matrix value also can be add in same region by similar way where each matrix element should be converted as above step ii.

In case of mono type song, the lower magnitude values can be add by separating specified positions with same channel.

2.3 Authentication

Embedding extra values in higher frequency region in specified manner creates a secure code that will use to identify the original song. The encoding data set in lower frequency set creates a unique relationship among magnitude values of song signal. The addition operation between window k^{th} and $(k+1)^{\text{th}}$ will create result window of same size which will equivalent of the original values of window that originally constitutes song signal.

Therefore, if any changes during processing, it will create a difference with the authenticating codes that present in the higher frequencies region of the song signal and changing a position will create difference with the hidden code in that region as well as linear coding relationship will break in lower frequency region.

2.4 Estimating Limit

Estimating limit of embedding data over song signal is one of the major issues when quality is a factor. For this purpose, an approach has been made for estimating the

boundary of adding impurities without compromising its audible quality is done with the help of channel coding theorem in modified form as follows.

Find the channel capacity with N_0W , where N_0 message embedding rate (here extra data) and W is the highest magnitude value of the song signal, here highest frequency value [i.e. 20,000 Hz]. Shannon limit may be considered for generating limit with help of associated formula.

$r = k/N$, where r is the channel transmission rate, k is number of added values (magnitude values), may also use for calculating embedding data limit with song signal.

Channel capacity can be expressed by following formula

$$C = W \log_2 \left(1 + \frac{P}{N_0 W} \right) \quad (1)$$

bits per second (here, magnitude values per sampled set of song signal)

Assigning the values of above variables with considering Shannon's limit are given as follows

$W = 20,000$ (approx) [considering audible range 20-20,000 Hz]

$C = 44,100$ [16-bit stereo audio signals sampled at 44.1 kHz]

$$P = E[X_k^2]$$

Where, X_k = frequency values of original song, $K=1, 2, 3, \dots, L$ [L = length of song signal]

Putting above values in the equation (1), we can easily find the noise value N_0 .

Spectrum density = $N_0/2$ [maximum noise] and ($N-k$) number of added impurities (noise) in sampled values of song signal.

Where $r \leq C$, according to channel coding theorem.

Therefore, we can conclude, if the above limit exceeds then song will lost its audible quality.

2.5 Extraction

The decoding is performed using similar calculations as encoding technique. The algorithm of the same is given below.

Algorithm:

Input: Modified song signal with embedded authenticating code in higher frequency range.

Output: Original song signal.

Method: The details of extraction of original song signal are given below.

Step 1: Apply FFT over x to get magnitude values in frequency domain of song signal, says $Y(n)$, n is the total range of frequencies of song signal.

Step 2: Find the 1st C matrix [as described in DHTL – ELF section] from higher frequencies region and using this matrix element find original sequence of magnitude values in lower frequency region. Then remove all the secret codes from higher frequencies region (above 20,000 Hz).

Step 3: Apply inverse FFT to get back the sampled values of original song signal.

3 Experimental Results

Encoding and decoding technique have been applied over 10 seconds recorded song, the song is represented by complete procedure along with results in each intermediate step has been outlined in subsections 3.1.1 to 3.1.4. The results are discussed in two sections out of which 3.1 deals with result associated with DHTL and that of 3.2 gives a comparison with existing techniques.

3.1 Results

For experimental observation, strip of 10 seconds classical song ('100 Miles From Memphis', sang by Sheryl Crow) has been taken. The sampled value of the song is given in table 1 as a two dimensional matrix. Figure 1 shows amplitude-time graph of the original signal. DHTL is applied on this signal and as a first step of the procedure which is performed DHTL over input song signal. The output generated in the process is shown in figure 2. Figure 3 shows the difference of frequency ratio of original and modified song after embedding secret code. From figure 3, it is seen that the deviation is very less and there will not affect the quality of the song at all.

3.1.1 Original Recorded Song Signal (10 Seconds)

The values for sampled data array $x(n,2)$ from the original song is given in table 1. Whereas the graphical representation of the original song, considering all rows (441000) of $x(n,2)$ is given in the figure 1.

Table 1. Sampled data array $x(n,2)$.

Sl no	$x(k,1)$	$x(k,2)$
...
0	0.0001	
0.0000	0.0000	
-0.0009	-0.0009	
-0.0006	-0.0007	
-0.0012	-0.0012	
-0.0014	-0.0014	
-0.0016	-0.0017	
-0.0023	-0.0022	
-0.0027	-0.0027	
-0.0022	-0.0021	
...

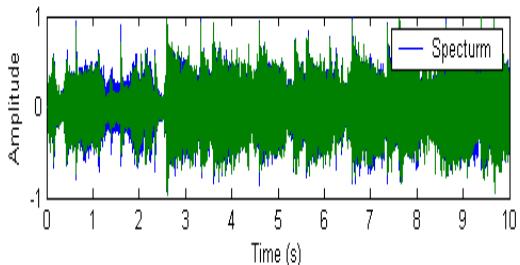


Fig. 1. Original song ('100 Miles From Memphis', sang by Sheryl Crow)

3.1.2 Modified Song after Encoding Lower Frequencies and Adding Secure Code (10 Seconds)

The graphical representation of the modified song signal is shown in the figure 2.

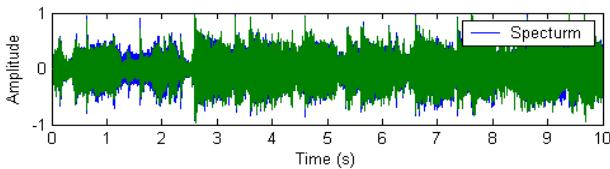


Fig. 2. Modified song with secure code

3.1.3 The Difference of Magnitude Values between Original and Modified Signals

The graphical representation of the magnitude differences of original and modified songs is shown in the figure 3.

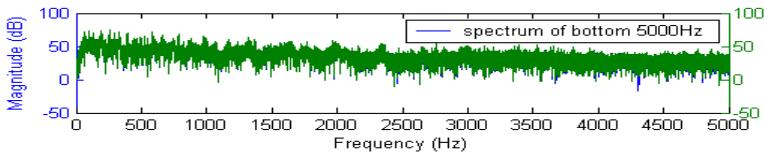


Fig. 3. The difference of magnitude values between signals shown in figure 1 and 2

3.1.4 Estimating Limit of Embedded Code

Applying equation (1) to above song [16-bit stereo type sampled at 44.1 kHz] , hidden extra data (noise) that added for authenticating original song is very less than the maximum noise (0.0128), will not affect over all song audible quality. Because, only about 900 positions have been altered from 441000 sampled values of taken song signal.

3.2 Comparison with Existing Systems

Various algorithms [6] are available for embedding information with audio signals. They usually do not care about the quality of audio but we are enforcing our authentication technique without changing the quality of song. A comparison study of properties of our proposed method with Data hiding via phase manipulation of audio signals(DHPMA)[4] before and after embedding secret message/modifying parts of signal (16-bit stereo audio signals sampled at 44.1 kHz.) is given in table 2, table3 and table4. Average absolute difference (AD) is used as the dissimilarity measurement between original song and modified song to justify the modified song. Whereas a lower value of AD signifies lesser error in the modified song. Normalized average absolute difference (NAD) is quantization error is to measure normalized distance to a range between 0 and 1. Mean square error (MSE) is the cumulative squared error between the embedded song and the original song. A lower value of MSE signifies lesser error in the embedded song. The SNR is used to measure how much a signal has been tainted by noise. It represents embedding errors between original song and modified song and calculated as the ratio of signal power (original song) to the noise power corrupting the signal. A ratio higher than 1:1 indicates more signal than noise. The PSNR is often used to assess the quality measurement between

the original and a modified song. The higher the PSNR represents the better the quality of the modified song. Thus from our experimental results of benchmarking parameters (NAD, MSE, NMSE, SNR and PSNR) in proposed method obtain better performances without affecting the audio quality of song.

Table 3 gives the experimental results in terms of SNR (Signal to Noise Ratio) and PSNR(Peak signal to Noise Ratio). Table 4 represents comparative values of Normalized Cross-Correlation (NC) and Correlation Quality (QC) of proposed algorithm with DHPMA. The Table 5 shows PSNR, SNR, BER (Bit Error Rate) and MOS (Mean opinion score) values for the proposed algorithm. Here all the BER values are 0. The figure 4 summarizes the results of this experimental test. It shows this algorithm's performance is stable for different types of audio signals.

Table 2. Metric for different distortions

Sl No	Statistical parameters for differential distortion	Value using DHTL	Value using DHPMA
1	MD	0.0132	3.6621e-004
2	AD	0.0017	2.0886e-005
3	NAD	0.0063	0.0063
4	MSE	5.7434e-006	1.4671e-009
5	NMSE	2.2066e+004	8.4137e-005

Table 3. SNR and PSNR

Sl No	Statistical parameters for Differential distortion	Value using DHTL	Value using DHPMA
1	Signal to Noise Ratio (SNR)	39.0012	40.7501
2	Peak Signal to Noise Ratio (PSNR)	52.4080	45.4226

Table 4. Representing NC and QC

S I N o	Statistical parameters for Corelation distortion	Value using DHTL	Value using DHPMA
1	Normalised Cross-Correlation(NC)	1	1
2	Correlation Quality (QC)	-0.0128	-0.5038

Table 5 Showing SNR, PSNR BER, MOS

Audio (10s)	SNR	PSNR	BER	MOS
Song1	40.9981	70.6543	0	5
Song2	39.0012	52.4080	0	5
Song3	37.0034	55.6543	0	5
Song4	42.2331	65.8650	0	5

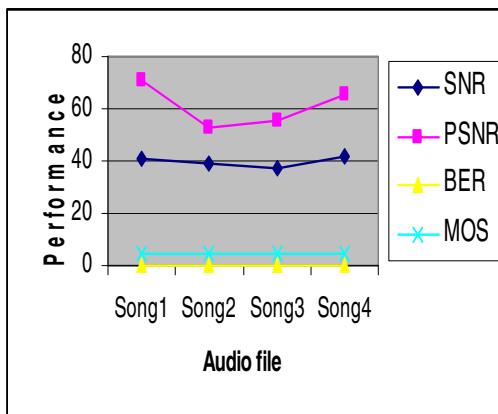
The quality rating (Mean opinion score) is computed by using equation (2).

$$Quality = \frac{5}{1 + N * SNR} \quad (2)$$

Where N is a normalization constant and SNR is the measured signal to noise ratio. The ITU-R Rec. 500 quality rating is perfectly suited for this task, as it gives a quality rating on a scale of 1 to 5 [7]. Table 6 shows the rating scale, along with the quality level being represented.

Table 6. Quality rating scale

Rating	Impairment	Quality
5	Imperceptible	Excellent
4	Perceptible, not annoying	Good
3	Slightly annoying	Fair
2	Annoying	Poor
1	Very annoying	Bad

**Fig 4.** Performance for different audio signals

4 Conclusion and Future Work

In this paper, an algorithm for encoding the lower frequency region with linear coding technique as embedding some secret code in higher frequency region has been proposed which will not affect the song quality but it will ensure to detect the distortion of song signal characteristics. Additionally, the proposed algorithm is also very easy to implement.

This technique is developed based on the observation of characteristics of different songs with human audible characteristics and an approach is also made for estimating the embedded extra data limit with the help of Shannon's limit in the channel encoding scheme. It also can be extended to embed an image into an audio signal instead of text and audio. The perfect estimation of percentage of threshold numbers of sample data of song that can be allow to change for a normal conditions will be done in future with all possibilities of errors in song signal processing.

References

1. Mondal, U.K., Mandal, J.K.: A Practical Approach of Embedding Secret Key to Authenticate Tagore Songs (ESKATS). In: Wireless Information Networks & Business Information System Proceedings (WINBIS 2010), vol. 6(1), pp. 67–74. Rural Nepal Technical Academy (Pvt.) Ltd, Nepal (2010) ISSN 2091-0266
2. Mondal, U.K., Mandal, J.K.: A Novel Technique to Protect Piracy of Quality Songs through Amplitude Manipulation (PPAM). In: International Symposium on Electronic System Design (ISED 2010), pp. 246–250 (2010) ISBN 978-0-7695-4294-2
3. Mondal, U.K., Mandal, J.K.: A Fourier Transform Based Authentication of Audio Signals through Alternation of Coefficients of Harmonics (FTAT). In: Nagamalai, D. (ed.) PDCTA 2011. CCIS, vol. 203, pp. 76–85. Springer, Heidelberg (2011) ISSN 1865-0929, ISBN 978-3-642-24036-2

4. Xiaoxiao, D., Mark, F., Bocko, Z.I.: Data Hiding Via Phase Manipulation of Audio Signals. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), vol. 5, pp. 377–380 (2004) ISBN 0-7803-8484-9
5. Erten, G., Salam, F.: Voice Output Extraction by Signal Separation. In: ISCAS 1998, vol. 3, pp. 5–8 (1998) ISBN 07803-4455-3
6. Katzenbeisser, S., Petitcolas, F.A.P.: Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, Norwood (2000) ISBN 978-1-58053-035-4
7. Arnold, M.: Audio watermarking: Features, applications and algorithms. In: IEEE International Conference on Multimedia and Expo., New York, NY, vol. 2, pp. 1013–1016 (2000)

A Novel DFT Based Information Embedding for Color Image Authentication (DFTIECIA)

J.K. Mandal and S.K. Ghosal

Department of Computer Science and Engineering, Kalyani University, Kalyani,
West Bengal, India, 741235

{jkm.cse,sudipta.ghosal}@gmail.com}

<http://www.klyuniv.ac.in>, <http://www.jkmandal.com/>

Abstract. In this paper, a novel two-dimensional Discrete Fourier Transformation based information embedding has been proposed for color image authentication (DFTIECIA). The DFT is applied on each 2 x 2 sub-image block of the carrier image. Based on channel and the absolute difference value of four most significant bits of the second frequency component in row major order, at most four bits are embedded in the LSB part of second, third and fourth frequency component based on the varying perceptibility of human eye for three channels viz. red, green and blue. The first component is used as indicator for further re-adjustment, if necessary. Inverse DFT (IDFT) is applied on each 2 x 2 mask after embedding to generate stego image. Experimental results conform that the proposed algorithm performs better than the Discrete Cosine Transform (DCT), Quaternion Fourier Transformation (QFT) and Spatio Chromatic DFT (SCDFT) based schemes.

Keywords: DFTIECIA, QFT, DFT, IDFT, DCT, SCDFT, and LSB.

1 Introduction

Information hiding is an idea in the field of information security which has received significant attention from both industry and academia [1]. Steganography is an important field of research to achieve the information hiding scheme. DFTIECIA has been proposed for color images. The application of steganographic technique can be broadly classified as operating in two different domains, such as spatial domain and frequency domain.

Frequency domain methods are widely used than spatial domain techniques. A number of works have been done for information hiding in this domain. Most common transformations are the Discrete Cosine Transformation (DCT), Quaternion Fourier Transformation (QFT), Discrete Fourier Transformation (DFT) and Discrete Wavelet Transformation (DWT). Here, hidden data are embedded into the frequency component of the image pixel in transform/frequency domain. I. J. Cox et al. [2, 3] developed an algorithm to inserts watermarks into the frequency components and spread over all the pixels. DCT-based image authentication is developed by N. Ahmadi et al. [4] using just noticeable difference profile [5] to determine maximum

amount of watermark signal that can be tolerated at each region in the image without degrading visual quality.

The Discrete Fourier Transformation is used to convert the image from spatial domain to frequency domain. The frequency components values are then chosen for embedding secret data which helps a lot to enhance the robustness as the secret data are embedded in both positive and negative frequency components. After embedding the secret data, Inverse Discrete Fourier Transformation is applied to get back into it in spatial domain.

So, the DFT of spatial $f(x,y)$ for the image of size $M \times N$ is defined in equation number (1) for frequency domain transformation.

$$F(u, v) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) e^{-j2\pi(ux/M+vy/N)} \quad (1)$$

Where, $u=0$ to $M-1$ and $v=0$ to $N-1$.

The variable u and v are the transform or frequency variables and x, y are the spatial or image variables and $f(x,y)$ are intensity values of pixels in spatial domain. Similarly, IDFT is used to convert frequency component to the spatial domain value, and is defined in equation (2).

$$f(x, y) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} F(u, v) e^{j2\pi(ux/M+vy/N)} \quad (2)$$

Where, $u=0$ to $M-1$ and $v=0$ to $N-1$.

The aim of DFTIECIA emphasizes on protection of secret information against unauthorized access in frequency domain to achieve a better tradeoff between robustness and perceptibility. Moreover, this technique provides image authentication process by embedding the secret data along with the message digest MD (which is generated from secret message/image) into the carrier image with a minimum change in visual pattern and improved security against visual attacks.

Problem motivation and formulation of transformation technique is given in section 2. The section 3 of the paper deals with the proposed technique. Results, comparison and analysis are given in section 4. Conclusions are drawn in section 5 and after that references are given.

2 Transformation Techniques

In this technique an image sub block of size 2×2 is taken and DFT is applied. The formulation of a 2×2 mask for four different image bytes in 2D-DFT is as follows:

$$F(a_{0,0}) = \frac{1}{2} \sum_{i=0}^1 \sum_{j=0}^1 a_{i,j} = c_{0,0} \text{ (say)}, \quad F(a_{0,1}) = \frac{1}{2} \sum_{i=0}^1 \sum_{j=0}^1 (-1)^j a_{i,j} = c_{0,1} \text{ (say)},$$

$$F(a_{1,0}) = \frac{1}{2} \sum_{i=0}^1 \sum_{j=0}^1 (-1)^i a_{i,j} = c_{1,0} \text{ (say)}, \quad F(a_{1,1}) = \frac{1}{2} \sum_{i=0}^1 \sum_{j=0}^1 (-1)^i (-1)^j a_{i,j} = c_{1,1} \text{ (say)},$$

Where $c_{0,0}$, $c_{0,1}$, $c_{1,0}$ and $c_{1,1}$ are all frequency components for $a_{0,0}$, $a_{0,1}$, $a_{1,0}$ and $a_{1,1}$ spatial domain values respectively. Embedding is done up to the last four least significant bits of $c_{0,1}$, $c_{1,0}$ and $c_{1,1}$. The first frequency component ($c_{0,0}$), is used as re-adjust phase to balance the quantum values between original and embedded data.

Similarly, by applying the inverse 2D-DFT, the 2×2 transformed masks can be converted back to spatial domain image bytes. After re-adjust phase, all 2D-IDFT values are non-fractional, non-negative and less than the maximum possible value of a byte.

3 Proposed Technique

In this paper a novel DFT based information hiding approach has been proposed for color image authentication (DFTIECIA) in frequency domain based on the two dimensional Discrete Fourier Transform (DFT). Initially, a 128 bit message digests (MD) and size of the hidden data is embedded using the proposed DFTIECIA scheme for authentication purpose. The DFT is applied on 2×2 sub-image block for converting the spatial domain values to frequency components. This process will be continued till the last sub-image block of the carrier/cover image in a row major order. The secret information will be embedded in three frequency components out of four frequency components (except the first frequency component) in each 2×2 sub-image matrix of the carrier image. In the proposed DFTIECIA scheme for each 2×2 sub-image block, the four most significant bits in an eight bit representation of the second frequency components are used as the indicator. Depending upon the channel we have chosen and the absolute difference value of number of one's and zero's (which is suppose, W) in the four most significant bits, data are to be embedded in three frequency components viz. the second, third and fourth for each 2×2 mask. For red channel the resultant number of bits are embedded whereas for green channel a maximum up to two bits are embedded which means the number of bits selection is based on the minimum value between the resultant number of bits and two i.e. $\text{MIN}(2, W)$. Similarly, in case of blue channel, a minimum of two and a maximum up to four bits are embedded i.e. $\text{MAX}(2, W)$. The first frequency component has not been used for embedding purpose because it has to be used for re-adjustment of frequency components whenever it violates the basic principles of pixel representation in spatial domain like non-fractional, non-negative pixel value and a value less than or equal to 255 for eight bit representation. In the proposed technique, if the value of frequency components becomes non fractional, then the LSB of the first frequency component is flipped. If the value becomes negative, then a multiple of absolute net value changes in each block is added whereas if the value becomes greater than 255, then a multiple of absolute net value changes in each block is deducted. Inverse DFT (IDFT) is applied on each 2×2 mask after embedding to transformed embedded image (sometimes, re-adjusted as well) in frequency domain to convert back into spatial domain. In the same manner, we can extract the secret data from the embedded image.

For example, let we have chosen the Baboon image as the cover/carrier image. At first, we have applied 2D-DFT on the first 2×2 sub-image block to convert it from spatial domain pixel value to frequency components value in transform domain where

R_1 , G_1 and B_1 are the sub-matrices for the first 2×2 sub-image block. The steps are as shown below:

$$R_1 = \{164, 63, 120, 135\}, G_1 = \{150, 57, 125, 97\}, B_1 = \{71, 31, 62, 33\}$$

After, applying DFT we get the transformed frequency component values as shown below:

$$F(R_1) = \{241, 43, -14, 58\}, F(G_1) = \{214, 60, -7, 32\}, F(B_1) = \{98, 34, 3, 5\}$$

Now, if we embed a letter A (where as the ASCII value of A is 65 and binary representation is 01000001) using the proposed scheme of size 10 bits (5 bits for width and 5 bits for height), then the transformed frequency components after embedding are as follows:

$$EF(R_1) = \{241, 40, -8, 56\}, EF(G_1) = \{214, 60, -7, 32\}, EF(B_1) = \{98, 37, 0, 1\}$$

Whereas, the bit string of hidden data is 000000000101000001 and the absolute difference value of number of one's and zero's for R, G and B sub-matrices are 3, 0 and 3.

So, the number of bits are embedded in each block for red channel is $(3 * 3) = 9$ bits, for green channel it is $(3 * \min(2, 0)) = 0$ bits and for blue channel it is $(3 * \max(2, 3)) = 9$ bits.

Again, applying IDFT we get back the values in pixel domain in the below forms:

$$F^{-1}(EF(R_1)) = \{164.5, 68.5, 116.5, 132.5\}, \quad F^{-1}(EF(G_1)) = \{150, 57, 125, 97\}, \quad F^{-1}(EF(B_1)) = \{68, 30, 67, 31\}$$

As we can see that the pixel values are fractional for red channel, we apply the re-adjust phase to make it non fractional just by flipping the LSB of first frequency component. As a result the embedded block is to be in the form shown below:

$$\text{Re-Adjust}(F^{-1}(EF(R_1))) = \{165, 68, 117, 133\}, \text{Re-Adjust}(F^{-1}(EF(G_1))) = \{150, 57, 125, 97\}, \\ \text{Re-Adjust}(F^{-1}(EF(B_1))) = \{68, 30, 67, 31\}.$$

Actually, the process of embedding is continued for each 2×2 mask till the last hidden bit.

We have categorized this section into three parts namely the Algorithm for Insertion, Re-Adjustment Phase and the Algorithm for Extraction.

3.1 Insertion

All insertion is made in frequency domain i.e. each byte of source image in each mask of size 2×2 is transformed to frequency domain using two dimensional DFT equations. The proposed scheme uses color image as the input to be authenticated by text message/image. All the three channels in a 24 bit color image have been chosen for embedding purpose. As we know, the perceptibility of green channel is comparatively high than red and blue channel, for that our intention is to embed less bit in green channel. Unlike, the perceptibility of blue channel is less as compared

to green and red channel, for that our tendency is to embed more bits in blue channel. Since, the perceptibility of red channel is higher than blue channel and less than green channel, we embed bits in red channel as per the proposed approach where we have no intention to embed much/less bit or more specifically, to incorporate a special condition to restrict the normal embedding process. The authenticating message/image bits size is $\beta * \mu * (m \times n) - (MD + L)$ where β is the average bit per byte, μ is the bit depth of the cover image which is 3 for color image, MD and L are the message digest and dimension of the authenticating image respectively for the source image size $m \times n$ bytes. The value of $MD + L$ is always to be less than $\beta * \mu * (m \times n)$.

Steps:

1. Obtain 128 bits message digest MD from the authenticating message/image.
2. Obtain the size of the authenticating message/image (($m + n$) bits, where m bits for width and n bits for height).
3. Read authenticating message/image data do:

- Read the source image matrix of size 2×2 mask from image matrix (I) in row major order and apply two dimensional DFT.
- Consider the second frequency component value to make the indicator of each 2×2 mask from transformed image matrix (I') in row major order.
- Convert the real value at position I' (0, 1) i.e., the second frequency component and take the binary form of the integer part.
- Calculate the number of 1's and 0's from the four most significant bits in that frequency component.
- Evaluate an absolute difference value of number of 1's and 0's and we can denote it by W.
- The numbers of bits are embedded in 2_{nd}, 3_{rd} and 4_{th} frequency component of each 2×2 mask depending upon the channels (RGB). If the channel is red it embeds W number of bits in each of the components except the first from a 2×2 transformed image matrix (I'), i.e., W number of bits is embedded in second, third and fourth component for each. But, if the channel is green, number of bits embedded in second, third and fourth frequency components are MIN(2,W). Similarly, if the channel is blue, bit embedded in second, third and fourth components are MAX(2,W).

[Embed authenticating message/image bit as per the above rules.]

4. Apply inverse DFT using identical mask.
5. Apply re-adjust phase if needed.
6. Repeat step 3 to step 5 for the whole authenticating message/image size, content and for message digest MD.
7. Stop.

3.2 Re-adjustment

In the proposed algorithm after embedding we have used inverse discrete Fourier transformation (IDFT) to get the embedded image in spatial domain. Applying IDFT on identical mask with embedded data of the frequency component value which may change and can generate the following situation:

- The converted value may be negative (-ve).
- The converted value in spatial domain may be a number with non zero fractional value i.e. pure non integer number.
- The converted value may be greater than the maximum value (i.e. 255).

The concept of re-adjust phase is to handle the above three serious problems by using the first frequency component of each mask. In this phase if the converted value is negative (-ve) i.e. for case A, the operation applied for each 2 x 2 mask is as follows:

$$\begin{aligned} F0 &= F0 + K * (AF2 + AF3 + AF4) \\ &= F0 + K * T \end{aligned} \quad (i)$$

Here, K is the multiple of T, takes values in the range, K=1, 2, 3, ..., n. That means K will be multiplied and incremented in each step till all the converted value in spatial domain value becomes positive. The AF2 specify the net absolute value of changes in second frequency component that means the net absolute difference value of the second frequency component before and after embedding secret data. Similarly, AF3 and AF4 specify the net absolute difference value of changes made after embedding in third and fourth frequency components. The letter T is the net value changes in each block which is the summation of absolute changes of second, third and fourth frequency components.

For case B, if after the IDFT operation the converted value (spatial domain value) becomes fractional then we will change the LSB of first frequency component value by the complement bit.

In this phase if the converted value exceeds the maximum value of a byte (i.e., 255) then for case C, the operation applied for each 2 x 2 mask is as follows:

$$\begin{aligned} F0 &= F0 - K * (AF2 + AF3 + AF4) \\ &= F0 - K * T \end{aligned} \quad (ii)$$

Here, K, T, AF2, AF3 and AF4 specifies the usual meaning used in eq. no. (i).

3.3 Extraction

The authenticated image is received in spatial domain. During extraction the embedded image has been taken as the input and the authenticating message/image size, image content and message digest MD are extracted from it. All extraction is done in frequency domain from frequency components.

Steps:

1. Read image matrix of size 2×2 mask from stego-image matrix (S) in row major order and apply DFT.
2. For each mask do the following operations:
 - Consider the second frequency component value as the indicator of each 2×2 mask from transformed stego-image matrix (S') in row major order.
 - Convert the real value at position S' (0, 1) i.e., the second frequency component and take the integer part.

- Calculate the number of 1's and 0's from the four most significant bits (MSB) from integer part of that frequency component.
 - Evaluate an absolute difference value of number of 1's and 0's and which is denoted by the alphabet W.
 - The number of bits is extracted from 2_{nd}, 3_{rd} and 4_{th} frequency component of each 2 x 2 mask depends upon the channel we have chosen. If the channel is red it extracts W number of bits from each of the components except the first from a 2 x 2 transformed image matrix (I'), i.e., W number of bits is extracted from second, third and fourth component for each. But, if the channel is green, number of bits extracted from second, third and fourth frequency components is MIN(2,W). Similarly, if the channel is blue, bit extracted from second, third and fourth components is MAX(2,W).
- [Extract authenticating message/image bit as per the above rules.]
- For each 8 (eight) bits extraction construct one alphabet/one primary (R/G/B) color image.
3. Repeat step 1 and step 2 to complete decoding as per size of the authenticating message/image.
4. Obtain 128 bits message digest MD' from the extracted authenticating message/image. Compare MD' with extracted MD. If both are same, then image is authorized, else, unauthorized.
5. Apply inverse DFT using identical mask.
6. Stop

4 Results, Comparison and Analysis

This section represents the results, discussion and a comparative study of the proposed technique DFTIECIA with the DCT based watermarking method, QFT based and Spatio Chromatic DFT based watermarking methods in terms of visual interpretation, peak signal to noise ratio (PSNR) analysis and histogram analysis. Benchmark (PPM) images [6] are taken to formulate results and are shown in Fig. 1. All cover images are 512 x 512 in dimension whereas the gold coin (i.e. the secret data) is embedded into the various source benchmark images. The experiment deals with ten different color images (i-x), where each pixel is represented by three intensity values RGB (Red, Green and Blue). Images are labeled as: (i) Lena, (ii) Baboon, (iii) Pepper, (iv) Airplane, (v) Splash, (vi) Earth, (vii) Sailboat, (viii) Foster City, (ix) San Diego, (x) Oakland. On embedding the the Gold-Coin image of varying sizes, the stego-image produces a good visual clarity.

Based on the proposed technique and experimental results, we can see that the technique can overcome many kinds of visual and statistical attacks and it is quite difficult for the observer to detect the difference between the original and embedded image. From Table 1, we can identify the payload for carrier images 'Lena', 'Baboon' and 'Pepper' are 235806, 212586 and 238614 bytes where the dimension of each original image is 512 x 512. After, embedding a large amount of hidden data the stego-images are also pertain a good visual clarity and produces value of 32 dB for peak to signal noise ratio in average case. Moreover, the histogram analysis shows the



Fig. 1. Cover images (512 x 512) and that of secret image (265 x 265 to 283 x 283).

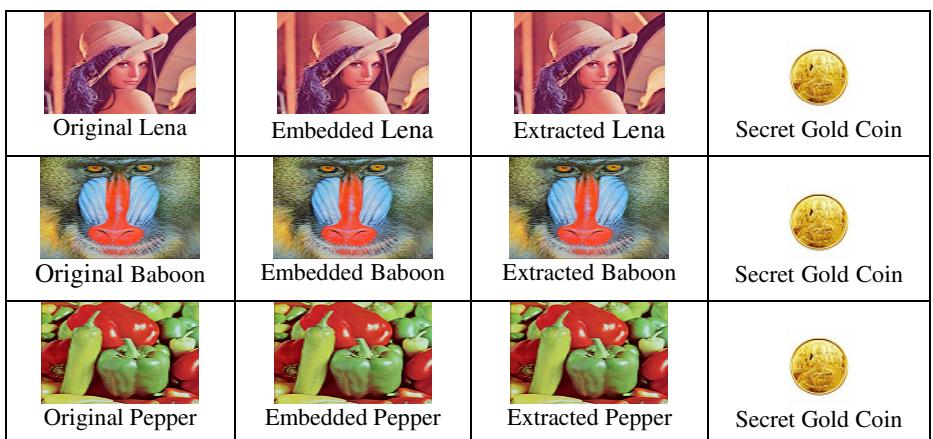


Fig. 2. Cover, Embedded, Extracted and Secret Images using proposed DFTIECIA scheme

changes made in the ‘Lena’ image are more stable after embedding hidden bits. The table also shows that the average number of bits embedded per byte (bpb) for ten images is 2.33. Fig. 2 shows different states of modifications (before and after) of three different images viz. Lena, Baboon and Peppers.

Also, a comparative study has been made among Discrete Cosine Transform (DCT), Quaternion Fourier Transformation (QFT) and Spatio Chromatic DFT (SCDFT) based scheme and our proposed DFTIECIA scheme based on the payload and the PSNR values. In the proposed scheme, the payload is much more as compared to the SCDFT, QFT and SCDFT technique, whereas the PSNR values is also perceptible in quality and produces good visual clarity stego-images.

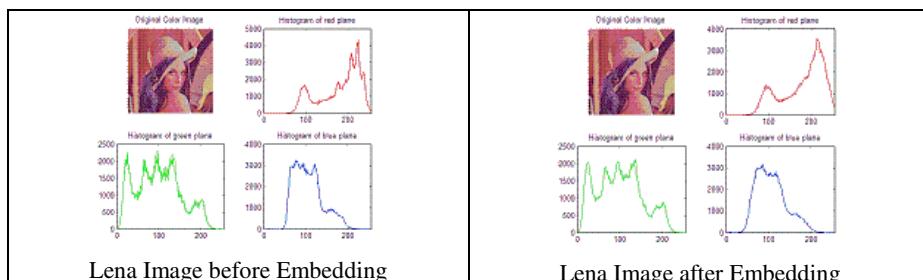
In Fig. 3, the histogram analysis of Lena image is shown before and after embedding secret information in a individual channel wise manner.

Table 1. Results of embedding a minimum of 210726 and a maximum of 240502 bytes of information in each Image of dimension 512 x 512

Carrier Image	Payload (byte)	PSNR (dB)	Bits Per Byte (bpb)
Lena	235806	31.96	2.39
Baboon	212586	34.08	2.16
Pepper	238614	31.89	2.42
Earth	238614	32.64	2.42
Sailboat	230774	33.12	2.34
Airplane	237174	31.86	2.41
Foster City	236034	32.56	2.40
Oakland	221730	33.18	2.25
San Diego	210726	34.58	2.14
Splash	240502	31.36	2.44
AVG	230256	32.72	2.33

Table 2. Comparison results of Payload Capacities and PSNR for Lena image in the existing technique namely SCDFT, QFT and DCT

Technique	Payload(bytes)	PSNR(dB)
SCDFT	3840	30.1024
QFT	3840	30.9283
DCT	3840	30.4046
DFTIECIA	235806	31.96

**Fig. 3.** Comparisons Results of Histograms between of Original and Stego Image of Lena

5 Conclusion

The DFTIECIA scheme is an image authentication process in frequency domain to enhance the security compared to the existing algorithm. Authentication is done by embedding secret data in a carrier image. Using the technique a maximum up to 30 bits can be embedded in 2 x 2 image block. Experimental results conform that the proposed algorithm performs better than the Discrete Cosine Transform (DCT),

Quaternion Fourier Transformation (QFT) and Spatio Chromatic DFT (SCDFT) based scheme and robustness is achieved by hiding data in both positive and negative frequency components.

References

1. Zhang, T., Ping, X.: A Fast and Effective Steganalytic Technique against JSteglike Algorithms. In: Proc. 8th ACM Symp. Applied Computing. ACM Press, New York (2003)
2. Cox, J., Kilian, J., Leighton, F.T., Shamoon, T.: Secure spread spectrum watermarking for images, audio and video. In: Proc. IEEE Int. Conf. Image Processing, Lausanne, Switzerland, September 16-19, vol. 111, pp. 243–246 (1996)
3. Cox, J., Kilian, J., Leighton, F.T., Shamoon, T.: Secure spread spectrum watermarking for multimedia. IEEE Trans. Image Processing 6(12), 1673–1687 (1997)
4. Ahmadi, N., Safabkhsh, R.: A novel DCT-based approach for secure color image watermarking. In: Proc. Int. Conf. Information Technology: Coding and Computing, vol. 2, pp. 709–713 (April 2004)
5. Chou, H., Li, Y.C.: A perceptually tuned subband image coder based on the measure of just-noticeable distortion profile. IEEE Trans. Circuits Syst. Video Technology 5(6), 467–476 (1995)
6. Weber, A.G.: The USC-SIPI Image Database: Version 5, Original release: October 1997, Signal and Image Processing Institute, University of Southern California, Department of Electrical Engineering (1997), <http://sipi.usc.edu/database/> (accessed on January 25, 2010)

A Real Time Detection of Distributed Denial-of-Service Attacks Using Cumulative Sum Algorithm and Adaptive Neuro-Fuzzy Inference System*

R. Anitha, R. Karthik, V. Pravin, and K. Thirugnanam

Department of Mathematics and Computer Applications,
PSG College of Technology, Coimbatore-641004, India

anitha_nadarajan@yahoo.com,
{karthy1988, pravinvenugopal, thirugnanam.tcs}@gmail.com

Abstract. Distributed denial-of-service (DDoS) is a very powerful attack on Internet resources as well as system resources. Hence, it is imperative to detect these attacks in real time else the impact will be irresistible. In this work we propose a new method of applying cumulative sum (CUSUM) algorithm to track variations of the attack characteristic variable $X(n)$ from the observed traffic (specific to different kinds of attacks) and raise an alarm based on threshold. But often a threshold based mechanism produces many false alarms. Adaptive Neuro Fuzzy Inference System (ANFIS) which is capable of removing the abrupt separation between normality and abnormality as well as appropriately select the membership function parameters has been used for detection of attacks based on CUSUM values. The detection mechanism is well corroborated by experimental results.

Keywords: CUSUM, ANFIS, Distributed denial-of-service.

1 Introduction

When networks grow in size and topologies become diverse, shielding them from attacks such as DoS/DDoS becomes intricate. In a typical DDoS setting, an attacker intends to down a service offered by a victim. A number of compromised computers are marshaled into an attack force. DDoS attacks generally aim the resources of a victim system or the bandwidth of the network. In all cases, the DDoS attack prevents legitimate clients from accessing system resources. A complete taxonomy of DDoS attacks and attack tools can be found in [1]. David Moore et.al in [2] provides quantitative estimates of internet wide DoS activity.

In general, the defense mechanisms for DDoS attacks can be classified into three different categories: detection, defense and prevention [3]. Detecting DDoS attacks is the first step to mitigate the DDoS attacks. Upon detection of the attacks, more sophisticated security mechanisms could be prompted to shield the victim server or system which is to be protected. Detection mechanisms should be constantly

* This work is a part of the CDBR-Smart and Secure Environment project sponsored by NTRD, New Delhi, India.

monitoring the target system with very less overhead. The above obligation for the detection mechanism makes the CUSUM algorithm more appropriate for the job. Many researchers have already used CUSUM in the detection of DDoS attacks. In this paper CUSUM algorithm is used to track variations of the attack characteristic variable { $X(n)$ } from the observed traffic and raises an alarm based on a threshold value. But often this fixed threshold mechanism produces many false alarms. Adaptive Neuro Fuzzy Inference Systems (ANFIS) which is capable of removing abrupt separation between normality and abnormality and setup appropriate membership function parameters based on learning is used. The rest of the paper is structured as follows: Section 2 addresses some of the related existing work. Section 3 describes the proposed model using cumulative sum and adaptive neuro-fuzzy inference system (CANFIS). Section 4 gives experimental results as well as a comparison of the proposed system with Snort, a well known intrusion detection system. Finally, Section 5 concludes the paper.

2 Related Work

DDoS defense mechanisms can be broadly categorized based on the attack detection strategy into pattern detection and anomaly detection. Further in anomaly detection there are two Normal Behavior Specifications: Standard and Trained [4]. In pattern detection mechanism, the signatures of known attacks are stored in a database and each communication is monitored for the presence of these patterns. The drawback here is that only known attacks can be detected, whereas attacks with slight variations of old attacks go unobserved. Jelenia Merkovic [6] proposed a technique called DWARD (DDoS Network Attack Recognition and Defense). It is a source end solution whose goal is to autonomously detect and stop outgoing attacks from the deploying network. It provides a dynamic response that is self adjusting. Ratul et. al. [7] proposed the aggregate congestion control (ACC) system, a mechanism which reacts as soon as attacks are detected, but does not give a mechanism to detect ongoing attacks. For both traffic monitoring and attack detection, it may suffice to focus on large flows. Multops [9] proposes a heuristic and a multilevel tree data structure that network devices maintain, storing data corresponding to subnet prefixes. The attack is detected by abnormal packet ratio values and offending flows are rate-limited. Methods presented in [6-9, 10-13] provide examples of anomaly detection approaches. The advantage of anomaly detection over pattern detection is that previously unknown attacks can also be discovered in anomaly detection.

Based on the specification of a normal behavior, we can segregate anomaly detection mechanisms into *standard* and *trained*. Mechanisms that employ standard specifications of normal behavior depend on some protocol standard or a set of rules. Some protocol stack enhancement approaches like SYN cookies and SYN cache are proposed in [14,15] respectively to mitigate SYN flood attacks. The advantage of a standard based specification is that it produces no false positives; all legitimate traffic must meet the terms of the specified behavior. The disadvantage is that attackers can still perform complicated attacks which, on the surface, seem amenable to the standard and hence unnoticed. Mechanisms that use trained specifications of normal behavior observe network traffic and system behavior and produce threshold values

for different parameters. One such widely used threshold based approach is CUSUM algorithm. Haining Wang et al. [16] proposed a detection mechanism very specifically for SYN flood attack. Tao Peng et al. [17] have proposed detection mechanism by monitoring Source IP addresses, Zaihong Zhou et al [18] proposed a detection based on CUSUM and space similarity at each host in a P2P network. Fang-Yie Leu et al. [19] proposed an intrusion prevention system named Cumulative Sum based Intrusion Prevention System (CSIPS). Though all the above mentioned work has its own novel ideas, but one common drawback which all the above work suffered from is the selection of threshold which is important since selection of low threshold leads to a lot of false positives, whereas a high threshold reduces the sensitivity of the detection. For this reason, in our proposed approach a more intelligent ANFIS engine is used.

3 Proposed Model

The proposed model combines CUSUM technique and ANFIS. The CUSUM for each attack is found and passed on to their respective ANFIS. As shown in the Fig 1 which is the block diagram of the proposed scheme, the real time traffic data is collected, the characteristics for modeling the attack using CUSUM are extracted (pre-processing) and then passed on to the corresponding ANFIS engine.

3.1 Attack Characteristic Variable Modeling Using CUSUM

The characteristics of different types of attacks differ significantly. Considering the fact that, it may not be possible to model the traffic in dynamic and complex systems like the Internet using simple parametric description, we chose the non-parametric version of the CUSUM algorithm [20]. CUSUM algorithm dynamically checks if the observed time series is statistically homogenous and, if not, it finds the point at which the change happens and responds accordingly [3, 20, 21]. Consider $\{X(n)\}$, the value of the attack characteristic variable during the n^{th} time interval and its corresponding weights $\{W(n)\}$, a derived value from $X(n)$. We have modeled the attack characteristic variable $X(n)$ for five different attacks where $X_{\text{SYN}}(n)$, $X_{\text{LAND}}(n)$, $X_{\text{SMURF}}(n)$, $X_{\text{UDP}}(n)$, $X_{\text{ICMP}}(n)$ represents $X(n)$ for SYN Flood, Land, Smurf, UDP Flood, and ICMP Flood attacks respectively. These variables are premeditated in such a way that it shows a sudden increase in value only during an attack and thus making CUSUM algorithm to detect sudden variations in traffic more accurately. Some of the existing DDoS detection techniques using CUSUM keep record of only the packet counts observed in a sample interval [19]. But in the proposed technique, CUSUM modeling is done in such a way that it exactly reflects the attack behavior. The variable $\{X(n)\}$ for a specific attack category keeps track of a unique quantity corresponding to the same. Its value will be large when there is a high attack. SYN flood attack exploits the vulnerability present in TCP/IP protocol. The attack causes buffer overflow because the attacking systems send packets requesting new connections at a rate faster than the victim system's speed of discarding the pending connections. So, the attack characteristic variable $X_{\text{SYN}}(n)$ is modeled as

$$X_{\text{SYN}}(n) = (N_{\text{RST}} + N_{\text{SYN}}) - N_{\text{SYN/ACK}} \quad (1)$$

where N_{RST} represents the number of incoming Reset packets, N_{SYN} refers to the number of incoming SYN packets and $N_{SYN/ACK}$ corresponds to the number of outgoing SYN/ACK packets.

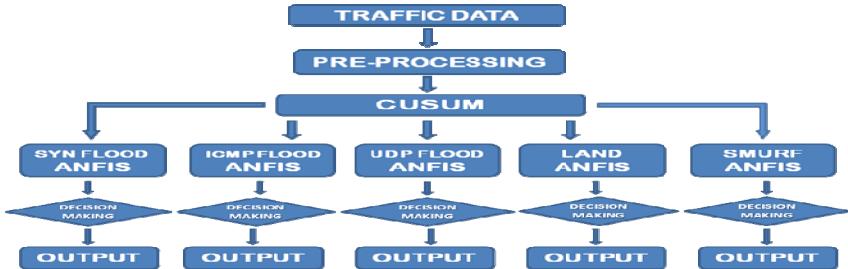


Fig. 1. Proposed model

The quantity $X_{SYN}(n)$ will show a sudden increase during an attack since if the IP address is spoofed, too many RST packets will be coming from the nodes which actually correspond to the spoofed IP. On the other hand if the IP address is not spoofed there will be no RST packets. In either way the proportion of the incoming SYN packets will be very much high than that of the outgoing SYN-ACK packets. In the last scenario where all the connections are legitimate, there will be a very few RST packets which is due to some error or loss of packets during the handshake. So the proportion of the incoming SYN packets will be almost equal to outgoing SYN-ACK packets. Thus, it makes the overall expression close to zero. Land is a special case of SYN flood attack where the strength of the attack is high on the victim machine since Land attack packets have the victim's IP address for both the source IP address and destination IP address, and thus making the victim to send SYN/ACK packets to it. Consequently whenever packets with this characteristic arrive in large number, sudden variation in traffic can be seen. Thus, the value of attack characteristic variable for this attack would be,

$$X_{LAND}(n) = N_{[(SRC_IP=DEST_IP) \& (SYNset)]} \quad (2)$$

where $N_{[(SRC_IP=DEST_IP) \& (SYNset)]}$ represents the number of incoming packets having same source IP address and destination IP address with its SYN FLAG set. Smurf attack exploits the vulnerability in the ICMP protocol. When an attacker sends too many ICMP packets to a network broadcast address with the victim's IP in source address of the packet header, all the machines in the network starts sending ICMP reply packet to the victim's machine which collapses the system. The traffic variation for this attack is modeled by $X_{SMURF}(n)$ as below:

$$X_{SMURF}(n) = N_{(DEST_ADDR=BADDR)} \quad (3)$$

where $N_{(DEST_ADDR=BADDR)}$ denotes the number of ICMP requests made to the broadcast address. In UDP flooding attack, the attacker floods too many UDP packets

to random ports of a victim machine. In this scenario there are two possibilities, either the ports to which packets are flooded are closed or opened. If it is closed an “ICMP Destination Port Unreachable Error” message is sent to the machine whose IP is spoofed by the attacker and if no service is running on that port, the packets will be dropped else the incoming packets get the service reply. The difference between the number of incoming UDP packets and number of outgoing UDP packets will always be a small quantity in legitimate cases whereas in attack scenario the number of incoming UDP packets will be very large and the number of outgoing UDP packets will be comparatively small. Considering all possibilities $X_{UDP}(n)$ is defined as

$$X_{UDP}(n) = (N_{(DEST_ADDR=HOST_IP)} - N_{(SRC_ADDR=HOST_IP)}) + N_{ICMP_error} \quad (4)$$

where $N_{(DEST_ADDR=HOST_IP)}$ denotes the number of incoming UDP packets, $N_{(SRC_ADDR=HOST_IP)}$ denotes the number of outgoing UDP Packets and N_{ICMP_error} denotes the number of ICMP Destination Port Unreachable Error packets. In the modeling of attack characteristic variable $X_{ICMP}(n)$, we are not taking into the consideration the number of ICMP packets since very few number of large ICMP packets are enough for observing an abrupt change in $X_{ICMP}(n)$ than the large number of small sized ICMP packets. Therefore,

$$X_{ICMP}(n) = S_{ICMP_Req} \quad (5)$$

where, S_{ICMP_Req} denotes the total payload size of the incoming ICMP request packets. The above characteristic variable keeps record of the summation of the sizes of incoming ICMP request packets since the time of observation. After modeling the attack characteristic variables, the values that are collected from the real time traffic are passed on to its corresponding trained ANFIS engine which is explained elaborately in the next section.

3.2 ANFIS Engines

Neuro fuzzy techniques have been developed by the blend of Artificial Neural Network (ANN) and Fuzzy Inference System (FIS) which is termed as ANFIS. Using ANFIS with CUSUM algorithm provides twofold advantage. First it helps in removing the crisp threshold based alarm raising mechanism of CUSUM by a more comprehensible fuzzy logic based mechanism and secondly, it fine tunes the membership function parameters involved in FIS using the neural network based learning technique. An adaptive network is a 5 layer feed-forward network in which each node performs a particular function (node function) on incoming signals as well as it has a set of fuzzy membership parameters pertaining to this node [22]. Fuzzy rule base of Mamdani type is modeled for each type of attack based on the corresponding values of $X(n)$.

Each node i in the layer 1 is a square node with node function $O_i^1 = \mu_{A_i}(X(n))$ where A_i is the linguistic label (LOW, MEDIUM, HIGH) associated with this node function. In other words O_i^1 is the membership function of A_i and it

specifies the degree to which the given $X(n)$ satisfies the quantifier A_i . In our case all the membership functions are Gaussian,

$$\mu_{A_i}(X(n)) = \exp\left(-\frac{(c_i - X(n))^2}{2\sigma_i^2}\right) \quad (6)$$

where $\{c_i, \sigma_i\}$ is the parameter set. Every node in the layer 2 is a circle node labeled II which multiplies the incoming signal and sends the product out but since in our model there is only signal coming, we get, $w_i = \mu_{A_i}(X(n)), i=1$. Each node output represents the firing strength of a rule. In layer 3, every node is a circle node labeled N. The i^{th} node calculates the ratio of the i^{th} rule's firing strength to the sum of all rule's firing strength:

$$\overline{w_i} = \frac{w_i}{\sum w_i} \quad (7)$$

Every node i in the layer 2 is a square node with a node function

$$O_i^4 = \overline{w_i} f_i = \overline{w_i} (p_i X(n) + q_i) \quad (8)$$

where $\overline{w_i}$ is the output of layer 3, and $\{p_i, q_i\}$ is the parameter set known as consequent parameters. The single node in the layer 5 is a circle node labeled Σ that computes the overall output as the summation of all incoming signals, i.e.

$$O_i^5 = \text{overall_output} = \sum_i \overline{w_i} f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \quad (9)$$

All the membership functions used are Gaussian functions whose parameters are trained using a network with a hybrid learning algorithm, more specifically in the forward pass of the hybrid learning algorithm, functional signal goes forward till layer 4 and the consequent parameters are identified by the least square estimate. In the backward pass, the error rates propagate backward and the premise parameters are updated by the gradient descent. In the proposed model, a separate Adaptive network as explained above is formed for each of the five different attacks. The training phase of ANFIS is carried out in offline mode over and over again to come out with a perfect training model and during the deployment of this system the testing is done. Thus the five ANFIS engines won't turn out to be a heavy computation because only the FIS is going to be evaluated during the detection phase and the number of ANFIS engines can be further increased if required.

3.3 Decision Making

After the evaluation of all the five ANFIS engines with the CUSUM measures, the output (defuzzified value) of each ANFIS engine is collected and a decision is made based on those defuzzified values. Depending upon the intensity of the attack, the final decision is given as LOW, MEDIUM or HIGH, which denotes the risk level of

the network being monitored. The intensity of each attack can be known from the output of its corresponding ANFIS engine. The risk level HIGH alarms the administrator to immediately take necessary actions.

4 Experimental Results

All the experiments were carried out in the eight nodes of Smart and Secure Environment network laboratories distributed across different geographic locations and connected through a MPLS VPN cloud with each lab consisting of around 9 workstations and one server. During a period of 10 days, 1000 sample sets S_i , $1 \leq i \leq 1000$ with each set consisting of variable number of entries S_{ij} , collected in a time interval of 5 seconds, $1 \leq i \leq 1000$, $1 \leq j \leq n$, (n is a finite number) for training. From the sample set, the attack characteristic variable value for each attack is calculated for each entry in it. The attack characteristic variable value of each attack calculated from each S_{ij} is then used for training its corresponding ANFIS engine. The training process is done many times with different S_i in order to fine tune the ANFIS engine corresponding to each attack, as mentioned in section 3.2. Once the ANFIS is trained, many trials of testing were done in the real time traffic against the trained model. For testing, 1100 sample sets were collected in which 200 samples sets each of SYN Flood, LAND, Smurf, ICMP flood and UDP flood attack traffic along with 100 samples sets of normal traffic. The detection delay, DP which is the time difference between the attack sample given as an input to the proposed scheme and its detection for each attack sample is observed.

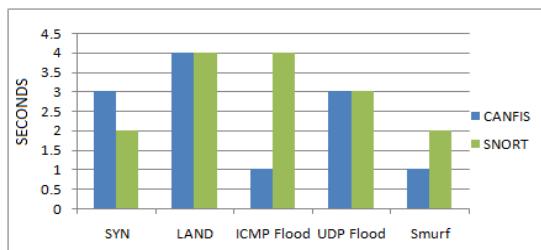


Fig. 2. Detection time in CANFIS and snort

Table 1 compares the maximum detection delay of CANFIS and Snort [5] by testing those using online samples. A sampling interval of 0.5 seconds is fixed for analyzing the samples. Fig 2 gives a pictorial representation of the detection time of CANFIS and Snort for a particular online sample. As we can note in Table 1 and Fig 2, the detection time of the proposed technique is better than that of Snort. Table 2 compares the detection time of CANFIS and Snort by testing them offline. For this purpose, a well known DDoS attack dataset CAIDA is used [23]. The specification of each sample regarding its average packets per second, the attack duration and the type of attacks involved are tabulated. From the tables, it is reasonable to infer that the proposed technique is faster when compared to Snort.

Table 1. Comparison of Detection time of CANFIS and snort when tested online

S.No	Sample Name	Avg. Packets /Sec	Attacks Injected	RESULT	Detection time - CANFIS (sec)	Detection time - SNORT (sec)
1.	Sample_17	1954.641	SYN Flood, ICMP Flood, Smurf	Stopped after detecting ICMP Flood	2.1	3.1
2.	Sample_238	1403.927	ICMP Flood, SYN Flood, Land, UDP Flood	Stopped after detecting UDP Flood	1.8	2.8
3.	Sample_394	1897.249	Land, Smurf	Stopped after detecting Land	2.6	NOT Detected
4.	Sample_431	1540.419	UDP Flood, ICMP Flood, Land	Stopped after detecting UDP Flood	1.8	3.1
5.	Sample_479	1478.397	SYN Flood, UDP Flood, Smurf	Stopped after detecting Smurf	2.1	3.1
6.	Sample_563	1562.240	ICMP Flood, SYN Flood, UDP Flood	Stopped after detecting SYN Flood	2.5	2.8
7.	Sample_704	1456.920	Smurf, Land, ICMP Flood	Stopped after detecting ICMP Flood	2.7	3.4
8.	Sample_859	1783.710	Smurf, Land, SYN Flood	Stopped after detecting SYN Flood	2.3	2.6
9.	Sample_947	1609.198	Smurf, SYN Flood, Land, ICMP Flood, UDP Flood	Stopped after detecting UDP Flood	2.8	3.8

Table 2. Comparison of Detection time of CANFIS and snort - Using CAIDA dataset (offline)

S. No.	Dataset Name	Avg. Packets /Sec	Attack duration (sec)	Attack Detected	Detection time - CANFIS (sec)	Detection time - SNORT (sec)
1.	ddos.20070804_1349_36.pcap	554.836	299.995	ICMP Flood	4.6	5.1
2.	ddotrace.20070804_135436.pcap	362.248	299.999	ICMP Flood	5.1	5.5
3.	ddotrace.20070804_143936.pcap	152619.178	300	ICMP Flood, SYN Flood	4.2	5.6
4.	ddotrace.20070804_142936.pcap	153180.224	300	ICMP Flood, SYN Flood	4.5	5.1
5.	ddotrace.20070804_143436.pcap	147740.123	300	ICMP Flood, SYN Flood, UDP Flood	4.2	4.4
6.	ddotrace.20070804_144436.pcap	165225.885	300	ICMP Flood, SYN Flood	4.8	4.8
7.	ddotrace.20070804_145436.pcap	172565.785	54.56	ICMP Flood, SYN Flood	4.4	4.5
8.	ddos.20070804_1424_36.pcap	164824.043	300	ICMP Flood, SYN Flood, UDP Flood	5.5	5.6

Table 3. Accuracy of CANFIS and snort

Attack		FP	FN	TP	TN	TPR	FPR	ACCURACY
SYN Flood	CANFIS	2	5	98	95	95.14	0.02	96.5
	SNORT	9	20	81	90	80.19	0.09	85.5
LAND	CANFIS	1	3	97	99	97.0	0.01	98.0
	SNORT	10	25	75	90	75.0	0.1	82.5
SMURF	CANFIS	3	6	92	99	93.87	0.02	95.5
	SNORT	8	26	82	84	75.92	0.08	83.0
ICMP Flood	CANFIS	4	1	98	97	98.98	0.04	97.5
	SNORT	11	17	92	80	84.40	0.12	86.0
UDP Flood	CANFIS	2	1	99	98	99.0	0.02	98.5
	SNORT	7	23	80	90	77.66	0.07	85.0

In Table 3, the accuracy of the proposed technique is shown. It shows the values of various evaluation parameters for each attack, namely True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), True Positive Rate (TPR), False Positive Rate (FPR) and the Accuracy. For calculating the various evaluation parameters as listed in Table 3, test sample containing 100 sample set of attack traffic of each type and 100 sample set of normal traffic is taken together and the various evaluation parameters are calculated. For all the attacks, the accuracy of CANFIS is quite better than Snort and the false alarms are also comparatively less when compared to it.

5 Conclusion and Future Work

A sequential analysis technique called CUSUM algorithm has been combined with ANFIS for a computationally light weight and real time detection of DDoS flooding attacks. With an innovative modeling of attack characteristic variables and by replacing the crisp threshold based CUSUM algorithm with a fuzzy logic based method, the false alarms are reduced. Also the proposed architecture is flexible in the sense that we can add more number of attack detection models using CUSUM and create appropriate ANFIS engines for the same. The experimental results show that the proposed technique is faster and has a better accuracy when compared to Snort IDS. This work could be extended further by adding some modules for detecting IP spoofing and IP blacklisting.

References

1. Specht, S., Lee, B.: Distributed denial of service: taxonomies of attacks, tools and countermeasures. In: Proc. of the 17th ICPADS, International Workshop on Security in Parallel and Distributed Systems, pp. 543–550 (September 2004)
2. Moore, D., Voelker, G.M., Savage, S.: Inferring Internet Denial-of-Service Activity. In: Proc. Usenix Security Symp., Usenix Assoc. (2001)
3. Wang, H., Zhang, D., Shin, K.G.: Change-Point Monitoring for the Detection of DoS Attacks. IEEE Transactions on Dependable and Secure Computing 1(4) (October–December 2004)

4. Mirkovic, J., Reiher, P.: Taxonomy of DDoS Attack and DDoS Defense Mechanisms. ACM SIGCOMM Computer Communication Review 34(2) (August 2004)
5. Sourcefire Snort: The Open Source Network Intrusion Detection System
6. Mirkovic, J.: D-WARD: Source-End Defense Against Distributed Denial-of-Service Attacks, PhD thesis, University of California Los Angeles (August 2003)
7. Mahajan, R., Bellovin, S., Floyd, S., Paxson, V., Shenker, S.: Controlling high bandwidth aggregates in the network. ACM Computer Communications Review 32(3) (July 2002)
8. Yan, J., Early, S., Anderson, R.: The XenoService: A Distributed Defeat for Distributed Denial of Service. In: Proceedings of ISW 2000 (October 2000)
9. Gil, T.M.: Poletto. M.: MULTOPS: a data-structure for bandwidth attack detection. In: Proceedings of 10th Usenix Security Symposium (August 2001)
10. Information Sciences Institute, Dynabone, <http://www.isi.edu/dynabone/>
11. Dittrich, D.: The Tribe Flood Network distributed denial of service attack tool, <http://sta@.washington.edu/dittrich/misc/tfn.analysis.txt>
12. Mazu Networks, Mazu Technical White Papers, <http://www.mazunetworks.com/whitepapers/>
13. BBN Technologies, Intrusion tolerance by unpredictability and adaptation, <http://www.bbn.com/infosec/itua.html>
14. Bernstein, D.J., Schenk, E.: Linux Kernel SYN Cookies Firewall Project, <http://www.bronzesoft.org/project/scfw>
15. Lemon, J.: Resisting SYN flood DoS attacks with a SYN cache. In: Proceedings of the BSDCon 2002 Conference, San Francisco, California, USA. USENIX Association (2002)
16. Wang, H., Zhang, D., Shin, K.: Detecting SYN Flooding Attacks. In: Proc. 21st Joint Conf. IEEE Computer and Comm. Societies (IEEE INFOCOM), pp. 1530–1539. IEEE Press (2002)
17. Peng, T., Leckie, C., Ramamohanarao, K.: Proactively Detecting Distributed Denial of Service Attacks Using Source IP Address Monitoring. In: Mitrou, N.M., Kontovasilis, K., Rouskas, G.N., Iliadis, I., Merakos, L. (eds.) NETWORKING 2004. LNCS, vol. 3042, pp. 771–782. Springer, Heidelberg (2004)
18. Zhou, Z., Xie, D., Xiong, W.: A Novel Distributed Detection Scheme against DDoS Attack. Journal of Networks 4(9), 921–928 (2009), doi:10.4304/jnw.4.9.921-928
19. Leu, F., Li, Z.: Detecting DoS and DDoS Attacks by Using an Intrusion Detection and Remote Prevention System. In: International Symposium on Information Assurance and Security, vol. 2, pp. 251–254 (2009); 2009 Fifth International Conference on Information Assurance and Security (2009)
20. Brodsky, B.E., Darkhovsky, B.S.: Nonparametric Methods in Change-Point Problems. Kluwer Academic (1993)
21. Basseville, M., Nikiforov, I.V.: Detection of Abrupt Changes: Theory and Application. Prentice-Hall (1993)
22. Shing, J., Jang, R.: ANFIS: Adaptive-Network-Based Fuzzy Inference System. IEEE Transactions on Systems, Man, and Cybernetics 23(3) (May/June 1993)
23. Hick, P., Aben, E., Polterock, J.: The CAIDA DDoS Attack 2007 Dataset (2007), http://www.caida.org/data/passive/ddos-20070804_dataset.xml

Encryption of Images Based on Genetic Algorithm – A New Approach

Jalesh Kumar¹ and S. Nirmala²

¹ Department of Computer Science and Engineering, J.N.N.C.E., Shimoga-577204,
Karnataka State, India
jalesh_k@yahoo.com

² Professor and Head, Department of Information Science and Engineering, J.N.N.C.E.,
Shimoga-577204, Karnataka State, India
nir_shiv_2002@yahoo.co.in

Abstract. The security of digital images has become increasingly more important in today's highly computerized and interconnected world. In this work a new image encryption technique is proposed which is based on genetic algorithm. The method comprises three stages. The first stage is selection of key sequence. Linear congruential pseudo random generator is used for key sequence generation. The crossover operation is performed in the second stage. In third stage, mutation operation is performed on the result obtained from the previous stage. The proposed method combines both transposition and substitution techniques to secure the data. The results obtained are analyzed through histograms. Performance of the proposed algorithm is measured in terms of correlation coefficient. The analysis carried out reveals that the proposed algorithm works successfully for all types of images.

Keywords: Image encryption, Genetic algorithm, Selection, Crossover, Mutation.

1 Introduction

Transmissions of digital images are needed in many applications, such as medical imaging systems, pay-per-view TV and confidential video conferencing. As digital information transmissions are increased, the problem of protecting data from illegal access has become a challenging task. In order to fulfill such a task, many image encryption methods have been proposed in the past. However, some of them have been known to be insecure [11]. As images form a main source of conveying information, novel encryption techniques are needed for secured transfer of information. Traditional data encryption techniques can be divided into two categories namely substitution and transposition, which are used individually or in combination in every cryptographic algorithm. In substitution technique, one symbol in the data is replaced with another symbol according to some algorithm. In transposition technique, positions of symbols are reordered according to some rule [10, 11]. The genetic algorithm is based on both substitution and transposition operations. It works on the basis of Darwinian evolution theory. Genetic operation is divided into three steps: selection, crossover

and mutation. The crossover process works like transposition and mutation like substitution technique.

In this paper, image encryption scheme based on crossover and mutation is presented. Rest of the paper is organized as follows. In Section 2, an overview of genetic process is described. A related literature survey is carried out in Section 3. In Section 4, the proposed method is explained along with an algorithm. Experimental results are discussed in Section 5. Statistical analysis of the results is presented in Section 6. Time complexity analysis is performed in Section 7. Conclusions are drawn in Section 8.

2 Genetic Process

The genetic algorithm belongs to the family of evolutionary algorithms along with genetic programming, evolution strategies and evolutionary programming. In general, a genetic algorithm is initiated with a randomly generated set of individuals. Once the initial population has been created, the genetic algorithm enters a loop and process iteratively. At the end of each iteration, a new population has been produced by applying a certain number of stochastic operators to the previous population. Each such iteration is known as a generation. Initially a selection operator is applied. This creates an intermediate population of n “parent” individuals. To produce these “parents”, n independent extractions of an individual from the old population are performed. Individuals which are chosen to form new individuals (offspring) are selected according to their fitness. An operator that involves one parent is called a mutation operator. When more than one parent is involved, then the operator is called recombination. The genetic algorithm uses two reproduction operators: crossover and mutation. To apply a crossover operator, parents are paired together. There are several different types of crossover operators. The types available depend on what representation is used for the individuals. For binary string individuals: one-point, two-point and uniform crossover is often used. Order, partially mapped and cycle crossover are used for permutation based individuals. The one-point crossover means that the parent individuals exchange a random prefix while creating the child individuals. Two-point crossover is an exchange of a random substring. In uniform crossover each bit in the child is taken arbitrarily from either parent. Order and partially mapped crossover are similar to two-point crossover. For order crossover, the section between the first and second cut points is copied from the first parent to the child. For partially mapped crossover, the section between the two cut points defines a series of swapping operations to be performed on the second parent. Cycle crossover satisfies two conditions - every position of the child must retain a value found in the corresponding position of a parent and the child must be a valid permutation. In each cycle a random parent is selected. After crossover, each individual has a small chance of mutation. The purpose of the mutation operator is to simulate the effect of transcription errors that can happen with a very low probability when a chromosome is mutated. A standard mutation operator for binary strings is bit inversion. Each bit in an individual has a small chance of mutating into its complement [1].

The evolutionary cycle could be summarized as follows:

```
generation = 0
seed population
while not (stop the operation) do
generation = generation + 1
calculate fitness
selection
crossover
mutation
end while
```

3 Literature Survey

Kumar [1] proposes a new approach based on genetic algorithms, with pseudorandom sequence to encrypt the data stream. The features of such an approach include high data security and feasibility for easy integration with commercial multimedia transmission applications. The concept of genetic algorithms is used along with the randomness properties of chaos. Encryption process using the crossover operator and pseudorandom sequence generator by Non-Linear Feed Forward Shift Register are discussed in [2]. The crossover point is decided by the pseudorandom sequence. Further, the work is extended using the concept of mutation [3]. The data encrypted is further hidden inside the stego-image. Tragha et al. [4, 5] describe a new symmetrical block ciphering system named Improved Cryptography Inspired by Genetic Algorithms (ICIGA) which generates a session key in a random process. The user can fix the size of the blocks as well as key length. The operation of ICIGA depends on the length of the secret key. The key length is used to divide the plaintext into parts of equal size. During the ciphering, the first part is broken up into blocks of uniform size which are used to generate the secret key. In [7], a novel image encryption algorithm is proposed which is called as Bit Recirculation Image Encryption (BRIE). The gray level of each pixel in the image is transformed based on a defined bit recirculation function and a binary sequence generated from a chaotic system. Two encrypted images are simulated and the fractal dimensions of the original and encrypted images are computed to demonstrate the effectiveness of the proposed algorithm. BRIE is not secure enough from strict cryptographic viewpoint [8]. It has been found that defects exist in BRIE and a known/chosen-plaintext attack can break BRIE with only one known/chosen plain-image. Experiments are carried on to verify the defects of BRIE and the feasibility of the attack [8]. Husainy [9] discuss a new Image Encryption technique using Genetic Algorithm based on mutation and crossover. Security of this method depends on the use of different vector lengths and number of crossover and mutation operations. Srikanth et al. [10] propose a new technique where image encryption is done using breaking and merging of bits. The image is first broken down into blocks, then the process similar to vernon cipher is used to locate the pixels and genetic algorithm is used to encrypt the images using one point crossover. In [11], a new effective method for image encryption which employs magnitude and phase manipulation using differential evolution approach is discussed. Linear feedback shift register is used to select the crossover points. In this approach, discrete fourier transform followed by differential evolution are used for image encryption.

From the literature survey it is evident that many security issues could be solved using genetic algorithms through modeling a simplified version of genetic processes. Various image encryption techniques are designed on the basis of genetic approach. The selection of crossover points and mutation operation plays a major role in encrypting the information. In this paper, a random selection of crossover points and mutation based on cipher feedback is proposed to enhance the security of images.

4 The Proposed Method

The approach proposed in this section comprises three stages namely selection, crossover and mutation. The block diagram of the proposed system is shown in Fig. 1. In selection process, crossover point is selected on the basis of key sequence. Linear congruential generator is used to generate pseudorandom key sequence. Starting seeds are chosen, which generates a large period [13]. In the proposed method the system clock is added with the random sequence, so that the sequence is non-reproducible [13]. The sequence of random numbers is generated using the equation (1).

$$X_{i+1} = (a X_i + c) \bmod m \quad (1)$$

Where 'm' is modulus $m > 0$

'a' is multiplier $0 < a < m$

'c' is increment $0 \leq c < m$

and 'X₀' is starting value $0 \leq X_0 < m$

After selecting the key sequence, crossover operation is performed on the input image. In the final stage, mutation operation is applied on the result obtained from second stage. Exclusive-OR operator is used in mutation operation. The cipher generated is fed back to generate a key sequence. After mutation the output image obtained will be completely encrypted image. Decryption is a reverse process of encryption.

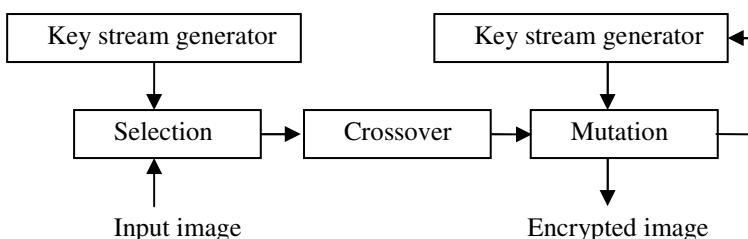


Fig. 1. Block diagram of the proposed system

The proposed algorithm is as follows.

Step 1: Input a color image 'I' of dimension $3 \times M \times N$

Step 2: Read each pixel of the image 'I' and store it in a Vector 'V'.

Step 3: Let each value of pixel in each color channel be in the range 0 to $((M \times N) - 1)$

- For $i = 0$ to $((M \times N) - 1)/2$, perform the following operations
- $v1 = V[i]$, $v2 = V[((M \times N)-1) - i]$
where 'v1' and 'v2' are variables to store pixel values.
 - Use a linear congruential pseudorandom generator for the selection of the crossover point, say 's'. According to selection point 's', divide 'v1' into 'v11' and 'v12', where $v11 = s$ and $v12 = v1 - s$
divide 'v2' into 'v21' and 'v22', where $v21 = s$ and $v22 = v2 - s$
 - Apply the crossover operation between 'v1' and 'v2', New generation obtained after crossover is stored in position

$$V[i] = v1 \text{ and } V[((M \times N)-1) - i] = v2$$

Step 4: Generate initial value using a congruential pseudorandom generator, say 'K'. Apply mutation operation using Exclusive-OR operation.
For $i = 0$ to $(M \times N) - 1$

$$V[i] = V[i] \oplus K$$

$$K = V[i]$$

Step 5: Write an encrypted image on the basis of vector 'V' obtained after Step 3 and 4.

5 Experimental Results

For the experimental study we have created an image corpus. The image corpus comprises 40 color images of different sizes. Image samples in the corpus are of uniform color and multi colors. Images in the corpus contain text information and non text/graphical information. The results of the proposed approach on sample images in image corpus are shown in Fig.2.

Fig.2 (a2) to (e2) show results after crossover operation. The cipher reflects the original image in all the cases. By observing the cipher, information present in the images could be easily recognized. Fig.2 (a3) to (e3) represents the cipher image obtained after crossover and mutation operation, respectively. From the results shown in Fig.2 (a3) to (e3), it is evident encrypted images will not reveal any identity of the original images. Encrypted images obtained are completely different from the original images. The algorithm is implemented in Java language using JDK 1.6 tool kit.

6 Statistical Analysis

From the literature, it is known that many ciphers have been successfully analyzed with the help of statistical analysis. Further, several statistical attacks [11] have been devised on them. To prove the robustness of the proposed method, statistical analysis is performed by plotting the histograms of the original and cipher images. The performance of the algorithm is measured in terms of correlation coefficient between original and cipher images.



Fig. 2. Results of the proposed method on sample color images in image corpus

6.1 Histogram Analysis

The histogram is a graphical representation showing a visual impression of the distribution of data. To prevent an attack, the cipher obtained should not give any clue about the original image. In the proposed approach, the cipher obtained does not give any clue about the original image which is analyzed through histograms. For sample image and corresponding encrypted images in Fig.2 (b1) to (b3), the histograms are as shown in Fig.3 (a) to (c), respectively. Levels of RGB channels of original image are unequally distributed which is shown in Fig.3 (a). Fig.3 (b) represents a histogram for the cipher obtained after performing crossover operation. Histogram obtained is almost similar to histogram of the original image, because crossover operation only alters the positions of the pixels. Histogram in Fig.3 (c) is for the cipher obtained after crossover and mutation operation, which shows that all pixels are uniformly distributed. It is revealed from the histogram shown in Fig.3 (c) that all pixels in R,G and B channels of sample image are distributed uniformly. Hence, the cipher image does not provide any clue to statistical attack.

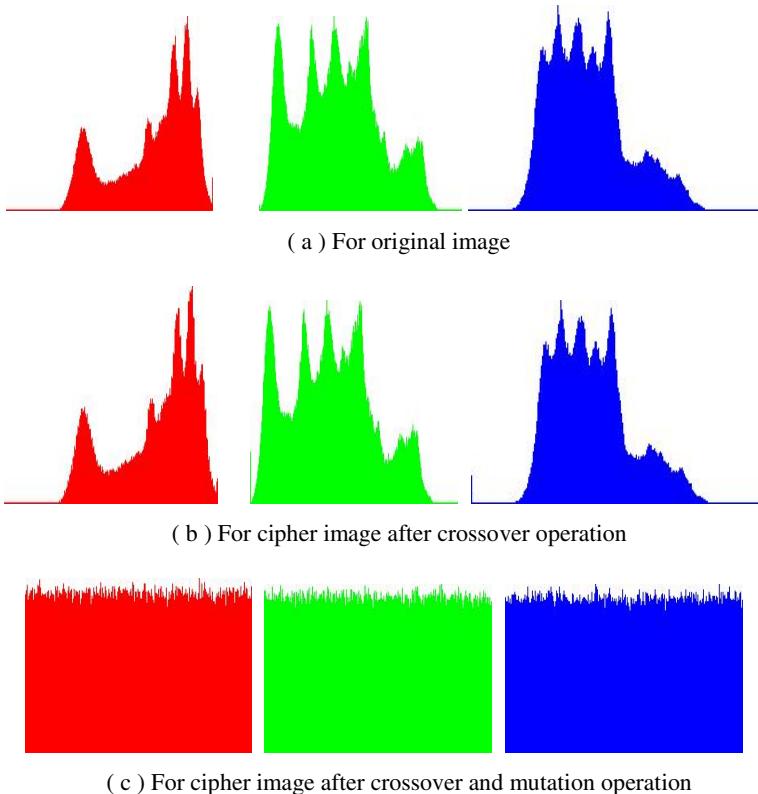


Fig. 3. Histogram of image shown in **Fig. 2. (b1)**

6.2 Correlation Coefficient Analysis

Correlation coefficient ‘r’ between two images is computed using an equation (2). In general, $r=1$ if two images are absolutely identical, they are completely uncorrelated if $r=0$ and if $r = -1$ then the images are completely anti-correlated [12].

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}} \quad (2)$$

where ‘n’ is the number of pairs of data.

We computed correlation coefficient of input color image and encrypted image in image corpus. The values falls in between - 0.002 to 0.006, which indicates encrypted images are completely uncorrelated with the original images. From these values, it can be concluded that the cipher images are completely different from the original images. This also proves that the cipher images are robust to statistical attack.

7 Complexity

The execution of proposed algorithm is computed on a system with dual core processor of 2.53 GHz and Microsoft Windows XP platform. Table 1 shows execution time for sample images. From the results in Table 1, it is evident that the computation time required for proposed method is very less, further the values reveals that, as the size of the image increases encryption time also increases.

Table 1. Encryption time for images of different size

Image size	Encryption Time
192 X 192	15 ms
256 X 256	31 ms
512 X 512	47 ms

8 Conclusions

In this work, a new image encryption scheme has been proposed which utilizes selection, crossover and mutation operations. The method proposed is tested on varieties of images. Statistical analysis is carried out through histograms. Performance of the approach is evaluated in terms of correlation coefficient. From the performance evaluation it is concluded that the algorithm proposed is robust to statistical attack.

In the proposed approach, it is hard to predict the key sequence as the key sequence generated depends on the input image and initial seed of the pseudorandom generator. This system works for still images. Encryption of video images is considered as further extension of the current study.

References

1. Kumar, A., Ghose, M.K.: Overview of Information Security Using Genetic Algorithm and Chaos. *Information Security Journal: A Global Perspective*, 306–315 (2009)
2. Kumar, A., Rajpal, N.: Application of Genetic Algorithm in the Field of Steganography. *Journal of Information Technology* 2(1), 12–15 (2004)
3. Kumar, A., Rajpal, N., Tayal, A.: New Signal Security System for Multimedia Data Transmission Using Genetic Algorithms. In: *NCC 2005*, IIT Kharagpur, January 28-20, pp. 579–583 (2005)
4. Tragha, A., Omari, F., Kriouile, A.: Genetic Algorithms Inspired Cryptography. In: *A.M.S.E Association for the Advancement of Modeling & Simulation Techniques in Enterprises. Series D: Computer Science and Statistics* (November 2005)
5. Tragha, A., Omari, F., Mouloudi, A.: ICIGA: Improved Cryptography Inspired by Genetic Algorithms. In: *International Conference on Hybrid Information Technology, ICHIT 2006* (2006)
6. Omari, F., Tragha, A., Bellaachia, A., Mouloudi, A.: Design and Evaluation of Two Symmetrical Evolutionist-Based Ciphering Algorithms. *IJCSNS International Journal of Computer Science and Network Security* 7(2), 181–190 (2007)
7. Yen, J.-C., Guo, J.-I.: A new image encryption algorithm and its VLSI architecture. In: *IEEE Workshop Signal Processing Systems*, pp. 430–437 (1999)
8. Liand, S., Zheng, X.: On The Security of An Image Encryption Method. In: *IEEE International Conference on Image Processing (ICIP 2002)*, Rochester, New York. Proceedings of ICIP 2002, vol. 2, pp. 925–928 (September 2002)
9. Husainy, M.: Image Encryption using Genetic Algorithm. *Information Technology Journal* 5(3), 516–519 (2006)
10. Srikanth, V., Asati, U., Natarajan, V., Pavan Kumar, T., Mullapudi, T., Iyengar, N.C.S.N.: Bit Level Encryption of Images using Genetic Algorithm. *TECHNIA-International Journal of Computing Science and Communication Technologies* 3(1) (July 2010) ISSN 0974-3375
11. Abuhaiba, I.S.I., Hassan, M.A.S.: Image Encryption using differential evolution approach in frequency domain. *Singal & Image Processing: An International Journal (SIPIJ)* 2(1) (March 2011)
12. El-Wahed, M.A., Mesbah, S., Shoukry, A.: Efficiency and Security of Some Image Encryption Algorithms. In: *Proceedings of the World Congress on Engineering, WCE 2008*, London, U.K, July 2-4, vol. I (2008)
13. Stallings, W.: *Cryptography and Network Security Principles and Practices*, 3rd edn. Pearson Education (2003)

Content Based Image Retrieval Using Normalization of Vector Approach to SVM

Sumit Dhariwal, Sandeep Raghuvanshi, and Shailendra Shrivastava

Smrat Ashok Technological Institute, Vidisha, Madhya Pradesh, India

{sumitdhariwal22, sraghuvanshi}@gmail.com,
shailendrashrivastava@rediffmail.com

Abstract. Semantically image has very meaningful categories. Classifying image using the low level feature is a challenging task. So far several methods has been used for automated machine learning in semantic image classification in this paper we have proposed a new and far more efficient method for semantic image classification using normalized vectors of WFSVM(weighted feature support vector machine). For image classification, the image data usually have a large data set on number of feature dimensions. Traditional image classification algorithms based on the SVM assign normalized automated weights to these features. The relevant and non relevant features of image are separated using these normalized vectors. Using normalized vector the efficiency and training time of SVM is improved to a greater extent. In this paper we proposed an approach to use weighted normalized vectors in place of normalized vectors. The Experiment is carried out on 256_categories database and result in better. The weighted normalization of vector has two advantages than the traditional SVM: the better performance of generalization ability and less training time.

Keywords: Semantic Classification, Support Vector Machine, automated weighted feature, Normalized Vector.

1 Introduction

In recent years there have been several activity done regarding development of image retrieval methods based on image content [1] [2]. Currently , many of the papers on image retrieval refers to the problem of semantic gap and semantic gap is the key obstacle in content based image retrieval.[3] The supervised machine learning technique is a effective way to reduce the semantic gap. The goal of supervised learning is to predict the outcome, based on the set of input metrics.[4,5].[6] proposes that a successful grouping of the database images into semantically meaningful classes will greatly enhance the performance of content- based image retrieval systems by filtering out images from irrelevant classes during matching. Supervised learning such as support vector machine (SVM) and the Image search technique with strong theoretical system available in supervised method, SVM has been used for object recognition and text classification, etc. and is considered a good and image retrieval system.

To improvise the classification adequacy, [7] normalize SVM margin and apply variance reduction technique to SVM pairwise classification result, [8] integrates two sets of support vector machines namely the multiple instance learning (MIL)-based

and global-feature-based SVM, for classification. The query point movement method is used to improve the estimate of the “ideal query point” by moving it towards good examples point and away from bad example points. On the other hand, the re-weighting method is also used to change the distance metric to make relevant images closer. This method tries to approximate the semantic concepts by mapping images to a new feature space. [9].[10] proposes a multimodal fusion framework using the information derived from the detected text as well as the low-level visual cues into pairwise SVM classifiers. [11]

However, all previous SVM algorithms on image classification have not distinguished the differences of different features for different object classification and assign the same weight to all low-level features. In fact, for high dimensional image data, many dimensions are less relevant or irrelevant for the task of classification. For example, color features are more relevant than shape or position features, while the relevant features for “ball” are shape as compared to the color feature. Thus, the relevant features or dominant features are very useful when we measure similarity between two images. Less relevant and irrelevant features will increase the calculation in noisy data. Hence, we think the calculation of each feature weight is needed when we use the SVM to classify images.

This paper proposes a method for enhancing the performance of support vector machine. In the proposed we try to find out the deviation of each element present in feature vector of image and then normalize those elements by maximum deviation of that particular element, hence each element in the vector now varies from 0 to 1, which reduces the domination of any particular element during training & classification and this section 3 we provide the results and the concluded in section 7.

2 Related Work

Many of the machine learning has been done on automatic semantic image classification, there are several other approaches for semantic image classification of images on the basis of SVM.

The semantic image classification method based on the WFSVM:

To classify the images automatically, first we need to extract the features from the training data set to calculate the weight w_{im} for each feature. Then apply the weighted feature to train the SVM, hence train SVM to classify the new images. The data set taken was consisting 1000 images. Each category including 100 images represents one distinct category of interest. Therefore, the data set has 10 thematically diverse image categories. All the images are in JPEG format with size 384x256 or 256x384. The keywords are assigned to describe each image category. The category names are: sun, food, flower, building, mountain, horse, dinosaur, elephant, beach and bus. And some randomly selected sample images from each category images of each class are randomly drawn as training samples. So 300 images are used for training and the remaining 700 are used for testing. Each image is represented as a 30 dimensional vector, which corresponds to 30 low-level features such as color, shape, size, texture, position. [14]

WFSVM outperforms the SVM system by 16.86% in the overall accuracy, it also improve the training speed by 5 times over the existing SVM system. [14]

3 Support Vector Machine

The SVM methodology comes from the application of statistical learning theory to separating hyper planes for binary classification problems. Given a set of cases which belong to one of two classes, training a linear SVM consists in searching for the hyper plane that leaves the largest number of cases of the same class on the same side, while maximizing the distance of both classes from the hyper plane.

Suppose the training set is linearly separable, we are given a set

$$T = \{(x_1; y_1), (x_2; y_2), \dots, (x_l; y_l)\},$$

Where $x_i \in \mathbb{R}^n$ are the input vector, Each point x_i belongs to either of two classes and thus is given a label $y_i \in \{-1, 1\}$. The goal is to establish the equation of a hyper plane that divides T . When the training set is not linearly separable, the optimal separating hyper plane is found, solving an optimization problem relaxed by introducing a set of slack variables and a penalization for cases that are misclassified or inside the margin. The task for finding the optimal hyper plane is to minimize the following objective function,

$$\begin{aligned} & \min \frac{1}{2} \|w\|^2 + c \sum_{i=1}^l \xi_i \\ & \text{s.t. } y_i [w \cdot (x_i) + b] \geq 1 - \xi_i \\ & \quad \xi_i \geq 0 \quad i = 1, 2, \dots, l \end{aligned} \quad \dots(1)$$

Parameter ‘C’ determines a trade-off between the error on the training set and the separation of the two classes. is ‘ ξ ’ a slack parameter that allows classification errors. The dual optimization problem is given as,

$$\begin{aligned} & \min \frac{1}{2} \sum_{i,j=1}^l a_i a_j y_i y_j K(x_i \cdot x_j) - \sum_{i=1}^l a_i \\ & \text{s.t. } \sum_{i=1}^l a_i y_i = 0 \\ & \quad 0 \leq a_i \leq C, \quad i = 1, \dots, l \end{aligned}$$

Where ‘ a_i ’ is the Lagrange multiplier.

4 Proposed Technique

We are proposing a method for enhancing the performance of support vector machine. In the proposed we try to find out the deviation of each element present in feature vector of image and then normalize those elements by maximum deviation of that particular element, hence each element in the vector now varies from 0 to 1, which reduces the domination of any particular element during training & classification.

Proposed algo-

- Step1:** Initialize the variables.
- Step 2:** Set the categories of image sample.
- Step 3:** Calculate the vector histogram information such as (Mean Red histogram value, Mean Green histogram, Mean Blue histogram).
- Step 4:** Calculate the vector of Standard deviation information such as (Standard deviation Red histogram, Standard deviation Green histogram. Standard deviation Blue histogram).
- Step5:** Calculate the Texture information such as (Texture Energy, Texture Entropy).
- Step6:** Calculate the image contrast and position for Information such as (Contrast Image, Horizontal vertical edge dens, X edge dens, Center X, Center Y).
- Step 7:** Check inputs.
- Step 8:** Check group is a vector. Group must be a Vector .
- Step 9:** Checking whether the equivalent vectors belongs to set of image vector.
- Step 10:** Calculate the maximum value of equivalent vector.
- Step 11:** Calculate the minimum value of image vector.
- Step 12:** Finally find the normalized vector.
- Step 13:** end. (And use normal vector train SVM are one against one $n(n-1)/s$ times.)

This technique is so efficient and proper worked, when we start the classification in this proposed technique we calculate the deviation (0 to 1) in particular image so in that case reduce the domination, so first find the total image of categories (picture categories) after that taking sample i,j (sample image) and calculate the equivalent vector formation used in training and classification likewise mean of red, blue, green histograms and now we are calculate standard deviation for red, blue, green histogram and after that we calculate the image of texture energy, entropy and contrast of image horizontal and vertical image dens last the calculate the center of x, y(Image) after that calculate the normalized vector of maximum and minimum and used to against $(n-1)/s$ times. Finally we taking a efficient result in compare to other so in that case this technique is useful to find the imaging system.

Acknowledgments. The heading should be treated as a 3rd level heading and should not be assigned a number.

5 Experiment Result

In this section, we evaluate our automatic weighted SVM for normalization image classification based on image from the 256_categories database. Taking 1000 images in our data set each category including 100 images represents one distinct topic of interest. Therefore, the data set has diverse image categories. All the images are in JPEG format with size 384x256 or 256x384. The category names are: Flags, Begs, Baseball bat, Tree, Building, Bus, Dinosaur, Rose, Horse, Basket ball hoop. 10 images of each class are randomly drawn as training samples. So 100 images are used for training and remaining 900 are used for testing. I am used a multi-class SVM, for classification constructed according to the strategy of “one per class”. The each class

category, an SVM is trained to separate that category from all the other categories. In this technique is used to each row lists of the average percentage of the images in one category classified into each of the 10 categories. The diagonal number shows the accuracy of categorization at each category.

From numbers in Table 1, we can see the classifier obtained using the weighted feature WFSVM is far accurate than the classifiers based on the SVM.

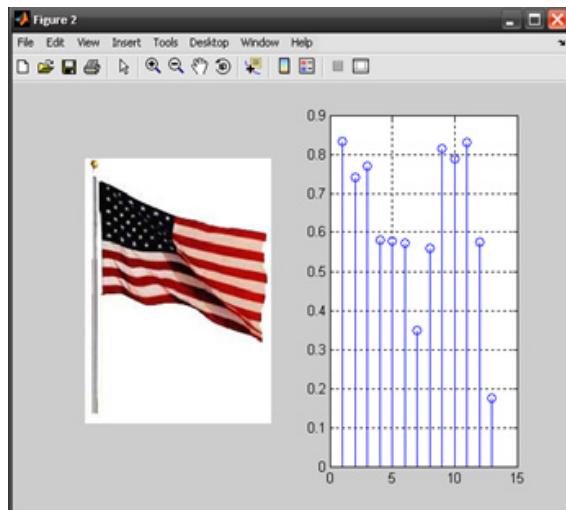


Fig. 1. Matlab figure- 2(Classifye Flage image).

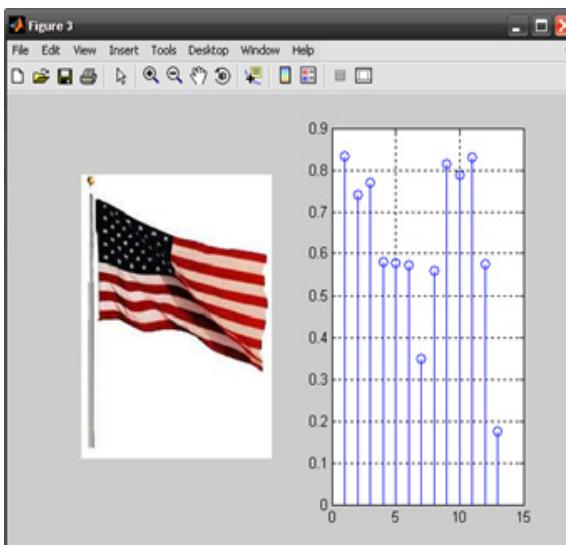


Fig. 2. Matlab figure- 3(Classifye Flage image).

Taking graph is to very efficient result find to compare to relevant database, first taking flag image and this flag image is compared to our standard database, the matlab image figure-2 is represent is efficient result because selected image is our site flag, so trained machine is 10 categories database image, so all image is compared after that it will gave the result graph . Matlab figure-3 represent also fine result and not give any other image in our database, so in that case this result is efficient and better performance results.

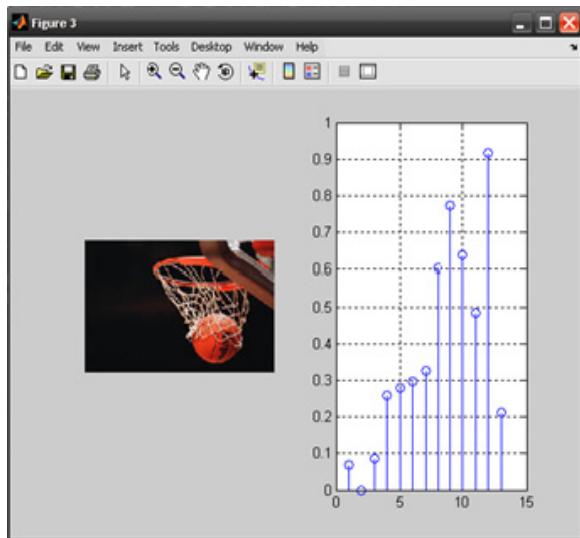


Fig. 3. Matlab figure- 2(Classifye basket ball hoop image).

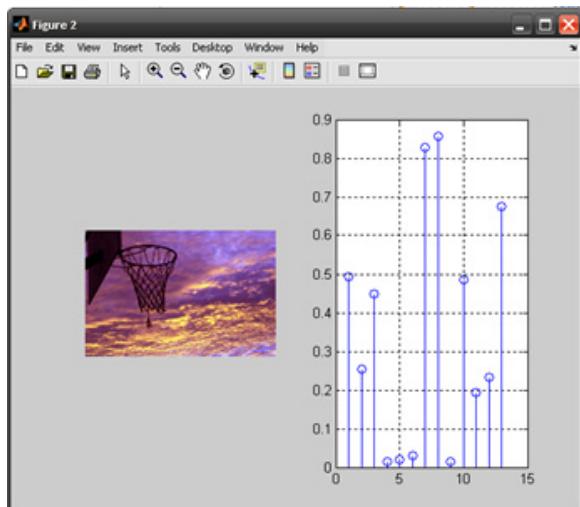


Fig. 4. Matlab figure- 2(Classifye basket ball hoop image).

Figure 3 and figure 4, taking basketball hoop image (any image in our database) after that trained our database and choose the basket ball hoop image so mostly time our system is give this image so in that case our system performance is too good ,and the better generalized ability and training time is better then other system.

In this proposed system is compared in to the format of Table 1 , Table 2 , the table one is the result are obtained the automatic weighted feature is more accurate result in the Classifiers is SVM based. The proposed system is compared in the general Support Vector Machine using image form the 10 categories. Table 2 , summarizes the performance of these two systems in terms of the overall classification accuracy and the approximate average training time in minutes for one binary SVM. It clearly shows that our proposed system performs the better. It outperforms the SVM system by **21.217%** in the overall accuracy. Our system also improves the training speed by 5 times over the SVM system that our approach not only improves the overall classification accuracy but also reduces the training time.

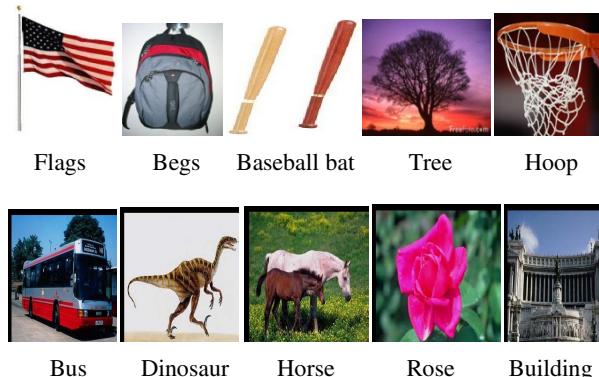


Fig. 5. Sample image chosen 10 image categories.

Table 1. The confusion matrix of the classification based on wfsvm

	Sun	Beach	Bus	Dinosaur	Horse	Elephant	Mountain	Flower	Building	food
Sun	0.9429	0.0000	0.0000	0.0286	0.0000	0.0286	0.0000	0.0000	0.0000	0.0000
Beach	0.0286	0.9714	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Bus	0.0000	0.0000	0.9571	0.0000	0.0000	0.0000	0.0000	0.0000	0.0286	0.0143
Dinosaur	0.1286	0.0000	0.0000	0.8714	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Horse	0.0143	0.0000	0.0000	0.0000	0.9857	0.0000	0.0000	0.0000	0.0000	0.0000
Elephant	0.0571	0.0000	0.0000	0.0000	0.0000	0.9429	0.0000	0.0000	0.0000	0.0000
Mountain	0.0571	0.0571	0.0000	0.0000	0.0000	0.0000	0.8571	0.0000	0.0286	0.0000
Flower	0.0571	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9429	0.0000	0.0000
Building	0.0429	0.0000	0.0143	0.0000	0.0000	0.0000	0.0000	0.0000	0.9429	0.0000
food	0.0000	0.0000	0.0143	0.0000	0.0000	0.0000	0.0000	0.0000	0.0143	0.9714

Table 2. The confusion matrix of the classification based on normalized svm

	Flags	Begs	Baseball bat	Tree	Basketball hoop	Bus	Dinosaur	Horse	Rose	Building
Flags	0.9834	0.0000	0.0166	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Begs	0.0000	0.9879	0.0000	0.0000	0.0121	0.0000	0.0000	0.0000	0.0000	0.0000
Baseball bat	0.0261	0.0000	0.9739	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Tree	0.0000	0.0000	0.0000	0.9823	0.0000	0.0000	0.0000	0.0000	0.0177	0.0000
Basketball hoop	0.0000	0.0147	0.0000	0.0000	0.9853	0.0000	0.0000	0.0000	0.0000	0.0000
Bus	0.0000	0.0000	0.0000	0.0000	0.0000	0.9894	0.0000	0.0106	0.0000	0.0000
Dinosaur	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9876	0.0000	0.0124	0.0000
Horse	0.0000	0.0000	0.0000	0.0432	0.0000	0.0000	0.0000	0.9568	0.0000	0.0000
Rose	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0235	0.9765	0.0000
Building	0.0000	0.0000	0.0000	0.0000	0.0000	0.0014	0.0000	0.0000	0.0000	0.9986

5 Conclusion

In this paper we have proposed a notion of normalized feature for computing the inner product and Euclidean distance in SVM, which is robust and has the better performance than the traditional SVM. The effective part of our algorithm is its automatic image classification and we compare our method with the traditional SVM. We took several images of different categories to develop our database, and selected 1000 images of 10 categories with 100 images from each category. Experiment was then conducted on those 1000 images from database. The result shows that the automatic weighted feature of SVM outperforms the traditional SVM in image classification and has better normalization ability and higher speed of calculation.

Acknowledgements. The success of this research work would have been uncertain without the help and guidance of dedicated group of people in our institute, Samrat Ashok Technological Institute (SATI) Vidisha. I would like to express my true and sincere acknowledgement as the appreciation for their contribution, encouragement and support. This research is to express gratitude and warmest guidance of them who in any way have contributed and inspired me.

References

1. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based imageretrieval: the end of the early years. *IEEE Trans. on Pattern. Analysis and Machine Intelligence* 22(12), 1349–1380 (2000)
2. Hare, J.S., Lewis, P.H., Enser, P.G.B., Sandom, C.J.: Mind the Gap: Another look at the problem of the semantic gap in image retrieval. In: *Proceedings of SPIE-The International Society for Optical Engineering*, vol. 6073 (2006)

3. Liu, Y., Zhang, D., Lu, G., Ma, W.-Y.: A survey of content-based image retrieval with high-level semantics. *Pattern Recognition* 40(1), 262–282 (2007)
4. Sethi, I.K., Coman, I.L.: Mining association rules Between low-level image features and high-level Concepts. In: Proceedings of the SPIE Data Mining and Knowledge Discovery, vol. III, pp. 279–290 (2001)
5. Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(3), 394–410 (2007)
6. Liua, Y., Zhanga, D., Lua, G., Mab, W.-Y.: A survey of content-based image retrieval with high-level semantics. *Pattern Recognition* 40(1), 262–282 (2007)
7. Zhang, L., Liu, F., Zhang, B.: Support vector machine learning for image retrieval. In: International Conference on Imag. Processing, pp. 7–10 (October 2001)
8. Qi, X., Han, Y.: Incorporating multiple SVMs for automatic image annotation. *Journal of Pattern Recognition* 40(2), 728–741 (2007)
9. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems 15, pp. 561–568. MIT Press, Cambridge (2003)
10. Zhu, Q., Yeh, M.-C., Cheng, K.-T.: Multimodal Fusion using Learned Text Concepts for Image Categorization. In: Proc. ACM International Conference Multimedia, pp. 211–220 (2006)
11. Vapnik, V.: The Nature of Statistical Learning Theory. Springer
12. Cortes, C., Vapnik, V.: Support-Vector Networks. *Machine Learning* 20(3), 273–297 (1995)
13. Wang, T., Tian, S., Huang, H.: Feature Weighted Support Vector Machine. *Journal of Electronics and Information Technology* 31(3), 514–518 (2009)
14. Wang, K., Wang, X., Zhong, Y.: A Weighted Feature Support Vector Mchines Method for Semantic Image Classification (2010)

Modified Grøstl: An Efficient Hash Function

Gurpreet Kaur¹, Vidyavati S. Nayak¹, Dhananjoy Dey², and S.K. Pal²

¹ Dept. of CE, Defence Institute of Advance Technology (DU), Pune-411 025, India

gurpreet.drd@gmail.com, vidyavatinayak@diat.ac.in

² Scientific Analysis Group, DRDO, Metcalfe House, Delhi-110 054, India

dhananjoydey@sag.drd.in, skptech@yahoo.com

Abstract. The cryptographic hash function Grøstl is one of the five finalists of SHA-3 competition organized by US National Institute of Standards and Technology (NIST). In this paper we propose a modified Grøstl-256 hash algorithm, which is 1.2 times faster than Grøstl-256 and as secure as Grostl-256. We further show that the modified Grøstl performs equally well as the original one when compared against standard metrics that are used to evaluate hash functions. A prototype tool developed to compare and evaluate the modified and the original Grostl-256 algorithm has been used for this purpose.

Keywords: Cryptographic hash function, fixed-point attack, Grøstl, length extension attack, SHA-3 Competition.

1 Introduction

A cryptographic hash function [1] H is an algorithm which processes an arbitrary length message into a fixed length digest or hash code. A hash function H is said to be pre-image resistant if for any given digest y of H , it is “computationally infeasible” to find a message x such that $H(x) = y$. A hash function H is said to be 2nd pre-image resistant if for any given message x , it is “computationally infeasible” to find another input message x' such that $x' \neq x$ and $H(x) = H(x')$. Also if it is “hard” to find any two messages x and x' such that $x \neq x'$ and $H(x) = H(x')$ then the hash function is said to be collision resistant. These three properties make the cryptographic one-way hash function suitable for achieving many security goals including authenticity, digital signatures and digital time stamping.

The most commonly used dedicated cryptographic hash functions are MD5 [2], SHA-1 [3] and SHA-2 [4]. After recent cryptanalytic attack [5] on MD5 and SHA-1, the security of their successor, SHA-2 family, against all kinds of cryptanalytic attacks has become an important issue. Although many theoretical attacks [6], on the reduced round of SHA-256 have been published during the period 2003 to 2008. In the mean time NIST announced the SHA-3 competition [7] in Nov 2007.

In SHA-3 hash function competition the five finalists selected in final round on 09 Dec 2010 are Blake [8], Grøstl [9], JH [10], Keccak [11] and Skein [12].

In this paper a modified Grøstl algorithm is proposed, which is around 1.2 times faster than Grøstl¹. The simulations have been carried out and extensive comparisons

¹ From now onwards we will mention Grostl-256 as Grøstl.

have been made on various factors like the avalanche effect, bit variance test, frequency test, runs test, auto correlation test and near collision test for testing the hash output with the Grøstl algorithm.

The organization of the paper is as follows. Section 2 presents the brief description of Grøstl hash function, In Section 3 we discuss our proposed modified Grøstl hash function. Section 4 gives the security analysis using various tests conducted on the proposed hash function and its performance along with the comparison with Grøstl algorithm.

2 Grøstl Overview

Grøstl [9] is SHA-3 finalist and is an iterated hash function with a compression function built from two fixed, large, distinct permutations. The design of Grøstl is based on principles very different from those used in the SHA-family. The two permutations are constructed using the wide pipe design strategy, which makes it possible to give strong statements about the resistance of Grøstl against large classes of cryptanalytic attacks.

2.1 Hash Function Construction

The Grøstl hash functions iterate the compression function f as follows.

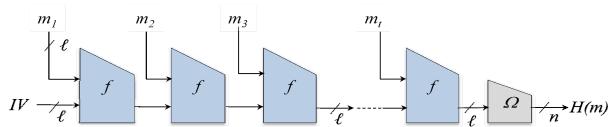


Fig. 1. The Grøstl hash function

The message M is padded and split into ℓ -bit message blocks $m_1 \dots m_t$, and each message block is processed sequentially. An initial ℓ -bit value $h_0 = IV$ is defined, and subsequently the message blocks m_i are processed as

$$h_i \leftarrow f(h_{i-1}, m_i) \text{ for } i = 1, \dots, t.$$

Hence, f maps two inputs of ℓ -bit each to an output of ℓ -bit as shown in Fig. 1.

2.2 The Compression Function

The compression function f is based on two underlying ℓ -bit permutations P and Q . It is defined as follows:

$$f(h, m) = P(h \oplus m) \oplus Q(m) \oplus h.$$

The construction of f is illustrated in Fig. 2 where P and Q are two fixed, large, distinct ℓ -bit permutations.

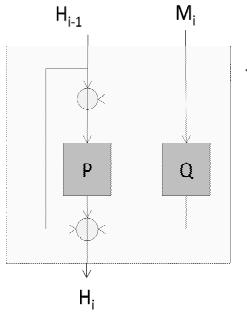


Fig. 2. The compression function f for Grøstl

2.3 The Output Transformation

Let $\text{trunc}_n(x)$ be the operation that discards all but the trailing n bits of x . The output transformation is defined by

$$\Omega(x) = \text{trunc}_n(P(x) \oplus x).$$

3 Proposed Algorithm

Modified Grøstl hash function can take arbitrary length ($< 2^{64}$ 1024-bit block) of input and gives 256 bits output. In the proposed design, we have modified the hash function construction, the padding procedure and the compression function.

3.1 Hash Function Construction

The modified Grøstl iterates the compression function f as follows. The message M is padded and split into 1024 bit (2ℓ -bit) message blocks $M_1 \dots M_t$, and each message block is processed sequentially.

Padding

The hash value of a message M of length $L=1024 \times (t-1) + 8r$ bits can be computed in the following manner [14]:

First we append 1 to the end of the message M . Let k be the number of zeros added for padding. 7-bit representation of r bytes is appended to the end of k zeros and at the last the 64-bit representation of total number of blocks t is placed. Now k will be the smallest non-negative integer satisfying the following condition:

$$8r + 1 + k + 7 + 64 \equiv 0 \pmod{1024}$$

$$\text{i.e., } k + 8r \equiv 952 \pmod{1024}$$

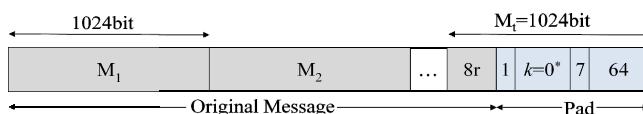


Fig. 3. Padding Procedure

The padding procedure is shown in Fig. 3. According to this padding procedure, we can compute the hash value of a message of length $\leq 2^{10} \times (2^{64} - 1)$ bits.

Parsing

Let L' be the length of the padded message. Divide the padded message into t ($= L'/1024$) 1024-bit block i.e. *thirty two* 32-bit words. Let M_i denote the i^{th} block of the padded message, where $1 \leq i \leq t$.

Initial values for IV and counter

The *initial value IV* is the 512-bits representation and is same as Grøstl.

$$IV = 00 \dots 00\ 01\ 00$$

The *initial counter* value C_1 is the 512-bits representation and is defined as

$$C_1 = 00 \dots 00\ 00\ 01$$

Where C_i will increment with each message block M_i processed for $1 \leq i \leq t$.

Hash Construction

An initial ℓ -bit value $H_0 = IV$ & ℓ -bit counter value C_1 is defined, and subsequently the 2ℓ -bit message blocks M_i are processed as

$$H_i \leftarrow f(H_{i-1}, M_i, C_i) \text{ for } 1 \leq i \leq t.$$

Hence, as shown in Fig. 4 the compression function f maps three inputs to an output of ℓ -bit. The first input is called the chaining variable; the second input is called the message block and the third input is block counter.

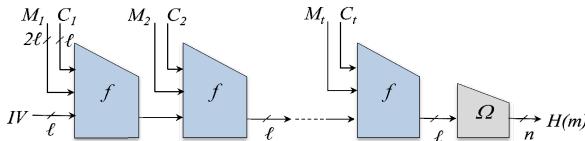


Fig. 4. The Modified Grøstl Construction

Modified Grøstl is designed for 256 bit output where ℓ is defined to be 512. We can also design the variants for 224, 384 and 512 bit message digest.

3.2 The Compression Function

The compression function f is based on two underlying ℓ -bit permutations P and Q. The function f is defined as follows:

Iteration

For every message block M_i is divided into *thirty two* 32-bit words and each 32-bit word is read in little endian format for $1 \leq i \leq t$. For example, to read an ASCII file with data ‘abcd’, it will be read as 0x64636261 [14].

Divide the input message block into two equal part, i.e., $M_i = L_i \parallel R_i$. Transform the message block in the following way:

$$Ptmp \leftarrow P(L_i \oplus C_i).$$

$$Qtmp \leftarrow Q(R_i \oplus H_{i-1}).$$

Where counter $C_i = i \bmod 2^{64}$ for $1 \leq i \leq t$. XORing C_i removes the fixed point attack, because as the number of blocks increases, counter C_i will be different for each block. $Ptmp$, $Qtmp$ are temporary 512-bit representation used for storing the intermediate values. The construction of compression function f for modified Grøstl is as shown in Fig. 5.

$$Ptmp \leftarrow P(Ptmp \oplus Qtmp).$$

$$H_i \leftarrow Ptmp \oplus Qtmp \oplus H_{i-1}.$$

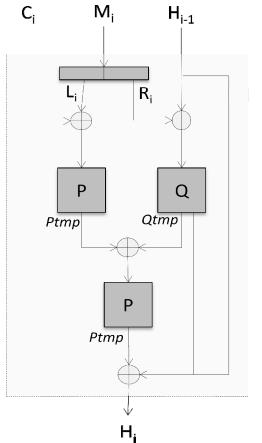


Fig. 5. The compression function f for modified Grøstl

3.3 The Output Transformation

The output transformation for modified Grøstl is same as Grøstl. Function $trunc_n(x)$ is the operation that discards all but the trailing n bits of x .

Test values of Modified Grøstl

Test values of the three inputs are given below:

$$\begin{aligned} H(a) &= 97A49C4C A3A4713B 93849EA0 42ECCBCE 937107A0 B1483212 \\ &\quad 5B692E4D 444A9FA8 \end{aligned}$$

$$\begin{aligned} H(ab) &= 808CA0ED F9616B5A B3B1156A D30B0D41 A84D2D34 4BBA162A \\ &\quad 155F0149 A31C8AB8 \end{aligned}$$

$$\begin{aligned} H(abc) &= 8BAC871F D9A13A08 858E953E 753F7E66 C837009A 4AB8C354 \\ &\quad DB1C4C73 D47B952A \end{aligned}$$

4 Analysis of Modified Grøstl Hash Function

This section discusses the experimental results, which show that the modified Grøstl is efficient (faster) than the Grøstl and resistant against the commonly known attacks for hashing algorithms retains the good hashing properties.

4.1 Efficiency of Modified Grøstl Function

A performance comparison of the both hash functions considered for message digest generation is reported in Table 1. The table gives a comparative study regarding the execution time of modified Grøstl with Grøstl on an Intel Core i3 with chipset M370 @ 2.40Gz processor with 3 GB RAM.

Table 1. Average execution time comparisons

File Size (in MB)	Grøstl (in ms.)	Modified Grøstl (in ms.)
1.6	173.09	124.47
5.4	458.65	361.47
10	894.44	671.06
12.6	1131.61	855.42
16.5	1447.14	1094.79
19.1	1605.09	1243.31
24.8	1977.77	1570.29

Here we observed that Modified Grøstl is 1.27 times or 21.3% faster than Grøstl. The performance comparison chart as shown in Fig 6.

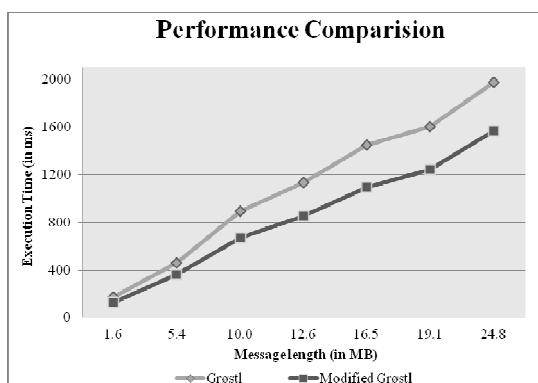


Fig. 6. Performance comparison

4.2 Near-Collision Resistance

A hash function H is near-collision resistant, if it is “hard” to find two messages with hash values that differ in only small number of bits. i.e., a near-collision [15] occurs if for two different messages $M \neq M'$, $H(M)$ differs from $H(M')$ by only a small number of bits.

To test whether our modified Grøstl is near-collision resistance; we have taken an input file (random file) M consisting of 552 bits. With some possible combination of the sequence we generated 314364 different messages M_i for $1 \leq i \leq 314364$. Then we compute hamming weight of $(H(M_i) \oplus H(M_j))$ for $1 \leq i \leq 314364$ and $i < j \leq 314364$, which gives the output differences of hash values for different input messages. We have observed (Table 2) that the minimum output difference is 76 bits (95.03% files have output difference ≥ 89 bits) and maximum difference is 182 bits (95.16% files have output difference ≤ 167 bits). Thus we can say that Modified Grøstl is near-collision resistant.

Table 2. Output differences for near collision test

	Grøstl		Modified Grøstl	
	Diff.	%files	Diff.	%files
Maximum	180	95.08% ≤ 167 bits	182	95.16% ≤ 167 bits
Minimum	76	95.13% ≥ 89 bits	76	95.03% ≥ 89 bits

4.3 Avalanche Effect

We have taken an input file M consisting of 1024 bits and computed $H(M)$. By changing the i^{th} bit of M , the files M_i have been generated, for $1 \leq i \leq 1024$. Thus hamming distance of each M_i from M is exactly one for $1 \leq i \leq 1024$. We then computed $H(M_i)$ for $1 \leq i \leq 1024$, computed the Hamming distances d_i between $H(M)$ and $H(M_i)$, for $1 \leq i \leq 1024$, i.e., number of ones for $H(M) \oplus H(M_i)$. The Table 3 shows the max, min, mode and the mean values of the above distances.

Table 3. Hamming Distances

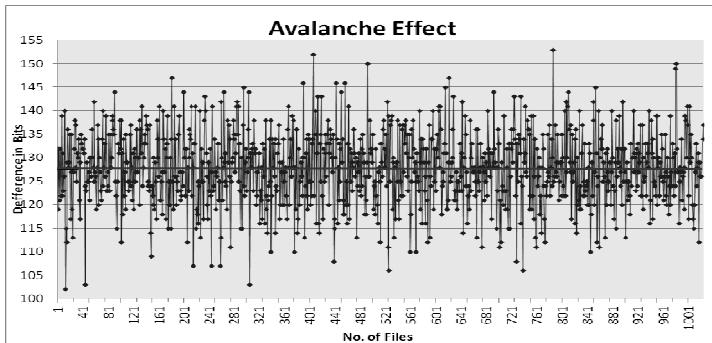
	No. of ones for $H(M) \oplus H(M_i)$	
	Grøstl	Modified Grøstl
Maximum	152	153
Minimum	101	102
Mode	126	128
Mean	128.82	127.73

To satisfy strict avalanche criterion, each d_i should be 128 for $1 \leq i \leq 1024$. But we have found (Table 3) that d_i 's were lying between 102 and 153 for the above files and in most of the cases $d_i = 128$. The observed deviation is acceptable so as to resist collision search using differential attack.

The Table 4 and Fig. 7 show the distribution of the 1024 files with respect to their differences (distance) in bits for modified Grøstl. The normal distribution of Hamming distances for modified Grøstl is shown in Fig. 8.

Table 4. Hamming Distances range of distances

Range of Distance	Grøstl		Modified Grøstl	
	No. of Files	%	No. of Files	%
128 ± 5	446	43.55%	457	44.63%
128 ± 10	802	78.32%	796	77.73%
128 ± 15	962	93.95%	956	93.36%
128 ± 20	1011	98.73%	1009	98.54%

**Fig. 7.** Distribution of the 1024 files with respect to their differences in bits

4.4 Randomness Test

To conduct randomness test, we have generated a file consisting of 1024512 bits by concatenating all the output of hash of the files 4002. After that, we have divided 1024512 bits into 204 blocks of length 5000 bits each, 102 blocks of length 10000 bits each, 40 blocks of length 25000 bits each, 20 blocks of length 50000 bits each, 10 blocks of length 100000 bits each, 2 blocks of length 500000 bits each and 1 block of the complete 1000000 bits. Thus we have generated 379 blocks in total and conducted five basic randomness tests in these blocks. The concise result is shown in the following Table 5 with level of significance of 0.01.

Table 5. Randomness test results at level of significance: 0.01

Name of Test	# Blocks	Grøstl	Modified Grøstl
		Passed %	Passed %
Frequency	379	100.00%	99.21%
Serial	379	99.74%	99.74%
Poker5	379	100.00%	100.00%
Runs	379	97.89%	99.74%
Auto-correlation	379	97.89%	99.74%

4.5 The Bit-Variance Test

The bit variance test consists of measuring the impact of change in input message bits on the digest bit. More specifically, given an input message, all the small changes as well as the large changes of this input message bits are taken and the bits in the corresponding digest are evaluated for each such change. Afterwards, for each digest bit the probabilities of taking on the values of 1 and 0 are measured considering all the digests produced by applying input message bit changes. If $P_i(1) = P_i(0) = 0.5$ for all digest bits $i = 1, \dots, 256$ then, the Modified Grøstl has attained maximum performance in terms of the bit variance test [14].

Since it is computationally difficult to consider all input message bit changes, we have evaluated the results for 1024 files, viz. $M, M_1, M_2, \dots, M_{1024}$ which we have generated for conducting avalanche effect, and found the following results:

Number of digests = 1025

Mean frequency of 1's (expected) = 512.50

Mean frequency of 1's (calculated) = 510.87

Results given in Table 6 shows for first 6 bit positions of messages for both the algorithms [16]. Plotting the probability (Fig. 9) of each of the bits (256-bit), we see that the average probability is approximately 0.50.

Table 6. Results Bit Variance Test

Bit positions of Message	Avg Probability for all digest bits	
	Grøstl	Modified Grøstl
b1	0.491707	0.485854
b2	0.495610	0.500488
b3	0.540488	0.454634
b4	0.487805	0.500488
b5	0.495610	0.510244
b6	0.488780	0.478049

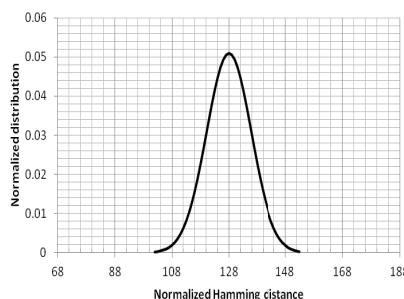


Fig. 8. The normalized Hamming distances

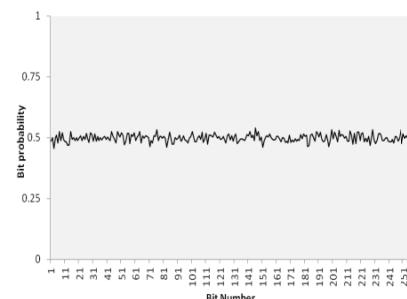


Fig. 9. The probability of a bit position

5 Conclusion

In this paper, we have proposed modifications of the SHA3 candidate Grøstl algorithm to make it more efficient. Analysis of proposed hash function viz. avalanche effect, near collision resistance test, bit variance test, randomness test are also described. From the experimental results, we can say that our modified Grøstl is 1.2 times faster than Grøstl and is as secure as Grøstl as the two permutation used are same as of original. Also the modified algorithm performs equally well as the original one when compared against standard metrics that are used to evaluate hash function.

References

1. Menezes, A., Oorschot, P., Vanstone, S.: *Handbook of Applied Cryptography*. CRC Press (1997), <http://www.cacr.math.uwaterloo.ca/hac/>
2. Rivest, R.: The MD5 Digest Algorithm, Network Working Group Request for Comments: 1321 (1992), <http://theory.lcs.mit.edu/rivest/Rivest-MD5.txt>
3. Federal Information Processing Standards Publication 180-1: Secure hash standard (1996)
4. Federal Information Processing Standards Publication 180-2: Secure hash standard (2002)
5. NIST Brief Comments: Recent Cryptanalytic Attacks on Secure Hashing Functions and the Continued Security Provided by SHA-1,
<http://csrc.nist.gov/news/highlights/NIST-brief-Comments-on-SHA1-attack.pdf>
6. Gilbert, H., Handschuh, H.: Security Analysis of SHA-256 and Sisters. In: Matsui, M., Zuccherato, R.J. (eds.) *SAC 2003. LNCS*, vol. 3006, pp. 175–193. Springer, Heidelberg (2004)
7. National Institute of Standards and Technology, Cryptographic Hash Project SHA-3 contest (2011), <http://csrc.nist.gov/groups/ST/hash/sha-3/index.html>
8. Aumasson, J.P., Henzen, L., Meier, W., Phan, R.: SHA-3 Proposal Blake. Candidate to the NIST Hash Competition (2011)
9. Gauravaram, P., Knudsen, L., Matusiewicz, K., Mendel, F., Rechberger, C., Schlaffer, M., Thomsen, S.: Grøstl - a SHA-3 candidate. Submission to NIST, Round-3 (2011), <http://groestl.info>
10. Wu, H.: JH. Candidate to the NIST Hash Competition (2011)
11. Bertoni, G., Daemen, J., Peeters, M., Assche, G.: Keccak. Candidate to the NIST Hash Competition (2011)
12. Ferguson, N., Lucks, S., Schneier, B., Whiting, D., Bellare, M., Kohno, T., Callas, J., Walker, J.: Skein. Candidate to the NIST Hash Competition (2008)
13. Daemen, J., Rijmen, V.: AES Proposal: Rijndael. AES Algorithm Submission (1999), <http://csrc.nist.gov/archive/aes/rijndael/Rijndael-ammended.pdf>
14. Dey, D., Shrotriya, N., Sengupta, I.: R-hash: Hash Function Using Random Quadratic Polynomials Over GF(2), <http://eprint.iacr.org/2011/450.pdf>
15. Bozhan, S., Wenling, W., Shuang, W., Dong, L.: Near-Collisions on the Reduced-Round Compression Functions of Skein and BLAKE,
<http://eprint.iacr.org/2010/355.pdf>
16. Karras, D., Zorkadis, V.: A Novel Suite of Tests for Evaluating One-Way Hash Functions for Electronic Commerce Application. IEEE (2000)

Iris Recognition Systems with Reduced Storage and High Accuracy Using Majority Voting and Haar Transform

V. Anitha and R. Leela Velusamy

Department of Computer Science and Engineering,
National Institute of Technology, Tiruchirappalli, India
anitha.v2000@gmail.com, leela@nitt.edu

Abstract. Reliable user authentication is becoming an increasingly important task. Biometric based authentication offers several advantages over other authentication methods. Iris based biometric authentication gained more popularity because of its greater accuracy and uniqueness. In this paper, a new method is proposed based on Haar transform and Majority Voting to improve the overall efficiency of existing iris recognition systems in terms of accuracy and storage space. An existing iris recognition algorithm proposed by Libor Masek is used to generate an iris template. Haar transform is applied on those templates to reduce the storage space. Majority voting technique is being performed with target class iriscodes to improve the accuracy of the recognition system. Iriscodes of various combinations are made using different levels of haar decomposition and each combination is represented as a method. Experiments on well known CASIA iris database show that the proposed technique is more efficient and promising.

Keywords: Biometrics, Iris Recognition, User authentication, reduced storage space, Haar transform, Majority voting.

1 Introduction

Security is of major concern nowadays in almost all fields. Personal identification / user authentication plays a major role in security. The consequences of an insecure authentication system may lead to loss of confidential information, denial of service, and compromised data integrity. Reliable user authentication has a wide range of application over banking, e-commerce, airport security, public safety and justice, access control of a certain building or restricted area, attendance monitoring, network access, and physical access control to computer resources, etc. Existing methods for user authentication such as password, PIN, signatures, ID cards have certain limitations of being stolen, imitated and forgotten. Moreover there is no way to link the usage of the system or service to the actual user. Biometric authentication overcomes these limitations because it has a direct relation with the person who has to be authenticated. Biometric data cannot be shared, stolen, forgotten or imitated. It is highly unique [1].

Biometric authentication methods based on fingerprints, palmprints, face, iris, voice, retinal blood vessel patterns, etc., can be used instead of non-biometric methods for more safety and reliability. Among all biometric methods, iris gained more popularity because of its greater accuracy and uniqueness. So the proposed algorithm is experimented on iris biometric recognition system.

Iris recognition is a biometric based technology which is used to recognize a person from his/her iris patterns. Iris patterns are characterized by high level of stability and uniqueness. Each individual has a unique iris pattern. The difference even exists between identical twins and between the left and right eye of the same person. Iris recognition system provides highly accurate, easy to use and fraud proof means to verify the identity of the customer.

In this paper, a new approach is proposed to make existing iris recognition systems efficient in terms of both accuracy and storage space. Storage space of the biometric templates in the database is of major concern because applications that employ biometric recognition system deal with a huge set of data. So the template which has to be created from the iris data for storage should occupy less space. In order to reduce the space, haar transform is applied on the template, which is created by the Libor Masek process [2] to reduce the number of bits in the template. Multiple levels of haar transform on the template yield various sizes of templates and accuracy. With increase in the combination levels of haar transform, the size of the database and the accuracy get reduced. But the desired requirement is high accuracy with less memory space. So the level of the haar transform should be chosen as a tradeoff between database size and accuracy as per the application requirement. Accuracy is further improved by means of majority voting technique.

The rest of the paper is organized as follows. In the next section, a brief discussion on the related work is given. The following section explains the proposed system in detail. The next section discusses the experiments and results and finally the concluding remarks and future work are mentioned in the last section.

2 Related Work

As demands on secure identification are constantly rising and the human iris provides a pattern that is unique for identification, extensive research have been done in the field of iris recognition. Daugman [3] presented many advances in iris recognition which employs efficient methods detecting the iris boundaries, statistical methods for detecting and excluding eyelashes, exploration of score normalization. A machine vision system for automatic iris recognition was proposed by Wildes et al. [4]. Both systems of Daugman and Wildes et al. employs carefully designed devices for image acquisition to ensure that the iris is located in the same location within the image, and the images have the same resolution and are glare free under fixed illuminations. However, these requirements are not always easy to be satisfied especially in practical applications. An accurate and fast method for iris segmentation is proposed by He et al.[5] using ada-boost-cascade iris detector, rank filter noise elimination and it also deals with the non-circular iris boundaries. Avila et al. [6] proposed iris recognition for biometric identification using dyadic wavelet transform zero-crossing and achieved a recognition rate of 98%. Li Ma et al. [7] proposed an iris recognition which uses circular symmetric filters for feature extraction and nearest feature line approach is used for iris matching. Tisse et al. [8] proposed iris based personal authentication technique based on gradient decomposed hough transform or integro-differential operators combination for iris localization and analytic image concept to extract information from iris texture. Naveen singh et al. [9] designed a iris recognition system using a Canny Edge Detection scheme and a

Circular Hough transform, to detect the iris boundaries in the eye's digital image. Libor Masek [2] developed an iris identification system which employs hough transform for iris segmentation and 1D Log-Gabor filters for feature extraction.

Simple and effective techniques such as majority voting and haar transform can be applied to the template / iriscode generated by those algorithms to reduce the storage space requirement and improve the accuracy. In this paper, these techniques are applied on Libor Masek algorithm [2] with the help of publicly available CASIA iris database [10] and the results are tested.

3 Proposed System

This section documents each stage of the proposed iris recognition system in detail. The idea is to use multilevel haar transform on biometric templates in order to reduce the storage space and to increase the accuracy by combining the important features of the templates after the haar transform. Various types of methods are proposed based upon various combinations of coefficients after the multilevel haar transform. Majority voting is employed finally which further improves the accuracy. The proposed technique can be applied to any biometric system in order to reduce the storage space and increase the accuracy. But in this paper the proposed system have been implemented in an existing iris recognition system (Libor Masek iris recognition). The proposed techniques are explained detailed in the following stages of Feature vector generation, iriscode generation and authentication using majority voting.

The overall system consists of the following stages.

3.1 Iris Template Generation Using Libormasek Algorithm [2]

The schematic diagram of the iris template generation is shown in Fig (1). In a real time iris recognition system the iris images can be got by scanning the user's iris by an iris scanner. But in our experiment the publicly available iris database CASIA [10] is used to test our results.

The generation of iris template from the iris image which is got in image acquisition module involves 3 main steps: Segmentation, Normalization, and Encoding.

The segmentation module based on circular hough transform is employed to isolate the iris region in the image from the sclera and other parts of the eye image. Segmentation module segments the iris and pupil by drawing an iris-sclera boundary and iris-pupil boundary and then the region of the annular iris ring that is not visible due to the overlap of eyelids and eyelashes over the iris region is marked. A mask is generated in order to find the corrupted bits.

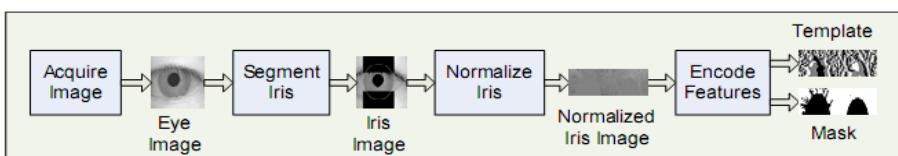


Fig. 1. Iris template generation

In-order to get rid of the inconsistencies in the image, the segmented iris image is then normalized into a rectangular grid of constant dimensions using polar-to-cartesian transformation and bilinear interpolation.

Finally, the normalized iris image is encoded into a binary image. Feature encoding is done by convolving the normalize iris image with 1D Log-Gabor wavelets. The result is a bit-wise biometric template of 4800 bits.

3.2 Feature Vector Generation Using Multi Level Haar Wavelet

Haar wavelet transform is the simplest useful energy compression process. Fig. 2(a)(b) shows how the haar wavelet transform breaks an image down into four sub-samples or images, named A, H, V, D which composed of coefficients. A is a set of approximation coefficients and H, V, D are sets of detail coefficients. The letters H, V, D corresponds to the horizontal, vertical and diagonal detail coefficients respectively. Fig. 2(c)(d) shows how an example image "lenna" is decomposed into 4 sub images when haar transform is applied to it. The results consist of one image that has been high pass in the horizontal and vertical directions, one that has been low passed in the vertical and high passed in the horizontal and vice versa, and one that has been low pass filtered in both directions. This transform is typically implemented in the spatial domain by using 1-D g convolution filters. This gives the reduced number of bits per pixel needed to represent the quantized data for each sub image, to a given accuracy. Each sub image provides various levels of compression.

Most of the energy will be stored in the approximation coefficients A. So the A sub-image of a level is used as the input to the next level of haar decomposition which is shown in Fig. 3. The template of 4800 bit can be decomposed using haar wavelet into a maximum of 5 levels. Each level of decomposition generates A, H, V, D. The detail coefficients H, V, D of each level are combined to give the feature vector. The bit rate of sub-images obtained from each level of decomposition is less than the size of sub image of the previous level [11].

Experimentation with maximum of 5 levels of haar wavelet decomposition generated various sizes of feature vectors. Various combinations of feature vector and the



Fig. 2. (a) Haar Tranform at decomposition level 1 (b) Haar transform at decomposition level 2 which decomposes the approximation coefficients A of level 1 into 4 subimages (c) Original "lenna" image (d) Level 1 haar transform which decomposes the original image into 4 subimages of different energy

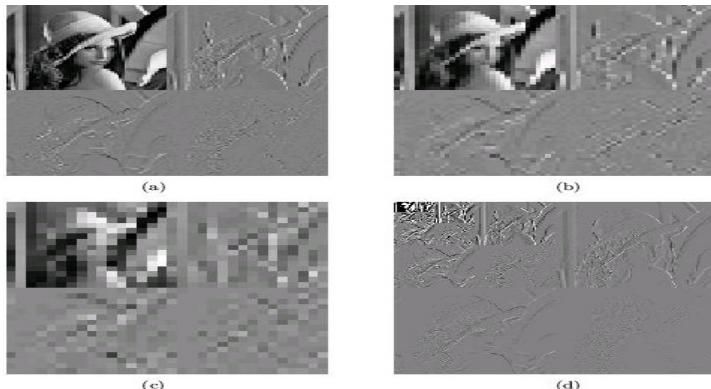


Fig. 3. (a) Haar 1st level decomposition of "lenna" image (b)(c)(d) Shows the result of applying haar transform to the A (approximation coefficients) subimage of previous figure

size of each are given in Table 1. The experimentation with the haar level is limited to level 5 because the database size of the 5th level combination itself exceeds the database size of the normal libormasek algorithm. In table1, H5, V5, D5 represents the H, V, D coefficients of haar decomposition level 5. Similarly the other terms such as H1, V1, D1, etc. represents the detail coefficients along with their decomposition level.

3.3 Iriscode Generation

Iriscode is the binary representation of feature vector. The obtained feature vector is represented in a binary form so that hamming distance calculation will become an easier task. Observation of the feature vector values shows that each coefficient of the vector ranges from -1 to +1.

Let FV_coeff be considered as each feature vector coefficient, and N be the number of coefficients in the feature vector, then a binary coding of the feature vector can be done using the following pseudo code:

```

for ( i = 1 to N)
    if ( FV_coeff ( i ) > 0 ) then
        IC ( i ) = 1;
    else
        IC ( i ) = 0;
    end
end

```

where IC represents the iriscode.

3.4 Storing Iriscode in Database

For a single user, during training phase a set of eye images can be taken e.g. 7 images. Iriscodes for those images are computed and is stored in the database against the user identity. When the user claims for authentication, the new iriscode generated will be compared with the stored iriscodes against the claimed identity.

3.5 Calculation of Hamming Distance between Templates

Hamming distance is a common metric which is used in the biometric data comparison. Hamming distance between 2 iriscodes indicate the variation between them. Let IC₁ and IC₂ be the two iriscodes to be compared. The hamming distance between them can be calculated as:

$$HD(IC_1, IC_2) = \frac{1}{N} \sum_{i=1}^N (IC_1 \oplus IC_2) \quad (1)$$

where N is the total number of bits in the iriscode and \oplus denotes an exclusive OR operator.

3.6 Authentication Using Majority Voting

Majority voting technique is being employed to improve the recognition rate of the system. If a user needs authentication to the system/service, a live eye image is taken from the user at the time of authentication by using the image acquisition model. From the raw eye image, iriscode is generated by employing the above iriscode generation procedure.

Once the iriscode is generated from the live eye image, it is compared with the iriscodes stored in the database against the claimed identity and hamming distances are calculated between them.

A threshold λ is set (by means of trial and error) for the hamming distance and HD_i be the hamming distance between the iriscode of the live iris image and the ith stored iriscode in that class. Here class indicates each person. The pseudo code for the authentication process is as follows:

```
for( i = 1 to n)
    if( HDi <= λ) -> Increment variable "yes"
    if( HDi > λ) -> Increment variable "no"
end
```

where n is the number of iriscode per class. If yes >= no, authenticate the user. Else, the user is not authenticated.

The above stages are implemented using MATLAB and tested with CASIA iris database. The outcome of the experiments will be discussed in detail in the next section.

Table 1. Feature vector combinations and sizes

Levels Combined	Combination	Feature vector size (bits)
5	H5,V5,D5	45
4,5	H4,H5,V4,V5,D4,D5	255
3,4,5	H3,H4,H5,V3,V4,V5,D3,D4,D5	756
2,3,4,5	H2,H3,H4,H5,V2,V3,V4,V5,D2,D3,D4,D5	2565
1,2,3,4,5	H1,H2,H3,H4,H5,V1,V2,V3,V4,V5,D1,D2,D3,D4,D5	9765

4 Experiments and Results

In this section, the database used for the implementation of the proposed work and the results obtained from the experiments done on that database is discussed.

4.1 Iris Database

We made use of CASIA-IrisV1 [10], a publicly available database, which includes about 756 images from 108 persons. For each person, 7 images are captured in two sessions (3 in the first session and 4 in the second session). The images in the database are in .bmp format with a resolution of 320×280. The segmentation algorithm of Libor Masek [2] gave a successful segmentation of only 83% among 756 images. Libor Masek made use of only the perfectly segmented iris and achieved an accuracy of 99.757%. But in our work we make use of all the 756 images which also include the un-segmented iris.

4.2 Evaluation Criteria

The accuracy of most of the biometric system is based on the accuracy/true recognition rate (TRR). Accuracy can be measured by calculating the total error rate (TER), which is the sum of false acceptance rate (FAR) and false rejection rate (FRR).

$$\text{Accuracy} = (100 - \text{TER}) \quad (2)$$

where

$$\text{TER} = (\text{FAR} + \text{FRR}) \quad (3)$$

False accept rate (FAR) is the probability that the system incorrectly matches the input pattern to a non-matching template in the database. It measures the percent of invalid inputs which are incorrectly accepted. False reject rate (FRR) is the probability that the system fails to detect a match between the input pattern and a matching template in the database. It measures the percent of valid inputs which are incorrectly rejected [12].

4.3 Results

Table 2 lists the method, size of each iriscode, storage space requirement, error rates, accuracy achieved and the corresponding threshold value λ with respect to the experiments done on CASIA database [10].

The comparisons of the methods over the main goals: size of the database and accuracy, are shown in Fig. 4(a)(b). The abbreviation MVIC in the figures denote (Majority voting with Iriscode) and the numbers along with it denote the haar level combinations. It clearly shows that the proposed methods have higher accuracy and reduced storage space compared to the existing Libor Masek method.

Various combinations of haar decomposition level along with the majority voting technique results in different storage requirements as well as accuracy rate. It is obvious from the graphs that as we keep including the haar levels, there is an increase in the database size as well as the accuracy. The desirable goal is to achieve a method which gives reduced storage and high accuracy. So the combination level should be chosen as a tradeoff between the database size and accuracy according to the application. MCIV2345 is considered as an efficient method because it yields a high accuracy of 99.8163% with the storage space usage of 2.62MB which is much lower than the storage space requirement of the existing libormasek method [2].

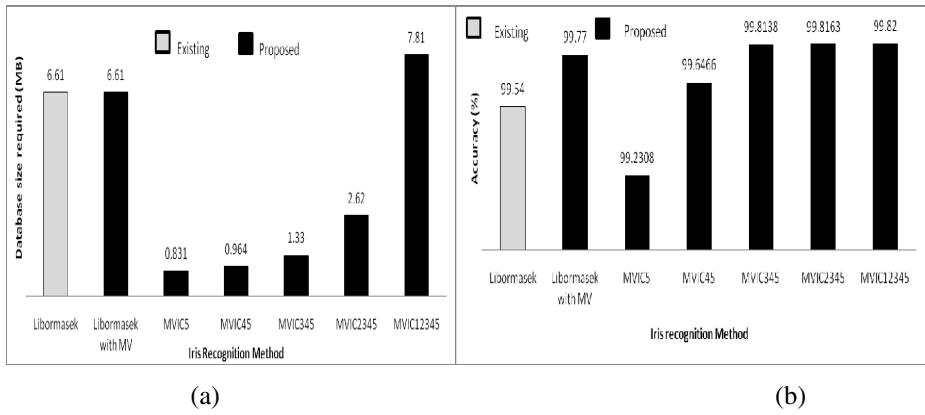


Fig. 4. (a) Comparison interms of storage space (b) Comparison interms of accuracy

Table 2. Comparisons of methods based on storage and accuracy

Method	Template/ IC size(bits)	Storage (MB)	FAR (%)	FRR (%)	TER (%)	TRR (%)	Threshold Λ
Libor Masek	4800	5.11	0.0623	0.3926	0.4549	99.5400	0.32
Libor Masek with MV	4800	5.11	0.0992	0.1298	0.2290	99.7710	0.3322
MVIC5	45	0.831	0.0625	0.7067	0.7692	99.2308	0.2667
MVIC45	225	0.964	0.0637	0.2621	0.3258	99.6166	0.2392
MVIC345	765	1.33	0.0233	0.1629	0.1862	99.8138	0.3240
MVIC2345	2565	2.62	0.0306	0.1531	0.1837	99.8163	0.4288
MVIC12345	9765	7.81	0.0404	0.1396	0.1800	99.8200	0.4557

The resultant efficient high accuracy of 99.8163 % (got from MVIC2345 method) shows that it is better than the accuracy of the methods developed by Avila et.al. [6], Ma et al. [7], Tisse et.al. [8] when implemented in the same CASIA V1 database by Vatsa et al. [13] and also better than the existing libormasek technique[2].

5 Conclusion and Future Work

In this paper, a method for reducing the storage space of the templates and to improve the accuracy of the iris recognition system is proposed. However the proposed method of majority voting and haar transform can be applied to any biometric recognition system which employs templates to reduce the storage space and to improve accuracy. But in this paper the proposed work is implemented on an Libor Masek iris recognition system [2], with the help of CASIA iris database and the experimental results show the effectiveness of the proposed scheme.

The accuracy of the implemented scheme can be further improved if the iris and pupil of the images are manually segmented, for the eye images we fail to segment, so that the successful segmentation rate can be improved from 83% to 100%.

In future work, we plan to implement our proposed work on more biometric recognition methods and variety of biometric databases with respect to time also. We expect that the proposed scheme will improve accuracy and reduce storage space in many of the existing recognition methods. Currently the proposed scheme is implemented only in the verification purpose of iris recognition system. We plan to implement our scheme in identification part of iris recognition system also.

References

1. Pato, J.N., Millet, L.I. (eds.): *Biometric Recognition: Challenges and Opportunities*. National Research Council, Whither Biometrics Committee (2010)
2. Masek, L.: *Recognition of human iris patterns for bio-metric identification*. Master thesis, The School of Computer Science and Software Engineering, The university of Western Australia (2003)
3. Daugman, J.: *New Methods in Iris Recognition*. IEEE Transactions on Systems, Man, Cybernetics 37(5), 1167–1175 (2007)
4. Wildes, R.P., Asmuth, J.C., Green, G.L., Hsu, S.C., Kolczynski, R.J., Matey, J.R., McBride, S.C.: *A machine-vision system for iris recognition*. Machine Vision and Applications 9, 1–8 (1996)
5. He, Z., Tan, T., Sun, Z.: *Towards Accurate and Fast Iris Segmentation for Iris Biometrics*. IEEE Transactions on PAMI 31(9), 1670–1684 (2009)
6. Avila, S.C., Reillo, S.R., Martin, I.D.: *Iris recognition for biometric identification using dyadic wavelet transform zero-crossing*. In: Proceedings of the IEEE 35th International Camahan Conference on Security Technology, pp. 272–277 (2001)
7. Ma, L., Wang, Y., Tan, T.: *Iris Recognition Using Circular Symmetric Filters*. In: Proceedings of the 16th International Conference on Pattern Recognition, pp. 414–417. IEEE press (2000)
8. Tisse, C.L., Torres, L., Robert, M.: *Person Identification Technique Using Human Iris*. In: Proceedings of the 15th International Recognition Conference on Vision Interface (2002)

9. Singh, N., Gandhi, D., Singh, D.P.: Iris recognition system using a canny edge detection and a circular hough transform. International Journal of Advances in Engineering & Technology 1, 221–228 (2011)
10. Chinese Academy of Sciences Institute of Automation (2004), CASIA Iris Image Database Version 1.0., <http://biometrics.idealtest.org/findTotalDbByMode.do?mode=Iris> (accessed October 2011)
11. Kingsbury, N.: The Multi-level Haar Transform. Connexions Web site (2005), <http://cnx.org/content/m11089/2.4/> (accessed October 2011)
12. SYRIS Technology Corporation, Technical document about FAR, FRR and EER, Version 1.0 (2004)
13. Vatsa, M., Singh, R., Gupta, P.: Comparison of iris Recognition Algorithms. In: Proceedings of International Conference on Intelligent Sensing and Information Processing, pp. 354–358 (2004)

Recovery of Live Evidence from Internet Applications

Ipsita Mohanty and R. Leela Velusamy

Department of Computer Science and Engineering, National Institute of Technology,
Tiruchirappalli, India

ipsita.mohanty689@gmail.com, leela@nitt.edu

Abstract. Advanced internet technologies providing services like e-mail, social networking, online banking, online shopping etc., have made day-to-day activities simple and convenient. Increasing dependency on the internet, convenience, and decreasing cost of electronic devices have resulted in frequent use of online services. However, increased indulgence in the internet by people has also accelerated the pace of digital crimes. The increase in number and complexity of digital crime cases has caught the attention of forensic investigators. The Digital Investigators are faced with the challenge of gathering accurate digital evidence from as many sources as possible. In this paper, an attempt was made to recover digital evidence from a system's RAM in the form of information about the most recent browsing session of the user. Four different applications were chosen for the experiment and it was found that crucial information about the target user such as, user name, passwords, etc., was recoverable.

Keywords: Digital forensic, Digital evidence, Live acquisition, Internet application.

1 Introduction

Digital forensics is a branch of forensic science encompassing the recovery and investigation of material found in digital devices, often in relation to computer crime [14]. It involves application of scientific methods within the regulations of law [6, 18]. At the most basic level, digital forensic is the process of acquiring, analyzing, and presenting the digital evidence [12]. Digital evidence is the information collected from digital media involved in crime, such as CDs, DVDs, flash drives, floppy disks, memory cards, mobile phones, network devices, RAM, etc., [6]. It is the basis upon which an assertion is established. Acquisition and Analysis of digital evidence has become an intensive area of research due to the increasing frequency of digital crimes across the world.

For crime investigation, the data stored in target user's system is of great significance. These data can be either static or live. Static data is stored in static storage devices such as hard disk, CDs, flash drives, etc., whereas live data is stored in RAM [6]. Live data, unlike static, changes continuously but contains the current information about the system. Any application used in a system gets loaded into RAM for operation. So, the content of RAM holds the key to information about the applications used by the user on the target system. Valuable information which can be obtained from the RAM includes the processes running, ports opened, files opened for each process, user names and

passwords of the user's accounts (created for different online applications and system log on), chat contents, e-mails, contacts, etc. Since the user names and passwords are recoverable, the investigator can log in to the respective accounts and collect more detailed information. A hit-and-trial method may be further adopted across multiple applications to check whether the same user name and password allows access or not. This enables the investigator to collect information from other online sources which had not been accessed on the target system. Thus, RAM is an important source for collection of live evidence in digital investigation and cannot be ignored.

Simon and Slay were able to retrieve live data such as communication content, communication history, contacts, passwords, and encryption keys for the application Skype [16]. In this paper, the work by Simon and Slay is extended for more diverse internet applications such as social networking, net banking, and online train reservation systems. The objective of this work is to collect relevant information about the target user from a number of websites that may have been accessed in a particular browsing session. With increasing focus to unveil digital crimes, the approach discussed in this paper acts as a potential tool for gathering live evidence from the target user.

The organization of the remaining portion of the paper is as follows. The next section provides a brief description about the basic concepts of Digital Forensic. This is followed by the Testing Procedure where the adopted methodologies are described. After methodologies, the results obtained and detailed analysis follow. The paper finally concludes with a brief discussion about the future scope.

2 Background

The process involved in Digital Forensic is split into three main phases namely Acquisition, Analysis and Report. Acquisition (imaging) is the process of creating the forensic duplicate i.e., a bit by bit copy of the digital media under investigation [3, 6, 15]. The goal of this phase is to save digital information from all sources possible [2]. However, a step which logically precedes acquisition is identification of various sources of data. Analysis, the second phase, can be defined as the in-depth systematic search of evidence [14]. The third phase, Report, involves complete description of all the actions taken in the first two phases and the conclusion drawn from analysis, so that a proper documentation of the investigation process can be submitted to the court of law. Digital Evidence, being a collection of bits, is very sensitive and can be easily altered [4, 6]. Any scientific procedure adopted during investigation should make no changes to the evidence in order to ensure its admissibility in the court. In case of any alteration due to forensic procedures, a proper explanation must be provided [12].

Acquisition can be done in two ways: Static and Live. Static acquisition involves halting the target system and making a forensically valid copy, or image, of all attached storage media whereas live acquisition involves gathering data while the system is in operation. Static acquisition has certain demerits, such as the need to shut down the system, incomplete evidence and inability to access the static media if encrypted or locked. Live acquisition makes it possible to get a running picture of the system involving information about opened applications, files, ports, running processes, user names, passwords, encryption keys, etc.; where static acquisition fails.

However, live acquisition has limitations such as need for administrator level of access, incorrect information from a compromised system, prior installation of hardware to be used such as Tribble and Firewire based devices, overwriting of some useful contents of RAM due to the software's own signature, inconsistent snapshots, and non-repeatable operations. The system state becomes a function of both user and investigator activities. In spite of such shortfalls, live acquisition cannot be avoided since it provides a plethora of information, which static acquisition cannot. Investigators should use softwares which cause as much less modification as possible because acquisition can be done only once though analysis of evidence is repeatable [3, 11]. Modifications can be accepted in critical situations as long as the investigator can clearly validate. Live acquisition is useful when the computer is on (or in standby or sleep mode or locked) and connected to a network [1]. As mentioned by Halderman et al. [9], in these situations RAM contents can be retrieved. But when the system is shut down, only static acquisition can be done. If the system is hibernated, the investigator can get RAM data by imaging the hard disk. Because after hibernation, the contents of RAM get stored in hard disk in a file called "hiberfil.sys". This file can be copied and analyzed to obtain the RAM contents [12]. Imaging of RAM can be done using different tools as discussed by Davis in [5].

The analysis of acquired evidence can be done either through live response or static memory dump analysis. The first approach involves querying the system using API-style tools such as Pslist, ListDLLs, Handle, Netstat, Fport, etc. The second approach is to gather useful data from the captured memory image in an isolated manner using different memory analysis tools such as volatility, hex editor and string extraction utilities [20]. Volatility provides command for determining the processes running, the DLLs associated with each process, the files opened for each process, the list of opened sockets etc. Hex editor can be used for manual string search. String extraction utilities can be used to extract strings from RAM image which can then be analyzed manually.

Report involves complete documentation of all processes and tools. It also summarizes the conclusions drawn in a layperson's terms [14]. Documentation cannot be considered to be an isolated or specific phase and should be done in every step of the investigation process in order to have a complete description of all steps involved and the results. The prepared document is used for verification and decision making in the court. This also helps a new investigator to understand the whole process quickly with less effort. Since only the investigator can know the evidence in raw level, the way of reporting is very crucial to ensure that others can understand the information from the report easily [4].

This was a brief description about the various steps involved in the digital forensic process. Following these steps, an attempt was made to recover and analyze useful information from browser based applications. Live acquisition was performed by collecting RAM images and analyzing them statically for evidence relevant to the applications used. The obtained information can act as a key to access the target user's profile in multiple sources and collect valuable information about the user's contact, messages exchanged, e-mails etc. In the following Section, the detailed procedure of our work is discussed.

3 Testing Procedure

The internet applications chosen for the testing were: Facebook, Gmail, IRCTC (Indian Railway Catering and Tourism Corporation Limited), and SBI(State Bank of India). The choice of the applications was based on popularity, frequency of usage and importance of contained data. The aim here was to recover vital information about the target user by leveraging on the RAM content for the most recent browsing session. The testing was carried out individually on each of the applications considered.

3.1 Test Overview

A fresh browsing session (after switching on the computer) was started with no remnant from previous sessions. Settings were modified not to save passwords and history. The application to be tested was opened in the browser and access to internet was obtained by logging into Sonicwall (a firewall interface). The next step was to take images of RAM at different time intervals, trying to cover all critical points without losing any valuable data. During acquisition only the application at issue was opened, in order to avoid alteration of relevant memory contents by other applications. This may not be the case in a real life scenario. The target user might have used more than one application and there is a probability of one application overwriting another application's data. But the testing had to be done in an isolated manner, so as to check for all the probable data that can be retrieved from the application being tested. After acquiring, the images were analyzed for contents specific to the application of concern.

3.2 Environment Setup

The system being used for testing was a Lenovo 0768 HBQ laptop with following specifications:

- OS: Windows XP Professional, Service Pack 2
- Processor: Intel Pentium Dual-Core Processor T2080 @ 1.73GHz 794MHz
- Physical Memory: 512 MB
- Hard Disk: 80 GB
- Page file size: 0MB
- Internet browser: Mozilla Firefox 6.0

The page file size was set as zero, in order to have all the contents in RAM, nothing being swapped out to the virtual memory, since the study involved taking image of only RAM.

3.3 Acquisition

For live memory acquisition, the tool 'Nigilant32' [5, 13] was used. This tool need not be installed in the target machine, but can be run from CD or external USB drive. It is just an exe file which needs to be run and has a small footprint, using less than 1 MB in memory, when loaded [5]. It took only 45 seconds to image 512MB of RAM. Although another tool (FTKImager [8]), was also available, Nigilant32 was preferred due to faster response time.

The steps followed in acquisition are:

1. Turn the system on
2. Take image of system memory-Img1.img
3. Start the browser (Mozilla Firefox)
4. Take image of system memory-Img2.img
5. Log in to Sonicwall
6. Take image of system memory-Img3.img
7. Open the application(e.g. Gmail) and log in
8. Take image of system memory-Img4.img
9. Keep the system idle for 1 minute
10. Take image of system memory-Img5.img
11. Keep the system idle for 5 minutes
12. Take image of system memory-Img6.img
13. Log out from the application
14. Take image of system memory-Img7.img
15. Close the browser
16. Take image of system memory-Img8.img
17. Keep the system idle for 1 minute
18. Take image of system memory-Img9.img
19. Log out from Sonicwall
20. Take image of system memory-Img10.img
21. Keep the system idle for 2 minutes
22. Take image of system memory-Img11.img
23. Keep the system idle for 3 minutes
24. Take image of system memory-Img12.img
25. Keep the system idle for 5 minutes
26. Take image of system memory-Img13.img
27. Shut down the system

The above sequence of steps was followed for all applications under test i.e., Facebook, Gmail, IRCTC, and SBI.

3.4 Analysis

In analysis phase, the images taken during acquisition were searched carefully to find information relevant to the application under concern. First, all strings were extracted from the images using Windows Sysinternals utility ‘Strings’ [17] and stored in different text files for different images. Then the text files were searched to find subtle hints pointing to relevant information like username, password etc. These text files were also used in the plug-in ‘strings’ of volatility [19] to know about the id of the processes, within which memory space, the strings were stored. The plug-in ‘strings’ of volatility takes as input an image file and the text file with lines of the form <offset>:<string>, usually created by Sysinternals utility ‘Strings’ for the same image, and creates a text file containing the corresponding process names (or id of the processes) and virtual addresses for the strings stored in the memory image [10]. The list of running processes while acquiring the image was generated using command ‘pslist’ of volatility and the pid associated with the searched string was matched to find out the process name.

The images can also be searched for strings using hex editor [7]. But the advantage of using Windows Sysinternals utility ‘Strings’ is that the output text file contains only printable characters, not the non-printable ones. So it is easy and clear to search.

4 Results

The primary data for search were user name and passwords, used for logging in to the applications. The results are summarized in Table 1. This table is followed by the detailed analyses with snapshots for individual applications. The user names and passwords are highlighted in each snapshot.

Table 1. Presence of Password

Application RAM image	⇒ ↓	Sonicwall	Facebook	Gmail	IRCTC	SBI
Img1		No	No	No	No	No
Img2		No	No	No	No	No
Img3		Yes	No	No	No	No
Img4		Yes	Yes	Yes	Yes	No
Img5		Yes	Yes	Yes	Yes	No
Img6		Yes	Yes	Yes	Yes	No
Img7		Yes	Yes	Yes	Yes	No
Img8		Yes	Yes	Yes	Yes	No
Img9		Yes	Yes	Yes	Yes	No
Img10		No	No	No	No	No
Img11		No	No	No	No	No
Img12		No	No	No	No	No
Img13		No	No	No	No	No

4.1 Sonicwall

The user name and password for logging into sonicwall was found in images: Img3 through Img9. The instances were in the memory space of firefox.exe. In Fig. 1, a snapshot of the text file img9.txt, created from Img9.img by ‘Strings’ utility is given, which contains the url of the website, the session id, user name and password of the user. Img3 was acquired after logging into the interface and Img9, just before logging out. After logging out from Sonicwall, the firefox.exe process gets closed. So the contents were not found in the images acquired after that i.e., Img10-Img13.

```

257728512:<HTML>
257728519:<HEAD><TITLE>Page Redirecting</TITLE>
257728557:<META HTTP-EQUIV="Pragma" CONTENT="no-cache">
257728603:<META HTTP-EQUIV="Expires" CONTENT="-1">
257728644:</HEAD>
257728652:<BODY onLoad="top.location.href =
    'http://192.168.20.1/userLogin.html';">
257728726:This page is redirecting! Click <A HREF=
    "http://192.168.20.1/userLogin.html">here</A>
257728812:</BODY>
257728820:</HTML>
257728828:on: keep-alive
257728844:Referer: https://192.168.20.1/auth1.html
257728866:Cookie: temp=temp; SessId=523518834; PageSeed=7e88fbfc81a9;
257728966:Content-Type: application/x-www-form-urlencoded
257729015:Content-Length: 122
257729038:param1=&param2=
    93BF844DF6D46F0F1453F46441968A46&sessId=523518834
    &id=a4&select=English&uName=306110003&pass=Nitt5008&digest=
257732617:t#hp

```

Fig. 1. Snapshot of text file from Img9 taken for Sonicwall

4.2 Facebook

After searching the images acquired for Facebook, it was found that the user name and password for log in were present in images: Img4 through Img9. The username and password were preceded by the words ‘email’ and ‘pass’ which can be used as keywords for search.

As in Fig. 2, the value set for user name is ‘ipsita.chinky@gmail.com’ and for password is ‘who678%2C%3B’. The actual password entered was ‘who678,;’. It could be concluded that the special symbols were converted into corresponding ASCII hex values, resulting in ‘,’ as ‘%2C’ and ‘;’ as ‘%3B’. Hence while searching for passwords; care should be taken for the ASCII values stored. If the password contains letters from A-Z, a-z and numbers, no special characters, it could be easily identified.

It was observed that the username and password were available after logging out from facebook (Img7.img) and also after closing the browser (Img8 and Img9). However, the username and password was not found in the further images (Img10-Img13). This can be attributed to the fact that, after logging out from Sonicwall, the internet access permission got aborted and the Firefox window used for showing the user status information for Sonicwall closed. So the process Firefox terminated completely resulting in the absence of the relevant data in images Img10-Img13.

Other useful information (except user name and password) like profile name, update dates of the target user’s friends, etc were also available for retrieval because the loaded pages were stored in RAM. The contents including user name and passwords were mostly in the memory space of firefox.exe and very few were in svchost.exe and kernel process.

```

img4.txt - Notepad
File Edit Format View Help
76140665:Location: https://www.facebook.com/checkpoint/
76140713:P3P: CP="Facebook does not have a P3P policy. Learn why here:
http://fb.me/p3p"
76140794:Pragma: no-cache
76140812:Set-Cookie: _e_41AX_0=deleted; expires=Thu, 01-Jan-1970 00:00
76140922:Set-Cookie: _e_41AX_1=deleted; expires=Thu, 01-Jan-1970 00:00
76141032:Set-Cookie: _e_41AX_2=deleted; expires=Thu, 01-Jan-1970 00:00
76141142:Set-Cookie: checkpoint=%7B%22u%22%3A100001759385636%
76141456:Set-Cookie: dat=URxWTwBwwh1_54FeaDLy-gi; expires=Sat,
24-Aug-2013 09:56:58 GMT; path=/; domain=.facebook.com; httpo
76141578:Set-Cookie: L=2; path=/; domain=.facebook.com; httponly
76141635:Set-Cookie: W=1314266218; path=/; domain=.facebook.com
76141691:Set-Cookie: wd=deleted; expires=Thu, 01-Jan-1970 00:00:01
GMT; path=/; domain=.facebook.com; httponly
76141794:Content-Type: text/html; charset=utf-8
76141834:X-FB-Server: 10.32.189.112
76141862:X-Connection: close
76141881:Date: Thu, 25 Aug 2011 09:56:58 GMT
76141918:Content-Length: 0
76141939:3Ow&locale=en_US;email=ipsita.chinky%40gmail.com
8pass=who678%2C%3B8 default_persistent=0&charset_test=0%E
76143592:8h,
76143624: h,

```

Fig. 2. Snapshot of text file from Img4 taken for Facebook

4.3 Gmail

The username and password were retrievable from Gmail in a similar fashion to facebook. The details were available in images Img4 through Img9. The username and password were preceded by the words ‘GAUSR=mail’ and ‘Passwd’ which can be used as keywords for search.

A string, ‘abc*%21123’, very much similar to the entered password, ‘abc*!123’, was obtained for Gmail (Fig. 3). It was found in images Img4 through Img9. In the password string, the special character ‘!’ was converted into its ASCII hex value ‘21’. Thus, it was observed that if there were two hexadecimal digits after ‘%’, the hexadecimal number should be converted to the associated special character.

In addition to username and password, other information like inbox contents and contacts were found. After logging into the account, the first page loaded contains inbox contents and some contacts available for chat. This ensured that the most recent inbox content and frequently used contacts could be retrievable. Similar to Facebook, the contents for Gmail were found in RAM while being logged into Sonicwall and stored in the memory space of firefox.exe, svchost.exe and kernel process.

4.4 IRCTC

For the application IRCTC, the user name and password were readily available and were found in images: Img4 through Img9. A snapshot highlighting user name and password is shown in Fig. 4. The username and password were preceded by the words ‘userName’ and ‘password’ which can be used as keywords for search. The explicit keywords made the search for username and password very easy. Moreover, since there were no special characters used in the password, the password was available exactly without any encoding. The instances were in the memory space of firefox.exe and very few in that of svchost.exe and kernel process.

```

File Edit Format View Help
377401745:qRW8I
377401842:=@)
377401847: SHKtU0htqqZw%26gausr%3Dipsita.chinky%254@gmail.com
377401900:Content-Encoding: gzip
377401924:Date: Tue, 23 Aug 2011 18:05:37 GMT
377401961:Expires: Tue, 23 Aug 2011 18:05:37 GMT
377402001:Cache-Control: private, max-age=0
377402036:X-Content-Type-Options: nosniff
377402069:X-XSS-Protection: 1; mode=block
377402102:Content-Length: 674
377402123:Server: GSE
377402142:Set-Cookie: LSID=mail1s.IN:DOAAAL0AAACNKnOxFIOEmQaAp
377402469:Set-Cookie: GAUSR=mail:ipsita.chinky@gmail.com
  Path=/accounts,secure
377402539:Location: https://accounts.google.co.in/accounts/SetsSID?
  sode=1&sid=ALWU2es0p2F1jKI0%2BfeR3yEy22NCywE05YSVI
  &Passwd=abc*%21123&rmShown=1&signIn=Sign+in&asts=
377410405: BI
377410485: BI
377410806:fff

```

Fig. 3. Snapshot of text file from Img4 taken for Gmail

```

File Edit Format View Help
83705907:Date: Fri, 26 Aug 2011 08:59:21 GMT
83705944:Server: Microsoft-IIS/6.0
83705971:X-UA-Compatible: IE=EmulateIE7
83706003:X-Powered-By: ASP.NET
83706026:Content-type: text/html
83706051:Location: https://www.irctc.co.in/cgi-bin/bv60.dll/
  irctc/booking/planner.do?screen=fromlogin&
  BV_SessionID=@@@@1286877822.1314349161@@@@&
  BV_EngineID=ccdgadfehghkgdjcefecihdfgmdfh.n
83706236:s://www.irctc.co.in/
83706258:Cookie: __utma=168397561.1365555935.1314349144.
  .1314349144.1314349144.1; __utmb=168397561.1.10.1314349
83706461:Content-Type: application/x-www-form-urlencoded
83706510:Content-Length: 59
83706532:screen=home&userName=ipsita689&password=iqd340
  &button=Login
83706944:m<zN
83707039;i:*

```

Fig. 4. Snapshot of text file from Img5 taken for IRCTC

As explained for Gmail and Facebook, there were no presence of user name and password in images: Img10 through Img13 i.e., after logging out from Sonicwall.

4.5 SBI

The same experimentation procedure was carried out for the internet banking site of State Bank of India. The password for logging into the application was not found in

any of the memory images taken. The user name was seen as an isolated string, with no element of the log in page present nearby. On the contrary, the user name and password for Gmail, Facebook and IRCTC were near the website address. With no string related to the application page loaded and no preceding keywords to help in identification of the user name, it was difficult to identify the string as user name. However the isolated string was in the memory space of Firefox.exe process opening up a possibility of association. But this is possible only when one application is opened. It will be a difficult task to figure this out if more than one application were used by the target user.

Apart from the user name, other useful information available was: account number, name of the account holder, bank branch code, name and branch of the bank. These contents were present only in images: Img4 through Img9. In the memory images taken after that, only the name of the website ‘www.onlinesbi.com’ was found. This shows that the online banking site is much better protected from acquisition as compared to the other websites.

5 Conclusion and Future Work

The approach followed in this paper is relevant to the existing global scenario where acquiring digital evidence holds primary importance in any investigation. Every browsing session of the target user leaves an imprint in the system memory and this has been exploited in this approach. It was possible to extract useful information from the memory images taken after the use of the application (without internet being severed). The application names or the words (different for different applications) preceding username and passwords, can be used as keywords for search. The information found out during analysis were username, password, list of contacts, mails, bank account number, name of the account holder etc.

However, it was observed that the information was not available in the memory images taken after logging out of the firewall. Information can be retrieved till the user is logged into the firewall and is connected to internet. This represents the case of Live Acquisition wherein plenty of information can be retrieved about the state of the system in the recent past while the system is still logged into the firewall. Despite some limitations of live acquisition, it is impossible to ignore the importance of the contents of RAM. The utility of the approach is definitely on the higher side and is likely to find applications in a number of cases.

In this paper, only a single browser is taken into consideration for conducting the experiments. However, it is probable that every browser would have its own default security and privacy settings which may give rise to different results. The experiment can be extended to include a number of browsers to provide a more comprehensive conclusion to the results that have been obtained for a single browser. Moreover, browsers are getting upgraded regularly and each version would have its own specific settings and features. The experiment can further include tests across various versions of every browser. This would ensure that the results are consistent across a number of versions of a number of browsers. That would give a higher probability of retrieval of evidence irrespective of the browser and version the target user is using.

References

1. Caloyannides, M.A.: Forensics Is So "Yesterday". *IEEE Security and Privacy* 7(2), 18–25 (2009), doi:10.1109/MSP.2009.37
2. Carrier, B.: Open Source Digital Forensics Tools: The Legal Argument, @stake Research Report (2002), http://www.digital-evidence.org/papers/opensrc_legal.pdf
3. Carrier, B.D.: Digital Forensics Works. *IEEE Security and Privacy* 7(2), 26–29 (2009), doi:10.1109/MSP.2009.35
4. Cohen, F.B.: Fundamentals of Digital Forensic Evidence. In: Stavroulakis, P.P., Stamp, M. (eds.) *Handbook of Information and Communications Security*, 1st edn., pp. 789–808. Springer (2010), doi:10.1007/978-1-84882-684-7
5. Davis, N.: Live Memory Acquisition for Windows Operating Systems: Tools and Techniques for Analysis. IA 328, Eastern Michigan University, USA, <http://www.emich.edu/ia/pdf/research/Live%20Memory%20Acquisition%20for%20Windows%20Operating%20Systems,%20Naja%20Davis.pdf>
6. Fei, B.K.L.: Data Visualization in Digital Forensics. Dissertation, University of Pretoria, South Africa (2007), <http://upetd.up.ac.za>
7. Free Hex Editor Neo Version 4.97.01.3661. HHD SOFTWARE (March 2011), <http://www.hhdsoftware.com/free-hex-editor>
8. FTK Imager Lite Version 2.9.0. AccessData (June 2010), <http://accessdata.com/support/adownloads#FTKImager>
9. Halderman, J.A., Schoen, S.D., Heninger, N., Clarkson, W., Paul, W., Calandrino, J.A., Feldman, A.J., Appelbaum, J., Felten, E.W.: Lest We Remember: Cold Boot Attacks on Encryption Keys. In: 17th USENIX Security Symposium (Sec 2008), pp. 45–60 (2008)
10. Command Reference: Example usage cases and output for Volatility commands, <http://code.google.com/p/volatility/wiki/CommandReference#strings>
11. Hay, B., Bishop, M., Nance, K.: Live Analysis: Progress and Challenges. *IEEE Security and Privacy* 7(2), 30–37 (2009), doi:10.1109/MSP.2009.43
12. Mrdovic, S., Huseinovic, A., Zajko, E.: Combining Static and Live Digital Forensic Analysis in Virtual Environment. In: XXII International Symposium on Information, Communication and Automation Technologies (ICAT 2009), pp. 1–6. IEEE Press (2009), doi:10.1109/ICAT.2009.5348415
13. Nigilant32: windows after dark forensic. Agile Risk Management LLC, http://www.agileriskmanagement.com/publications_4.html
14. Reith, M., Carr, C., Gunsch, G.: An examination of digital forensic models. *International Journal of Digital Evidence* 1(3), 1–12 (2002)
15. Sammes, T., Jenkinson, B.: *Forensic computing: a practitioner's guide*. Springer (2000)
16. Simon, M., Slay, J.: Recovery of Skype Application Activity Data From Physical Memory. In: Fifth International Conference on Availability, Reliability and Security (ARES 2010), pp. 283–288. IEEE Press (2010), doi:10.1109/ARES.2010.73
17. Russinovich, M.: Strings - Windows Sysinternals (2009), <http://technet.microsoft.com/en-us/sysinternals/bb897439>
18. Vacca, J.R.: *Computer forensic: computer crime scene investigation*. Charles River Media (2002)

19. Volatility 2.0, The Volatility Framework: Volatile memory artifact extraction utility framework (2011), Volatile Systems,
<https://www.volatilesystems.com/default/volatility>
20. Waits, C., Akinyele, J.A., Nolan, R., Rogers, L.: Computer Forensics: Results of Live Response Inquiry vs. Memory Image Analysis. Technical Note, CERT Program, United States: Software Engineering Institute, Carnegie Mellon University (2008),
<http://www.sei.cmu.edu/library/abstracts/reports/08tn017.cfm>

Face Detection Using HMM –SVM Method

Nupur Rajput, Pranita Jain, and Shailendra Shrivastava

Smrat Ashok Technological Institute, Vidisha, Madhya Pradesh, India
{nupurrajput, pranita.jain}@gmail.com,
shailendrashrivastava@rediffmail.com

Abstract. This paper proposes a method for face detection and recognition using Modified Hidden Markov Model (HMM) and Support Vector Machine (SVM). It is a two layer architecture system that identifies all image regions which contain face or non-face. At the first stage, the Kernel HMM classifies input pattern into three classes: a face class, undecided class or non-face class. In the final stage, SVM detects the face class or non-face class if any sub-image falsely judged as undecided class. This system alleviates the problem of false positive rate. The experimental result shows that the proposed approach outperforms some of the existing face detection methods and we have compared various face detection method.

Keywords: Face detection; Support Vector Machine; Hidden Markov Model, Kernel HMM.

1 Introduction

The face detection and recognition system has drawn considerable importance from researchers for many decades. The SVM is a learning technique based on statistical learning theory that has been applied successfully in variety of application and regression problems [3, 5, 15, 14]. Support vector machine with a binary tree recognition strategy is proposed to tackle the multi-class face recognition problem that gives the pair wise discrimination capability of SVM to the multi class scenario [16]. The stability of SVM in classification task is dealt by decomposing the average prediction error of SVM into the bias and variance terms [2]. The modified kernel based FDA method applied for face detection with low confidence of classification [18]. Hanumantha Reddy, Karibasappa K and Damodaram A presented a distribution based FDA-SVM method for face detection, but the amplitude projections of face and non face resembles the same thus reducing the detection accuracy [10]. The hierarchy quarantine method is basis for face and non face classification and hierarchy classifiers analyze the whole image for face class identification and reject the non face image [12, 13]. Character recognition with SVM by using kernel and feature extraction on pair of letters is proposed in [4, 9]. In the paper [11], the visual features of the face like eye, nose etc is modeled for facial expression recognition using SVM. The combined global and local approach for feature extraction is used for face recognition with SVM [17]. It uses linear subspace projection and Euclidean norm for computational speed up for large database. The Bayesian Expectation inference algorithm is proposed for binary classification in the paper [11] by Daniel L and Herandez Lobatto. The component and

two global approaches are presented in this paper [10] for face recognition. In component system, the facial features extracted and combined them into single vector which is used by SVM for recognition. In global approaches, the face image is feature vector which is used by SVM for recognition. The main motivation of this paper is to diminish the false positive rate of face detection system. The two parameters that decides the efficacy of the system are high positive detection rate and low or nearest to zero percentage of false positive rate. In this paper, a novel Kernel HMM method is proposed for face detection by kernel based Hidden Markov Model (HMM) and Support Vector Machine (SVM) method which reduces the dimensionality of the input image by preserving the intrinsic information. The method statistically estimates the sub image is face or non-face class. The non-face or undecided class is given as input to the SVM for classification of face or non-face class. Finally, HMM-SVM classifies the undecided class as face or non-face class. The SVM classifier is used to recognize the face with optimal hyper plane but recognition accuracy is hindered for large database. Section 2 describe the Modified HMM, SVM and architecture of the proposed face detection system. Experimental results are described in the section 8. Conclusions are given in the section 4.

2 Face Detection Using Modified Kernel Based HMM and SVM

The flow graph of the face detection system is shown in the Fig.1. The input image is first processed by Kernel based Hidden Markov Model (HMM). The multivariate distribution

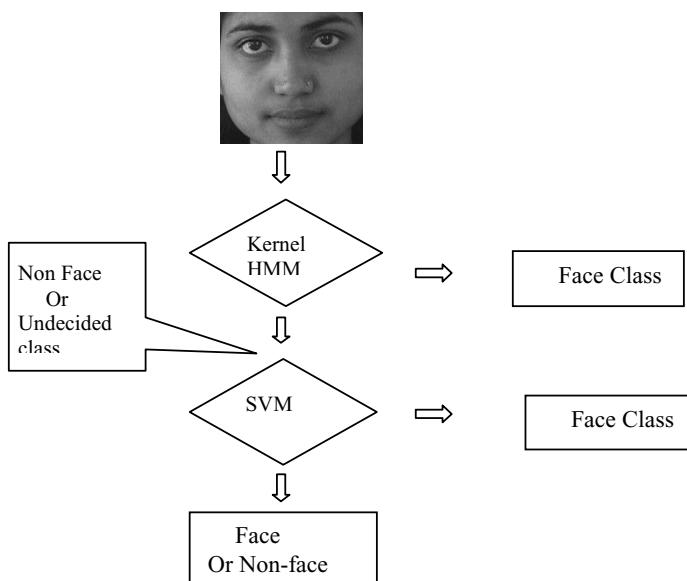


Fig. 1. Flow graph to detect the face in the input Image

measure of the feature space is used for classifying the input image as face or non face class. In the final stage, the HMM-SVM classifies the undecided class of the input image as face or non face image. This system alleviates the problem of false positive rate.

2.1 Kernel Hidden Markov Model (K-HMM)

In face detection, it is important to design a two category classifier which can detect whether the given input sub images is face or non face class. The classifier is trained with face and non-face images. It is proved that the Kernel based Hidden Markov Model cannot be applied to classification task because it cannot compensate high dimensional space. In this paper we have used the HMM as classifier independently and also in combination with Support Vector Machine. The Modified Kernel HMM is proposed to apply it to classification task. Hidden Markov Model (HMM) are a set of statistical models used to characterize the statistical properties of a signal [9]. HMM consist of: (1) An underlying un-observable Markov chain with a finite number of states, (2) a state transition probability matrix and an initial state probability distribution. (3) A set of probability density functions associated with each state. here was problem occurred in simple HMM and in KFDA of complexity. KFDA is the modification of FLDA by using the Kernel function that also performed well as a non linear classifier. Now we have applied the kernel function with the HMM and that performed very well. In which we have passed the HHM from the SVM Kernel.

Elements of HMM are:

1. N is the number of states in the model. If S is the states , then $S = \{S_1, S_2, \dots, S_N\}$ The state of the model at time t is given by $q_t \in S$, $1 \leq t \leq T$, where T is the length of the observation sequence (Number of frames).
2. \prod the initial state distribution , i.e. $\prod = \{\pi_i\}$ where:

$$\pi_i = P[q_1=i], 1 \leq i \leq N$$

3. State transition probability matrix, i.e.

$$A = \{a_{ij}\} \text{ where}$$

$$a_{ij} = P[q_t = S_j | q_{t-1} = S_i] \quad 1 \leq i, j \leq N$$

With the constraint

$$0 \leq a_{ij} \leq 1$$

And,

$$\sum_{j=1}^N a_{ij} = 1, 1 \leq i \leq N$$

4. the state probability matrix , i.e. $B = \{b_i(O_t)\}$ in a continuous observation functions. The most general representation of the model probability density function (pdf) is a finite mixture of the form :

$$b_i(O_t) = \sum_{k=1}^N c_{ik} (N o_t, u_{ik}), 1 \leq i \leq N$$

Where C_{ik} is the mixture coefficient for the Kth mixture in state i.without loss of generality $N(O_t, \mu_{ij}, U_{ik})$ is assumed to be Gaussian pdf with mean vector μ_{ij} and covariance matrix U_{ik} .

Using short hand notation, a HMM is defined as triplet-

$$\lambda = (A, B, \Pi)$$

2.2 Support Vector Machine (SVM)

SVM is a two class realization of statistical realization statistical learning theory. It describes an approach known as structural minimization [3, 8]. The main idea of SVM comes from the mapping of input space to high dimensional space and designing optimal hyper plane in terms of margin.

Let $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k) \in \mathbb{R}^n$ and $y_i \in \{-1, +1\}$ be the k training samples in the input space, where y_i indicates the class membership of x_i . Let ϕ be non-linear mapping

$$X \rightarrow \phi(x).$$

The optimal hyper plane is defined as below.

$$W_0 \Phi(x) + b_0 = 0$$

It is proved [6] that the vector w_0 is linear combination of weight vector which are vectors x_i that satisfy

$$y_i (w_0 \cdot \Phi(x_i)) + b_0 = 1$$

$$W_0 = \sum_{\text{support vectors}} y_i \alpha \Phi(x_i)$$

The linear decision function,

$$f(x) = \text{sign}(\sum_{\text{support vectors}} y_i \alpha \Phi(x_i) \cdot \Phi(x) + b_0)$$

To increase the efficiency of the operation the dot product is replaced by kernel Gaussian function

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

$$K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2)$$

3 Face Image HMM

Hidden Markov Model has been successfully used for speech recognition and more recently in action recognition and face recognition also. Where data is one dimensional over time. In this paper we investigate the face detection and recognition performance of one dimensional Kernel based HMM with SVM with gray scale face images. For frontal face images the significant facial regions (hair, forehead, eyes, nose, and mouth) come in a natural order from top to bottom even if the images go small rotation in the image plane and/or rotations in the plane perpendicular to the image plane. Each of these facial

regions is assigned to a state in a left to right 1D continuous HMM. The state structure of the ace model and non zero, transition probabilities a_{ij} are shown in figure 2.

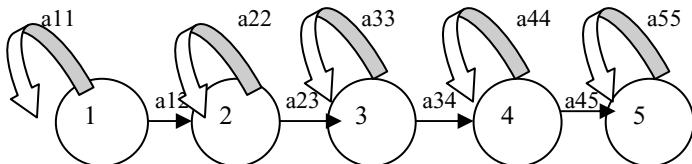


Fig. 2. Left to Right HMM for race recognition

4 Feature Extractions

In this paper we used clustering to extract the observation vector. We have taken 5 clusters (hair, forehead, eyes, nose, and mouth). In the [7] observation vectors consist of all the pixel values from each of the blocks, and therefore the dimension of the observation vector is $L \times W$. The use of pixel values as observation vectors has two important disadvantages: First, pixel values do not represent robust features, shift or changes in illumination. Second, the large dimension of the observation vector leads to high computational complexity of the training and detection/recognition system is used for real time applications. In [6], the 2D-DCT coefficients extracted from each block were used to obtain an efficient set of observation vectors. The co-efficient inside a rectangular window over the lowest frequencies in the DCT domain, which concentrate most of the block energy, were used as observation vectors. The choice of the 2D-DCT coefficients reduced dramatically the size of the observation vectors and, therefore decreased the complexity of the system. In this paper, the observation vectors consist of the Discrete Wavelet Transform (DWT) coefficients. The DWT compression properties as well as de-correlation properties make it an alternative technique for the extraction of the observation vectors. From the images in the training set, the blocks are extracted and arranged column wise to form vectors. The eigenvectors corresponding to the largest Eigen-values of the covariance matrix of these vectors from the DWT basis. To obtain the observation vectors, the mean vector μ is subtracted from each of the vectors corresponding to a block in the image. The resulting vectors are then projected onto the eigenvectors of the covariance matrix and the resulting coefficient from the observation vectors.



Fig. 3. Feature Extraction

5 Training the Face Model

For face detection, a set of face images is used in the training of one HMM. The images in the training set represent frontal faces of different people taken under different illumination conditions.

For face recognition, each individual in the database is represented by an HMM face model. A set of images representing different instances of the same face are used to train each HMM.

In this paper, we have used Kernel HMM as we have kept HMM code in the SVM kernel by which we have outperformed the existing method by making new method KHMM-SVM. For extracting each image in the training set, the observation vectors (DWT coefficient) are obtained and used to train each of the HMMs. First, the HMM $\lambda = (A, B, \Pi)$ is initialized as follows. The training data is uniformly segmented from top to bottom in $N = 5$ states, and the observation vectors associated with each state are used to obtain initial estimate of the observation probability matrix B. The initial values for A and Π are set given the left to right structure of the face model.

6 Experimental Results

The accuracy of the face detection depends on the coefficient chosen for the system. We have Discrete Wavelet Transform (DWT) for the feature extraction. The face detection/ Recognition system has been tested on the ORL database (400 images of 40 individual, 10 face images per individual at the resolution of 92×112 pixels). Half of the images were used in training, and the other half were used for testing. We have also used non-face database to separate it from face database and have tested successfully. The database contains face images showing different facial expressions, hair style, and eye wear (glasses or no-glasses). On the same database the recognition performance of the method [2] was 86% using HMM and in [1] was 93% using FDA-SVM. The accuracy of the system presented in this paper is increased to 95%.

The comparison of the face detection rate of the different rates of the different methods on ORL database to the proposed system is shown in the table 1.

$$\text{Accuracy} = \text{No. of face detected} / \text{total no. of face} * 100$$

Table 1. Comparison of the different face detection system

Si. No.	Method	Detection Rate %	False Positive
1	HMM	86%	4
2	FDA-SVM	93%	6
3	HMM-SVM	95%	4



Fig. 4. Showing Face detection Result

The HMM-SVM method successfully identifies the faces in the probe image. Fig.3 and fig.4 shows the sample of facial and non-face images used in the experiment and the face detected by propose method.

7 Conclusions

This paper describes Kernel HMM approach for face recognition and detection that uses an efficient set of observation vectors based on the extraction of the DWT coefficient. This method reduces the computational complexity of simple HMM and KFDA. The accuracy of this method with respect to variation in lighting conditions and its complexity efficiency, suggest that this method may be a promising approach for face detection. The Kernel HMM modelling method for face recognition and face detection under a wider range of image orientation and facial expression.

References

1. Hanumantha Reddy, T.: Face detection using modified FDA-SVM method. International Journal of Machine Intelligence 1(2), 26–29 (2009) ISSN: 0975-2927
2. Face detection and recognition using hidden Markov model: central for Image prosseing model school of electrical and computer science engineering. Ara. v nafin GA30332
3. Burges, C.J.C.: Data Mining Knowledge Discovery 2, 121–167 (1998)

4. Malon, C., Uchida, Suzuki, M.: PRL 29, 1326–1332 (2008)
5. Tax, D.M.J., Duin, R.P.W.: PRL 20(11), 1191–1199 (1999)
6. Nefian, A.V., Hayes, M.H.: Hidden Markov models for face recognition. In: ICASSP 1998, pp. 2721–2724 (1998)
7. Samaria, F., Young, S.: HMM based architecture for face identification. Image and Computer Vision 12, 537–583 (1994)
8. Tang, F., Chen, M., Wang, Z.: JSEE 17(1), 200–205 (2006)
9. Rabiner, L., Huang, B.: Fundamentals of speech recognition. Prentice-Hall, Englewood Cliffs (1993)
10. Heisele, B., Ho, P., Wu, J., Poggio, T.: CVIU 91, 6–21 (2003)
11. Geetha, A., Ramalingam, V., Palanivel, S., Palanippan, B.: Expert System with Applications 36(1), 303–308 (2009)
12. Sahbi, H., Boujemaa, N.: 16th International Conference on Pattern Recognition in IEEE Xplore, pp. 359–362 (2002)
13. Ng, J., Gong, S.: Image and Vision Computing 20(5-6), 359–368 (2002)
14. Muller, K.R., Mika, S., Ratsch, M., Tsuda, K., Scholkopf, B.: IEEE Trans. on NN 12(2), 181–201 (2001)
15. Tang, F., Chen, M., Wang, Z.: JSEE 17(1), 200–205 (2006)
16. Sugiyama, M.: 23rd International Conference on Machine Learning, pp. 905–912 (2006)
17. Abeni, P., Baltatu, M., D'Alessandro, R.: 3rd Canadian Conference on Computer and Robot Vision, vol. 42 (2006)
18. Shih, P., Liu, C.: PR 39(2), 260–272 (2006)

High Capacity Lossless Semi-fragile Audio Watermarking in the Time Domain

Sunita V. Dhavale, R.S. Deodhar, and L.M. Patnaik

Defence Institute of Advanced Technology, Girinagar, Pune-411025, India
sunitadhadavale75@rediffmail.com, {rsdeodhar, lalit}@diat.ac.in

Abstract. A blind high capacity lossless semi-fragile audio watermarking algorithm based on the statistical quantity related to the correlation among the audio sample values is proposed. Time domain embedding is used to reduce the computational time in searching the synchronization codes. The watermark is embedded into the non-silent high energy frames (HEF) to take advantage of the perceptual properties of the Human Auditory System (HAS) and to improve the transparency of the digital watermark. The Offset value used for embedding is made adaptive to the required SNR for the final watermarked audio signal. The watermark can be removed using a secret watermarking key with only minimal remaining distortion. The method proposed is media format independent and it can be used with lossy compression. Both subjective and objective tests reveal that the proposed watermarking scheme maintains high audio quality and is simultaneously highly robust to pirate attacks, including MP3 compression, cropping, time shifting, filtering, resampling, and re-quantization., and re-quantization.

Keywords: Audio watermarking, digital rights management, Lossless, Reversible, Self Synchronization, Time Domain, Semi-fragile, Blind.

1 Introduction

Due to outstanding progress of digital audio technology, ease of reproducing and retransmitting digital audio has been greatly facilitated. Hence there is a need for the protection and enforcement of intellectual property rights for digital media. Digital watermarking is one of the promising ways to meet this requirement. The primary objective of digital watermarking is to hide the copyright information (e.g. owners/company name, logo etc.) into a multimedia object, without disturbing the perceivable quality of the content [1].

Watermarking of audio signals is more challenging compared to the watermarking of images or video sequences due to wider dynamic range of the human auditory system (HAS) in comparison with the human visual system (HVS). Two properties of the HAS dominantly used in audio watermarking algorithms are frequency masking and temporal masking.

According to the International Federation of the Phonographic Industry (IFPI), SNR of watermarked audio signal should be always greater than 20 dB. The embedded watermarks should not be removed or degraded using common audio processing techniques. The watermark embedding process should be faster, so that integrated

watermarking functionality can be enabled in the delivery of an audio over a network. Also it should support fast watermark detection in order to authenticate audio objects, delivered over the networks. According to the IFPI, there should be more than 20 bits per second (bps) data payload for watermark. These requirements present great challenges. Existing audio watermarking techniques are broadly categorized into time domain and transform domain techniques [2]. Time domain techniques [4] are simple to realize, but they are less robust compared to transform domain techniques. In transform domain techniques, the host signal is transformed into the transform domain and the watermark is embedded into the transform domain coefficients [3]. Both techniques are further classified as blind or non-blind schemes. In blind audio watermarking methods, watermark is extracted directly from the watermarked audio signals, without having the knowledge of original host audio signals.

Although reversible watermarking has been widely researched [5-7], the research has mainly been focused on imperceptible reversible watermarking of images only. Reversible watermarking enables the embedding of useful information in a host signal without any loss of host information. In [7], the watermark is embedded into selected FFT coefficients' magnitudes of the cover audio using frequency hopping method. However, the embedded watermark is perceptible unless it is removed and original audio is extracted using watermarking secret key.

As an extension of our previous work [8] in order to retrieve original audio back after extraction of embedded watermark data, we propose a high capacity removable semi-fragile watermarking method for an audio copyright protection. In case of semi-fragile authentication if the marked audio does not change at all, the hidden data can be extracted out, and the original audio can be recovered exactly. On the other hand, if the marked audio goes through compression to some extent, the hidden data can still be correctly extracted for semi-fragile authentication. Semi-fragile authentication may be more practical than fragile authentication since it allows some incidental modification like mp3 compression, re-sampling or Amplitude changing up to certain extent [5-7]. Audio is segmented in small sized frames in order to achieve efficient processing as well as high embedding capacity. This also makes the system useful for wide variety of watermarking applications [10]. The watermark is embedded into the non-silent high energy frames (HEF) to take advantage of the perceptual properties of the Human Auditory System (HAS) and to improve the transparency of the digital watermark. The Offset value used for embedding is made dependent on the mean value of energy of a given audio signal in order to increase imperceptibility. The watermark can be removed using a secret watermarking key and original audio data can be retrieved with only minimal remaining distortion. Also in order to resist synchronization attack effectively watermark is embedded along with synchronization codes in the time domain directly. This also reduces search time for detection of the synchronization codes. The proposed watermarking method does not require the use of the original signal for watermark detection. Thus, the copyright owner does not have to perform data search in his archives prior to detection. Such searches are very time consuming and render watermarking useless for audio monitoring in digital broadcasting. The binary watermark logo image is first permuted using Arnold Transform in order to increase the secrecy of embedded watermark. To achieve better error detection and correction capability, the combined watermark bit-stream is replicated according to the redundancy factor (r) specified.

The outline of the paper is as follows. Section 2 provides the outline of the proposed algorithm for embedding and extraction of binary image in an audio signal. Experimental results are compared with the results of previous works in Section 3; followed by the conclusions in Section 4.

2 Proposed Scheme

The proposed scheme consists of watermark processing stage and audio processing stage as shown in Figure 1.

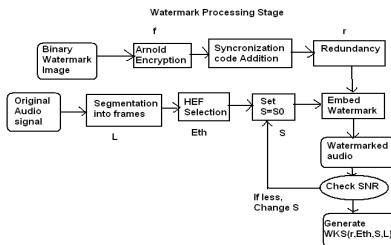


Fig. 1. Proposed system for audio watermarking

The detailed procedure in case of watermark embedding process is as follows,

2.1 Watermark Embedding Process

Step 1: In watermark processing stage, the binary logo image is first permuted by the Arnold Transform in order to enhance the security of the system. Arnold Transform is the image transformation technique where the pixels of the image are scrambled [8]. Due to the periodicity of the transform, the image can be recovered after transformation. If watermark image is of size $N \times N$ and (x, y) is the coordinate of the image pixel, then (x', y') is the coordinate after applying the transform and given by,

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \pmod{N} \quad (1)$$

Scrambling the watermark before embedding guarantees the embedded watermark will be robust against attacks like clipping, resampling etc. So even if an attacker detects the watermark, he cannot recover the original watermark without the knowledge of scrambling algorithm and the parameters used for the scrambling.

Let W denote permuted binary watermark of size $m \times n$ representing the watermark logo;

$$W = \{w(i, j), 0 \leq i \leq m, 0 \leq j \leq n\} \quad (2)$$

where $w(i, j) \in \{0, 1\}$ are the respective pixel intensities of the watermark image at coordinate (i, j) .

Step 2: The resulting permuted binary image is converted into one dimensional array containing series of 0's and 1's (bits). To enhance the error detection capability of the system, this bit stream is replicated as specified by the redundancy factor (r).

Step 3: The proposed scheme embeds 16 bit Barker code as synchronization code to locate the position of hidden informative bits, thus resisting the cropping and shifting attacks [3]. Barker codes are subsets of PN sequences and used for frame synchronization in digital communication systems. They have low correlation side lobes. A correlation side lobe is the correlation of a codeword with a time-shifted version of itself [3]. The correlation side lobe C_k for a k-symbol shift of an N-bit code sequence $\{x_j\}$ is given by,

$$C_k = \sum_{j=1}^{N-k} x_j x_{j+k} \quad (3)$$

where x_j is an individual code symbol taking values +1 or -1 for $j = 1, 2, \dots, N$ and the adjacent symbols are assumed to be zero.

Step 4: In audio processing stage, the original host audio is first segmented into non-overlapping audio frames of size L samples. For convinience, we take L as an even number. If X denotes the original audio .wav signal having size N, then

$$X = (x(1), x(2), \dots, x(N)) \quad (4)$$

where, $x(i) \in (-1.0, +1.0)$ are respective sample values normalized in the given range.

Then the audio segments are given as,

$$Y_k = (X_{L(k-1)+1}, X_{L(k-1)+2}, \dots, X_{Lk}) \quad (5)$$

where $k = 1, 2, \dots, N_s$ and N_s =total number of audio segments=N/L.

The embedding capacity also depends on this N_s , as each frame can embed one bit of watermark

Step 5: As distortion becomes noticeable in silent parts of the signal, skip all silent audio frames that are present in the audio signal.

Step 6: Calculate the energy of each frame as,

$$E_k = \frac{1}{N_k} \sum_{i=1}^{N_k-1} \|X_k\|^2 \quad (6)$$

Where, E_k is energy of k^{th} frame and N_k is total number of samples in each frame. If $E_k > E_{\text{th}}$ then mark that frame as High Energy Frames (HEF) and select HEFs only for further embedding process where E_{th} is an energy threshold factor used . Decide the Offset value value used for the embedding stage $S=S0$ initially. This S is made

adaptive to the minimum SNR value required to be maintained for final watermarked signal. This also retains transparency of the embedded watermark.

Step 7: For a given HEF, we split it into two different sub-sets A and B as shown in Figure 2, i.e., subset A = { $a_1, a_2, \dots, a_{L/2}$ } containing of all samples marked by '+', the other subset B = { $b_1, b_2, \dots, b_{L/2}$ } containing of all samples marked by '-'.

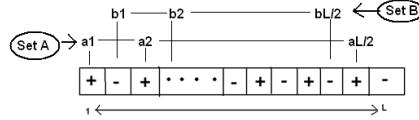


Fig. 2. Difference pair pattern

Each sub-set has $L/2$ samples each. For each HEF calculate the robust difference parameter ' α ' which is defined as an arithmetic average of differences of sample pairs within the frame. as,

$$\alpha = \frac{1}{L/2} \sum_{i=1}^{L/2} (a_i - b_i) \quad (7)$$

The distribution of the difference value α of audio frames is shown in the Fig. 3.

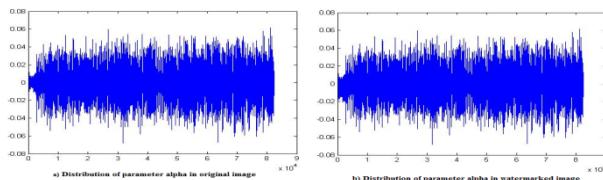


Fig. 3. Distribution of α Parameter

Most values of ' α ' are very close to zero (or the mean value of this distribution is zero). As the value ' α ' is based on the statistics of all samples in the HEF, it has certain robustness against attacks (such as mp3 compression and other slight alteration) so that we can select this ' α ' as a robust quantity for embedding information bit.

Step 8: Bit embedding strategy

The threshold value ' t ' is chosen according to the distribution of ' α ' in the original audio signal.

Case 1: If the difference value ' α ' is located within a defined threshold then,

a) If bit='1' is to be embedded, we shift the difference value ' α ' to the right side beyond a threshold, by adding a fixed step offset ' S ' from each sample value within subset A and subtracting same ' S ' value from each sample value within subset B.

b). If bit='0' is to be embedded, the frame is kept intact.

Case 2: If the difference value ‘ α ’ is located outside the threshold

No matter bit ‘1’ or ‘0’ is to be embedded, we always embed bit=’1’, thus shifting the value ‘ α ’ farther away beyond the threshold. We then rely on the error correction to correct the bit error introduced in this case.

While adding or subtracting ‘S’, care is taken such that the sample values should not lead to overflow/underflow problem. The introduced error bit due to this strategy can be corrected by using error correction.

Step 9: Reconstruction

Reconstruction of a watermarked audio is achieved by combining all frames sequentially, after embedding. The PSNR of this watermarked audio with respect to the original audio signal is calculated and compared with the expected value of the PSNR and Offset value ‘S’ is changed adaptively. The experimentation shows that after embedding the 1024 bits of watermark bits the stego audio gives PSNR value more than 40dB.

Step 10: Watermarking Secret Key Generation

The Unique Watermarking Secret Key (WSK) is generated and distributed along with the watermarked audio signal. Different parameters used during embedding process like Energy Threshold Factor ‘ E_{th} ’, Redundancy factor ‘r’ , Offset value ‘S’ are chosen, Arnold Transform frequency parameter etc. are used for generating WSK. Only authenticated user holding valid WSK can able to remove watermark from the watermarked audio and can able to recover original audio with minimal distortion.

2.2 Watermark Detection Process

The extraction algorithm consists of all the audio processing steps that are carried out at the time of embedding the audio frames. All required parameters are extracted from WSK by the receiver. First stego audio signal is segmented into non-overlapping frames of size L samples each. Then calculate the energy of each frame and select all non-silent HEF audio frames using ‘ E_{th} ’.

The difference value ‘ α ’ is calculated for each HEF using equation (3.1.1). If the difference value ‘ α ’ is outside the threshold, then bit ‘1’ is extracted and the sample value of one sub-set can be restored back to its original value. If the difference value ‘ α ’ is within the threshold, then bit ‘0’ is extracted and nothing is done on the audio sample value of that frame. In this way, we can extract the watermark bits along with the synchronization codes. Once all the bits are extracted, the watermark logo image can be reconstructed by first detecting the synchronization codes. The original audio is not required in the extracting process and thus the proposed algorithm is blind. Also the watermarked audio can be restored back as original audio with minimum distortion. Thus the technique is lossless and reversible.

The distortion caused is not perceptually audible as only HEFs are modified. Experimental results show that a maximum Offset value $S=’0.005’$ can be used without affecting the perceptual audio quality of the host signals. For the stereo audio signals, dual-channel signals are available for watermarking, while in case of a mono audio signals; only one single-channel signal is available for watermarking.

3 Experimental Results

3.1 Experimental Setup

To assess the performance of the proposed audio watermarking scheme, several experiments are carried out on different types of 250 mono audio signals of length 20 seconds each. These CD Quality audio signals are sampled at sampling rate 44.1 KHz with 16 bit resolution. These audio signals are categorized into following categories; the rock music (A1) that has very high signal energy, classical music (A2) and speech signal (A3) that has moderate signal energy. The ownership information is represented by a set of three different 32x32 binary logo images (w_1 , w_2 and w_3 respectively). The logo image is first permuted using Arnold Transform and converted into a one dimensional bit stream of 1's and 0's. After adding synchronization codes, it is replicated using redundancy factor $r=3$.

The data payload refers to the number of bits that are embedded into the audio signal within a unit of time. For a frame containing 32 samples, the estimated data payload is 1378 bps. Without applying any redundancy, it needs an audio section about 0.7546 seconds in order to embed a 16 bit synchronization code (1111100110101110) along with a 32x32 binary watermark. With $r=3$, it needs an audio section about 2.2639 seconds. To achieve good robustness against different attacks, the value of offset value/embedding strength (S) can be increased upto 0.005 maximum for all audio files.

3.2 Perceptual Quality Measures and Robustness Test

To measure imperceptibility, we use signal-to-noise ratio (SNR) as an objective measure and a listening test as a subjective measure. SNR is based on the difference between the undistorted original audio signal and the distorted watermarked audio signal. The SNRs of watermarked audio calculated are 35.49dB for A1, 32.22dB for A2 and 34.09dB for A3. The SNR is calculated only on the portion of an audio signal, where actual watermark bits are embedded. After extraction of watermark bits, the watermarked audio is restored back and corresponding SNR of restored audio are 47.90dB for A1, 45.23dB for A2 and 46.21dB for A3. The proposed algorithm is also compared with the algorithm proposed by Mikko et. al. [9] where watermark is embedded in FFT domain using frequency hopping. In [9], the average SNR of watermarked audio is 15dB and average SNR of restored audio is 40dB which is very less compared to our proposed algorithm in both the cases. Also the computational complexity of our algorithm is very less compared to Mikko's Algorithm [9].

Although SNR is a simple way to measure the noise introduced by the embedded watermark, it does not take into account the specific characteristics of the human auditory system (HAS). Therefore, we employ a subjective measure called Mean Opinion Score (MOS), whose grading scale is as shown in Table 1. Ten listeners were involved in the actual listening test to estimate the subjective MOS grade of the watermarked audio signals. After presenting with the pairs of original audio signal and the watermarked audio signal, each listener was asked to report any difference

detected between the two signals. The average grade for each pair from all listeners is taken as the final grade for that pair.

Table 1. MOS Grading Scale.

MOS Grade	1	2	3	4	5
MOS Grade Description	Very Annoying	Annoying	Slightly annoying	Perceptible, not annoying	Imperceptible

Table 2 lists the corresponding SNR values, along with MOS grades obtained by conducting listening test.

Table 2. SNR and MOS values for audio signals

Watermark Used	W1		W2		W3	
	SNR (dB)	MOS	SNR (dB)	MOS	SNR (dB)	MOS
A1	47.90	5.0	46.10	5.0	47.95	5.0
A2	45.23	5.0	45.71	4.9	45.99	5.0
A3	46.21	5.0	45.88	5.0	45.87	5.0

We have checked the effect of varying Offset value S along with different Frame Size Fs on the quality of final watermarked audio. As S increases, SNR decreases also as Fs increases, SNR decreases. But at the same time robustness increases.

So a good tradeoff among different parameters is needed in order to achieve good transparency of embedded watermark along with sufficient robustness against common intentional attacks.

Both Normalized Correlation (NC) and Bit Error Rate (BER) between the original watermark and the extracted watermark are used as an objective measure for the robustness and calculated. The proposed algorithm gives moderate SNR values along with good amount of embedding capacity and lower bit error rates. The NC values are always above 0.8 for most of the common audio processing attacks.

The watermarked audio signal is subjected to several standard audio processing attacks, in order to assess the robustness of the proposed watermarking scheme. The results are summarized in Table 3. We adopt the audio editing and attacking tool called Stirmark for Audio, in order to carry out variety of different attacks.

From the results, it can be seen that the proposed audio watermarking scheme is robust to most of the common audio processing attacks. Both rock audio signals (A1) and speech audio signals (A3) perform well compared to classical music (A2). In case of stereo audio signal, same watermark information or the meta data used during the embedding process (redundancy factor, length of watermark, Arnold Transform Frequency, value of S etc.), can be embedded in the second audio channel. This can also help the receiver to retrieve and verify the watermark blindly.

Table 3. NC and BER values along with corresponding extracted watermarks for various attacks

Attack	Robustness	
	BER (%)	NC (%)
No attack	0.0	1.00
White noise (awgn)	10.40	1.00
Cropping (20%)	13.48	0.67
Re-sampling (22.05 kHz)	5.78	0.85
Re-sampling (11.025 kHz)	10.45	0.71
Re-sampling (8.00 kHz)	12.89	0.63
Re-quantization (8 bit)	0.0	1.00
Re-quantization (24 bit)	0.0	1.00
LPF (6 order Butterworth, 22.05 kHz)	9.54	0.74
MP3 compression with the rate of 48 kbps	15.80	0.69
MP3 compression with the rate of 32 kbps	20.41	0.61
Amplitude (increased by 0.2)	5.11	0.87
Amplitude (increased by 0.4)	9.06	0.74

4 Conclusions

In this correspondence, we propose a blind high capacity lossless semi-fragile audio watermarking algorithm based on the statistical quantity related to the correlation among the audio sample values. To increase the secrecy of watermark, the proposed scheme first permutes the watermark using Arnold Transform. To resist cropping attacks, watermark is embedded along with synchronization codes directly in the Time domain. This also improves the efficiency in searching synchronization codes. To take advantage of the perceptual properties of the Human Auditory System, watermark is embeded in non-silent HEF audio frames only. Further adding redundancy to the watermark information provides error detection and correction capabilities. The experimental results show that the embedded watermark is perceptually transparent and the proposed scheme is robust against different types of attacks. For most of the attacks, the normalized correlation coefficient is more than 0.8. In addition, after extracting the watermark from watermarked audio with WSK, watermarked audio signal can be restored back to original audio signal with minimul distortion. Experimental results are analyzed by both the subjective listening test using MOS values and objective test using SNR values. Further research will focus on achieving more robustness towards other intentional attacks.

References

1. Cvejic, N., Seppanen, T.: Audio watermarking: requirement, algorithms, and benchmarking. In: Digital Watermarking for Digital Media, pp. 135–181. IGI Global Information Science Publishing, Pennsylvania (2005)
2. Acevedo, A.: Audio watermarking: properties, techniques and evaluation. In: Multimedia Security: Steganography and Digital Watermarking Techniques for Protection of Intellectual Property, pp. 75–125. IGI Global (Idea Group Publishing), Pennsylvania (2005)

3. Wu, S., Huang, J., Huang, D., Shi, Y.Q.: Self-Synchronized Audio Watermark in DWT Domain. In: Proceedings of the International Symposium in Circuits and Systems, ISCAS 2004, vol. 5, pp. 712–715 (2004)
4. Bassia, P., Pitas, I., Nikolaidis, N.: Robust Audio Watermarking in the Time domain. IEEE Transactions on Multimedia 3(2) (June 2001)
5. Feng, J.B., Lin, I.C., Tsai, C.S., Chu, Y.P.: Reversible watermarking: current status and key issues. International Journal of Network Security 2(3), 161–171 (2006)
6. Van der Veen, M., Bruekers, F., Van Leest, A., Cavin, S.: High capacity reversible watermarking for audio. In: Proc. SPIE, vol. 5020, pp. 1–11 (2003)
7. Loytynoja, M., Cvejic, N., Seppanen, T.: Audio Protection with Removable Watermarking. In: ICICS (2007)
8. Dhavale, S.V., Deodhar, R.S., Patnaik, L.M.: Walsh Hadamard Transform based Robust Blind Watermarking for Digital Audio Copyright Protection. In: International Conference on Computational Intelligence and Information Technology, CIIT, Pune (November 2011)
9. Kang, H., Jung, S.-H.: An Efficient Audio Watermark Extraction in Time Domain. International Journal of Information Processing Systems 2(1) (2006)
10. Mariko, R.M.N., Kurkoski, N.B., Yamaguchi, K.: High Payload Audio Watermarking: toward Channel Characterization of MP3 Compression. Journal of Information Hiding and Multimedia Signal Processing 2(2) (2011)

Key Management Protocol in WiMAX Revisited

Noudjoud Kahya¹, Nacira Ghoualmi¹, and Pascal Lafourcade²

¹ Networks and Systems Laboratory (LRS); Badji Mokhtar University, Annaba, Algeria

² VERIMAG Laboratory; Joseph Fourier University, Grenoble, France

Abstract. Without physical boundaries, a wireless network faces many more vulnerabilities than a wired network does. Compared to Wi-Fi, security has been included in the design of WiMAX systems at the very start. IEEE802.16 standard (WiMAX) provides a security sublayer in the MAC layer to address the privacy issues across the fixed BWA (Broadband Wireless Access). After the launch of this new standard, a number of security issues were reported in several articles. Ever since the beginning, work has been in progress for the neutralization of these identified threats.

In this paper, we first overview the IEEE802.16 standard, especially the security sublayer, and then authorization protocol PKM in WiMAX has been analyzed. We found that PKM (Privacy and Key Management) is vulnerable to replay, DoS, Man-in-the-middle attacks and we propose a new methodology to prevent the authorization protocol from such attacks.

We also give a formal analysis of authentication protocol (PKMv2) and for the proposed protocol; we conclude that our proposition prevent the attacks like Denial of service (DOS), Man-in-the-middle and replay. The formal analysis has been conducted using a specialized model checker Scyther, which provides formal proofs of the security protocol.

1 Introduction

Scyther tool were developed by Cas Cremers in 2007 [1]. Scyther, is a formal protocol analysis tool, for the symbolic automatic analysis of the security properties of cryptographic protocols (typically confidentiality or variants of authenticity). It assumes perfect cryptography, meaning that an attacker gains no information from an encrypted message unless he knows the decryption key. Scyther takes as input a role-based description of a protocol in which the intended security properties are specified using claims. Claims are of the form claim (Principal, Claim, Parameter), where Principal is the user's name, Claim is a security property (such as 'secret'), and Parameter is the term for which the security property is checked.

The aim of this paper is using Scyther tool to verify the security properties and discover the vulnerabilities in Wimax (Worldwide Interoperability for Microwave Access).

Wimax is a broadband wireless system which offers packet switched services for fixed, nomadic, and mobile accesses. Wimax utilizes many advanced technologies and mechanism in the physical and medium access control (MAC) layers to provide high spectrum efficiency and protect the traffic confidentiality and integrity and to prevent different network security attacks. The 802.16 standard (Wimax) specifies a security sublayer at the bottom of the MAC layer. This security sublayer provides SS

(Subscriber Station) with privacy and protects BS (Base Station) from service hijacking. There are two component protocols in the security sublayer: an encapsulation protocol for encrypting packet data, and a PKM (Privacy and Key Management) protocol for providing the secure distribution of keying data from BS to SS as well as enabling BS to enforce conditional access to network services.

The contribution of this work is twofold: first, we formally and analyze PKMv2 protocol with scyther tool to extract holes or threat that might exist. Second, we propose a new protocol and we also use the formal method to verified if our proposed revision resolute the security problems of the PKMv2 protocol.

This paper is organized as follows. In Section 2, we provide background and detailed information about Wimax architecture and Privacy and Key Management (PKM) protocol. Section 3, we describe the designs of scyther tool and we performing an evaluation the security objectives. In Section 4, we model and analyze PKMv2 with Scyther tool. Section 5, covers the proposed solution and modified authentication model. Finally, we conclude in section 6.

2 Security Requirements

IEEE 802.16 is the standard to specify the air interface of fixed BWA. IEEE standard 802.16-2001 [2] was first designed to provide the last mile for Wireless Metropolitan Area Network (WMAN) with line-of-sight (LOS) working at 10-66GHz bands. The latest version, IEEE standard 802.16-2004 [3], which consolidates previous standards, also supports non-line-of-sight (NLOS) within 2-11 GHz bands and mesh nodes. The recently released amendment, IEEE 802.16e [4], aims to provide mobility in WMAN.

The protocol architecture of Wimax is structured into two main layers of OSI model: the Medium Access Control (MAC) layer and physical layer. The MAC layer consists of three sublayers: the service-specific convergence sub-layer (CS), MAC common part sub-layer (MAC CPS), and security sub-layer [5]. Security sub-layer has two goals, one is to provide privacy across the wireless network and the other is to provide access control to the network. By encrypting connections between the SS and the BS, privacy is accomplished by enforcing encryption of service flows across the network, the BS protects against unauthorized access. The base station uses a Privacy and Key Management (PKM) protocol to control the distribution of secret data to subscriber stations. We will focus on PKM because it is the main part of security.

PKM provides the authorization process and secure distribution of keying data from the BS (base station) to SS/MS (mobile station). BS uses the protocol to enforce conditional access to network services.

The IEEE 802.16 PKM protocol uses X.509 digital certificates, RSA public-key algorithm, and strong encryption algorithm to perform key exchanges between SS and BS, at client/server model. IEEE 802.16 PKM employs two-tier key systems. The Authentication Protocol first authenticates SS to BS, establishing a shared secret (Authorization Key, AK) via public-key cryptography; then via Key Management Protocol, SS registers to the network, during which AK is used to secure the exchange of Transport Encryption Keys (TEK) [6].

3 Security Property

3.1 Scyther Tool

Scyther is an automatic tool for the verification and falsification of security protocols. Scyther provides a graphic user interface which incorporates the scyther command line tool and python scripting interface. Scyther tool takes protocol description and optimal parameters as input, and output a summary report and display a graph for each attack. The description of a protocol is written in SPDL language [7]. Security properties are modeless as claim events. Claim (Principal, Claim, and Parameter), where Principal is the user's name, Claim is a security property, and Parameter is the term for which the security property is checked.

For the protocol verification, Scyther can be used in three ways [7]:

- *Verification claim:* Scyther verifies or falsifies security properties.
- *Automatic claims:* if user does not specify security properties as claim event the scyther automatically generates claims and verifies them.
- *Characterization:* each protocol role can be characterized. Scyther analyses the protocol and provides a finite representation of all traces that contain an execution of the protocol role.

Scyther generates attack graph for counter example, and represents individual attack graph for each claim.

3.2 Security Propriety

All security solution for WiMAX network should satisfy the requirements as follows.

Property 1- Confidentiality: This claim is fulfilled if the MS/SS has the guarantee that all exchanged user data is secret. The exchanged user data messages between the MS and the BS is called Msg. Every information (α) in Msg should remain secret. The formalization of information confidentiality is given below.

$$\forall \alpha \in \text{Msg}(\text{claim}(\text{SS}, \text{Secret}, \alpha)) \quad (1)$$

Property 2- Authenticity: This claim is fulfilled if an outsider, who keeps track of the communication, cannot relate the traffic to a specific MS. In order to fulfill authenticity the MAC address of the MS which identifies it must remain secret. The MAC address is included in the MS's certificate (MsCert). The formal definition of pseudonymity is given below.

$$\text{claim}(\text{SS}, \text{Secret}, \text{SSCert}) \quad (2)$$

Property 3- Integrity: This claim is fulfilled if the BS and the SS have the guarantee that all exchanged keys (described as key) are secret and unique. We have included an additional restriction that only claims concerning sessions between trusted agents are evaluated. Its formal definition is shown as follows:

$$\forall \text{key}(\text{claim}(\text{BS/SS}, \text{Secret}, \text{key})) \quad (3)$$

Property 4- Access control: A WiMAX network should have a correct mechanism to verify that a given user is authorized to use a particular service[8]. Furthermore, access control can guarantee that only authorized users are allowed to connect to a given network and get access to the offered services. A service should always be bound to an authenticated user. Its formal definition is given as follows:

$$\forall \alpha \in \text{Msg}(\text{claim}(\text{BS}, \text{Secret}, \alpha)) \quad (4)$$

Property 5- Freshly of messages: An important part of security protocols is the generation of fresh values which are used for challenge-response mechanisms (often called nonces), or as session keys. This claim is fulfilled if the BS and MS/SS have the guarantee that the session key is fresh.

$$(\text{claim}(\text{BS/SS}, \text{Fresh}, \text{key})) \quad (5)$$

4 Modeling and Analyzing Pkmv2 Protocol

In this section, we model PKMv2 protocol in Scyther tool and we verify if the five properties (claim events) are respected.

4.1 PKMv2 Protocol

The latest standard, IEEE 802.16e-2005, includes a new version (PKMv2). The major security problems were solved in PKMv2. It makes authorization procedure secure enough to prevent attacks. After initial authorization, PKMv2 also checks for reauthorization periodically. Complete authorization procedure has been defined by David and Jesse in [9].

PKMv2 supports two different mechanisms for authentication: the SS/MS and the BS may use RSA-based authentication or Extensible Authentication Protocol (EAP) - based authentication. We will focus in this paper on RSA based authentication for PKMv2 authentication protocol. The flow of messages exchange in RSA-based authentication is shown as follows:

- msg1.** SS→BS: *Mancert (SS)*
- msg2.** SS→BS: { N_{ss} , $SSCert$, *Capabilities*, $BCID$ } $sk(SS)$
- msg3.** BS→SS: { N_{ss} , N_{bs} , {*prePAK*, $SSID$ } $pk(SS)$, *SAIDlist*, *prePAKSeq*, *prePAKlifetime*, $BSCert$ } $sk(WS)$;
- msg4.** SS→BS: N_{bs} , $SSaddr$, { N_{bs} , $SSaddr$ } AK ;

SS/MS sends its MCerSS (manufacturer's certificate) and then sends its own CerSS which is X.509 certificate along with a nonce; a 64 bit random number generated by the SS/MS, BC-Identity and cryptographic Capb (capabilities). BC-Identity is assigned to SS/MS when it enters in a network and requests for ranging. After receiving the authorization request message from SS/MS, BS responds by sending some information and a nonce; one generated by the BS and one that SS/MS sends in its request's message. BS also attaches its certificate (CerBS) in response to SS/MS for mutual authentication. BS also includes its signatures for validity in response message to SS/MS. A 256 bit key (Pre-Au-K) with the SS's identifier (SSID) is encrypted by the BS with the public key of SS/MS. A 4 bit sequence number for the authorization key (Seq_No) and its life time with the SAID's List (SAIDL) are sent by the BS.

After validating the message from BS, the SS/MS sends the acknowledgement message with nonce created by BS and MAC address (MAC_{SS}) of the subscriber station.

Authorization Key (AK) transmitted by BS to SS/MS in previous message is used to encrypt the Nonce_{BS} (BS generated random number) and MAC_{SS} [10].

4.2 Modeling PKMv2

In scyther, a protocol is described in SPDL language in which agent defined a role. PKMv2 can be modeled as follows:

```
// The protocol description
protocol pkmv2(SS,BS,CA)
{
role SS
{
const Mancert,cap,SAID: Data;
var CerSS,CerBS:Data;
const Ns:Nonce;
var Nb:Nonce;
var SAIDlist,AKSeq,AKlifetime:Data;

send_1(SS,BS,Mancert (SS));
send_2(SS,CA,SS);
read_3(CA,SS,{SS,{CerSS,pk(SS)}sk(CA)})sk(CA));
send_4(SS,BS,{CerSS,pk(SS)}sk(CA));
send_5(SS,BS,{cap,SAID,Ns,SS});
read_8(BS,SS,{CerBS,pk(BS)})sk(CA));
read_9(BS,SS,{(preAK)pk(SS), AKSeq,AKlifetime, SAIDlist,Ns,Nb}sk(BS));
send_10(SS,BS,{Nb,SS}AK);

}
role BS
{
var CerBS,CerSS,Mancert,cap,SAID: Data;
const Nb:Nonce;
var Ns:Nonce;
const SAIDlist,AKSeq,AKlifetime:Data;

read_1(SS,BS,Mancert (SS));
read_4(SS,BS,{CerSS,pk(SS)}sk(CA));
read_5(SS,BS,{cap,SAID,Ns,SS});
send_6(BS,CA,BS);
read_7(CA,BS,{BS,{CerBS,pk(BS)}sk(CA)})sk(CA));
send_8(BS,SS,{CerBS,pk(BS)})sk(CA));
send_9(BS,SS,{(preAK)pk(SS), AKSeq,AKlifetime, SAIDlist,Ns,Nb}sk(BS));
read_10(SS,BS,{Nb,SS}AK);
}
role CA
{
const CerSS,CerBS: Data;
read_2(SS,CA,SS);
send_3(CA,SS,{SS,{CerSS,pk(SS)}sk(CA)})sk(CA));
read_6(BS,CA,BS);
send_7(CA,BS,{BS,{CerBS,pk(BS)}sk(CA)})sk(CA));
}
```

4.3 Analysis of PKMv2

This model is going to be challenged with the following requirements using the Scyther tool.

1. Property 1: Scyther identified problems in the confidentiality protocol. It is a passive attack on confidentiality. An intruding entity eavesdrops the second message (Auth-REQ) and he is able to read the information that is sent from the SS/MS to the BS, gathering information about the trusted SS/MS (cryptographic capabilities and security association identifier (SAID)).

2. Property 2: Scyther detected a possible Authenticity attack. Message2 is sent in plaintext so an intruder eavesdrops this message and obtains the SS's certificate (MsCert). BS may face a replay attack from a malicious SS who intercepts and saves or modified the authentication messages sent by a legal MS/SS previously.

Property 3: it is proved that the authorization key exchanged in the authentication protocol is secret.

Property 4 and 5: It is proved that an adversary cannot obtain the pre-PAK, which will be used to extract the AK and the session key is fresh, as it is encrypted with the public key of the SS.

As seen in the formal analysis, the *secrecy of the keying* material distributed claim is valid in PKMv2. However, *Authenticity*, *integrity* and *information confidentiality* are broken, PKMv2 still vulnerable to replay, DoS and Man-in-the-middle attacks.

5 The Proposed Revised Authentication Protocol

As discussed in the previous section, the PKMv2 protocol does not fulfill the claims pseudonymity and information confidentiality because it still vulnerable to replay, DoS and Man-in-the-middle. In related works the nonce is used to prevent replay and man-in-the middle attacks, Nonce indicate that the requests were not used before, but he will not give any information about the time that was sent. Nonce is also not sufficient to tell the BS that it is the current message received from the SS/MS. In our revised protocol to assure synchronization between SS/MS and BS both nonce and timestamp are use. So the revised protocol has the timestamp attached with the SS/MS message to the BS along with the nonce. The protocol will be described as follows.

- msg1.** SS→CA: SS
- msg2.** CA→SS:{SS,{CertSS,pk(SS)}sk(CA)}sk(CA)
- msg3.** SS→BS:{(CertSS,N_s)pk(CA)}sk(SS)
- msg4.** BS→CA: BS
- msg5.** CA→BS:{BS, {CertBS, pk(BS)}sk(CA)}sk(CA)
- msg6.** BS→CA: {{(CerSS, N_s)pk(CA)}sk(SS), CertBS, N_b}sk(BS)
- msg7.** CA→BS:{(CerSS, N_s, N_b)pk(BS), (CerBS, N_s, N_b)pk(SS)}sk(CA))
- msg8.** BS→SS:{(CerBS, N_s, N_b)pk(SS)}sk(CA)
- msg9.** SS→BS:{(Ts, N_b, cap, SAID)pk(BS)};
- msg10.** BS→SS:{(prePAK(BS)}sk(BS), SAIDlist, Ts, Tb, N_s, preSeq, prePAKlifetime}pk(SS)
- msg11.** SS→BS:{Tb, N_b}sk(SS)

The new protocol can be divided into four main stapes:

1-Certificates Register: SS/MS and BS send a message to find an X.509 certificate and it own public key information onto the server CA. This first step contained 1), 2) and 4), 5) messages: CA is only as a certification center which does not participant in the session key exchange.

2-Certificates Exchange: SS/MS and BS exchange their certificates through the trusted server CA in order to decide if etch particular is a trusted device or not. This step contained 3), 6), 7), 8) messages

3-Authorization request message: SS/MS sends a message contains the SS/MS certificate (SsCert) and a nonces (N_s) used for registration and exchange certificates, it also contains the timestamp of SS/MS along with SAID and its security capabilities. Authorization request message is encrypted with the public key of the BS $pk(Bs)$, the timestamp addition could bring an extra layer of security since the BS could identify the message as current one. The timestamp could avoid the intruders who are trying to synchronize time with either BS or SS/MS.

4- Authorization reply message: If BS determines that the MS/SS is authorized it replies with a message authorization reply message. BS sends nonce (N_s) which was sent by the SS. That could ensure SS/MS that message 10 is the reply of the request send by SS/MS itself. BS Nonce ensures the SS about the authentication of BS. This mutual authentication gives extra layer of security. BS sends a pre-AK encrypted with the private key of BS $sk(Bs)$. From pre-PAK, the SS generates AK. If AK is used correctly, then SS gains the authorization to access the WIMAX channel. The message contained also Lifetime of Pre-AK a Sequence number of pre-AK. BS sends his Timestamp (T_b) to grant that is not copied by adversaries, the timestamp and the nonce of BS previously received to confirm authorization access. BS encrypted the message with his public key.

5- Verification the information integrity: The last message ensures that the message is from the actual BS. Two layers of assurance are provided in this message: the nonce (N_b), and time stamp sent by BS (T_b). SS use it signature to ensure that message is from an actual SS and to assure the information integrity.

5.1 Modeling New Protocol in SPDL language

The new version of the PKM protocol can be modeled in SPDL as follows:

```
// The protocol description
protocol new version(SS,Bs,CA)
{
role SS
{
const cap,SAID:Data;
var prePAKSeq,prePAKLifetime,CerSS,CerBS, SAIDlist: Data;
var Ns:Nonce;
const Nb:Nonce;
var Ts:TimeStamp;
const Tb:TimeStamp;
var prePAK:SessionKey;
```

```

send_1(SS,CA,SS);
read_2(CA,SS, {SS,{CerSS,pk(SS)}sk(CA)}sk(CA));
send_3(SS,BS,{{CerSS,Ns}pk(CA)}sk(SS));
read_8(BS,SS, {{CerBS,pk(BS),Ns,Nb}pk(SS)}sk(CA));
send_9(SS,BS,{Ts,Nb,cap,SAID}pk(BS));
read_10(BS,SS,{{prePAK}sk(BS), SAIDlist,Ts,Tb,prePAKSeq,prePAKlifetime}pk(SS));
send_11(SS,BS,{Tb,SS}pk(BS));

claim_ss1(SS, Secret,CerSS);
claim_ss2(SS, Nisynch);
claim_ss3(SS, Niagree);
claim_ss4(SS, Secret,Data);
claim_ss5(SS,Secret,prePAK);
claim_ss8(SS,Secret,Ns);
claim_ss11(SS,Empty,(Fresh,prePAK));
}

role BS
{
const prePAKSeq,prePAKlifetime, SAIDlist: Data;
var Ns:Nonce;
const Nb:Nonce;
const Ts:TimeStamp;
var Tb:TimeStamp;
var cap,SAID,CerSS,CerBS:Data;
const prePAK:SessionKey;

read_3(SS,BS, {{CerSS,Ns}pk(TS)}sk(SS));
send_4(BS,CA, BS);
read_5(CA,BS, {BS,{CerBS,pk(BS)}sk(CA)}sk(CA));
send_6(BS,CA,{{{{CerSS,Ns}pk(CA)}sk(SS),CerBS,Nb}pk(CA)}sk(BS));
read_7(CA,BS,{{CerSS,pk(SS),Ns,Nb}pk(BS)}sk(CA),{{CerBS,pk(BS),Nb,Ns}pk(SS)}sk(CA));
send_8(BS,SS, {{CerBS,pk(BS),Ns,Nb}pk(SS)}sk(CA));
read_9(SS,BS,{Ts,Nb,cap,SAID}pk(BS));
send_10(BS,SS,{{prePAK}sk(BS), SAIDlist,Ts,Tb, prePAKSeq,prePAKlifetime}pk(SS));
read_11(SS,BS,{Tb,SS}pk(BS));

claim_bs1(BS, Secret,CerBS);
claim_bs2(BS, Nisynch);
claim_bs3(BS, Niagree);
claim_bs4(BS, Secret,Nb);
claim_bs8(BS,Secret,prePAK);
claim_bs11(BS,Empty,(Fresh,prePAK));
}

role CA
{
const Nb,Ns:Nonce;
const CerBS: Data;
const CerSS: Data;
    read_1(SS,CA, SS);
    send_2(CA,SS, {SS,{CerSS,pk(SS)}sk(CA)}sk(CA));
    read_4(BS,CA, BS);
    send_5(CA,BS,{BS,{CerBS,pk(BS)}sk(CA)}sk(CA));
    read_6(BS,CA,{{{{CerSS,Ns}pk(CA)}sk(SS),CerBS,Nb}pk(CA)}sk(BS));
}

send_7(CA,BS,{{CerSS,pk(SS),Ns,Nb}pk(BS)}sk(CA),{{CerBS,pk(BS),Nb,Ns}pk(SS)}sk(CA));
}
}

```

5.2 Analysis the New Version

This model is going to be challenged with the following requirements using the Scyther tool.

1. *Property 1 and 2:* In the formal analysis it is proved that an intruder cannot obtain the SS/MS certificate (MsCert) and data exchange between SS and BS.
2. *Property 3:* In the formal analysis it is proved that the authorization key exchanged in the authentication protocol is secret and not broken.
3. *Property 4:* It is proved that unauthenticated user cannot access the services provided, and cannot impersonate another user. Also, it is not possible to modify the data by an unauthorized individual.
4. *Property 5:* It is proved that an adversary cannot obtain the unique pre-PAK. Time-stamp and nonce are used in the revised protocol to prevent replay and man-in-the-middle attack. The SS/MS appends the time stamp and nonce. This helps the BS to identify the request as a newer one. The nonce will wipe out the possibility of replay attack.

The nonce helps the BS to identify successive requests and it enhances the BS capacity to reject those requests which was sent by the intruders or adversaries so to prevent DOS attack. BS, thus, can identify the latest requests and it is able to filter out samples of replay attacks. In stapes authorization reply message, the BS sends the timestamp and nonce of SS/MS. That helps in preventing an adversary from forging a BS. This protocol also provides mutual authentication. The nonce value sent by the BS helps in preventing the man-in-the middle attack. The revised protocol helps SS/MS and BS exchange their certificates through the trusted server CA in order to decide if etch particular is a trusted device or not; hence it avoids the possibility of the DoS attack.

Second, the timestamp helps the BS in identifying the latest requests, which prevents reply attacks. It also helps the SS/MS to identify the recent messages, and hence it can identify the AK used by the SS/MS as new or not. The addition of nonce from the BS helps the SS/MS to identify whether the message which he received with pre-AK is a newer one or not. It is better to add more buffers to carry the used nonce values in the previous sessions. This gives more security to the BS and user SS/MS.

6 Conclusion

The paper analyzes the vulnerabilities in the basic authentication protocol PKMv2. As seen in the formal analysis, we formally verified the key management protocol of PKMv2 in terms of the secure session key establishment and distribution, confidentiality, authenticity, integrity, access control.

The *secrecy of the keying material distributed claim* is valid. However, *Authenticity, integrity* and *information confidentiality* are broken in PKMv2.

A revised authentication protocol is proposed by using nonce and timestamp together. The new solution is efficient to tackling the various security threats such as replay, man in the middle and DOS attacks. The revised authentication protocol is expected to provide better secure platform for IEEE 802.16(e).

References

- [1] Cremers, C.: Scyther-Semantics and verification of security protocols. PhD dissertation; Eindhoven University of technology (2006)
- [2] IEEE Std. 802.16-2001, IEEE Standard for Local and Metropolitan Area Networks Part16: Air Interface for Fixed Broadband Wireless Access Systems, IEEE 2002 (2002)
- [3] IEEE Std. 802.16-2004, IEEE Standard for Local and Metropolitan Area Networks Part16: Air Interface for Fixed Broadband Wireless Access Systems, IEEE 2004 (2004)
- [4] IEEE Std. 802.16e-2005, IEEE Standard for Local and Metropolitan Area Networks Part16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, IEEE 2006 (2006)
- [5] Abbaci-kahya, N., Ghoualmi, N.: Security in Wimax. In: International Conference on Information Technology and e-Services, Tunisia (2011) ISBN 978-9938-9511-03
- [6] Xu, S., Huang, C.T.: Attacks on PKM protocols of IEEE 802.16 and its later versions. In: Proceedings of 3rd International Symposium on Wireless Communication Systems (ISWCS 2006), Valencia, Spain (2006)
- [7] Cremers, C.J.F.: The Scyther Tool: Verification, Falsification, and Analysis of Security Protocols. In: Gupta, A., Malik, S. (eds.) CAV 2008. LNCS, vol. 5123, pp. 414–418. Springer, Heidelberg (2008)
- [8] Lang, W.-M., Wu, R.-S., Wang, J.-Q.: A Simple Key Management Scheme based on WiMAX. In: International Symposium on Computer Science and Computational Technology, IEEE 2008 (2008)
- [9] Johnston, D., Walker, J.: Overview of IEEE 802.16 security. IEEE Security and Privacy Magazine 2(3), 40–48 (2004)
- [10] Altaf, A., Younus Javed, M., Ahmed, A.: Security Enhancements for Privacy and Key Management Protocol in IEEE 802.16e-2005. College of Signals, NUST. In: Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, IEEE 2008 (2008)

Image Authentication Technique Based on DCT (IATDCT)

Nabin Ghosal¹, Anirban Goswami², Jyotsna Kumar Mondal³, and Dipankar Pal⁴

¹ Dept. of Engineering and Technological Studies, University of Kalyani,
Kalyani, Nadia-741235

² Dept of Information Technology, Techno India, EM 4/1 Salt lake, Sec-v, Kolkata-700091

³ Dept. of Computer Science and Engineering, University of Kalyani, Kalyani, Nadia-741235

⁴ Dept. of Computer Science and Engineering, Techno India, EM 4/1 Salt lake, Sec-v,
Kolkata-700091, West Bengal, India
nabin_ghoshal@yahoo.co.in, {an_gos, mail2dpal}@yahoo.com,
jkm.cse@gmail.com

Abstract. In this paper a novel steganographic technique based on Discrete Cosine Transform (DCT) is demonstrated for image authentication in frequency domain. The transformation is implemented on sub-image block called mask of size 2×2 of spatial components in row major order for the entire image. Single bit of authenticating secret message/image is fabricated in the real part of the frequency component of 2nd and 3rd carrier image byte of each sub-image block. A minor re-adjustment is incorporated only in the first component of each sub image block after embedding to keep the pixel values positive and non fractional in spatial domain. Robustness is achieved through embedding secret bits in variable positions of the carrier image byte determined by random function and subsequent masking. Experimental results depict enhanced performance of the proposed watermarking technique.

1 Introduction

Steganography is basically a technique of art and science which hides message/image in such a way that no one, apart from the sender and intended recipient, suspects the existence of the secret message/image, i.e. a form of security through obscurity. Generally, message may be secret images, articles, shopping lists, or some other source text. Plain visible encrypted messages no matter how unbreakable may arouse suspicion. So cryptography protects the contents of a message, where as Steganography [8, 9] can be said to protect both messages and communicating parties. With millions of Web sites flooding the Internet, it may seem easy for on-line pirates to copy and paste information, images, video and audio from one Web site to another, deliberately infringing on someone else's copyright [4,5]. However, owners of copyrighted materials on the Web can take advantage of digital watermarking [2, 3] in order to protect their materials from being duplicated without their permission. Digital watermarks cannot be removed or altered, making them a very important tool when fighting copyright infringement on the Web.

Digital watermarking allows users to legally use content, while adding security to the content to prevent illegal usage and are robust, that is they are able to survive

attacks from potential hackers or any type of manipulation. In case of invisible watermarks, the locations in which the watermark is embedded are secret, only the authorized persons extract the watermark with some mathematical calculations. These kinds of watermarks are not viewable by human eye and are more secure and robust than visible watermarks. In general there is a tradeoff between the watermark embedding [6, 7] strength and its quality. Increased robustness emphasizes on strong embedding, which in turn increases the visual degradation of secret image/message. The proposed watermarking scheme is applicable for gray image authentication by hiding secret data using some random mathematical calculation. The result of the proposed technique IATDCT is compared with the existing Reversible data hiding based on block median preservation (RDHBBMP [13]) watermarking method in terms of visual interpretation, MSE, PSNR in dB and IF. Fig. 1 and Fig. 2 shows the insertion and extraction process of IATDCT.

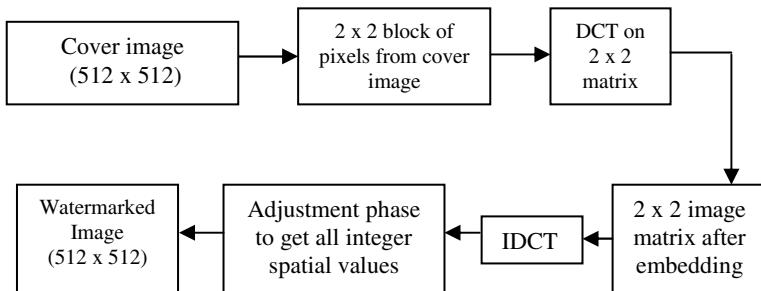


Fig. 1. The process to embed the Secret data into the source image

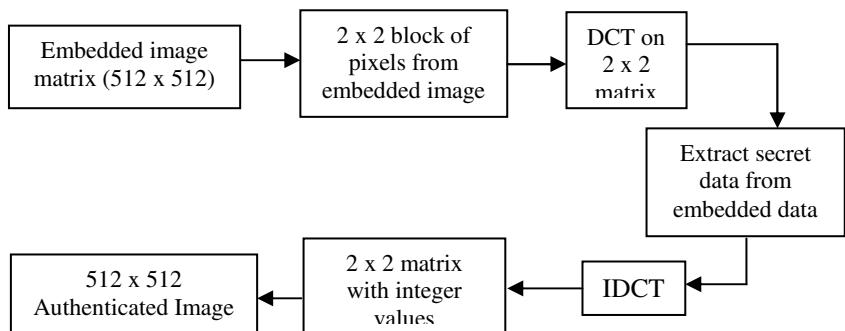


Fig. 2. The process to extract Secret data from the watermarked image

In sec. 1.1 and 1.2 two dimensional Discrete Cosine Transform and Inverse Discrete Cosine transform have been represented. In sec. 2.1 and 2.2 the insertion and extraction algorithms of IATDCT are stated. The results of the proposed technique

are depicted in terms of MSE, PSNR in dB and IF in sec. 3 followed by conclusion in sec. 4.

The proposed technique uses two dimensional Discrete Cosine Transform and Inverse Discrete Cosine Transform represented as:-

1.1 Two Dimensional Discrete Cosine Transform

The DCT represents an image as a sum of sinusoids of varying magnitudes and frequencies. The property of DCT for a typical image is that significant information about the image is concentrated in just a few coefficients of the DCT. For this reason, the DCT is often used in image compression applications. The two-dimensional DCT of an M x N matrix is defined as follows:

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} \cos \frac{\pi(2m+1)p}{2M} \cos \frac{\pi(2n+1)q}{2N}, \quad 0 \leq p \leq M-1 \\ 0 \leq q \leq N-1$$

$$\alpha_p = \begin{cases} \frac{1}{\sqrt{M}}, & p = 0 \\ \sqrt{2}/M, & 1 \leq p \leq M-1 \end{cases} \quad \alpha_q = \begin{cases} \frac{1}{\sqrt{N}}, & q = 0 \\ \sqrt{2}/N, & 1 \leq q \leq N-1 \end{cases}$$

The values B_{pq} are called the DCT coefficients of spatial value A_{mn} . M x N denotes the dimension of source image.

1.2 Two Dimensional Inverse Discrete Cosine Transform

The IDCT is an invertible DCT transform, and is given by,

$$A_{mn} = \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} \alpha_p \alpha_q B_{pq} \cos \frac{\pi(2m+1)p}{2M} \cos \frac{\pi(2n+1)q}{2N}, \quad 0 \leq m \leq M-1 \\ 0 \leq n \leq N-1$$

$$\alpha_p = \begin{cases} \frac{1}{\sqrt{M}}, & p = 0 \\ \sqrt{2}/M, & 1 \leq p \leq M-1 \end{cases} \quad \alpha_q = \begin{cases} \frac{1}{\sqrt{N}}, & q = 0 \\ \sqrt{2}/N, & 1 \leq q \leq N-1 \end{cases}$$

The IDCT equation can be interpreted as any M x N matrix say A and can be written as a sum of M x N functions of the form:

$$\alpha_p \alpha_q \cos \frac{\pi(2m+1)p}{2M} \cos \frac{\pi(2n+1)q}{2N}, \quad 0 \leq m \leq M-1, 0 \leq n \leq N-1$$

These functions are called the basis functions of the DCT. The DCT coefficients B_{pq} , can be regarded as the *heights* applied to each basis function.

2 The Technique

The Insertion of secret data is performed into frequency values. The frequency values are obtained after performing Discrete Cosine Transform on 2 x 2 sub-image matrix one by one taken from the original image matrix. Hence, in order to perform the

insertion operation of the authenticating message/image into frequency values of converted original image, bits from authenticating image are embedded in single bit position under each byte of the source image. Location of insertion of a bit is obtained dynamically by executing a random function and subsequent masking. Secret bits are fabricated only into the 2nd and 3rd image byte of each sub image block. The insertion and extraction algorithm are as follows:

2.1 Insertion Algorithm

Input : A source image and authenticating message/image.

Output : A watermarked image.

Method : Embedding is being performed only in the integer part of frequency value, whereas the fractional part remains intact. The fractional part is re-added after embedding the secret bits. The algorithm is as follows:

1. Read the header information (source image type, dimensions and maximum intensity) from source image and write into the output image.
2. Repeat the following steps until all pixels have been read from the source image file,
 - 2.1 Take 2 x 2 blocks of pixels from the source image matrix in row major order.
 - 2.2 Apply DCT on the current block of pixels.
 - 2.3 Compute a random number ipos (between 0 - 3) using random function.
 - 2.4 Perform subsequent masking [1] (i.e. only value at 1st position of 2 x 2 matrix is incremented by 1).
 - 2.5 Read the authenticating message/image i.e. secret data.
 - 2.6 Embed the watermark bits in the source image byte where the position of embedding is defined by the variable ipos.
 - 2.7 Apply Inverse DCT on current block of pixels.
 - 2.8 If any pixel is found to be of negative value, then
 - 2.8.1 Only the value at 1st position of the current sub block matrix is subsequently incremented by 1.
 - 2.8.2 Repeat steps 2.2 to 2.7 until the negative value is eliminated.
 - 2.9 Write the modified sub block into the output image in row major order.
 - 2.10 Repeat steps from 2.1 to 2.9 until all the pixels of authenticating image/message are embedded.
3. Stop.

2.2 Extraction Algorithm

The extraction of the authenticating message/image is performed on the frequency component of watermarked image bytes. The frequency values are obtained after performing the DCT operation on the watermarked image. A masking based detection scheme has been proposed to retrieve the embedded watermark from a gray carrier image. In case of retrieval of the authenticating message/image, we extract one bit each from 2nd and 3rd pixel of each sub image block of the watermarked image. Location of extraction of a bit is obtained by executing the random function and subsequent masking.

Input : A watermarked image.

Output : An authenticating message/image.

Method : Extraction is performed on the integer values only while the floating point part remains intact. The floating part is added after extracting the embedded bits from the integer part of the pixels values of the input image. The algorithm is as follows:

1. Read from the watermarked image the image type and maximum intensity and write into the output image.
2. Repeat the following steps until all pixels have been read from the input image,
 - 2.1. Take 2×2 blocks of pixels from the embedded image matrix and perform DCT on the current block.
 - 2.2. Compute a random number ipos (between 0 - 3) using random function and subsequent masking [1].
 - 2.3. Extract an embedded bit from the integer part of the frequency values in the watermarked image from a position specified by ipos.
 - 2.4. Combine and convert each 8 bits of 0's and 1's into a decimal value and write the value in the output image.
 - 2.5. Repeat the steps from 2.1 to 2.4 until all pixels of the authenticate message/image has been extracted from the watermarked image.
3. Stop.

3 Result Comparison and Analysis

This section represents the results, discussion and a comparative study of the proposed technique IATDCT with other watermarking methods are DCT, QFT and Spatio-Chromatic DFT-based. The parameters involved in the comparative study are visual interpretation, image fidelity (IF), peak signal-to-noise ratio (PSNR) analysis and mean square error (MSE). In order to test the robustness of the scheme IATDCT, the technique is applied on more than fifty PGM grayscale images from which it can be concluded that the algorithm may overcome any type of attack like visual or statistical. Experimental set up for preparing result is any type of PC with 2.00 GHz or above processor speed, 1 GB or higher primary memory and Unix/Linux OS with Gimp (GNU Image Manipulation Program) application. Distinction between the carrier and embedded images under human visual scrutiny is somewhat difficult. In this section some statistical and mathematical analysis has been presented. The original carrier images ‘Airplane’, ‘Baboon’, ‘Lenna’ and ‘Fruits’ are shown in fig 3a, 3b, 3c and 3d. The dimension of each carrier grayscale image is 512×512 and the dimension of the authenticating grayscale image ‘Earth’ (Fig. 3q) is 127×127 . The resultant embedded grayscale images are shown in Fig 3e, 3f, 3g and 3h respectively obtained using IATDCT. Single bit of secret data is embedded in each 2^{nd} and 3^{rd} image byte of the sub image block of the carrier image. 3i, 3j, 3k, and 3l are magnified version of source images and 3m, 3n, 3o, and 3p are magnified version of embedded images respectively.

Peak signal-to-noise ratio (PSNR) is used to evaluate qualities of the stego-images. Table 1 and 2 shows single level of authenticating data byte embedding which is defined by $\text{EL}=0$ based on PSNR values. Table 2 shows the PSNR values for comparative studies of IATDCT and Reversible data hiding based on block median preservation (RDHBBMP) and also the enhancement in terms of hiding capacity of

secret data and PSNR in dB. The average enhancement of secret data embedding is 23553 bits in IATDCT than the existing technique RDHBBMP and also 1.30 dB of PSNR in EL=0. In all the existing technique the PSNRs are low, means bit-error rate are high but in the proposed scheme more bytes of authenticating data can be embedded and the PSNR values are significantly high, means bit-error rate is low.

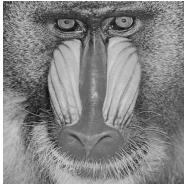
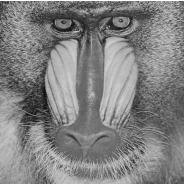
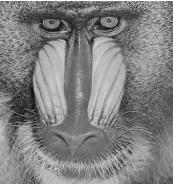
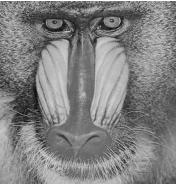
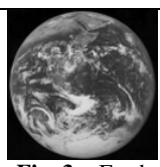
Source Images	Embedded Images using IATDCT	Magnified Source Images	Magnified Embedded Images
			
Fig. 3a. Airplane	Fig. 3e. Airplane	Fig. 3i. Airplane	Fig. 3m. Airplane
			
Fig. 3b. Lenna	Fig. 3f. Lenna	Fig. 3j. Lenna	Fig. 3n. Lenna
			
Fig. 3c. Baboon	Fig. 3g. Baboon	Fig. 3k. Baboon	Fig. 3o. Baboon
			
Fig. 3d. Fruits	Fig. 3h. Fruits	Fig. 3l. Fruits	Fig. 3p. Fruits
			
Fig. 3q. Earth			

Fig. 3. Visual interpretation of embedded images using IATDCT and corresponding magnified images before and after embedding

The average improvement is shown in Table 2. Table 3 shows the better PSNR values than other exiting techniques like DCT-based [10] watermarking, QFT-based [11] watermarking, and SCDFT-based [12] watermarking in frequency domain. In DCT based watermarking scheme do not embed watermarks in every single block of image. Here selectively pick the regions that do not generate visible distortion for embedding, thus decreasing the authenticating data size. In QFT based watermarking compensation mark allows the watermark to be undetected even if the strength of it is high. For low compression factor it can not completely recover the embedded message. Capacities of existing techniques are 3840 bytes and the PSNR values are 30.1024 dB, 30.9283 dB, and 30.4046 dB in SCDFT, QFT, and DCT respectively. Whereas the capacity of IATDCT is 16129 bytes and PSNR is 50.1338 dB and which is fully recoverable. 12289 bytes more secret data embedding is possible in IATDCT technique than existing techniques with average 20 dB more PSNR values.

Table 1. Capacities and PSNR values of IATDCT

Test images	Indicator	EL=0
Baboon	C(bits)	129032
	PSNR	44.92
Fruits	C(bits)	129032
	PSNR	43.33
Peppers	C(bits)	129032
	PSNR	44.97
Average Image	C(bits)	129032
	PSNR	44.41

Table 2. Results and comparison in capacities and PSNR of IATDCT and RDHBBMP

Test images	Indicator	EL=0	
		RDHBBMP	IATDCT
Lena	C(bits)	26,465	54,896
	PSNR	49.68	51.16
Airplane	C(bits)	36,221	54,896
	PSNR	49.80	50.91
Average Image	Δ Ca	23553	
	Δ PSNRa	1.30	

Table 3. Results and comparison in capacities and PSNR of IATDCT and DCT, QFT, SCDFT [12]

Technique	Capacity (bytes)	PSNR in dB
SCDFT	3840	30.1024
QFT	3840	30.9283
DCT	3840	30.4046
IATDCT	16129	50.1332

4 Conclusion

IATDCT is proposed for increasing the security of data hiding as compared with the existing algorithms. Authenticity is incorporated by embedding secret data in specific carrier image byte in randomly generated position. As compared with Reversible data hiding based on block median preservation, proposed IATDCT algorithm is applicable for any types of grayscale image under the umbrella of security and authenticity. The watermarked image in this algorithm is very difficult to detect due to unknown insertion position of the authenticating image bits in the carrier image. Hence, the proposed technique IATDCT is quite secured from almost any possible attacks.

Acknowledgement. The authors express their deep sense of gratitude to the faculty members of the Dept. of Engineering and Technological Studies, University of Kalyani, West Bengal, India, where the work has been carried out.

References

1. Radhakrishnan, R., Kharrazi, M., Menon, N.: Data Masking: A new approach for steganography. *Journal of VLSI Signal Processing* 41, 293–303 (2005)
2. EL-Emam, N.N.: Hiding a large Amount of data with High Security Using Steganography Algorithm. *Journal of Computer Science* 3(4), 223–232 (2007) ISSN 1549-3636
3. Amin, P., Lue, N., Subbalakshmi, K.: Statistically secure digital image data hiding. In: *IEEE Multimedia Signal Processing MMSP 2005*, Shanghai, China, pp. 1–4 (October 2005)
4. Pavan, S., Gangadharpani, S., Sridhar, V.: Multivariate entropy detector based hybrid image registration algorithm. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Philadelphia, Pennsylvania, USA, pp. 18–23 (March 2005)
5. Al-Hamami, A.H., Al-Ani, S.A.: A New Approach for Authentication Technique. *Journal of Computer Science* 1(1), 103–106 (2005) ISSN 1549-3636
6. Ker, A.: Steganalysis of Embedding in Two Least-Significant Bits. *IEEE Transaction on Information Forensics and Security* 2(1), 46–54 (2008) ISSN 1556-6013
7. Yang, C., Liu, F., Luo, X., Liu, B.: Steganalysis Frameworks of Embedding in Multiple Least Significant Bits. *IEEE Transaction on Information Forensics and Security* 3(4), 662–672 (2008) ISSN 1556-6013
8. Wu, H.C., Wu, N.I., Tsai, C.S., Hwang, M.S.: Image steganographic scheme based on pixel-value differencing and LSB replacement methods. *Proc. Inst. Elect. Eng., Vis. Images Signal Processing* 152(5), 611–615 (2005)
9. Yang, C.H., Weng, C.Y., Wang, S.J., Sun, H.M.: Adaptive Data Hiding in edge areas of Images With Spatial LSB Domain Systems. *IEEE Transaction on Information Forensics and Security* 3(3), 488–497 (2008) ISSN 1556-6013
10. Ahmadi, N., Safabkhsh, R.: A novel DCT-based approach for secure color image watermarking. In: *Proc. Int. Conf. Information Technology: Coding and Computing*, vol. 2, pp. 709–713 (April 2004)

11. Bas, P., Biham, N.L., Chassery, J.: Color watermarking using quaternion Fourier transformation. In: Proc. ICASSP, Hong Kong, China, pp. 521–524 (June 2003)
12. Tsui, T.T., Zhang, X.-P., Androullos, D.: Color Image Watermarking Using Multidimensional Fourier Transformation. IEEE Trans. on Info. Forensics and Security 3(1), 16–28 (2008)
13. Luo, H., Yu, F.-X., Chen, H., Huang, Z.-L., Li, H., Wang, P.-H.: Reversible data hiding based on block median preservation. Information Sciences 181, 308–328 (2011)

Survey on a Co-operative Multi-agent Based Wireless Intrusion Detection Systems Using MIBs

Ashvini Vyavhare, Varsharani Bhosale, Mrunal Sawant, and Fazila Girkar

B.Tech Information Technology

Department of Computer Engineering And Information Technology,

College of Engineering, Pune-5, MS, India

{ashviniv9, varsharanibhosale145, mrunal08, fazilagirkar04}@gmail.com

Abstract. In emerging technology of Internet, security issues are becoming more challenging. In case of wired LAN it is somewhat in control, but in case of wireless networks due to exponential growth in attacks, it has made difficult to detect such security loopholes. Wireless network security is being addressed using firewalls, encryption techniques and wired IDS (Intrusion Detection System) methods. But the approaches which were used in wired network were not successful in producing effective results for wireless networks. It is so because of features of wireless network such as open medium, dynamic changing topology, cooperative algorithms, lack of centralized monitoring and management point, and lack of a clear line of defense etc. So, there is need for new approach which will efficiently detect intrusion in wireless network. Efficiency can be achieved by implementing distributive, co-operative based, multi-agent IDS. The proposed system supports all these three features. It includes mobile agents for intrusion detection which uses SNMP (Simple network Management Protocol) and MIB (Management Information Base) variables for mobile wireless networks.

Keywords: Multi- agent, MIB, SNMP, Security.

1 Introduction

Security is important in any environment. As large information is available on the network and it is possible to share this data through it, it should be secure. It is somewhat defined in wired network but in wireless there is great challenge of different attacks. People and organizations have been protecting their data from harmful activities using rules that identify and block such things. However current and future threats require development of more adaptive defensive tool.

Attack is an assault on system security that derives from an intelligent threat. It can be mainly classified as Active attacks and Passive attacks. Active attacks are in the nature of eavesdropping on, or monitoring of, transmissions while passive attacks involves some modification of the data stream or creation of false stream.[6]

Intrusion detection is the act of identifying intruders who attempt to compromise the integrity, confidentiality or availability of resource. It is used to secure the systems in the networks.[1] There is a common misunderstanding that firewalls do the same thing of detecting and blocking of attack by shutting off everything and then turning back on only some well-chosen items.[8] It just restricts access to the designated points. Securing the computer networks with firewalls or using strong encryption algorithm keys are longer effective. This leads to the development of new architecture and mechanisms to protect wireless and mobile networks.

An IDS is a software or hardware tool that monitors traffic on network looking for and logging threats. The purpose of IDS is to monitor the computer networks, detect intrusions and alert the concern person. Network based (NIDS) and Host based (HIDS) are types of IDS. In NIDS traffic flowing through network is analyzed. In HIDS activities on each individual computer are examined.[3,8]

There are two ways on which basis we can implement IDS. The first one is signature based in which the attacks have unique signature that can be detected. Known attacks can be detected by looking for these signatures. Second approach is anomaly based in which a system develops a base line what it considers a normal traffic. Any activity which is recorded beyond this traffic is considered as anomaly and alert is generated.

2 Wireless Intrusion Detection System

The organizations invest in wireless networks as compare to traditional wired LANs because of its low cost and relative ease of use. Although the wired-IDS are powerful systems, unfortunately they do little for the wireless world. The main difference between wired and wireless networks is their nature of transmission medium, different protocol specification in lower layer, different lower layer functionality of the intruders etc. [4]

The rapid proliferation of wireless networks and mobile computing applications has changed the landscape of network security. The nature of mobility creates new vulnerabilities due to open medium, dynamically changing network topology, co-operative algorithms, lack of centralized monitoring and management points that do not exist in a fixed wired network, and yet many of the proven security measures turn out to be inactive. This has led to the development of new architecture and mechanism to protect the wireless networks and mobile computing applications. [5]

The various approaches like IDS using Neural Networks, Artificial Immune Systems, MANET based, Clustering, System calls based, co-operative agent based etc. are developed and implemented so far.

3 Problems Related to Wireless Networks

Till now little research has been done in area of wireless IDS. Because of its structural and behavioral differences, IDS designed for wired networks is not that applicable to the wireless network.[7]

In case of wireless networks communication is done through an open air environment and the medium is not well protected .So it is impossible to monitor network traffic at bottlenecks. It is necessary to do monitoring at each and every network node. But it is inefficient due to high network bandwidth requirement and increased power resources that are not easily available.[7]

Ad hoc wireless networks are very dynamic in structure, giving rise to apparently random communication patterns, thus making it challenging to build a reliable behavioral model. Misuse detection requires maintenance of an extensive database of attack signatures, which in the case of ad hoc network would have to be replicated among all the hosts.[2] This will result in an extended initial setup time and decrease in useful computational power of each host.

Another problem is monolithic IDS design. Each node must have an IDS client and should take part in global detection process. To get rid of this problem modular IDS should be implemented using mobile agents. By this we have many advantage of increase in fault tolerance, reduced communication cost, and improved performance of whole network and scalability.[7]

4 Architectures of Wireless IDS

The structure for the wireless mobile network can be configured depending on various applications. The optimal IDS architecture for a mobile network will totally depend on the network infrastructure itself. In a flat infrastructure, all the nodes are to the same level of priorities and in multi-layered network infrastructure nodes may be separated into different clusters for communication.

4.1 Stand-Alone IDS

As its name suggests data is collected on each and every independent node. Depending on this information, decision is taken for detecting intrusions. In this architecture, each node runs separate IDS.No information or message is passed to each other. Even though restricted by its limitations, it is more adaptable in situation when each node can run an IDS on their own or have IDS installed it is much more preferred for a flat network architecture which will unfortunately not suitable for wireless mobile network.[2]

4.2 Co-operative and Distributed IDS

It is mentioned earlier that stand-alone IDS does not work properly. So, co-operative and distributed intrusion detection system architecture should be implemented. For this there are IDS agents running on each node. IDS agent has complex design but by analyzing it properly and closely it can be viewed as having six different modules.

In this type every single node plays an important and critical role. Each node contributes individually or on entire network for the process of detection. It scans for any sign of intruder. Using six different modules such as local data

collection, local intrusion detection, cooperative message passing, secure communication etc. the above explained approach can be achieved. The IDS also triggers response if intrusion is detected. The isolated IDS agents are entirely linked together to form the IDS system defending the mobile wireless network.[2]

5 Proposed Work

Intrusion detection systems are burglar alarms of computer security field. Here other than alerting the host system on which IDS is installed, this system forwards information of detected intruders to other nodes in the network. The aim is to warn other nodes prior about the intruder so that they can take precaution.

5.1 System Design

Fig. 1 shows overall system design of Co-operative Multi-agent Based Wireless IDS Using MIBs. IDS agent is present on each system. Internal of this IDS agent can be shown as above diagram which has four modules in it.

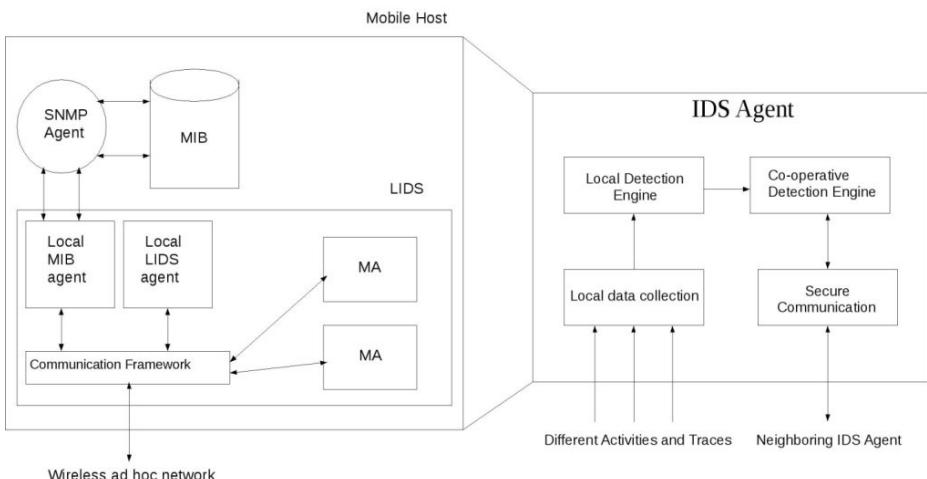


Fig. 1. System Design

- Local data collection module collects audit traces and activity logs.
- Local detection engine module uses information collected by local data collection module for detection of anomaly.
- Co-operative Detection Engine module alerts other nodes when any node detects intrusion locally.
- Secure communication module provides communication channel between two IDS agents.

IDS agent can be viewed as mobile host shown in the above diagram. This includes LIDS, SNMP agent, mobile agent and MIB. LIDS, mobile agent and MIB agent come under Local Detection engine, co-operative detection engine and Local data collection module respectively.

- MIB agents collect information from MIB variables.
- LIDS and mobile agents use information collected by MIB agents for their specific work.
- LIDS detects attack locally at that particular system and reacts to alerts by other IDS nodes.
- LIDS hands over the special task of transporting information about the intruder to other IDS, to mobile agents in that network.

5.1.1 SNMP

In complex network environment it is very difficult task to manage all the devices like routers, switches and servers. They should be up and perform optimally. SNMP helps to do this. It is application level protocol and set of rules that allows computer to get statistics from another computer across the network. This is a standard for managing Internet Protocol (IP) devices. Here a few manager stations control a set of agents. The manager station is a host that runs the SNMP client on it and agent, which is a managed station, is a router that runs the SNMP server program on it.

Computers keep track of information present in routers like information about packets, number of bytes and errors that are transmitted and received through each interface. All this information is kept in a database called MIB. For management tasks SNMP uses this MIB along with Structure of Management Information (SMI).

An agent has a list of objects that it is tracking. This list contains all the information that Network Management System (NMS) can use to determine health of the device on which this agent resides. These objects that agent tracks are managed in MIB defined above. The objects in MIB are categorized under 10 different groups as system, interface, address, translation, ip, icmp, tcp, udp, egp, transmission and snmp. The information from MIB variables can be read by using languages like JAVA.

5.1.2 MIB

The SMI provides a way to define managed objects and their behavior. An agent has in its possession a list of the objects that it tracks. One such object is the operational status of a router interface. This list collectively defines information that NMS can use to determine the overall health of the device on which the agent resides.

The MIB can be thought of as a database of managed objects that the agent tracks. Any sort of status or statistical information that can be accessed by the NMS is defined in a MIB. The SMI provides a way to define managed objects, while the MIB is the definition of the objects themselves. MIB creates a collection

of named objects, there types, and their relationships to each other in an entity to be managed. MIB creates as set of objects defined for each entity similar to a database.

The various values that can be retrieved from a MIB are called MIB variables. These variables are defined in the MIB for a device. Each MIB variable is named by an Object Identifier (OID), which usually has a name in the form of numbers separated by periods ("."), like this: 1.3.6.1.xxxx.x.x.x.x... e.g. the MIB-II has a variable that indicates the number of interfaces (ports) in a router. It's called the "ifNumber", and its OID is 1.3.6.1.2.1.2.1.0. Network monitoring tools will query a device for the MIB variables and display the results. When a device receives a SNMP Get-Request for this ifNumber OID, it will respond with the count of interfaces.

- **Querying MIB variable by Intermapper:** There are two kinds of MIB variables: scalar values and table entries. Scalars have a single value, such as the interface number shown above. For example, the ifNumber MIB variable of a router is a single number that represents the total number of its interfaces. Table values, on the other hand, provide the same pieces of information for different items, such as the traffic for each of a router's ports, or information about each of the TCP connections in a device.
- **Specifying and Accessing MIB variables:** SNMP MIB variables are referenced by an OID, a sequence of digits and dots. This specifies the position of the variable in the MIB tree. Almost all the MIB variables you see commercially will start with 1.3.6.1 (iso.org.dod.internet) and will then either take the proprietary limb of the tree (.4.1: private.enterprise) the standard limb (2.1: mgmt.mib).

5.1.3 System Functions

Fig. 2 gives the static implementation of the system. It shows how the various classes are related to each other using relationships like association and dependency. On larger scale system functions for this co-operative agent based WIDS can be divided into two main functions as intrusion detection and message passing to other nodes. Intrusion detection includes local data collection and local

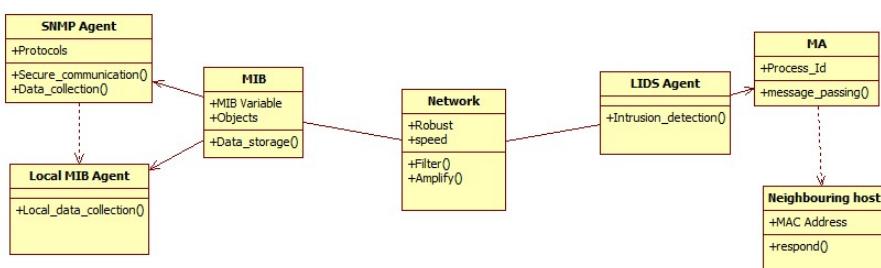


Fig. 2. UML Class Diagram

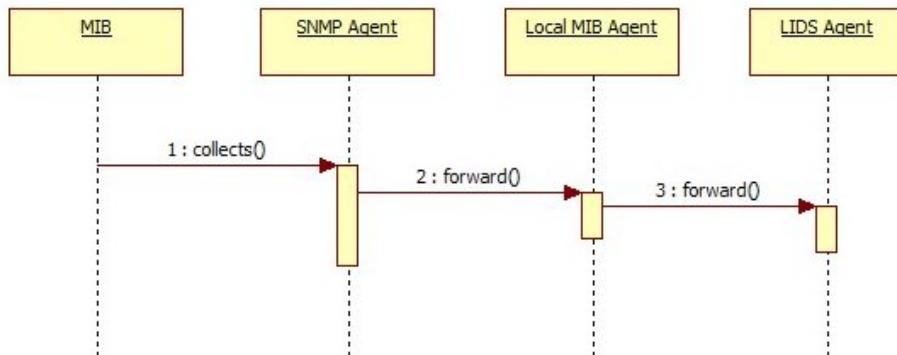


Fig. 3. UML Sequence Diagram 1

detection of intrusion as sub functions. Message passing contains secure communication channel and transfer of messages.

UML sequence diagram in Fig. 3 shows first main function of local intrusion detection. For this first data gets collected from MIB forwarded to local MIB agent and then to LIDS. UML sequence diagram in Fig. 4 shows further main

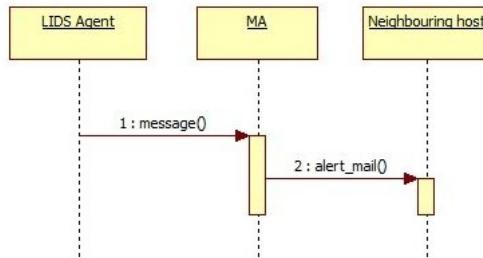


Fig. 4. UML Sequence Diagram 2

function of message passing to other neighboring nodes in the network. Responsibility of this transfer of message is given to MA. LIDS forwards message to MA which in turn passes it to other nodes.

6 Experimentation and Results

Initial requirement will be the installation of SNMP on all the nodes present in ad-hoc network. SNMP agent will extract the information from MIB variables. MIB includes network related information. Thus approach is network based IDS. This information will be analyzed by LIDS agent using either Misuse Based Detection module or Anomaly Based Detection module.

If the intrusion is detected by LIDS, then it will generate alarm locally and then message will be passed to all the nodes present in ad-hoc network using mobile agent. Whenever any such type of message arrives at any node, then corresponding LIDS will generate alarm or will display warning message.

7 Conclusion

All networks are vulnerable to different attacks. Regardless of whether the network is wired or wireless, network security and integrity should always be preserved. As it said that wired IDS are not capable of taking care of things in wireless there is strong demand of effective IDS for wireless networks. Due to dynamic nature of wireless networks it is a challenging topic of research. We have shown that architecture for wireless IDS should be distributed and cooperative in nature. So, in the proposed system using information from MIB variables intrusion will get detected and mobile agents will alert other nodes by passing the message. It works in co-operative way.

We propose to use SNMP data located in MIBs, as an audit source for LIDS. Such a data source provides several advantages:

- It is independent from the operating system.
- It can be extended in order to collect and store additional data relative to network activities, operating system or applications.
- If an SNMP agent runs on a node, the cost of the collection of local information needs no additional resources.
- The standard representation of the data collected on each node facilitates co-operation between LIDS.

Acknowledgement. We owe a great thanks to the people who helped and supported us while writing this survey paper. Our deepest thanks to Prof. V. K. Khatavkar, the guide of the project for guiding and correcting us with attention and care. He has taken efforts to go through the project work and make necessary correction as and when needed.

We would also thank our institute and faculty members without whom this would have been a distant reality.

References

1. Arokia Renjit, J., Shunmuganathan, K.L.: Distributed and cooperative multi-agent based intrusion detection system. Indian Journal of Science and Technology 3(10) (October 2010) ISSN: 0974-6846
2. Sansurooah, K.: Intrusion Detection System (IDS) Techniques and Responses for Mobile Wireless Networks. Edith Cowan University
3. Hijazi, A., Nasser, N.: Using Mobile Agents for Intrusion Detection in Wireless Ad Hoc Networks (2005) 0-7803-9019-9/05
4. Zhang, Y., Lee, W., Huang, Y.-A.: Intrusion Detection Techniques for Mobile Wireless Networks, pp. 3–4 (2003)

5. Haddadi, F., Sarram, M.A.: Wireless Intrusion Detection System Using a Lightweight Agent (2010) 978-0-7695-4042-9/10
6. Rao, A.A., Srinivas, P., Chakravarthy, B., Marx, K., Kiran, P.: A Java Based Network Intrusion Detection System (IDS), Session ENG 206-118 (2006)
7. Kachirski, O., Guha, R.: Effective Intrusion Detection Using Multiple Sensors in Wireless Ad Hoc Networks (2002) 0-7695-1874-5/03
8. Nakkeeran, R., Aruldoss Albert, T., Ezumalai, R.: Agent Based Efficient Anomaly Intrusion Detection System in Adhoc networks (2010) ISSN: 1793-8236
9. <http://tools.ietf.org/html/rfc2248> (accessed November 15, 2011)
10. <http://lyberty.com/encyc/articles/snmp.html> (accessed November 30, 2011)
11. http://docstore.mik.ua/orelly/networking_2ndEd/snmp/ch01_04.htm (accessed November 30, 2011)
12. <http://tools.ietf.org/html/rfc1351> (accessed November 17, 2011)
13. <http://www.opennet.ru/base/cisco/monitor.txt.htm> (accessed November 30, 2011)

A Binary Vote Based Comparison of Simple Majority and Hierarchical Decision for Survivable Networks^{*}

Charles A. Kamhoua¹, Kevin A. Kwiat¹, and Joon S. Park²

¹ Air Force Research Laboratory, Information Directorate, Rome, NY 13441

² Syracuse University, Syracuse, NY 13244

{charles.kamhoua.ctr, kevin.kwiat}@rl.af.mil, jspark@syr.edu

Abstract. Nodes are replicated in fault-tolerant networks not only to increase the aggregate decision reliability but also to survive the failure of a subset of those nodes. A simple majority rule is the most common aggregate decision rule. One may believe that a simple majority rule may not be optimal when node replication is performed in organization following a hierarchical structure like a corporation or a military command. This research shows that if the node's observations are better than random, then a simple majority rule is better than a hierarchical decision. Moreover, even though there are a few compromised nodes that falsify their vote, a simple majority rule will still be superior. However, a hierarchical decision process is more scalable and the vote can be aggregated faster. This paper also proposed a technique based on the law of diminishing marginal utility to calculate the optimum number of nodes in a decision process.

Keywords: binary voting, fault-tolerant Network, hierarchical decision process, network security, reliability, survivability.

1 Introduction

Researchers have recently investigated the internet topology and concluded that it has a hierarchical structure [1]. Further, new developments in cloud computing make it possible to run applications using numerous computer nodes or virtual machines distributed around the world. This advancement in cloud computing facilitates the design of fault-tolerant network. In fact, one approach to fault-tolerant network is node replication. Using replicated nodes, the system-of-nodes can tolerate the failure of a few replicas while guarantying critical functionality.

One specific technique of fault-tolerant network is binary voting. Binary voting is of great interest when the system administrator wants to make a binary decision from the monitoring of a binary event. The binary event of interest may be distributed in the internet, the cloud or in a large organization that has branch around the world. Moreover, most civilian and military organizations have a hierarchical structure.

* This research was performed while Charles Kamhoua and Joon Park held a National Research Council Research Associateship Award at the Air Force Research Laboratory. This research was supported by the Air Force Office of Scientific Research (AFOSR). Approved for Public Release; Distribution Unlimited: 88ABW-2011-6296 Dated 05 December 2011.

Let us consider that each soldier in a battle field is equipped with a sensor that monitors a binary event. Soldiers are partitioned in subset under the control of a captain. The sensor in each soldier directly reports its observation in the form of a binary vote to a minor decision center commanded by a Captain. Each Captain reports as a single binary vote its soldier's majority opinion to a Colonel. Further, each Colonel sends to the General a single binary vote consistent with its Captains' majority vote. Only the General has the power to decide and make its binary decision based uniquely on its Colonel's majority vote.

The main contribution of this paper is to analyze what constitutes the optimum vote aggregation mechanism between simple majority and hierarchical decisions. We find that in most circumstances, the simple majority rule is more robust than hierarchical decision. However, the hierarchical vote aggregation method is faster and more scalable. Further, a special consideration is given to intelligent malicious nodes that attempt to strategically defeat the aggregate results. The chance that the aggregate decision survives node compromising is analyzed in both hierarchical decision and simple majority. In addition, we use the law of diminishing marginal utility to show how to calculate the optimum number of nodes that participate in the decision process.

This work is organized as follows. Section 2 is dedicated to the related works. Section 3 shows how to calculate the optimum number of nodes. After the optimum number of nodes is calculated, we will analyze in Section 4 the optimum nodes' arrangement. Section 5 exhibits our numerical results and Section 6 concludes the paper.

2 Related Works

In recent years, several researches have focused in binary voting. Kwiat *et al.* [2] analyzed the best way to aggregate the node observations given the nodes reliability. The nodes are assumed to be homogeneous. The reliability of a single node p is its probability to make the correct decision. They showed that Majority Rule (MR) performs better if the node' observations are highly reliable (p close to 1). But for low value of p , $(p < \frac{1}{2})$ choosing a Random Dictator (RD) is better than MR. Random Troika (RT) combines the advantage of those two strategies when the node reliability is unknown ($0 \leq p \leq 1$). Generally, it can be shown that if a small proportion of nodes are compromised and nodes are highly reliable, assuming that an odd number of nodes is used, we will have $\text{MR}=\text{Random N} > \dots > \text{Random 5} > \text{RT} > \text{RD}$. However, if the majority of nodes are compromised, the previous inequality is reversed. That is because, with a majority of compromised nodes, increasing the size of the subset of deciding nodes also increases the likelihood of compromised nodes taking part to the decision.

Following the previous research, Wang *et al.* [3] analyzed the nodes decision in a cluster. There are n clusters of m nodes, with a total of $n*m$ nodes. The attacker chooses the number of clusters to attack while the defender chooses how many nodes participate in the decision in each cluster. They formulated a zero-sum game in which the defender maximizes the expected number of clusters deciding correctly while the attacker minimizes that number. They proposed a general framework to find the Nash equilibrium of such game. However, the cluster structure is assumed to be fixed. This research will show that the defender has a better optimization strategy just by changing the cluster structure.

Malki and Reiter [4] analyze Byzantine quorum systems. They propose a masking quorum system in which data are consistently replicated to survive an arbitrary failure of data repositories. Their work also proposes a disseminating quorum system. Faulty server can fail to redistribute the data but cannot alter them.

Bhattacharjee *et al* [5] use a distributed binary voting model in cognitive radio. To compensate their noisy observation of channel utilization by primary spectrum users, each secondary user requests their neighbor opinion (vote). Those interactions are repeated and the Beta distribution is used to formulate a trust metric. Nodes with low trust are eliminated to have a more accurate channel evaluation. Replica voting for data collection in active environment is investigated in [6-7].

Park *et al.* [8-9] proposed a trusted software-component sharing architecture in order to support the survivability at runtime against internal failures and cyber attacks in mission critical systems. They defined the definition of survivability using state diagrams, developed static and dynamic survivability models, and introduced the framework of multiple-aspect software testing and software-component immunization.

Alongside the research above, there is a large mathematic literature about binary voting starting with Condorcet [10]. Simply stated, the Condorcet Jury Theorem (CJT) shows that if a group of homogeneous and independent voters, with voter competence better than random, uses the simple majority rules to choose among two alternatives having equal a priori probability, then the group's decision accuracy monotonically increases and converges to one as the number of voter increases. Owen *et al.* [11] generalized the CJT while considering any distribution of voter competence. The original CJT was restricted to a uniform distribution of voter competence or reliability p . A mathematical survey of binary voting is provided in [12].

3 Calculation of the Optimum Number of Nodes

The optimum number of replicated nodes has attracted less attention in the literature. The implicit assumption is that the number of nodes that participate in the decision is given. However, we believe that that number strongly contributes to optimizing the decision center's reactions. We are proposing an optimization approach based on the law of diminishing marginal utility.

Without loss of generality, we assume in this paper that the nodes are homogeneous and that each node's reliability is p . We also consider that $0.5 < p \leq 1$. Therefore, in the framework of Condorcet [10], using a simple majority and without malicious nodes, the reliability of the decision monotonically increases and converges to one as the number of voter grows to infinity. This is valid for either the simple majority rule or the hierarchical decision process.

In the democratic political system that Condorcet advocated, the government organizes the election and does not pay its citizens to vote. Thus, a larger electorate increases the result accuracy at no fee to the government. Accordingly, a larger electorate is always better in term of vote accuracy. However, in fault-tolerant networks, there is a system designer's cost associated with any additional voter (node, sensor). Precisely, there is a tradeoff between costs and accuracy in fault-tolerant networks. We will show that above the optimum number of nodes, any increase is inadequate.

Let C be the cost of a node and V be the value of the target being protected by a mission. A binary voting mechanism is implemented to aggregate the decision of the N nodes. Let us take $m = \frac{N+1}{2}$. The probability $P_N(p)$ that N nodes reach the correct decision in a majority rule can be calculated as:

$$P_N(p) = \sum_{k=m}^N \binom{N}{k} p^k (1-p)^{N-k}, \text{ with } m = \frac{N+1}{2}. \quad (1)$$

When we increase two nodes, the new decision accuracy becomes:

$$P_{N+2}(p) = \sum_{k=m+1}^{N+2} \binom{N+2}{k} p^k (1-p)^{(N+2)-k}. \quad (2)$$

We will proceed in two steps. In the first step, we repeat the CJT to show that P_N monotonically increases. The second step will show that the rate of that increment decreases. Those two steps are enough to validate a diminishing marginal utility.

Theorem 1: The sequence P_N increases with N when $0.5 < p \leq 1$. (CJT, [10])

Proof: The following recursion formula holds:

$$P_{N+2} = P_N + p^2 \binom{N}{m} p^{m-1} (1-p)^m - (1-p)^2 \binom{N}{m} p^m (1-p)^{m-1}. \quad (3)$$

In fact, two additional voters can influence a binary election using the simple majority rules if and only if one alternative has one vote more than the other. The second term of the right hand side (RHS) of (3) is the probability that the incorrect alternative has one more vote than the correct one and the two new voters vote correctly. The third term of the right hand side (RHS) of (3) is just the reverse or the probability that the correct alternative has one more vote than the incorrect one and the two new voters vote incorrectly. After a few algebraic manipulations, we have:

$$P_{N+2} - P_N = (2p - 1) \binom{N}{m} [p(1-p)]^m > 0 \text{ if } p > 0.5. \quad (4)$$

Theorem 2: The sequence $W_N = P_{N+2} - P_N$ decreases with N when $0.5 < p \leq 1$. ■

Proof:

$$\begin{aligned} \frac{W_{N+2}}{W_N} &= \frac{P_{N+4} - P_{N+2}}{P_{N+2} - P_N} = \frac{(2p - 1) \binom{N+2}{m+1} [p(1-p)]^{m+1}}{(2p - 1) \binom{N}{m} [p(1-p)]^m} = \frac{2p(1-p)(N+2)}{m+1} \\ &= \frac{4p(1-p)(N+2)}{N+3} < 4p(1-p) < 1. \end{aligned} \quad (5)$$

Theorem 2 shows that the marginal reliability value of two additional nodes diminishes. Then increasing the number of nodes yields concave utility. Thus, ■

applying the law of diminishing marginal utility, the optimum number of nodes to use in the decision process should be the larger number N such that:

$$(P_{N+2} - P_N)V \geq C. \quad (6)$$

Figure 1 provides a numerical example. We use $p = 0.9$. We can see that $P_1 = p = 0.9$, $P_3 = 0.91944$, $P_5 = 0.925272$. Thus, the increase in precision from the addition of the first two nodes (2%) is higher than that of the last two (0.5%).

We have provided an approach to calculate the optimum number of nodes when using the simple majority rule with uncompromised nodes. However, this approach can be generalized to the case of RD, RT, or when nodes are arranged in cluster or hierarchically and in the presence of compromised nodes.

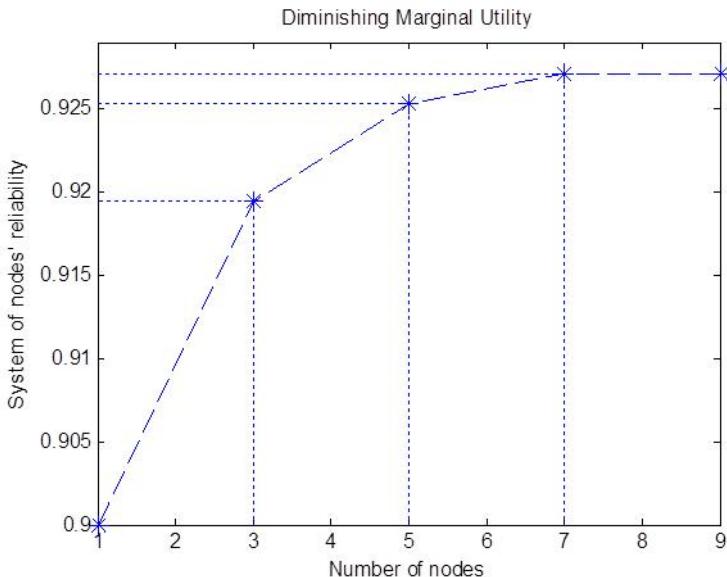


Fig. 1. Decision reliability as a function of N

4 Optimum Node Arrangement

In this Section, we discuss the optimum nodes' structure to maximize their fault tolerance. For the purpose of the discussion in this section, we define the *tenacity* of a structure of nodes as the minimum proportion of nodes that an attacker must compromise to have a total control over the aggregate decision. Therefore, using a simple majority with N nodes (N odd), the *tenacity* will be $m = \frac{N+1}{2}$, or close to 50%.

4.1 Clustering

Presently, let us consider a rational agent (the defender) at the decision center that believes more than 50% of its nodes have been compromised by an attacker. The

attacker will then have full control over the decision outcome if using a simple majority rule to aggregate the votes. To respond to the situation, the defender may simply arrange its nodes in cluster as in Table 1.

In Table 1, the C represents the compromised nodes and R the regular nodes. In the first three columns, R is the majority. Thus, we may have a correct decision in the majority of columns or clusters. Let us also consider that the defender aggregate's result is that of the majority of nodes in the majority of clusters. In this case, if we consider highly reliable nodes, 25 nodes can survive the failure of up to 16 nodes as illustrated in Table 1. This is a clear improvement compare to simple majority rule that can only survive the failure of 12 out of 25 nodes. However, the defender can take advantage of the cluster structure if and only if that structure is unknown to the attacker. For instance, an attacker that knows the cluster structure just needs to compromise 9 nodes out of 25 as presented in Table 2. In this case, using a simple majority rule is a superior solution. In fact, the attacker's optimum strategy is to compromise a bare majority of nodes in a bare majority of clusters.

Table 1. 25 Nodes Illustration

C	C	C	C	C
C	C	C	C	C
R	R	R	C	C
R	R	R	C	C
R	R	R	C	C

Table 2. 25 Nodes Illustration

R	R	R	R	R
R	R	R	R	R
R	R	C	C	C
R	R	C	C	C
R	R	C	C	C

Using $N = (2k + 1)^2$ nodes for instance, we can see that the attacker just needs to compromise $(k + 1)$ nodes in $(k + 1)$ clusters for a total number of $(k + 1)^2$ nodes out of the $(2k + 1)^2$ nodes (see Table 2). The ratio $\frac{(k+1)^2}{(2k+1)^2} = \frac{k^2+2k+1}{4k^2+4k+1}$ converges to 0.25 when k grows.

In contrast, when the defender knows the cluster structure while that structure is unknown to the attacker, the optimum attacker's strategy is to randomly attack the nodes. As a consequence, taking the ratio, the cluster structure can survive the failure of $1 - \frac{(k+1)^2}{(2k+1)^2}$ nodes (see Table 1), or 75%. By definition, the maximum tenacity of a cluster structure is 75%.

In short, the clustering strategy in two dimensions (table 1) cannot protect the aggregate decision when the number of compromised nodes is higher than 75%. To survive, the compromising of more than 75% of nodes, clustering should be applied in three dimensions or higher. We see in the next section that arranging the nodes hierarchically (which is higher dimension clustering) can possibly survive the compromising of more than 75% of nodes.

4.2 Hierarchical Troika

Figure 2 shows 27 nodes hierarchically arranged in subset of three nodes (hierarchical troika). The 27 nodes make their decision in three layers. The resulting decision of the higher layer is that of at least 2 out of 3 nodes in the lower layer. Since $27=9*3=3*3*3$, the first layer has 27 nodes, the second layer has 9 sub results, the third layer has 3 sub results, and the final aggregate result is obtained from the last three results.

The higher layers are not nodes but materialize the logical aggregate decision from the lower layers. Recall that in our scenario, only the soldiers on the ground are equipped with sensors. Thus, logical connection materializes the hierarchical structure: a Captain sending the partial aggregate vote to the Colonel that in turn will partially aggregate the vote at his layer to send it to the General. In Corporations, logical connection should be Supervisor sending the vote to the Manager that will report to the Director.

A careful analysis of this decision process shows that the 27 nodes can tolerate the failure of up to 19 nodes (see Fig. 2). Figure 2 shows 19 red nodes (compromised) and 8 green nodes (regular). The green line shows the transmission of a correct vote to the higher layer while a red line shows the transmission of an incorrect vote. We can see that the final decision is correct because two out of three votes are correct in the upper layer.

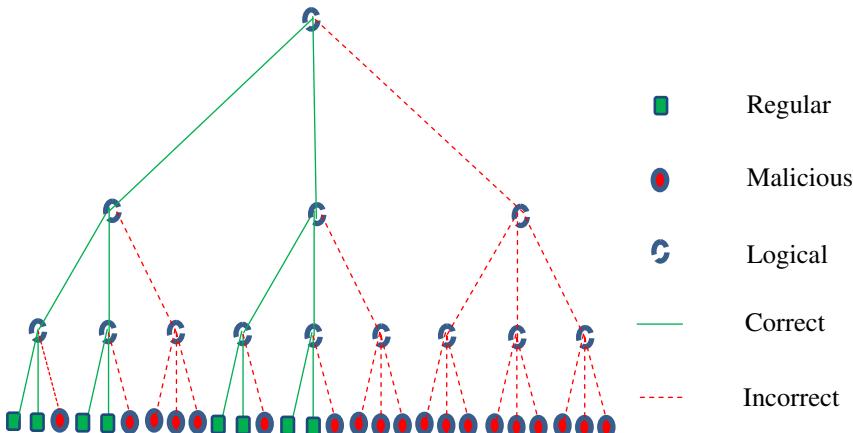


Fig. 2. Hierarchical Troika nodes arrangement

Extending the process with 81 nodes, we see that the system tolerates the failure of up to 65 nodes or 80% of nodes. The truthful node can win an election with only 20% of the vote. In general, 3^n nodes require n layers decision process and can tolerate the failure of $(3^n - 2^n)$ nodes. Thus, the tolerance ratio is $\frac{(3^n - 2^n)}{3^n} = 1 - \left(\frac{2}{3}\right)^n$ which converges to 100% as n grows. Therefore, we can see that a hierarchical node arrangement maximizes the *tenacity* of the network when the attacker does not know the nodes assignment into

the hierarchical structure. The analysis in this subsection remains valid in other hierarchical structure such as hierarchical 5, 7 ...or a combination.

4.3 Structure Comparison

In summary, we have shown that clustering the nodes or arranging them hierarchically is a valuable strategy to the defender if and only if it is hard for the attacker to infer the cluster structure. One way to insure that is for the defender to randomly and periodically reassign the nodes in different clusters or troika. Moreover, we can see that there is a tradeoff between resisting the compromising of a large number of nodes and the risk to be exposed to the compromising of a few nodes. Therefore, the defender's belief about the distribution of the number of compromised nodes is the most important factor that determines the best structure to use.

Generally speaking, at a given time, if the defender believes that the attacker has compromised only a minority of nodes, the defender may choose among simple majority rules, clustering, or hierarchical troika. However, if the defender believes that the attacker has compromised between 50% and 75% of nodes, the defender must avoid simple majority and use clustering or hierarchical troika. When more than 75% of nodes are compromised, the only solution left is hierarchical troika.

We can perform a similar comparison when the defender does not know the exact number of compromised nodes but his belief about that number has a specific probability density function (PDF). Table 3 provides a summary. Definitely, more specific results will depend on the exact PDF (uniform, normal, exponential...), the shape of the PDF (symmetric or skewed), a node's reliability, and the number of nodes.

Table 3. Comparison of the Structures

		Majority	Cluster	Troika
Tenacity		Low	Medium	High
Scalability		Low	Medium	High
Speed of vote aggregation		Low	Medium	High
Proportion of compromised nodes	0-50%	Good	Good	Good
	50%-75%	Poor	Good	Good
	75%-100%	Poor	Poor	Good
Probability density function of defender belief's on the number of compromised nodes	Uniform, Symmetric	Good	Good	Good
	Positive skew	Good	Good	Good
	Negative skew	Poor	Medium	Good

Note that we can also have a rectangular cluster. In a rectangular cluster, the number of rows should be as close as possible to the number of columns to maximize the structure *tenacity*. Moreover, additional precautions can be taken to arrange the nodes hierarchically even though the number of nodes is not a specific power of an integer. Section 5 will reveal more structure comparisons.

5 Simulation Results

This section shows some simulation results to support the different techniques analyzed in this paper. Those simulation results are generated from MATLAB. We organize the results in two sets of experiment. Each set examines a specific factor. The first factor we examine is the system of nodes reliability. We compare the system of nodes reliability when there is no malicious node in two structures: hierarchical troika and simple majority. This experiment helps to characterize the nodes structure when malicious nodes' interventions are not the first concern but the possibility of mistake from truthful nodes when observing the state of nature. The second factor we consider is the impact of a malicious node in a system of nodes. Again, we consider hierarchical troika and simple majority rule.

5.1 System-of-Node Reliability Comparison

We can observe that hierarchical troika and simple majority are identical in the case of 3 nodes. Figure 3 and 4 show how the group of nodes decision reliability varies with individual node reliably using 9 and 27 nodes respectively.

The result is that hierarchical troika outperform simple majority when $0 \leq p < 0.5$ and the reverse is true when $0.5 < p \leq 1$. We forecast that this result holds for any number of nodes above 27. Thus, if nodes reliability is such that $0.5 < p \leq 1$, and the defender's main concern is to increase the collective decision reliability while not considering the malicious nodes' action, a simple majority should be used.

However, the main concern of this research is the malicious nodes' action. We deal with malicious nodes' action in the next subsection. We can also see that hierarchical troika is also consistent with CJT. If we have $p = 0.5$, the system of node reliability stays at 0.5. When we have $p > 0.5$ ($0 \leq p < 0.5$ respectively) the system of node reliability is above 0.5 (below 0.5 respectively). Looking at the difference between Fig. 3 and 4, we see a fast convergence to one when $0.5 < p \leq 1$ (zero respectively if $0 \leq p < 0.5$) as the number of nodes increases. We can also see that the convergence is faster as we move away from $p = 0.5$. Also, as the number of nodes increases or the node's reliability increases, the difference between simple majority and hierarchical troika becomes negligible.

To generalize our analysis above 27 nodes, we have already shown that the aggregate decision reliability using simple majority increases according to the sequence (3) and (4). A similar sequence can be derived using hierarchical troika. First, we need to observe that using three nodes, the aggregate decision reliability in a troika is:

$$P^T_3 = 3p^2 - 2p^3. \quad (7)$$

Given the recursive structure of hierarchical troika, we have

$$P^T_{3N} = 3(P^T_N)^2 - 2(P^T_N)^3. \quad (8)$$

Equation (4) and (8) allow a direct comparison of simple majority for any number of nodes that is a power of 3.

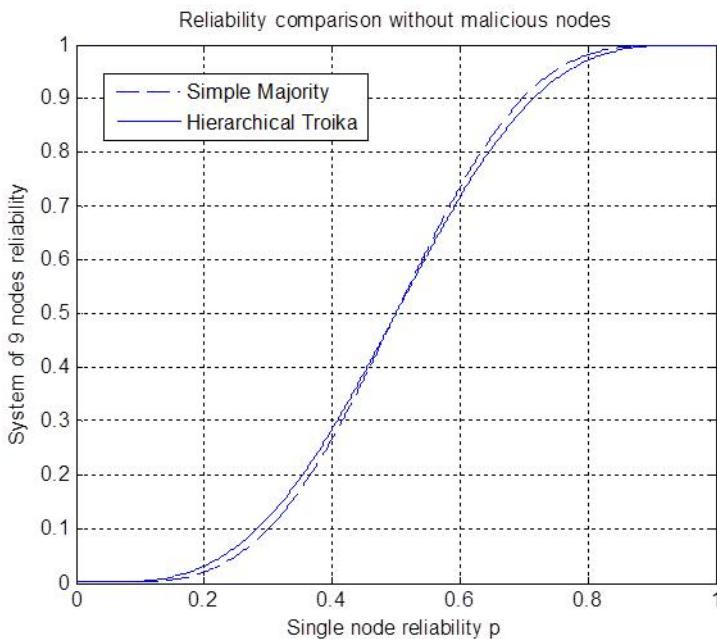


Fig. 3. Troika vs. Majority group of 9 nodes decision reliability

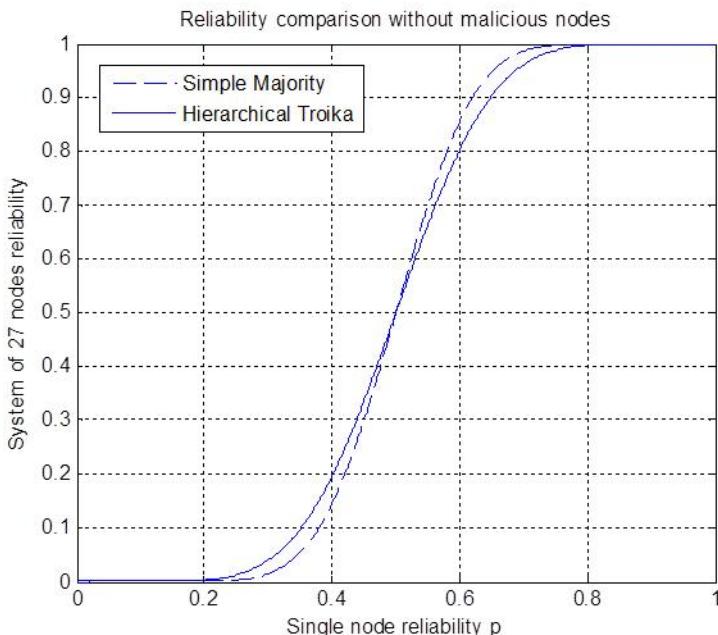


Fig. 4. Troika vs. Majority group of 27 nodes decision reliability

5.2 Malicious Nodes Influence Comparison

One metric we can use to compare the malicious nodes influence is the probability that a single vote changes the aggregate decision. Again, we compare troika and simple majority using 9 and 27 nodes because they are powers of 3.

Using simple majority with an odd number of nodes, a single malicious node can change the aggregate decision if and only if the vote from other nodes breaks even. The probability of a tie vote in a simple majority is:

$$C_N = \binom{N-1}{m-1} p^{m-1} (1-p)^{m-1}, \text{ with } m = \frac{N+1}{2}. \quad (9)$$

Using hierarchical troika, the process is different. First, we can see that if there are only three nodes, a malicious node changes the aggregate decision if the two other nodes have different votes, that happen with probability

$$C^T_3 = 2p(1-p). \quad (10)$$

Second, with 9 nodes, there are two decision layers (see Fig. 2). A single node can influence the 9 nodes decisions if at the first layer the two other nodes have different votes (which happen with probability $C^T_3 = 2p(1-p)$) and, at the second layer, the two logical connections have different results (which happen with probability $2P^T_3(1-P^T_3)$). Thus, we have

$$C^T_9 = [2p(1-p)][2P^T_3(1-P^T_3)]. \quad (11)$$

More generally, we have

$$C^T_N = \prod_{k=1}^{\log_3 N} 2P^T_k(1-P^T_k), k \text{ a power of 3}. \quad (12)$$

Figure 5 shows that a malicious node has a stronger influence on simple majority than on hierarchical troika if $\frac{1}{3} < p < \frac{2}{3}$. The reverse is true elsewhere. Further, if we take the integral for all value of p ($0 \leq p \leq 1$), hierarchical troika and simple majority have equal results. Figure 6 shows a similar result to Fig. 5 but with the interval in which hierarchical troika is more effective than simple majority is reduced to $0.4 < p < 0.6$. We foresee that this interval will continue to be reduced as the number of nodes increases.

Recall that in Section 4 we considered highly reliable nodes and showed that hierarchical troika will outperform simple majority if a high proportion of nodes (approximately more than 50%) are compromised. Thus, we anticipate that when nodes are highly reliable ($\frac{2}{3} < p \leq 1$ for 9 nodes), if there is a small proportion of compromised nodes, simple majority should be used. However, as the number of compromised nodes increases, there must be a critical proportion above which hierarchical troika is superior to simple majority.

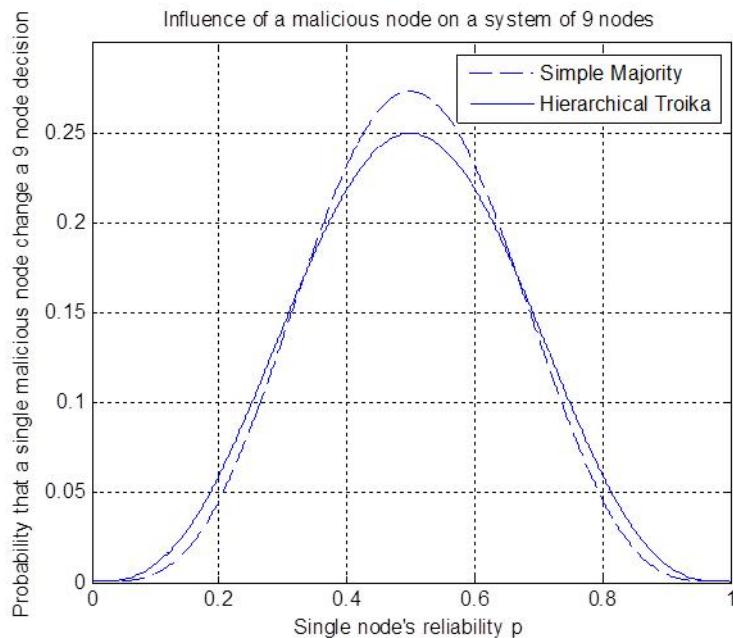


Fig. 5. Troika vs. Majority in the mitigation of malicious nodes' vote

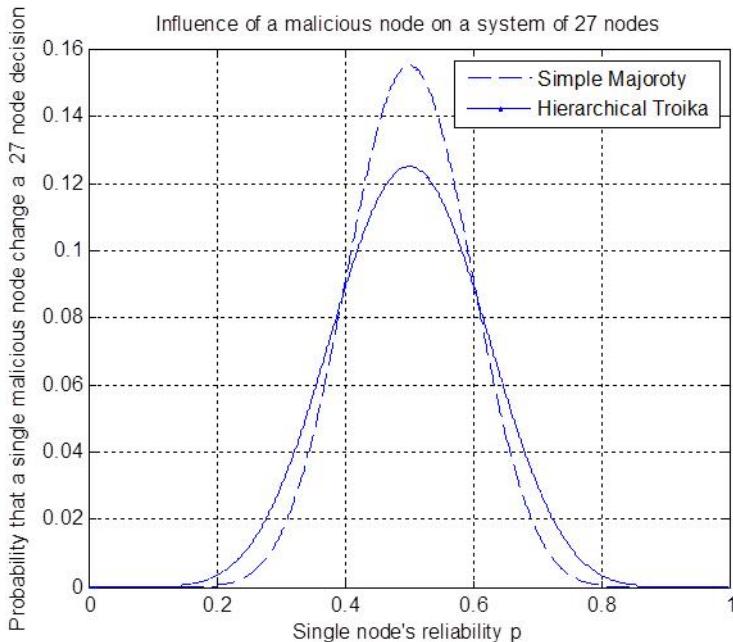


Fig. 6. Troika vs. Majority in the mitigation of malicious nodes' vote

In this comparison, we have focused our attention to simple majority and hierarchical troika. In fact, those are the two extreme structures. For instance, the cluster of Table 1 and 2 can be represented as hierarchical 5. We may also have hierarchical 7, 9...Furthermore, simple majority can be represented as hierarchical N with the entire vote aggregated in one step. The properties of hierarchical 5, 7, and other can be inferred from the comparison of simple majority and hierarchical troika. For instance, in section 5.1, we can easily infer that if $0.5 < p \leq 1$, the aggregate decision reliability must be such that: simple majority $> \dots >$ hierarchical 5 $>$ hierarchical troika.

6 Conclusion and Future Works

This research has compared the simple majority rule with a hierarchical decision process in a strategic environment. We have shown that a simple majority decision rule is generally preferable to a hierarchical decision. A hierarchical decision process should only be chosen when fast vote aggregation in a large scale network is the primary metric to consider. We will investigate in more detail the scalability and speed of troika in our future research.

References

1. Ge, Z., Figueiredo, D., Jaiswal, S., Gao, L.: Hierarchical structure of the logical Internet graph. In: The Proceeding of SPIE, vol. 4526, p. 208 (2001)
2. Kwiat, K., Taylor, A., Zwicker, W., Hill, D., Wetzonis, S., Ren, S.: Analysis of binary voting algorithms for use in fault-tolerant and secure computing. In: International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt (December 2010)
3. Wang, L., Li, Z., Ren, S., Kwiat, K.: Optimal Voting Strategy Against Rational Attackers. In: The Sixth International Conference on Risks and Security of Internet and Systems, CRiSIS 2011, Timisoara, Romania (September 2011)
4. Malki, D., Reiter, M.: Byzantine quorum systems. *Distributed Computer System*, 203–213 (1998)
5. Bhattacharjee, S., Debroy, S., Chatterjee, M., Kwiat, K.: Trust based Fusion over Noisy Channels through Anomaly Detection in Cognitive Radio Networks. In: 4th International Conference on Security of Information and Networks (ACM SIN 2011), Sydney, Australia (November 2011)
6. Ravindran, K., Rabby, M., Kwiat, K., Elmetwaly, S.: Replica Voting based Data Collection in Hostile Environments: A Case for QoS Assurance With Hierarchical Adaptation Control. *Journal of Network and Systems Management* (2011)
7. Ravindran, K., Rabby, M., Kwiat, K.: Data Collection in Hostile Environments: Adaptive Protocols and Case Studies. In: The Second International Conference on Adaptive and Self-Adaptive Systems and Applications (2011)
8. An, G., Park, J.S.: Cooperative Component Testing Architecture in Collaborating Network Environment. In: Xiao, B., Yang, L.T., Ma, J., Muller-Schloer, C., Hua, Y. (eds.) ATC 2007. LNCS, vol. 4610, pp. 179–190. Springer, Heidelberg (2007)

9. Park, J., Chandramohan, P., Devarajan, G., Giordano, J.: Trusted component sharing by runtime test and immunization for survivable distributed systems. In: Sasaki, R., Qing, S., Okamoto, E., Yoshiura, H. (eds.) *Security and Privacy in the Age of Ubiquitous Computing*, pp. 127–142. Springer (2005); Proceedings of the 20th IFIP TC11 International Conference on Information Security (IFIP/SEC), Chiba, Japan, May 30-June 1 (2005)
10. Condorcet: *Essai sur l'application de l'analyse a la probabilite des decisions rendues a la pluralite des voix*. Paris: Imprimerie Royale (1785)
11. Owen, G., Grofman, B., Feld, S.: Proving a Distribution-Free Generalization of the Condorcet Jury Theorem. *Mathematical Social Sciences* 17, 1–16 (1989)
12. Grofman, B., Owen, G., Feld, S.: Thirteen Theorems in Search of the Truth. *Theory and Decision* 15, 261–278 (1983)

A Novel Way of Protecting the Shared Key by Using Secret Sharing and Embedding Using Pseudo Random Numbers

P. Devaki¹ and G. Raghavendra Rao²

¹ Associate professor, Department of Information Science and Engineering
The National Institute of Engineering, Mysore-8
Karnataka, India

p_devaki1@yahoo.com

² Professor, Department of Computer Science and Engineering
The National Institute of Engineering, Mysore-8
Karnataka, India
grrao57@gmail.com

Abstract. This work is modified work of Anil kumar and Navin Rajgopal, where they have considered that dealer encrypts the secret and share the secret with 2 participants. Only the dealer can encrypt and decrypt the secret. Even the sharing and reconstruction of secret is performed by only the dealer. Due to this no other user will be able to reconstruct the secret. There are few drawbacks in this method. This works with only 2 participants as the second image is dependent on the first share and the image. For reconstruction both the shares are must. The dealer encrypts and decrypts the secret which takes more time. No other participants can reconstruct the secret even though they have the secret.

We have considered this work and made some changes so that no encryption decryptions are performed. The secret can be shared among any number of participants, and a few shares are sufficient to reconstruct the secret. The reconstruction can be done by any participant unlike the work mentioned above, where only the dealer can reconstruct the secret. We are also considering the images to cover the secret shares based on the pseudo random numbers, but this is different from the way the numbers are used in the above method.

Keywords: Cover image, Secret key/password, Secret Sharing, Embedding, Pseudo Random Sequence.

1 Introduction

Secret sharing is a method to protect the secret information like encryption key, password, or any short messages. Sharing of key among several authorized users is necessary when the system can be accessible by multi users in any organization. To maintain the secrecy and also the accountability it is better to divide the key by using some method so that the key shares can be given to multiple users instead of giving the whole key to a single user. This sharing protects from unauthorized users getting the key. It also avoids a single user, compromising the key for some reasons. It also avoids single point of failure.

There are many methods to share a secret key. The simplest and first one was introduced by shamir [2]. He has given threshold secret sharing. Followed by Asmuth bloom[15], Blakely[5], Thein and Lin [12].

Threshold secret sharing is the one where the key will be divided in to number of shares based on the number of authorized users. When the key is required then any authorized user can collect the shares from a set of authorized users and reconstruct the key. (m, n) indicates that total number of authorized users is n and out of n number of shares, only m number of shares is sufficient to reconstruct the key. This overcomes the drawback of single user owning the key. The importance of this may be in banking, medical field, military, and business field where important sensitive information is being maintained.

Shamir's threshold secret sharing is based on the polynomial interpolation. He uses Lagrange's interpolation for reconstruction of the key. When it is necessary to distribute the key among multiple users, this secret sharing method can be used to generate the shares using the polynomial of the order $m-1$.

The polynomial is

$$f(x) = S + C_1x + C_2x^2 + \dots + C_{m-1}x^{m-1}$$

Where S is the secret , to be shared among the users. C_1, C_2, C_3 are the coefficients whose values are randomly selected. Using this polynomial and selecting unique values for x , number of shares can be generated.

$$f(x_1) = y_1$$

$$f(x_2) = y_2$$

....

....

....

$$f(n) = y_n$$

Where y_1, y_2, \dots, y_n are the shares generated. Now the shares along with the x values will be distributed to the authorized users. The dealer is not responsible for reconstruction of the key. So each user will get a pair of values x_i and y_i .

When it is necessary for a user to reconstruct the key, he requests other users to send their shares. After receiving at least m number of shares the user can reconstruct the key. The significance of this secret sharing method is, no single share will reveal any information about the key, and with less than m number of shares also it is not possible to reconstruct the key. Only if a user gets minimum m number of shares from the authorized users, he can reconstruct the key. This will ensure that even if an attacker gets a share or few shares which is less than m , the attacker will not be able to reconstruct the key. This gives protection against any attacker getting the key, and also protects against any authorized user miss using the key.

When it is required to reconstruct the key, the Lagrange's interpolation method can be used to solve for the coefficients especially S .

The interpolation formula is:

$$f(x) = \frac{(x-x_1)(x-x_2)y_0}{(x_0-x_1)(x_0-x_2)} + \frac{(x-x_0)(x-x_2)y_1}{(x_1-x_0)(x_1-x_2)} + \frac{(x-x_0)(x-x_1)y_2}{(x_2-x_0)(x_2-x_1)} + \dots$$

2 Embedding

In [1] they use 2 cover images X1 and X2 for 2 participants and based on the pseudo random values generated, they select the pixels from X1 and perform some operations and combine that with X2 to hide the key. Here both X1 and X2 are required while reconstructing the secret.

We are considering the multiple cover images I1, I2, I3 , ...In, based on the number of users . The cover images can be same for all the users or it can be different for different users. It would be better to use different images for different users, as the difference can not be calculated for different images and not possible to guess about the shares.

Using any good pseudo random number generator we can embed the shares in these images.

For a user 1, I1 is selected and the pseudo random sequence is also selected, according to this random number, the share y1 will be embedded in the cover image I1.

Like this all the shares y1 y2 y3....yn will be embedded in to I1 I2 I3...In based on the random sequence generated for each participant. Unlike the normal embedding method in stenography, where a specified bit of every pixel in the cover image, will be replaced by a bit of the secret.

If the size of the secret is much smaller than the cover image, then instead of replacing a single bit of a pixel, one pixel can be replaced by a pixel of the secret. This replacement is based on the random number sequence.

3 Replacement of Pixels Based on Pseudorandom Number

Any good random number generator can be used to generate the numbers from 1 to n. where n is the number of pixels in the cover image.

If the random sequence is:

$$R = \{3, 9, 20, 50, \dots\}$$

And if the share for the user 1 is

$$S1 = \{8, 3, 50\}$$

Then the third pixel in the cover image will be replaced by the value 8, 9th pixel will be replaced by 3, 20th pixel will be replaced by 50 and so on.

Like this for every user the pixels will be replaced by the share values based on the random numbers.

Since only a few pixel values are going to be changed, it is not going to affect the quality of the cover image and it is impossible for an unauthorized user to make out that there is some data in the cover image.

```
I= 10000101      p1
    11010101      p2
    00110101      p3
    01010101      p4
    01010000
    11111111
    00000011
    11111000
    01010111
    00111010      p10

.....
.....
1000000      p20
.....
.....
.....
11110101      p50
.....
.....
```

Based on the above random number sequence the pixels of the cover image will be changed to

```
I= 10000101      p1
    11010101      p2
    00001000      p3
    01010101      p4
    01010000
    11111111

    00000011
    11111000
    00000011
    00111010

.....
.....
1000000      p20
.....
.....
.....
00110010      p50
```

The shaded part shows the replaced pixels. This method will not affect the overall appearance of the cover image. This is repeated for all the cover images of the different users. When it is necessary for a user to reconstruct the key, the user requests the other users for their shares.

When m number of images is collected by the user, he can reconstruct the key, by extracting the shares based on the random number sequence. After extracting from all the m images, using the Lagrange's interpolation method the key can be reconstructed.

4 Proposed Work

We are considering the key of text data type. And the cover image is the gray image, for simplicity we are considering $(2, 4)$ threshold secret sharing.

Algorithm for generating the shares and embed the same in cover images

1. Convert the secret key S into its equivalent ASCII values
2. Generate 4 random number sequences $R1, R2, R3, R4$ where the sequence is from 1 to n . n is the number of pixels in the cover image.
3. Using the Shamir's threshold secret sharing method, divide the S into 4 shares.
4. Embed these 4 shares into 4 cover images as explained in section 3.
5. Send these cover images to the respective users.
6. Also send the random sequences to all the users in a secured manner.

Algorithm for reconstructing the key

1. User, who needs the key, sends a request to all the authorized users.
2. When he receives minimum $m-1$ images from different users along with their random sequences, he starts the reconstruction
 - a) Extract the pixel values based on the random number sequence from each of the images provided by the users.
 - b) Collect the shares from the images
 - c) Using the Lagrange's interpolation formula solve the equations for S which is the ASCII equivalent of the key.
 - d) Convert ASCII values back to text data which is the key.

Secret key $S = \text{benz}$

Random number sequences for 4 users are

$$\begin{aligned} R1 &= \{8, 4, 90, 200\} & R2 &= \{70, 180, 300, 1\} \\ R3 &= \{250, 5, 89, 2\} & R4 &= \{33, 83, 200, 1\} \end{aligned}$$



Fig. 1. Cover Image

The ASCII values of the key is {66, 69, 78, 90}

These values are divided in to 4 shares using Shmair's threshold secret sharing.

Since we are using (2,4) threshold sharing each ASCII value need to be divided in to 4 shares by considering the following polynomial.

$$\begin{aligned}
 f(x) &= S + C_1x \\
 f(1) &= 66 + 2x = 68 \\
 f(2) &= 66 + 4 = 70 \\
 f(3) &= 66 + 6 = 72 \\
 f(4) &= 66 + 8 = 74
 \end{aligned}$$

Like this calculate $f(x_i)$ for all the values 69, 78, 90.

After calculating for all the 4 values create the shares S1 S2 S3 S4

$$\begin{aligned}
 S1 &= [68, 71, 80, 92] \\
 S2 &= [70, 73, 82, 94] \\
 S3 &= [72, 75, 84, 96] \\
 S4 &= [74, 77, 86, 98]
 \end{aligned}$$

Now embed these shares in the above grey level image based on the random numbers generated for 4 users.

Here we have considered only one grey image. Based on the random number for the users, the corresponding share will be embedded.

Instead of using the same image for all the users, we can use different images of different sizes. But according to the size of the images we need to generate the random numbers. This makes the attacker impossible to guess any data hidden in the image.

The images I1, I2, I3, and I4 after embedding the shares look as if it were the original image with out any visible difference.

Reconstruction

Assume that user 2 wants to reconstruct the key. He sends a request to all the users 1, 3 and 4.

Assume that user 2 receives the image from user 4, since he is having 2 images one belonging to himself and another belonging to user 4, user 2 can start reconstructing the key without waiting for the other 2 users.

Steps:

1. Using the random number sequence of user 2 and user 4, user 2 extracts the bits from the 2 images.
2. After extracting he gets 2 shares

$$S2 = [70, 73, 82, 94]$$

$$S4 = [74, 77, 86, 98]$$

3. Now he uses Lagrange's interpolation method to recover the key values 66, 69, 78, 90
4. Then those integer values will be converted to the corresponding characters, which is nothing but "benz".

5 Security Analysis

Our method provides protection against any kind of attacks due to following reasons.

a) As the secret key "benz" is converted to ASCII equivalent and then each value is divided into 4 shares, it is very difficult for an attacker to guess or compute the secret. b) Each share belonging to different user has been embedded in the cover image based on the random number sequence for that user. c) Since all the images sent to different users are of same type except for few pixel values, attacker will not be able to guess that the image carries any secret data. Even if the attacker comes to know that the share has been embedded in the image, he may not guess that the pixel values have been replaced with the share value based on the random number.

d) No single image reveals any information about the key, it is necessary that minimum m number of images are required to reconstruct the key. To make it more secured, we can choose different cover images for different users. Apart from these, the processing time has also been significantly reduced, since we are not performing the encryption and decryption operations. Otherwise one more key for encryption and decryption is required. Thus the confidentiality of the key has been enhanced and at the same time the processing time has been reduced.

6 Conclusions

With the experimental results we have shown that the confidentiality of the shared key can be provided without performing encryption and decryption. Also the embedding of data in an image is not based on replacement of LSB of each pixel which is a very common method, which can be easily guessed by an attacker. Also our method works for any number of users, and not all shares are required for reconstruction of the key. Thus our method is better than [1].

References

1. Kumar, A., Rajgopal, N.: Secret Image Sharing Using Pseudo Random Sequence. IJCSNS 6(2B) (February 2006)
2. Shamir, A.: How to share a secret. Comm. ACM 22, 612–613 (1979)
3. Tang, S.: Simple Secret Sharing and Threshold RSA Signature Schemes. Journal of Information and Computational Science 1, 259–262 (2004)
4. Cryptography and network security by William stallings, 3rd edn. Pearson education
5. Blakley, G.: Safeguarding Cryptographic Keys. In: AFIPS Conference Proceedings, vol. 48 (1979)
6. Tsai, C.S., Chang, C.C., Chen, T.S.: Sharing multiple secrets in digital images. J. Syst. Software 64, 163–170 (2002)
7. Gennaro, R., Jarecki, S., Krawczyk, H., Rabin, T.: Secure Distributed Key Generation for Discrete-Log Based Cryptosystems. In: Stern, J. (ed.) EUROCRYPT 1999. LNCS, vol. 1592, pp. 295–310. Springer, Heidelberg (1999)
8. Beimel, A., Chor, B.: Interaction in Key Distribution Schemes. In: Stinson, D.R. (ed.) CRYPTO 1993. LNCS, vol. 773, pp. 444–455. Springer, Heidelberg (1994)
9. Chor, B., Goldwasser, S., Micali, S., Awerbuch, B.: Verifiable Secret Sharing and Achieving Simultaneity in the Presence of Faults. In: Proc. 26th Annual Symposium on Foundations of Computer Science, Portland, OR, October 21-23, pp. 383–395 (1985)
10. Charnes, C., Pieprzyk, J., Safavi-Naini, R.: Families of Threshold Schemes. In: Proc. IEEE International Symposium on Information Theory, Trondheim, Norway, p. 499 (July 1994)
11. Thien, C.C., Lin, J.C.: Secret image sharing. Comput. Graphics 6(5), 765–770 (2002)
12. Blakley, G.R.: Safeguarding cryptographic keys. In: Proc. AFIPS 1979, National Computer Conference, vol. 48, pp. 313–137 (1979)
13. Carpentieri, M.: A Perfect Threshold Secret Sharing Scheme to Identify Cheaters. Designs, Codes and Cryptography 5(3), 183–187 (1995)
14. Tompa, M., Woll, H.: How to Share a Secret with cheaters. Journal of Cryptology 1(3), 133–138 (1989)
15. Asmuth, C., Bloom, J.: A modular approach to key safeguarding. IEEE Trans. Informat. Theory 29(2), 208–210 (1983)

Security Assurance by Efficient Non-repudiation Requirements

S.K. Pandey¹ and K. Mustafa²

¹ Department of Information Technology, Board of Studies
The Institute of Chartered Accountants of India (Set up by an Act of Parliament),
Noida- 201 309, India

santo.panday@yahoo.co.in
² Department of Computer Science

Jamia Millia Islamia (A Central University), New Delhi-110 025, India
kmfarooki@yahoo.com

Abstract. Security is an age long dream in all the walks of our social life. In software industry, security is regarded as wheels on which the entire system can move smoothly. Various tools/techniques have been deployed for developing secure software, but, on the other hand, attackers are continuously exploiting vulnerabilities to compromise security. Firewalls, intrusion prevention/detection and antivirus systems cannot simply solve this problem to the desirable extent. Only a rigorous effort by the software development community for building more secure software can foil attackers and allow users to feel protected from such exploitations. Research studies reveal that security cannot be added in developed software rather it should be introduced *right from the beginning* in the SDLC. To achieve this objective, security measures must be embedded throughout the SDLC phases and starting from the requirements phase itself. Non-Repudiation requirement is globally accepted as one of the prominent security requirements. Appropriate level of non-repudiation may well enforce security features and hence, ensure security for deployed software. A checklist is proposed, in this paper, which may enable assessment of the appropriateness of non-repudiation requirements and lead to counter/additional measures for security assurance.

Keywords: Software Security, Security Assurance, Non-Repudiation, Non-Repudiation Checklist.

1 Introduction

Software security suggests the idea of engineering the software to function correctly even under malicious attack(s). Most of the critical infrastructures, which all of us take for granted, are highly complex, interconnected and interdependent systems of sophisticated information processing, command, control, and communication. It is noteworthy that a single programming or design flaw in today's complex software system can undermine an entire system. In 1990, failure due to a single line of buggy code in AT & T's 4ESS switch caused systems to drop roughly 50% of long distance over a period of nine hours and \$60 million loss. Another incident of computer security reported to the CERT coordination center in recent years due to a single class of programming flaw of buffer overrun (Kouns

& Minoli, 2010). Software security is a foremost concern for modern information enterprise. Designing highly dependable security systems to ensure secure access to a distributed software and information has been recorded as the need of the hour. Software security is about designing such software that are to be secured; making sure that software is secure, and guiding software developers, architects and users about how to build secure software (Mustafa et al., 2008).

Requirements phase is the foundation of the entire software development life cycle. With proper requirements management, a project can deliver the right solution within the time and budget (Pandey et al., 2008). Requirement elicitation, requirement specification, and requirement validation are important activities to assure the quality of requirements. As the vulnerabilities of software increases, system needs an additional requirement for the security aspect, which protects the software from vulnerabilities and makes software more reliable. The requirements team's overall perspective of security goals, challenges, and plans need to be incorporated in the SRS that is produced during the requirement's phase.

Following are the major security requirements traceable in the literature and reported practices (Gilliam et al., 2011):

- Authentication,
- Access Controls and Rights,
- Confidentiality,
- Non-Repudiation,
- Data Classification Procedures,
- Business Continuity and Disaster Recovery,
- Virus Protection,
- Event Log and Audit Trails,
- Backup & Recovery, and
- Incident Management, Intrusion Detection and Forensic Analysis.

In our previous work, mechanisms for the assurance of first three requirements have been covered up to some extent. To extend this series one step further, in this paper, we address on Non-Repudiation requirements exclusively. A Non-Repudiation requirement specifies the extent to which a business, application, or component shall prevent a party to one of its interactions (e.g., message, transaction etc.) from denying having participated in all or part of the interaction. A checklist is hereby proposed for the verification of major facts related with Non-Repudiation in further sections.

Beyond this introduction on the background details, the remainder of this paper is organized as follows. Section 2 describes ‘Non-Repudiation’. The ‘Checklist Approach’ is discussed in Section 3, while a Checklist for Non-Repudiation is proposed in Section 4. ‘Implementation Mechanism’ is discussed in Section 5. ‘Tryout Results and Discussion’ is provided in Section 6 and ‘Conclusions and Future Works’ are given in Section 7.

2 Non-repudiation

Non-Repudiation denotes ‘Not denying or reneging’. Digital signatures and certificates offer non-repudiation as they guarantee the authenticity of a document or message

(Encyclopedia; and McCullagh & Caelli, 2000). The basic concept behind this requirement is to provide guidance for the usage of Digital Signatures for electronically signing any document that may need to uphold validity for the purpose of non-repudiation, as a manual signature or thumb impression on a physical document, in the court of law, as per the applicable Act, e.g. Information Technology Act (Amended), 2008 in India. All the legal documents, in electronic format should be signed using Digital Signatures. Any other electronic document that may require validity and non-repudiation in the court of law should be signed using Digital Signature. The Digital Certificate, which is used for digitally signing any such documents are treated as a valid certificate in any court of law, issued by a ‘Licensed Certifying Authority’.

3 The Checklist Approach

There is no doubt that to ensure better process, atomic checklists are generally used and have been found to be handy and quite fruitful (Pandey & Mustafa, 2010). Further, it becomes evident through the explanation of the researchers that a little work has been reported, hence, it is viable to have a checklist for Non-Repudiation process, which should be atomic in nature and can be easily usable for secure development process. Taking into account the need and significance of Non-Repudiation checklist for building secure software, an integrated and prescriptive checklist is hereby proposed. Items of the checklist have been derived from the reported and well-verified practices of the literature and industry, as evident from the item-wise references in most of the cases.

4 A Non-repudiation Checklist

The typical objective of a Non-Repudiation requirement is to ensure that adequate tamper-proof records are kept to prevent parties to interactions from denying that they have taken place. Non-Repudiation requirement minimizes any potential future legal and liability problems that might result from someone disputing one of their interactions. Here, a checklist for Non-Repudiation is developed based on the existing literature and industry best practices. Non-Repudiation requirements can be well implemented, which should have approved solution and may meet all or most of the following checklist items:

S. No.	Attribute	Check point Description	Status (Y/N/NA)
1.	<i>Storage of Information</i>	Is there any process that typically involves the storage of a significant amount of information about each interaction including the ‘authenticated identity of all parties involved in the transaction’, ‘date and time that the interaction was sent/received/acknowledged (if relevant), and ‘significant information that is passed during the interaction’ (National Thermal Power Corporation Ltd., 2006) ?	

2.	<i>Authentication by Digital Signature</i>	Is there any tool which can check that every electronic record authenticated by a digital signature, using asymmetric crypto system and hash function (Ponder, 1999)?	
3.	<i>Identity Management Capability</i>	Is there any capability in the system to associate the identity of the producer with the information (Locke & Gallagher, 2009)?	
4.	<i>Compliance with IT (Amended) Act, 2008</i>	Is the signing of electronic document with any government or its agency performed as per guidelines in the “Electronic governance” of the applicable Act e.g. IT (Amended) Act, 2008 in India (TIMEIS. IT technologies)?	
5.	<i>Information Accessibility and Usability</i>	Is information accessible and usable for a subsequent reference for any electronic document that needs to be retained for a later reference for the purpose of validity and non-repudiation (e.g. Electronic file, e-mails, etc.) (National Thermal Power Corporation Ltd., 2006)?	
6.	<i>Accuracy of Information</i>	Is the original format able to demonstrate the accuracy of the Information for any electronic document that need to be retained for a later reference for the purpose of validity and non-repudiation (e.g. Electronic file, e-mails, etc.) (Glaessner et al., 2002)?	
7.	<i>Metadata of Document</i>	Are details able to give the identification (of origin, destination, etc.), date, time and receipt of document for any electronic document that need to be retained for a later reference for the purpose of validity and non-repudiation (e.g. Electronic file, e-mails, etc.) (Rwanda Information Technology Authority, 2006)?	
8.	<i>Uniqueness of Digital Signature</i>	Is there uniqueness in digital signature with respect to the party affixing digital signature (National Thermal Power Corporation Ltd., 2006)?	
9.	<i>Capability of Digital Signature Integrity</i>	Is digital signature capable of identifying the user according to the level of assurance and accorded to the digital certificate (American Bar Association, 1996)?	
10.	<i>Private Key</i>	Is the private key accessed by any other person not authorized to affix the digital signature on any electronic documents (Controller of Certifying Authorities)? <i>/* If the Private Key corresponding to the public key has been compromised, it should be immediately brought to the notice of the Certifying Authority. */</i>	

11.	<p><i>Certifying Authority</i></p> <p>Are all the facts required to be represented for obtaining a Digital Certificate from the Enterprise Certificate Authority correctly represented to the issuing Certifying Authority (Hamel, 2004), (Cornell University, 2008)?</p> <p><i>/* This check must be performed since misrepresentation of facts for obtaining a Digital Certificate is punishable under the Information Technology Act, 2008. */</i></p>	
-----	---	--

Based on these checkpoints, their corresponding attribute have been identified, which are shown in Fig. 1.

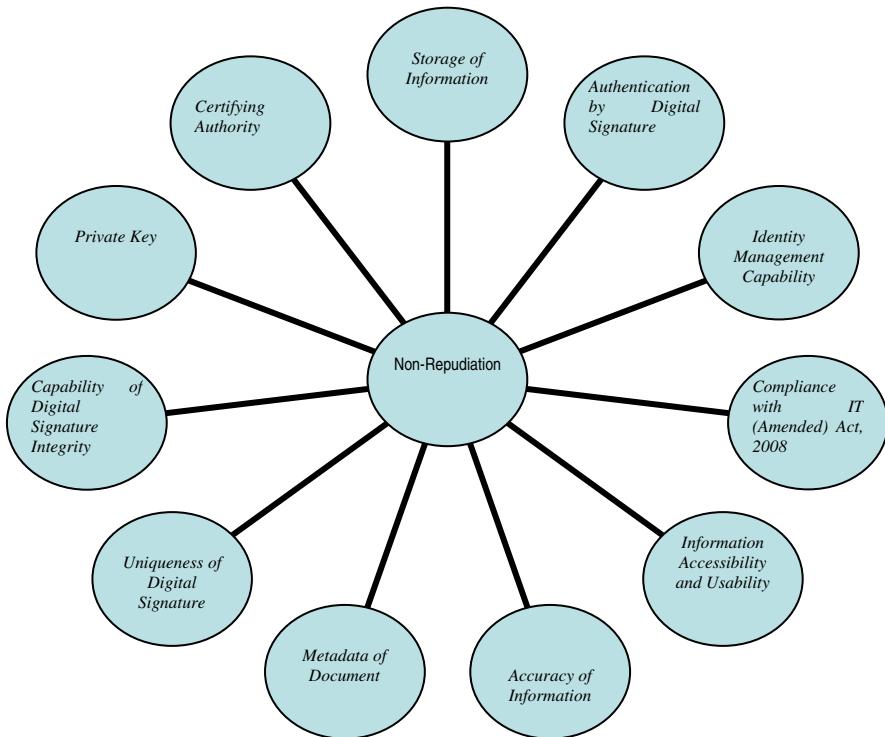


Fig. 1. Attributes of Non-repudiation

5 Implementation Mechanism

Following are the guidelines/steps for implementation of the proposed checklist:

- First step will be the structured walkthrough by checklist filtering of the SRS, in which various checkpoints are provided for verification of the Non-Repudiation requirement.

- If any checkpoint is not pertinent to the project, it may be identified as 'NA'. This will not be taken into consideration.
- For all the applicable checkpoints, requirement engineer/s may assess the compliance/ noncompliance of the checkpoints.
- Further, compute the overall compliance status of the checklist in % with the help of all the compliance/non-compliance checkpoints. This will provide the exact status of the incorporation of the Non-Repudiation requirement.
- Based on the project need and other relevant factors like cost, effort etc., further course of action may be decided.

6 Tryout Results and Discussion

Proposal of any process/methodology is subject to the experimental validation and analysis of the results. There must be some experimental data, which should show the utility of the proposal. Keeping in mind the above fact, proposed checklist was applied to a real life project obtained from a software development company (on the request of the company, identity is concealed), and the final result of checklist assessment is computed on the basis of the total compliant, non-compliant and 'N/A' checkpoints as per prescribed implementation mechanism given in the above section. The assessment results of the checklist are given as follows:

Table 1. Assessment Results of the Checklist

Total Check points	Not Applicable (NA) Check points	Total Available Check points	Non-Compliant Checkpoints	Compliant Checkpoints	Overall Compliance Status
11	0	11	7	4	36.36 %

For the comparison of results, we demanded the results from the SRS provider. But as we know that in industry, this is highly informal; they were unable to provide such type of details. They could only provide a general opinion as saying that 'we rate the SRS highly insecure with reference to the Non-Repudiation requirements'. Their informal revelation about the final result confirms our formal results. From these evidences, the utility of the checklist is automatically ascertained up to some extent. However, it may not be enough/sufficient to conclude so strongly about the effectiveness of the proposal but certainly, up to some extent.

7 Conclusion and Future Work

A checklist is proposed for the implementation of the Non-Repudiation requirements. The system will be stronger if it satisfies all or most of the checklist items given in the proposed checklist. A complete process of Non-Repudiation is described for the

security assurance of the SRS. Being prescriptive in nature, the checklist is highly implementable and it is a concrete step for the implementing security *right from the inception itself.*

Future work may include the standardization of the results by strong validation of the proposed checklist on a large sample size. In addition, the weights of each attribute given in the checklist may also be decided to provide more reliable results. In future, we are also trying to develop some more checklists for the implementation of the other security requirements, based on the same pattern. This will help software developers and security experts for building secure software.

References

- American Bar Association. Digital signature guidelines: Legal infrastructure for certification authorities and secure electronic commerce, USA (August 1, 1996),
<http://www.abanet.org/scitech/ec/isc/dsg.pdf> (retrieved June 2, 2008)
- Controller of Certifying Authorities. Security procedure for electronic records and digital signature, http://www.cca.gov.in/faq_it.jsp (retrieved July 12, 2009)
- Cornell University. Baseline IT security requirements, version 1.2. (October 17, 2008),
<http://www.cit.cornell.edu/security/depth/requirements/>
(retrieved June 12, 2008)
- Encyclopedia. Non-repudiation. PC Magazine Encyclopedia,
http://www.pc当地.com/encyclopedia_term/0,2542,t=nonrepudiation&i=48067,00.asp (retrieved March 2, 2010)
- Gilliam David, P., Kelly John, C., Powell John, D., Matt, B.: Development of a Software Security Assessment Instrument to Reduce Software Security Risk. In: The Proceedings of the WETICE, pp. 144–149 (2011)
- Thomas, G., Tom, K., Valerie, M.: Electronic security: Risk mitigation in financial transactions public policy issues. The World Bank (2002),
http://info.worldbank.org/etools/docs/library/83592/eseecurity_risk_mitigation.pdf (retrieved June 12, 2008)
- Linda, H.: MUETA: What every public sector lawyer should know, MUETA: What Every Public Sector Lawyer Should Know (December 2004),
http://www.mass.gov/Eoaf/docs/itd/guidance/legal/mueta_for_public_sector_lawyers.ppt#256 (retrieved June 12, 2008)
- Jake, K., Daniel, M.: Information technology risk assessment in enterprise environments. John Wiley & Sons (2010)
- Locke, G., Gallagher, P.D.: Recommended security controls for federal information systems and organizations, NIST Special Publication 800-53 (August 2009)
- Adrian, M., William, C.: Non repudiation in the digital environment. First Monday 5(8) (August 7, 2000), http://www.firstmonday.org/issues/issue5_8/mccullagh/
(retrieved May 3, 2008)
- Mustafa, K., Pandey, S.K., Rehman, S.: Security assurance by efficient access control and rights. CSI Communication 32(6), 29–33 (2008)
- National Thermal Power Corporation Ltd, Information security policies & procedures. [Technical report] Final V. 1.0 (July 2006)
- Pandey, S.K., Rehman, S., Mustafa, K., Ahson, S.I.: Security assurance: The requirements way (January 21 2008), http://www.stickyminds.com/s.asp?F=S13426_ART_2
(retrieved June 30, 2009)

- Pandey, S.K., Mustafa, K.: Security Assurance: An Authentication Initiative by Checklist. International Journal of Advanced Research in Computer Science 1(2), 110–113 (2010)
- Ponder, P.J.: Professionals' electronic data delivery system (PEDDS). Tallahassee, Florida: CADD Systems Office, Department of Transportation Engineering (June 1999)
- Rwanda Information Technology Authority. Technical standards and guidelines for e-government [Final report]. Kampala, Uganda (February 2006),
<http://www.rita.gov.rw/docs/Egovernance%20Standards%20-%20Final%20Report%20presented.pdf> (retrieved June 12, 2008)
- TIMEIS. IT technologies, <http://www.techno-preneur.net/cgovt/it-act.htm> (retrieved June 12, 2008)

Poor Quality Watermark Barcodes Image Enhancement

Mohammed A. Atiea, Yousef B. Mahdy, and Abdel-Rahman Hedar

Computer Science Department,
Faculty of Computers and Information
Assuit University, Egypt

m_ali_atiea@hotmail.com, {mahdy, hedar}@aun.edu.eg

Abstract. The one dimensional (1D) barcode was developed as a package label that could be swiftly and accurately read by a laser scanner. It has become ubiquitous, with symbologies such as UPC used to label approximately 99% of all packaged goods in the US [1]. The two-dimensional (2D) barcode has improved the information encoded capacity, and it also has enriched the applications of barcode technique. Recently, there are researches dealing with watermark technique on barcode to prevent it from counterfeited or prepensely tampered. The existent methods still have to limit the size of embedded watermark in a relatively small portion. Furthermore, it also needs to utilize original watermark or other auxiliary verification mechanism to achieve the barcode verification. In this paper, we propose a novel watermarking barcode reading enhancement method. The proposed method can fight most of reading challenges of watermarking barcode. Experiments with challenging barcode images show substantial improvement over other state-of-the-art algorithms.

Keywords: Barcode, digital watermark, barcode verification.

1 Introduction

Digital watermarking is a concept that emerged in the digital signal processing community in early 1990. Although defined for digital data, it is closely related to article watermarking, a mechanism invented about 700 years ago in Fabriano, Italy [2]. The problem our ancestors tried to solve was how to label an article in an invisible manner. They came up with the idea to slightly thin the article at some locations. Holding the treated article up against a strong light source and looking at it, one could then perceive an image produced by the thinner parts of the article. The method to threat the article this way was called watermarking and the inserted image was called a watermark because the perceived images looks like watery areas on the article. Nowadays we try to solve very similar problems with the sole difference that they are in the digital world. The concept of Watermarking was therefore adapted to the digital world and the new concept was coined digital watermarking. The word watermarking was chosen for digital data because the inserted watermark cannot be seen by simply looking at the digitally watermarked data. However, when given to a computer the watermark can be detected. Digital watermarking is defined as the imperceptible insertion of information into multimedia data. It means the digital data is modified in an imperceptible way to insert the watermark. Depending on the application, the watermark itself may be a string of characters, a number, an image, a piece of sound, or just a 1-bit

piece of information to indicate if the data has been watermarked. The data into which the watermark is inserted was given various names, such as host data, source data, and cover data. The last name, cover data, has been borrowed from steganography, the art of information hiding. In addition to the cover data and the watermark, some schemes use a digital key. The key is used to embed the watermark; the same key is required to extract or detect the embedded watermark. The watermarked data also has different names, such as stego data, signed data, and of course watermarked data. In the watermark detection process, the embedded watermark is detected, or extracted. It is clear that if the embedded watermark is a string of characters, then we would like to extract the information during the watermark detection process. However, if the embedded information was only a single bit, then it is enough if the detector just gives a yes or no answer to indicate if the data has been watermarked. In addition to extracting or detecting embedded information, many schemes provide a confidentiality measure to indicate how reliable the extracted information is [2].

Recently, there are researches dealing with watermark technique for barcode data authentication [3, 4], but the existent methods have to limit the size of embedded watermark. Furthermore, they also need to utilize original watermark or other auxiliary verification mechanism to achieve the barcode verification. The appearance of PDF417 2D barcode significantly increases the barcode information hiding capacity [3]. Much important information can be stored into barcode without any extra storage media. The most obviously application is used on identification card. With PDF417 2D barcode, the personal data even photo on ID card can be encoded into a 2D barcode which is then printed on the back of ID card. Then, the identity authentication can be verified automatically by simply scanning the barcode. With the variety of applications, the request of 2D barcode information hiding capacity becomes critical [3]. Unfortunately, 2D barcode has to be limited in its printed area on ID card. Moreover, ID card usually applies to a lot of key application such as financial transaction, medical transaction and border crossing [3, 5]. So, it is often concerned the problems about counterfeited and preensibly tampered to cause danger of ID card owner. For this severity problem, some experts utilize high information hiding capacity of barcode to design a watermarking technique to make contributions for these problems [3]. Unfortunately, the poor quality of the barcode images extracted from attacked watermarked media makes it surprisingly difficult to correctly decode barcodes and these will be the Challenge we meet using barcode to watermarking any digital media.

This paper proposes a new algorithm for watermark barcode image enhancement that produces excellent results even for poor quality watermark images that are extracted from attacked watermarked media.

2 Barcode

A barcode is an optical machine-readable representation of data, which shows certain data on certain products. Originally, barcodes represented data in the widths (lines) and the spacings of parallel lines, and may be referred to as linear or 1D barcodes or symbologies. They also come in patterns of squares, dots, hexagons and other geometric patterns within images termed 2D matrix codes or symbologies. Although 2D systems use symbols other than bars, they are generally referred to as barcodes as well. Barcodes can be read by optical scanners called barcode readers, or scanned from an

image by special software. Virtually every packaged good is labeled with at least one form of barcode, generally a flavor of either the EAN or the UPC standards. The success of barcode technology for identification, tracking, and inventory derives from its ability to encode information in a compact fashion with very low associated cost [6]. The first use of barcodes was to label railroad cars, but they were not commercially successful until they were used to automate supermarket checkout systems, a task in which they have become almost universal. Their use has spread to many other roles as well, tasks that are generically referred to as Auto ID Data Capture (AIDC). Other systems are attempting to make inroads in the AIDC market, but the simplicity, universality and low cost of barcodes has limited the role of these other systems.

Types of barcodes are Linear (1D) barcodes and Matrix (2D) barcodes. There are many linear barcodes such as UPC, Code 25, Code 93 and Code 128. A matrix code, also known as a 2D barcode or simply a 2D code, is a two-dimensional way of representing information. It is similar to a linear barcode, but has more data representation capability. There are many Matrix barcodes such as 3-DI, Datamatrix, Code 16K, UltraCode and WaterCode. Some examples of Barcodes Generation are shown in Fig. 1.



Fig. 1. Examples of Barcode

3 The Proposed Technique

Given an image containing a barcode, two distinct operations are needed for accessing the information contained in the barcode: localization and reading. Localization typically relies on the strong textural content of the barcode. Reading can be performed on one or more scan lines extracted by the localization step.

The proposed work focuses on barcode reading, because our image all the time contain code 128 barcode only, we implemented a simple and fast localization algorithm that finds the highest energy region in the map generated by subtracting the vertical from the horizontal gradient. This algorithm is by no means optimal but it works reasonably well in our studies, as it only serves as a necessary pre-processing stage to enable the experiments of Sec. 4.

There will be some Challenges, when we working with barcode images extracted from attacked watermarked media as shown in Fig.2. There will be noises and some of line bares get damaged.

The proposed algorithm analyzes a single scanline contained in the detected barcode area. The only requirement is that the beginning and the end of the barcode pattern in the scanline are detected with certain accuracy. For example, in our implementation we assume a localization tolerance in either end point equal to 3 times the width of the narrowest bar. Note that this task is much simpler than localizing all bars



Fig. 2. Barcode Images Extracted from Attacked Watermarked Media

in the code, an operation that we avoid altogether. The Code 128 standard, in fact, requires that the initial and final bars be separated from the edge of the barcode by a quiet area of a width equal to at least 10 times the width of the narrowest bar, usually referred to as Quiet Zone.

3.1 Algorithm 1: Reading Algorithm

The steps in the reading algorithm are as follows:

1. Apply color filtering for all R=0, G=0, and B=0 (Black Color) with the use of fill color R=255,G=255 and B=255 (White Color) and the fill type outside (Background).
2. Extract Y Channel (YCrCb color space) from the image.
3. Apply median filter.
4. Invert Colors and save a copy of this image as image A.
5. Scan a single pixel line from top to down searching for white pixel if found go to Step-6 else if no white pixel in line get new line and repeat Step-5 until image width finish then go to Step-7.
6. Check all the next pixels of the found white pixel in the same line and based on the number of other white pixels keep the pixel white or Convert it to black then go to Step-5.
7. Find the average height of the line bars and convert all bars to that height then save it as image B.
8. Use edge detection to find Left and Right edges of each line bare.
9. Repeat Step-8 for image A.
10. For each line bare compare edge location and its width for both images A and B.
11. From the results of Step-10 decide if founded edges in image B are true edges come from image A of false edges come from Step-5 and Step-6 then convert any false edges to black color.
12. Invert Colors and save the result.

4 Experimental Results

Extensive simulations are carried out to prove that the proposed technique is effective under different conditions. We implemented and tested our algorithm using C# language. A simple localization algorithm was used to provide a scanline segment input to

our reader. In order to assess the performance of the system, we tested it on a variety of extracted watermarking images. We apply different types of attacks on watermarked media such as lossy compression, Rescaling and Gaussian noise. Figure. 3 gives the original barcode image before embedded in media, Fig. 4, Fig. 5 and Fig. 6 are gives some examples of extracted watermarking images after attacks and Fig. 7 gives the output image of our algorithm.

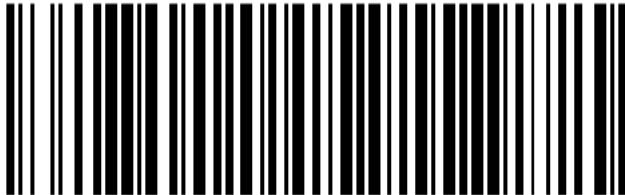


Fig. 3. The Original Barcode Image



Fig. 4. Law Noise Barcode Image

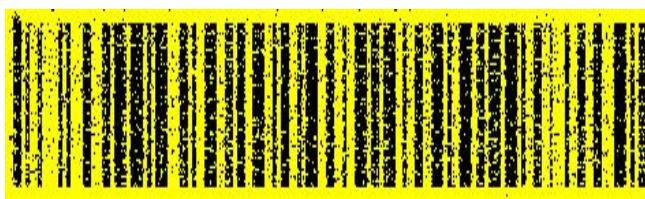


Fig. 5. Heavy Noise Barcode Image



Fig. 6. Some of Barcodes are Broken

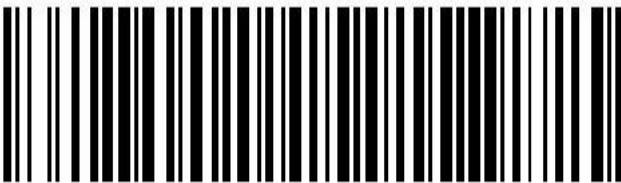


Fig. 7. The Result is The Same Readable Image for The Three Input Images with the Information AUN12345678 Encoded

5 Conclusion

We have presented a new algorithm for poor quality watermark barcodes image enhancement. The proposed algorithm can fight most of reading challenges of watermarking barcode and able to give a clear barcode image, that can be read with any other online barcode reading program such as [7, 8]. Experiments with challenging barcode images show substantial improvement over other state-of-the-art algorithms.

References

1. Tekin, E., Coughlan, J.: A Bayesian Algorithm for Reading 1D Barcodes. In: Sixth Canadian Conference on Computer and Robot Vision (CRV), Kelowna, British Columbia (2009)
2. Hernandez Martin, J.R., Kutter, M.: Information retrieval in digital watermarking. IEEE Communications Magazine 39(8) (2001)
3. Shen, J.-J., Hsu, P.-W.: A Fragile Associative Watermarking on 2D Barcode for Data Authentication. International Journal of Network Security 7(3) (2008)
4. Afzel, N., Nikhil, T., Max, M.H.: Embedding biometric identifiers in 2D barcodes for improved security. ScienceDirect, Computers & Security 23(8) (2004)
5. Ross, D.: Back on the cards. IEE Review 49 (2003)
6. Gallo, O., Manduchi, R.: Date of Current Version 2010, Reading challenging barcodes with cameras. In: Applications of Computer Vision (WACV) (2010)
7. DataSymbol Barcode Recognition SDK, <http://www.datasymbol.com/> (last retrieved December 07, 2011)
8. DTK Software, <http://www.dtksoft.com/> (last retrieved December 07, 2011)

Hiding Data in FLV Video File

Mohammed A. Atiea, Yousef B. Mahdy, and Abdel-Rahman Hedar

Computer Science Department,
Faculty of Computers and Information
Assuit University, Egypt
m_ali_atiea@hotmail.com, {mahdy, hedar}@aun.edu.eg

Abstract. Video Frame quality and statistical undetectability are two key issues related to steganography techniques. In this paper, we propose a novel flash video file (.flv file extension) information-embedding scheme in which the embedded information is reconstructed without knowing the original host flash video file. The proposed method presents high rate of information embedding and is robust to lossless and lossy compression. The characteristic of the proposed scheme is to use a weak point in the header information of flash video file to assist compression process. Experimental results have indicated that the method is robust against lossless and lossy compression.

Keywords: Steganography, FLV, lossless and lossy compression.

1 Introduction

Steganography is the art of hiding the existence of a message, the word Steganography comes from the Greek words steganos (secret) and graphy (writing) [1]. An example could be a letter written with two different inks, when the letter is submerged in water, one of the inks dissolves while the other remains on the letter, thus revealing the secret message. The original message on the letter is just a cover to hide the existence of the secret message, so we can say steganography is the art of hiding communication, referring to the process of embedding a message or any kind of information that is wished to be hidden in a medium (stego-medium) usually a picture, an audio file, or a video file, in such a way that no one apart from the sender and intended recipient even realizes there is a hidden message. After the embedding process, extraction must be possible. The basic steganography process of embedding [2] is shown in Fig. 1. The advantage of steganography over cryptography is that messages do not attract attention to attackers and even receivers. Steganography and cryptography are often used together to ensure security of the secret messages [3]. For example, many previous steganography approaches [3-5] use the secret key (i.e., idea borrowed from cryptography) to produce better protection of the information if the stego-object arouses suspicion. Therefore, these approaches can become more secure and can be potentially useful to some security-demanding applications such as military intelligence.

Recently, more and more people communicate with each other by surfing on the internet. However, it is not very secure when we transmit information through the internet. Everyone can peek, copy even alter our information easily in this wide-open environment. Thus, people don't want to transmit the important information without

any protection in the public network unless a secure channel is provided for the transmission. The cryptography technique can protect the message content from a peeper. But the cryptography technique will cause the message content to be meaningless random codes. It is easy to guess something important in the transmitted information even the receiver do not know what is inside. They may cut, hack or break these meaningless random codes. Therefore, information-hiding technique is needed to help in solving the problem of transmitting important data in an absolute secure channel. Thus, it is not only difficult to decrypt the data, but also difficult for attackers to detect the hidden data.

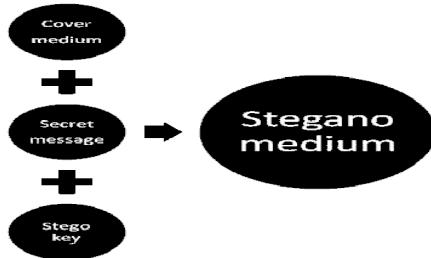


Fig. 1. The Basic Steganography Process

People usually hide data or information into one medium. It can be a text article, image, music, or video. The medium that hides data is named stego-medium. It can keep the confidential data secretly, and call the original medium that does not hide data the cover medium. It is difficult for the unauthorized people to detect hidden data from a stego-medium. They can use it to hide important data. Then, when they transmit the medium, the peepers will not find the secret data in it. They will think that the transmitted medium is an unimportant data or a data that they don't want to collect or peep [6].

One of the widely used Steganography methods is LSB (least significant bit) [6] method. This method changes the last bit of each pixel bit to hide the data, because the last bit of pixel changes little pixel's color, the image quality would not be affected much. In fact, human vision will not be aware of the difference of image.

In frequency domain [7], people will employ the feature of frequency parameters to hide data. The methods in this domain can fight against more attacks and raise the robustness. In frequency domain, good image quality can be retained. But it will lose some embedded data, after performing the lossy compression process to the stego-medium. Thus, users cannot hide text information in it. Besides, the process of compression and hiding has a higher complexity than spatial domain. The information hiding technique common methods are based on discrete cosine transformation and discrete wavelet transformation [8].

Nowadays, the newest form of steganography has become the target of researches to find new ways to embed hidden messages of larger sizes. Steganography on video files answer these needs for larger spaces in hiding or embedding data. Videos are generally just collections of images and sound files making some of the effective methods of steganography on images and audio files possible on hiding data in video files. Larger space for embedding and having small unnoticeable distortions make video steganography a reliable method in hiding data.

The rest of this paper is organized as follows. In section 2, flash video file format is described. The proposed method presented in detail in section 3. In section 4, experimental results are described and conclusions are given in section 5.

2 Flash Video File Format

Flash video is a container file format used to deliver video over the internet using adobe flash player versions 6–10. Flash video content may also be embedded within SWF (Small Web Format) files. There are two different video file formats known as flash video: FLV and F4V. The audio and video data within FLV files are encoded in the same way as they are within SWF files. The latter F4V file format is based on the ISO (International Organization for Standardization) base media file format and is supported starting with flash player 9 update 3. Both formats are supported in adobe flash player and currently developed by adobe systems [9]. FLV was originally developed by Macromedia. The format has quickly established itself as the format of choice for embedded video on the web. Notable users of the flash video format include YouTube, Google Video, Yahoo Video, Metacafe and Reuters.com, and many other news providers.

This research started with the idea of using a video file format which is prevalent, in a way that it is primarily the one everyone has seen, used, and even edited a lot. In this way, the application of our research would be more meaningful and innovative. After discussions and comparisons of different video formats, there is much promise and hope for concentrating particularly on flash video files or FLV files. A part of a FLV file (Header) opened in a Hexadecimal Editor is shown in Fig 2. FLV files are widely used that they can be easily included in websites. In addition, FLV files usually have smaller file sizes compared to all the other formats.

0000:	46 4C 55 01	01 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	D1	FLV	.	.	.
0010:	40 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	D2	A6	@	.	.
0020:	SE 19 0F 03 00	E0 DC 03 02 00 60 0E 3E 0C 00 00 88	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	B_>>	.	.	.
0030:	30 02 A2 57 81	89 73 06 01 BF D3 D4 49 FD F0 0	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	W S I	.	.	.
0040:	30 01 A0 C0 09	2A 56 23 17 83 00 E0 24 FA 5A 3D 0	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	*V# S Z=	.	.	.
0050:	FC 97 E3 E0 60	03 47 DA 01 E1 0E 4D C1 20 48 B4	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	G M H	.	.	.
0060:	20 80 52 F0 30	0E 00 C0 38 84 10 60 1C 82 09 78	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	R O 8 : x	.	.	.
0070:	30 02 C0 C0 08	17 81 E5 63 EB E8 0C 03 60 30 0D 0	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	C 0	.	.	.
0080:	B2 45 2E 06 00	47 DF F4 52 02 00 42 9E 9E 54 3F E	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	G V B T?	.	.	.
0090:	06 01 B0 4B 2E	FC FA A5 5F 9F 2E 04 01 F7 FC 10	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	K V	.	.	.
00A0:	42 1F BE 7E 24	09 00 D0 3C AE 04 20 0C 08 00 C0	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	B_>S <	.	.	.
00B0:	08 04 2F 04 00	0E 2E 03 C5 DE BF FD B6 D5 E3	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	/ . e	.	.	.
00C0:	F1 FA BA A4 BC	4B 12 47 EA EF CB CD B1 24 7F FD	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	K G s	.	.	.
00D0:	56 3F 12 CB FF	F2 F5 7B 7F FA FD 31 28 4A 54	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	V? { 1(JT	.	.	.
00E0:	24 7B CA C4 BF	0F 4B F5 27 CB 84 8B B2 4F 5E F8	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	{ K O	.	.	.
00F0:	BC 49 CC B3 DF	9D 2E 1F 4B 3E AE 02 8F F7 C3 E1	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	I K >	.	.	.

Fig. 2. Part of an FLV File

The file format of FLV is very simple. It starts with the headers then metadata tag (data that describe the FLV), then interleaved audio and video tags (actual data). Each tag type in an FLV file constitutes a single stream. There can be no more than one audio and one video stream, synchronized together, in an FLV file [9]. You cannot define multiple independent streams of a single type. Unlike SWF files, FLV files store multiple byte integers in big-endian byte order [9]. For example, as a UI16 in SWF file format, the byte sequence that represents the number 300 (0x12C) is 0x2C 0x01; as a UI16 in FLV file format, the byte sequence that represents the number 300 is 0x01 0x2C. Also, FLV files use a 3-byte integer type, UI24 that is not used in SWF files.

3 The Proposed Technique

There are some characteristics of an FLV [2], based on the results after several experiments that have been done on the FLV, such as deleting certain areas in the header, metadata, video and audio tags, a video steganography C# program has been developed and it incorporates the following characteristics of an FLV [2]:

1. It is possible to delete tags in whole without corrupting the FLV.
2. It is possible to delete audio and video tags scattered throughout the entire FLV without corruption.
3. It is possible to add data successfully at the end of tags by changing the appropriate Body Length and Previous Tag Size.
4. It is possible to add an extremely large amount of data at the end of a video tag as compared to an audio tag.
5. It is possible to add data at the end of the metadata as long as we incorporate that addition to the Body Length.

All these characteristics already give the idea on how the Video Steganography program works.

The new research find there is also weak point in the header of FLV file, which gives us the ability to hide data there. In Table 1, the FLV header, all FLV files begin with the same header structure, the DataOffset field usually has a value of 9 for FLV version 1. This field is present to accommodate larger headers in future versions [9].

The DataOffset field is the weak point, as shown in Table 1 the size of DataOffset is 32 bits which can give us 4 GB as displacement, so we can use this space to hide any new file in it. This may be possible with the nature of the FLV file itself.

The FLV file appears to allocate more memory or more space for the video – with video requiring more bytes to code compared to audio. Thus it is possible to add huge amounts of data at the header of video file without changing the quality of the FLV.

There is no need to worry about the quality of the FLV since technically no actual video and audio tag data were changed or even removed. Some data were simply added and the only thing to be done is to incorporate the additional data in the Header and DataOffset in which the data were hidden.

Table 1. The FLV Header [9]

FLV header		
Field	Type	Comment
Signature	UI8	Signature byte always 'F' (0x46)
Signature	UI8	Signature byte always 'L' (0x4C)
Signature	UI8	Signature byte always 'V' (0x56)
Version	UI8	File version (for example, 0x01 for FLV version 1)
TypeFlagsReserved	UB[5]	Must be 0
TypeFlagsAudio	UB[1]	Audio tags are present
TypeFlagsReserved	UB[1]	Must be 0
TypeFlagsVideo	UB[1]	Video tags are present
DataOffset	UI32	Offset in bytes from start of file to start of body (that is, size of header)

Figure.3 and Fig.4 are gives an example of information hiding in FLV Header. As shown in Fig.3 the original FLV file header with DataOffset equal 9 and after we hide the word “Hello”, Fig.4 gives us the stego file header with DataOffset equal E (14 in Decimal) and the word “Hello” in the new space of the FLV file; of course we will not hide data without encryption.

```
00000000 46 4C 56 01 05 00 00 00 09 00 00 00 00 12 00 03 FLV.....
00000010 15 00 00 00 00 00 00 00 02 00 0A 6F 6E 4D 65 74 .....onMet
00000020 61 44 61 74 61 08 00 00 00 1C 00 0C 68 61 73 4B aData.....hasK
00000030 65 79 66 72 61 6D 65 73 01 01 00 09 63 75 65 50 eyframes....cueP
00000040 6F 69 6E 74 73 0A 00 00 00 00 00 0D 61 75 64 69 oints.....audi
00000050 6F 64 61 74 61 72 61 74 65 00 40 48 76 D7 B3 6E odatatype.@Hv..
00000060 35 3A 00 08 68 61 73 56 69 64 65 6F 01 01 00 06 5...hasVideo....
```

Fig. 3. Original File Header

```
00000000 46 4C 56 01 05 00 00 00 0E 48 65 6C 6C 6E 00 00 FLV....[Hello]
00000010 00 00 12 00 03 15 00 00 00 00 00 00 00 02 00 0A .....
00000020 6F 6E 4D 65 74 61 44 61 74 61 08 00 00 00 1C 00 onMetaData....
00000030 0C 68 61 73 4B 65 79 66 72 61 6D 65 73 01 01 00 .hasKeyframes...
00000040 09 63 75 65 50 6F 69 6E 74 73 0A 00 00 00 00 00 .cuePoints.....
00000050 0D 61 75 64 69 6F 64 61 74 61 72 61 74 65 00 40 .audiodatarate.@
```

Fig. 4. Stego File Header

However, an obvious difference of the modified FLV with respect to the original is that its file size has increased by the value of the hidden file. If it does not matter for users that the file size of their FLVs increase, nothing else needs to be done. This is especially the case for FLVs that are not so popular so people do not really remember their original size. If someone needs to be sure that others would not get suspicious at the changed file size, it would be better if large FLVs are embedded on since the small additional data would be insignificant compared to the large FLV. Adding an image of about 50 KB to a 5 MB FLV would result in a 5.05 MB modified FLV which is not that much different.

However, there is an additional feature to the steganography program to make it a more efficient program. Banking on the idea that it is possible to delete whole tags without corrupting the FLV and also in such a way that the deleted whole tags can be scattered throughout the FLV, it is also possible to compress any FLV [2]. The number of tags that is needed to be deleted would be dependent on how much data is to be embedded on that same FLV. This would ensure that the original file size is maintained. Of course, there are negative results and one of them is its effect on the FLV's quality. But since any whole tag can be deleted, the user is given the flexibility to choose which tags he wants to remove to be able to see afterwards how the modified FLV has changed. If the distortions are too great and noticeable, he can simply choose

other tags to omit and check later on if the new modified FLV is much better. This feature gives the user the freedom to choose between video-quality and size discrepancy. Whether the video itself will still be comprehensible or if will only be used for file storage is completely up to the user.

3.1 Algorithm 1: Embedding Algorithm

Algorithm 1. Pseudo-code description of embedding algorithm

- 1 Read the total size of the secret data.
- 2 Read DataOffset from the cover FLV file.
- 3 Increment DataOffset by the total size of the secret data.
- 4 Write the header of cover FLV file to Stego FLV.
- 5 Write all secret data to the header of Stego file.
- 6 Read the body of cover file and write it to Stego file.

3.2 Algorithm 2: Extraction Algorithm

Algorithm 2. Pseudo-code description of extraction algorithm

- 1 Read DataOffset from the Stego FLV file.
- 2 Decrement DataOffset by nine then you have secret data size.
- 3 Read secret data From the Header of Stego file.

4 Experimental Results

Extensive simulations are carried out to prove that the proposed technique is effective under different conditions. The tests on the FLV files were done in three parts: playback, upload, and retrieval. Two particular files were used: beach.flv and Car.flv. From these two files, a control group is used consisting of the original versions of both files. And from the original files, four more groups of embedded FLV files were made: high-compression group, medium-compression group, low-compression group, and uncompressed group. A major difference in these files is their size. The uncompressed file is the original FLV plus the data but without any compression performed, thus this file has the largest out of the five. The three versions that are compressed indicate three levels of compression with the high-compressed being the highest and its size being closest to the original FLV file. Altogether, there are a total of 10 FLV files used for the testing.

Playback testing involved playing the FLV files themselves and making observations. And in order to perform this test, research was made on at least 10 different FLV players. Playing the files in different FLV players ensures that the files would work on multiple platforms since FLV players are coded differently. Doing the test itself, all the FLV files worked on the different FLV players.

Upload testing involved uploading the FLV files onto video-hosting sites and testing the files for streaming. This was done to ensure that the FLV files would be uploaded and streamed properly across different platforms, as video-hosting sites are also coded differently. Observations are then made if changes will happen once the files were uploaded to the internet. After creating user accounts for the different sites, upload was successful for all the FLV files. After upload, streaming was observed and it was successful for all the FLV files as well. Observations made on the picture and sound quality were consistent with the results obtained from the playback test.

Retrieval testing involved retrieval of the uploaded FLV files from the video-hosting sites. This was done with the use of a downloader, which is an external site or program that is unrelated to the video-hosting site used. However, a limitation was seen in that the downloader sites mostly supported one or two video hosting sites. However, for the available sites that was able to download from, the FLV versions were all intact and were played normally with the FLV players. Performing the last test of extracting the embedded data was successful as the program was able to recognize the file and was able to extract the information within.

5 Conclusion

The proposed technique is successfully able to extract embedded information from stego-video without using the original video. Besides, the embedded data is robust against lossless and lossy compression without any perceptual distortion. The proposed technique enables high rate of information embedding more than 1 Giga bits that is several times more than the proposed technique in [1], and the information extracted with high efficiency, even after any numbers of frames dropped, more than what produced by the proposed technique in [2].

References

1. Zhang, X.: Efficient Data Hiding With Plus-Minus One or Two. *IEEE Signal Processing Letters* 17(7) (July 2010)
2. Mozo, A.J., Obien, M.E., Rigor, C.J., Rayel, D.F., Chua, K., Tangonan, G.: Video Steganography using Flash Video (FLV). In: I2MTC 2009 IEEE - International Instrumentation and Measurement Technology Conference (May 2009)
3. Chao, M.-W., Lin, C.-H., Yu, C.-W., Lee, T.-Y.: A High Capacity 3D Steganography Algorithm. *IEEE Transactions on Visualization and Computer Graphics* 15(2) (March/April 2009)
4. Cheng, Y.-M., Wang, C.-M.: A High-Capacity Steganographic Approach for 3D Polygonal Meshes. *The Visual Computer* 22(9) (2006)
5. Cayre, F., Macq, B.: Data Hiding on 3-D Triangle Meshes. *IEEE Trans. Signal Processing* 51(4) (2003)
6. Alia, M.A., Yahya, A.A.: Public-Key Steganography Based on Matching Method. *European Journal of Scientific Research* 40(2) (2010) ISSN 1450-216X
7. Chen, T.S., Chang, C.C., Hwang, M.S.: A Virtual Image Cryptosystem Based upon Vector Quantization. *IEEE Transactions on Image Processing* 7(10) (1998)
8. Langelaar, G.C., Lagendijk, R.L.: Optimal Differential Energy Watermarking of DCT Encoded Images and Video. *IEEE Transactions on Image Processing* 10(2) (2001)
9. Adobe Systems Incorporated. Video File Format Specification, Version 10. Adobe Systems Incorporated, http://download.macromedia.com/f4v/video_file_format_spec_v10_1.pdf (last retrieved 2011-12-20)

Taxonomy of Network Layer Attacks in Wireless Mesh Network

K. Ganesh Reddy and P. Santhi Thilagam

Department of Computer Science and Engineering, NITK Surathkal, India
`{guncity11, santhisocrates}@gmail.com`

Abstract. Wireless Mesh Networks (WMNs) have emerging application because of its ad-hoc features, high internet bandwidth capability, and interoperable with various networks. However, all features of WMNs vulnerable due to their inadequate security services, and most of the existing techniques protect WMNs only from single adversary node, but these techniques are failed to protect against multiple colluding attacks, and also have same reputation value for all types of attacks. To overcome these problems for future solutions, we have done clear analytical survey on network layer attacks. Eventually, we have come up with taxonomy of network layer attack.

Keywords: colluding attacks, intrusion detection, wireless mesh, network layer.

1 Introduction

Wireless mesh networks (WMNs) have been emerged as a key technology for providing fast and hassle free services to users and inspiring numerous applications. In recent years, wireless mesh networks have been becoming more popular because of its ubiquitous broadband wireless internet connectivity in a sizable geographic area and cost effective network deployment. WMNs also support features such as dynamic self-organization, self-configuration and self-healing.

Fig. 1 depicts wireless mesh network architecture. Here all wireless radio nodes are connected in mesh to form infrastructure mesh and client mesh in which nodes are ordered hierarchy: gateway, router, and mesh client. WMNs can also interoperate with other wireless networks such as high-speed metropolitan area mobile networks, back-haul connectivity for cellular radio access networks, intelligent transport system, network defense system and citywide surveillance systems.

The study shows that wireless mesh networks are more vulnerable especially in Network layer followed by MAC layer and Physical layer because of open medium, multihop wireless network, heterogeneous networks, dynamic topology and physical threat [6][9][15]. This paper, we classify the network layer attacks and their interdependencies. Network layer attacks are mainly classified into two types: control plane and data plane. Control plane adversaries affect the route discovery and maintenance phase of reactive, proactive, etc. routing protocols.

Here, adversary node creates attacks by itself such as blackhole, rushing attacks, or combine with other adversaries such as wormhole and colluding attacks. Moreover,

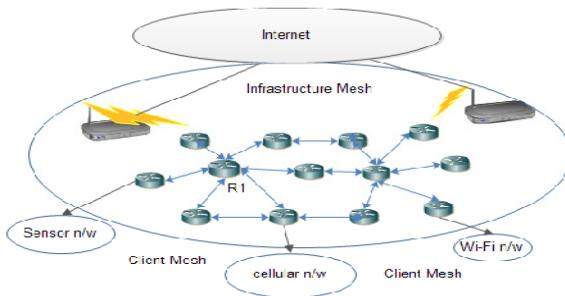


Fig. 1. Wireless Mesh Network Architecture

all the adversaries are internal attackers, and to prevent these attacks existing prevention techniques are ineffective. Other alternative for this problem is Intrusion Detection (ID), many ID techniques have been proposed for single adversary attacks (attack specific) [4][2]. However, very few existing ID techniques are available to protect colluding attacks. However, these solutions are inadequate protect against the all network layer attacks because lack of clear classification of attacks and their interdependencies. Furthermore all these ID techniques follows same reputation value for all type of attacks, it is because lack of available attacks classification. To overcome these problems we have designed taxonomy of network layer attacks, which mainly concentrates on attacks and their interdependencies. In the following, section 2 describes the network layers attacks classification, Section 3 describes conclusion.

2 Network Layer Attacks

WMN lacks robust standard security frameworks, due to this, network layers are more vulnerable to various types attacks. Since WMN supports all wireless networks, it inherits the vulnerability of the protocols supporting that networks. In survey, we found that, there is no in-depth classification of attack on network layers. As a result, existing security solutions are attacks specific, and these solutions cannot detect more than one attack effectively. To overcome these problems, in this section, we have classified all possible attacks on network layers of WMNs.

Fig.2 depicts the taxonomy of network layer attacks. Network layer attacks are classified into two types: Control plane and Data plane attacks. In control plane attacks, adversary intention is to disturb the routing functionalities and/or gain the network traffic of the targeted node. In data plane attacks, adversary (selfish or compromised node) intention is to drop the data packets, injects the false packets, delays the packets etc. In both cases, the adversary may be either internal or external attacker. Internal adversary node is more harmful compare to the external adversary node because it has enough privileges to participate in routing and data forwarding phases. Whereas, external adversary node waste more time to gain the knowledge of target node. We classified the control plane and data plane attacks in the following:

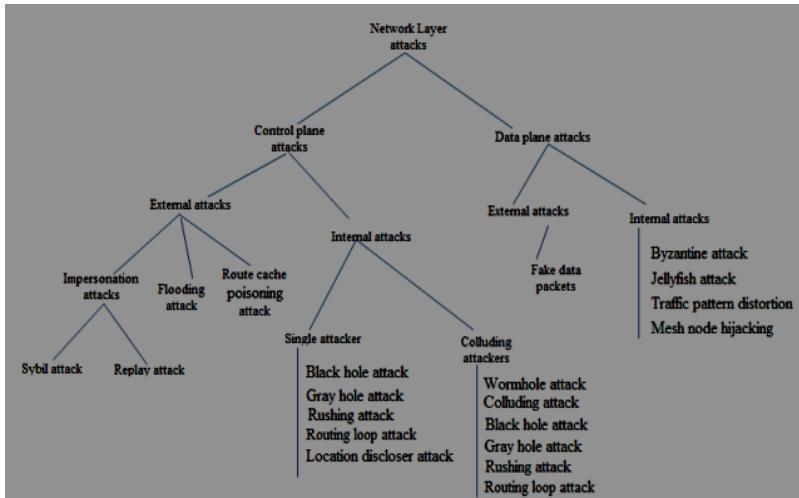


Fig. 2. Taxonomy of Network Layer Attacks

2.1 Control Plane Attacks

External Attacks

In external attacks, unauthorized (external) node creates DoS attacks by flooding or route cache poisoning attacks, and sends authentication routing message by doing impersonation attacks.

Flooding attack: In Flooding attack, the attacker's intention is to exhaust the network resource, such as bandwidth and consumes node's resources like battery and computational power etc [3]. This attack appears in network layer by flooding request, replay, hello packets. Adversary node takes this is an advantage and often creates DoS attacks.

Routing cache poisoning attack: In routing cache-poisoning attack, the legitimate node's important routing data is disturbed by adversary node. Here, the attacker sends excessive false or stale route updates or error packets to the neighbor nodes, due to this neighbor nodes cache often is disturbed and important routing updates are replaced or dropped by unused routing updates or error packets.

Impersonation Attacks: In impersonation attack, the prerequisite condition of an adversary node is to steal the legitimate user's ids and then misuse these ids for different authentication attacks with respect to layers[15] [10].The following replay and Sybil attacks comes under this attack.

Replay attack: In network layer, replay attacker is initially in passive mode to gain authenticated routing request (RREQ), and route reply (RREP). Once, the adversary node gains the authentic information of target nodes then it sends the RREQ or RREP packet on behalf target nodes to gain the network access [15].

Sybil attack: Sybil attack disrupts the network topology and multi-path routing protocol functionality. Here, the attacker appears with multiple identities in the network, which are taken from the compromised node. To disturb network topology, adversary often changes the locations with different legitimate node ids [11]. This attack, mainly disturbs the multipath routing protocols by adversary appearing in most of the node disjoint paths. Fig. 3 shows the Sybil attack. In this scenario, malicious node X has three identities M1, M2 and M3 and all these identities are spoofed in passive mode. A and F are source and destination nodes which need to multipath between them. Malicious node will appear in all multiple paths with different identities such as (A,C,D,M1,F), (A,B,G,M2,F) and (A,B,G,H,L,M3,F).

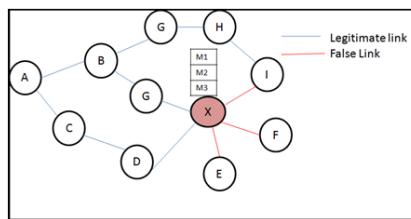


Fig. 3. Sybil attack

Internal Attacks

Internal attackers easily do all external attacks, apart from that these attackers can do more severe attacks. These attacks mainly classified into two groups: Single attacker, Colluding attacker. Single attacker can do the following attacks

Blackhole attack: In this attack, adversary node drops all the packets passed through it. In order to do this, the adversary node attracts the neighbor node with false route reply with less hop count and greater sequence number [13]. Once, route is established through that node then the neighbor node starts sending packets and eventually all packets will be dropped at adversary. This attack scenario in on-demand routing protocol such as DSR [5] is depicted in Figure 4. Here, nodes A and F are source and destination. At route discovery phase, A disseminates route request to find destination path. On-behalf of node A, RREQ packet has been broadcasted by intermediate nodes until it reaches to F node, once, RREQ received by F through I and E nodes. F sets the reverse path to A which is F,E,D,C,A (less hopcount). In Fig. 4 this path is disturbed by the malicious node X at D. X traps D with less hopcount and high sequence number then D will drop the actual route and forwards the X route (F,X,D,C,A) to A. When A starts sending data packets to F, X receives and drops them.

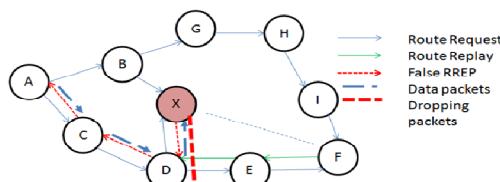


Fig. 4. Black hole attack

Grayhole attack: Grayhole attack is similar kind of blackhole attack, but more sophisticated attack comparative to blackhole attack [1]. Here, adversary node participates without any malicious functionality in the route discovery stage, but when the packets start sending through this node it drops the packets in selective intervals. Detecting this attack is more complex than blackhole because packet drops in wireless networks often happens due to communication errors, hardware error and buffer overflow etc.

Location discloser attack: Location discloser attack is easily caused by internal network node in which it reveals the network topology or location of the nodes [3]. This information is gained by external attacker, and then starts deploying passive attacks or active attacks on the target node or adjacent nodes.

Rushing attack: Rushing attack is a zero delay attack. On-demand routing protocols like AODV [14] / DSR [5] are more vulnerable to this attack, because whenever source node floods the route request packet in the network, an adversary node receives the route request packet and sends without any delay into the network. Whenever the legitimate nodes receive the original source request packets, they are dropped because legitimate nodes, has already received packet from the attacker and treats them as a duplicate packet. Eventually adversary is included in active route and damages at data forwarding phase. Rushing attack can take place at source side or destination side or at the middle [7], [9]. Fig. 5 explains this scenario with attacker at destination side. Here, the attacker X receives the RREQ packet from G then it is broadcasted immediately (no verification process is done). This packet will be received by intermediate nodes I, E and destination node F. Intermediate nodes I and D suppress the actual RREQ packet received from H, E nodes due to their RREQ staleness or duplicity. Destination node F will receive the RREQ packet from I and E on behalf of X. Eventually destination node F includes adversary node X as an intermediate node in source to destination path.

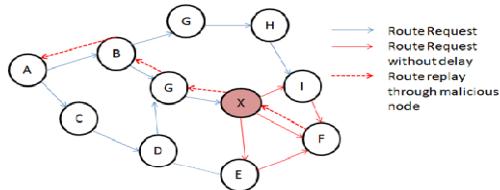


Fig. 5. Rushing attack

Routing loop attack: Routing loop attack intention is to devastate the network resource [12]. Adversary effectively creates the loops when it knows network topology at the route discovery phase. Fig. 6 shows the routing loop attack, in the example scenario node S disseminates the route request (RREQ) packet and received by adversary X and node A. Here adversary X selectively sends RREQ packet to C, not to E and I nodes. C receives RREQ from B and adversary X. C drops B RREQ and takes X RREQ only because less hopcont. Here, adversary do not send RREQ packet to E node intentionally. Then C node broadcast RREQ and it is received by node D then D again broadcast RREQ packet. Eventually E receives RREQ packet then E broadcast

the RREQ packet that is received by node D and node Adversary node. Node D drops this duplicate packet but the adversary X selectively forwards RREQ packet to node I. Then the route is like this S,X,C,D,E,X,I.

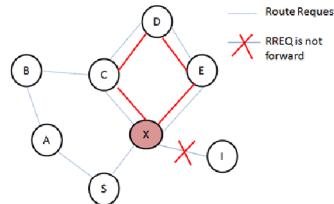


Fig. 6. Routing loop attack

All these attacks are possible by single attacker. Whereas, group of attackers combine to do wormhole, colluding attacks and most of single attacker attacks that we specified in taxonomy. These attackers are called colluding attackers and to detect or prevent these attackers is a challenging issue in WMNs.

Wormhole attack: Wormhole attack is formed by two colluding nodes in the network. To create wormhole attack any two mutual understanding malicious nodes form a tunnel with low latency and broadcast this information into the network. All overheard neighbor nodes send data packets through the tunnel, and then malicious nodes extract the important data from the data packets or drop the packets [8]. This attack is more effective when these nodes coverage more area in WMN. Fig. 7 depicts the wormhole attack formed by Colluding nodes are M1 and M2. Here, M1 coverage area is a_1 and M2 coverage area represented is a_2 respectively. These two nodes form a tunnel with low latency. Moreover, these two nodes cover entire network, thus all the neighbor nodes give more priority to send their data packets through either M1 or M2. Once, the colluding nodes (M1 and M2) get the control over the network then possible attacks are jellyfish and byzantine attacks.

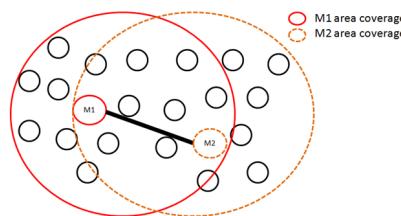


Fig. 7. Wormhole attack

Colluding attack: Colluding attack is a super set of all single node attackers, and more severe attack also. In which, group of malicious nodes work together to isolate the target nodes. Here, colluding nodes send wrong information about target node to isolate it from the legitimate users in the network [3].

Colluding attackers easily do the blackhole, grayhole, rushing routing loop attacks. Moreover these attackers' attacks are more effective than single attacker attacks. To detect or prevent these attacks is more complex due to mutual understanding between attackers.

2.2 Data Plane Attacks

External Attacks

Fake data packets attack: Fake Data packets attack depletes the networks resource by using brute force mechanism. Here, the attacker blindly injects fake or error packets into the network. It creates the interference when the legitimate nodes are in the data transmission phase. Existing Routing paths could be disturbed by this attack.

Internal Attacks

Byzantine attack: In byzantine attack, malicious node intension is to degrade the network performance by doing malicious functionalities such as packet dropping, packet modification and injecting false packets [11]. Here, the internal attacker compromises the legitimate node and this compromised node follows the instructions that are given by the attacker. Eventually this attack, severely degrades the network performance by implementing the following control plane attacks : blackhole, grayhole, wormhole, rushing attacks.

Jellyfish.attack: In jellyfish attack, attacker perception is to decrease the goodput of owns to near-zero and leads to DoS attack [3]. To do this attack, the attacker behave trustworthy at control plane but when the data packets come through this node, more delay is created for selected packets or reordered the packets. Eventually, attacker can increase end-to-end delay, and drastically reduce the goodput of the network and the following control plane attacks implement this attack : wormhole, rushing attacks.

Traffic pattern distortion: Traffic pattern distortion creates the resource depletion such as channel jamming [15]. The attacker easily acquires the data forwarding information due to the broadcast nature of wireless communication. To do this, attacker overhears the communication channels of the neighbor nodes and analyzes the traffic patterns, and sends excessive packets when high traffic relies on any of its two neighbor nodes. The original data packet is jammed by the attacker's false data packet and the following control plane attacks implement this attack : flooding and route cache poisoning attacks.

Mesh node hijacking: Mesh node hijacking is mainly caused by unfair greedy nodes, which always look to send their traffic in with high priority. Here, a greedy network owner may attempt to leverage other owners' mesh nodes for forwarding its own traffic. A hostile network owner may attempt to leverage neighbor owners' mesh nodes for forwarding its own traffic and take one-step further by protecting its own mesh nodes by proprietary means [15].

In the above network layer attacks classification, we explain the interdependencies of control plane and data plane attacks such as wormhole attack followed by byzantine and jellyfish attacks, traffic pattern distortion attack followed by flooding and

route cache poisoning attacks wormhole attack followed by byzantine and jellyfish attacks (WFJBA) [16]. Base on this work, we have proposed technique a technique called Intrusion Detection Technique for Wormhole and Following Jellyfish and Byzantine attacks in Wireless Mesh Network. This approach mainly depends on initial end-to-end packet delay, average end-to-end packet delay, and worst-case end-to-end packet delays are the major parameter to detect these attacks.

3 Conclusion

This paper, we have done an exclusive survey WMN network layer attacks. Based on this survey, we classify and draw the taxonomy of the all network layer attacks in WMN. Moreover, we explain the interdependencies of attacks and their severity. This taxonomy helps us to identify and isolate the multiple colluding attacks with respective to different reputation values of each attack by using intrusion detection of WFJBA approach.

References

- [1] Gerkis, A.: A survey of wireless mesh networking security technology and threats. In: OTM workshops, pp. 1–17 (2006)
- [2] Hoang, L., Uyen, T.: Secure routing in wireless sensor networks: attacks and counter measures. Ad Hoc Networks, pp. 113–127 (2003)
- [3] Khan, A.-S.: Security of self-organizing networks. In: MANET, WSN, WMN, VANET. CRC Press (2010)
- [4] Ganesh, K., Khilar, P.: Routing misbehavior detection and reaction in MANETs. In: ICIIS 2010 International Conference, pp. 80–85 (July 2010)
- [5] Marshall, J.: An analysis of srp for mobile ad hoc networks. IEEE Internet Computing 12, 30–36 (2002)
- [6] Muhammad, S., Choong, S.: Security issues in wireless mesh networks. In: IEEE/IPSJ International Symposium on Applications and the Internet, pp. 717–722 (2009)
- [7] Palanisamy, V., Annadurai, P.: Impact of rushing attack on multicast in mobile ad hoc network. International Journal of Computer Science and Information Security, IJCSIS 4(1&2) (2009)
- [8] Papadimitratos, P., Haas, Z.J.: Secure routing for mobile ad hoc networks. In: Proceedings of the SCS Communication Networks and Distributed Systems Modeling Simulation Conference (CNDS 2002), pp. 193–204 (2002)
- [9] Ping, Y., Yue, W.: A survey on security in wireless mesh networks. IETE Technical Review 27(1), 6–14 (2010)
- [10] Redwan, H., Ki-Hyung, K.: Survey of security requirements, attacks and network integration in wireless mesh networks. Frontier of Computer Science and Technology 2, 3–9 (2008)
- [11] Sahil, S., Anil, G.: Current state of art research issues and challenges in wireless mesh networks. In: IEEE Second International Conference on Computer Engineering and Applications, pp. 199–203 (2006)
- [12] Shariful, M., Hamid, A.: Shwmp: a secure hybrid wireless mesh protocol for ieee802.11s wireless mesh networks, pp. 95–114. Springer, Heidelberg (2009)

- [13] Tamilselvan, L., Sankaranarayanan, V.: Prevention of blackhole attack in manet. In: Tamilselvan, L., Sankaranarayanan, V. (eds.) International Conference on Wireless Broadband and Ultra Wideband Communications, pp. 16–21 (2007)
- [14] Trong, H., Dai, T.: Adaptive algorithms to enhance routing and security for wireless pan mesh networks. In: OTM Workshops, pp. 585–594
- [15] Zhang, Y.: Security in Wireless Mesh Networks. CRC Press (2008)
- [16] Reddy, K.G., Thilagam, P.S.: Intrusion Detection Technique for Wormhole and Following Jellyfish and Byzantine Attacks in Wireless Mesh Network. In: Thilagam, P.S., Pais, A.R., Chandrasekaran, K., Balakrishnan, N. (eds.) ADCONS 2011. LNCS, vol. 7135, pp. 631–637. Springer, Heidelberg (2012)

Implementing Availability State Transition Model to Quantify Risk Factor

Shalini Chandra and Raees Ahmad Khan

Department of Information Technology, BBA University, Lucknow, U.P., India

Abstract. In IT era, every organization depends on computer and internet for its daily routine works. A major objective of an information security policy is to ensure that information is always available to support critical business processing. This is a great challenge to develop secure software to meet its requirements and to satisfy security requirements i.e. Confidentiality, Integrity, and Availability (CIA) against identified risks. To prevent sensitive data, creating session mechanism is used which is helpful in reducing denial of service attack. In this paper, a methodology has been proposed and validated to assess the availability risk at design level using methods and classes.

1 Introduction

Security is main concern for both developers and end users. Normally, there is a lack of in depth security knowledge to both developer and users. Several security mechanisms including digital signature, timestamp, encryption etc. are available. These security mechanisms generally do not stop malicious attacks completely. In Sep 2009, customer service's sixteen-year-old EDI system was taken out of services when a communication cable snapped. It was estimated that exporters were losing about one-million dollars per day [15].

Availability is the fraction of time that a produced system is functioning acceptably. Denial of service attack makes system's service unavailable for unauthorized users. Availability is about ensuring that services are available and operational when they are needed. It is suggested that security metrics can be used during coding phase to eliminate vulnerabilities, by measuring adherence to secure coding standards, identifying vulnerabilities that may exist and analyzing security flaws that are eventually discovered [6, 7].

In [9, 10] three basic security requirements CIA has been considered. To satisfy these requirements it is essential to protect our data from unauthorized information disclosure and information alteration [8]. In order to achieve the objective it is required that the services must be provided to only authorized user for specific time duration. It will never be open for unlimited time. Confidentiality and integrity has been taken in [16, 17]. This part of the work is focused to measure availability risk at design stage of software development.

2 Review Work

In 2004, B. B. Madan. et. al. presented an approach for quantitative assessment of security attributes for an intrusion tolerant system. Using the approach, they developed a generic model that enables the study of variety of intrusion tolerant strategies as well as the impact of a security attack. They had computed the probability of security failure due to the violations of security attributes [3]. H. Song and C. Lengsuksun developed a framework for cluster availability specification and evaluation in 2005. Authors concluded that for providing continuous services in cluster computing environment it is required to include high availability features. They also predicted that using the framework availability analysis of cluster computing systems could be done in the early stage of software development [11].

Rap tool is developed to evaluate reliability and availability of software, which supports architecture level. The tool performs in three phases: the first phase is to define reliability and availability goals; second phase is to transform the goals to architectural elements and third phase is to represent these elements in architectural models and perform evaluation in order to verify that the resultant architecture is satisfying the requirements or not [1, 2]. In literature survey, no such methodology is available to assess and ensure that composition of software architecture satisfies security requirements. No technique is available to predict and verify that design of software systems satisfies the property [5].

3 Terminologies Used

Before implementation of the methodology some prerequisite requirements need to be fulfilled such as set of sensitive methods, validity check methods, session methods, don't care classes, sensitive classes etc. The following terminologies are used in the proposed methodology.

Sensitive Methods: Methods displaying, altering, or processing any sensitive information, is called sensitive method. Such method must be processed only after its validity check. If there are n classes in a class diagram then set of the methods of class C_i is represented as $SM(C_i)=\{SM_{i1}, SM_{i2}..\}$, where $i=1....n$. In order to identify the methods, some suggestive checks are proposed in table 1.

Validity Check Method: A method verifying the authenticity for accessing sensitive data by implementing a conditional check is known as validity check method. Set of the methods of class C_i is represented as $VM(C_i)=\{VM_{i1}, VM_{i2}..\}$. In order to identify these methods, some suggestive checks are proposed in table 2.

Session Method: Methods used for running a session for specific time duration. Session is created whenever sensitive services need to provide to authorize user. Session method supports series of sensitive methods. Set of session methods of class C_i is represented as $TM(C_i) = \{TM_{i1}, TM_{i2}...\}$, where $i=1....n$.

Table 1. Checks to identify sensitive methods (Availability perspective)

S.No.	Checks	Y/N
1.	The method has any attribute displaying, modifying or altering any sensitive information (e.g. user id, user name, card id, card no. etc.)	
2.	The method has any attribute displaying or altering encrypted code, password / passcode etc.	
3.	The method has any attribute displaying or altering authorization code	
4.	The method displaying or altering conditional check (e.g. expiry date).	
5.	The method displaying or alters any transaction id etc.	

Table 2. Checks to identify validity check methods

S.No.	Checks	Y/N
1.	The method process attributes having sensitive information (e.g. user id, user name, card id, card no. etc.)	
2.	The method process any attribute carrying encrypted code, password / passcode etc.	
3.	The method process any attribute having an authorization code	
4.	The method process any conditional check (e.g. expiry date).	
5.	The method process transaction id etc.	

Sensitive Class: A class having sensitive methods and needs protection or having availability violation is said to be sensitive class. The methods of the class may not be public or modifying information without any authentication, authorization, and validity check methods. Set of sensitive class are represented as SC.

Safe Class: A class having sensitive methods with corresponding validity checks method of its own inherited or coupled before allowing modification of sensitive information, is considered as safe class and represented as SF.

Definitions of don't care, risky class, super class, and sub class have been considered same as already discussed in [16] [17]. During availability check of class hierarchy, these terminologies have been used. As limitation of the methodology identification of sensitive methods and validity check methods are required. In order to find out these methods some prescriptive checks have been proposed. These checks are developed for a specific application. In broader perspective, UMLsec and SPARK's annotations may be used for identification of sensitive methods and validity check methods. Focus of the methodology is to produce quantitative results in order to assess security level of software among different versions of software design and provide basis for ranking software.

4 The Methodology

The methodology describes, to check availability of sensitive methods is maintained or not. In a class, if any method display, alter or process, any sensitive data, then the

class may be considered as a sensitive class. In the design phase, a class having any self-checking method or inheriting any validity check method and session method corresponding to the sensitive method may be verified. It is not permitted to use or extract sensitive information for unlimited time. In order to maintain availability it is required to check that sensitive methods have its corresponding validity check method. Additionally, it is also required to check that sensitive methods or verification methods have its corresponding session method or not. If yes, then availability of methods will be maintained, if not then there is availability risk.

Availability risk quantification is possible through quantifying its constructs that affect the security attribute. During development of security metric, some terminologies have been discussed. These terminologies are used to measure the constructs. During the development of metric, security attribute is required to check at the method level, using check process. To check the security attribute at class level and hierarchy level an algorithm has been developed. Additionally, ASTM has been developed to make it easy to understand the dynamic behavior of availability [4]. As an outcome of the algorithm and the transition model, we get values of constructs. Using the constructs, security metrics have been developed to measure availability risk. Development of security metrics consist four parts. First, method check process (ARCP), second class hierarchy check algorithm (AC^2H), third state transition model (ASTM), and fourth security metric (CHARF and SPARF).

4.1 Availability Risk Check Process (ARCP)

In order to check availability of class hierarchy every sensitive method of class need to go through ARCP. ARCP checks two constraints for all sensitive methods of class a). For every sensitive method corresponding validity check method (authentication, authorization or validation, depending on the security context), is must. b). Corresponding session method is existing or not? (Corresponding validity check method or session method may be defined in the same class or it may be coupled or inherited to another class.) If both constraints are satisfied only then it is considered safe otherwise it may be risky itself and for its subclasses also.

Classes having sensitive methods have been considered as sensitive class. For every sensitive method in $SM(C_i)$, its corresponding validity check method and session method should be there. It may be in the same class or in any of its superclass $SP(C_i)$. If any method of $SM(C_i)$ does not have its corresponding validity check method and session method, the method may be exploited easily, and making the class risky.

Let, $X=\{SM_{ij}|SM_{ij} \in SM(C_i)\}$, where $SM(C_i)=\{SM(C_i) \cup SM(SP(C_i))\}$, $SM(C_i)$ is the set of sensitive methods of a class C_i , $SM(SP(C_i))$ is the set of sensitive methods of all super classes of C_i where $0 < i \leq n$, n is no. of classes in a class hierarchy. SM_{ij} is a sensitive method of Class C_i , need protection from unauthorized access, where, $0 < j \leq m$ m is no. of methods in class C_i . $Y=\{VM_{ik}|VM_{ik} \in VM(C_i)\}$, Where $VM(C_i)=\{VM(C_i) \cup VM(SP(C_i))\}$, $VM(C_i)$ is the set of validity check methods of a

class C_i . $VM(SP(C_i))$ is the set of validity check methods of all super classes of C_i where $0 < i \leq n$, n is no. of classes in a class hierarchy. VM_{ik} is the validity check methods for class C_i , where, $0 < k \leq p$ p is no. of methods in class C_i . $Z = \{TM_{ik} | TM_{ik} \in TM(C_i)\}$, where $TM(C_i) = \{TM(C_i) \cup TM(SP(C_i))\}$, $TM(C_i)$ is the set of session methods of a class C_i where $0 < i \leq n$, n is no. of classes in a class hierarchy. TM_{ik} is the session methods for class C_i , where $0 < k \leq p$, p is no. of methods in class C_i .

$$\begin{aligned} \text{Set } A &= \left\{ \begin{array}{ll} 1, & \forall SM \in X \\ 0, & \text{otherwise} \end{array} \right\}, \quad B = \left\{ \begin{array}{ll} 1, & \text{if } Y \text{ exists corresponding to } X \\ 0, & \text{otherwise} \end{array} \right\} \\ C &= \left\{ \begin{array}{ll} 1, & \text{if } Z \text{ exists corresponding to } Y \\ 0, & \text{otherwise} \end{array} \right\} \\ L &= \left\{ \begin{array}{ll} 1 & \text{if } A = 1 \& B = 1; \text{method is safe} \\ 0 & \text{if } A = 1 \& B = 0; \text{method is risky} \end{array} \right\} \end{aligned}$$

4.1.1 Implementation

If $L=1$, means there is no availability risk. In that particular class for sensitive method corresponding validity check method and session method both exist. If $L=0$, means there is availability risk. In that class for any sensitive method, its corresponding validity check method or session method is missing. This method may act as loop hole or leakage channel (or used by any other means).

Availability risk check of method is based on the presence / absence of sensitive method, validity check method and session method either its own or used by any means. If result of any input combination is false, then the whole class will be considered as risky class making the complete hierarchy risky.

4.1.2 An Example

As shown in fig 1, session class is running user session method. It starts automatically when user login. If system is in steady state for long period it automatically timeout. To avail sensitive method user need to login again. Account class is safe because get balance() method is inheriting start session() method and user login() method both and satisfying condition (presence of validity check method as login() and session method as start session() method). Class client is risky class because change password is also a sensitive method and user must be login.

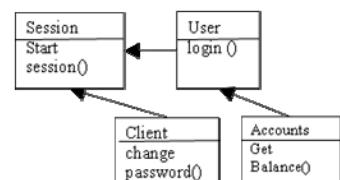


Fig. 1. Class hierarchy

4.2 Availability Check for Class Hierarchy (AC²H) Algorithm

ARCP process is used to check availability of a class. For quantitative assessment of availability risk, it is required to go through ARCP for all sensitive classes and implement it on complete class hierarchy. Due to sharing of methods between classes it

becomes mandatory to check its associated classes also. In order to satisfy above requirement an algorithm has been proposed.

Procedure begin

```

Step 1: For each class  $C_i \in TC$ 
Step 2:  $C_i$  [status] =null
Step 3: Do if  $C_i$  is sensitive
Step 4:           Then Enqueue (SC, Ci)
Step 5:           Else Enqueue (DC, Ci)
Step 6: For each class  $C_i \in SC$ 
Step 7: If  $C_i$  [status]! =checked
Step 8:            $C_i$  [status] =checked
Step 9:           Do cflag=null
Step 10:          Derive  $SMC_i$ 
Step 11:          For each method  $SM_{ij} \in SMC_i$ 
Step 12:            Do if ARCP=false
Step 13:            Then cflag=false
Step 14:            If cflag=false
Step 15:              Then verify  $C_i$  with checklist
Step 16:              Enqueue (RC,  $C_i$ )
Step 17:               $Q \leftarrow 0$ 
Step 18:              Enqueue (Q,  $C_i$ )
Step 19:              While  $Q \neq \emptyset$ 
Step 20:                Do  $C_i \leftarrow Dequeue (Q)$ 
Step 21:                For each class  $C_k \in SBC_i$ 
Step 22:                  Enqueue (Q,  $C_k$ )
Step 23:                  Repeat step 7 for  $C_k$ 
Step 24: For each class  $C_i \in DC$ 
Step 25: If count ( $SPC_i \cap RC$ ) >0
Step 26:           Then Enqueue (RC, Ci)
End

```

The algorithm first identifies set of SC and DC. In order to check the availability risk, it is required to check all sensitive methods and valid check methods of the same class. If, for any sensitive method, ARCP returns false ($L=0$) then after verifying with checklist, the class and its all subclass will be recognized as a risky class. Moreover, even if any of super-class of DC is risky class then the class will be included in set of RC. From the ARCP it gets clear that don't care class does not require any protection but if it uses or shares method or attribute from any sensitive class, safe class or risky class, in the absence of validity check method, then it may also act as a leakage channel or risky class.

4.3 Availability State Transition Model (ASTM)

ASTM already discussed in [4]. As shown in fig 2, first state is don't care state. A class at don't care state may come to second state i.e. sensitive state if it adds any

sensitive method add to it or if it inherits sensitive method. Class at sensitive state may come to third state i.e. validated state if its corresponding validity check method adds or shares from any other class. Class of validated state is comparatively secure to sensitive class but not completely without session methods. If class adds or inherits session method corresponding to validity check method then it comes to fourth state i.e. safe state. If there is any loss of validity check method and session method then it comes to fifth state i.e. risky state. If any class have any loss of validity check method and session method they comes to risky state directly. Sharing of methods from lower states class may introduce unwanted interrupts that may encounter at any state.

4.4 Security Metrics

Two security metrics have been established, Class Hierarchy Availability Risk Factor and Security Pattern Availability Risk Factor. Availability of software can be easily disturbed if sensitive data is computed incorrectly. Probability of Availability risk violation increases with the increase of availability risk factor.

4.4.1 Class Hierarchy Availability Risk Factor (CHARF)

This metric measures the overall class hierarchy availability risk factor. The metric is defined as the ratio of total number of risky classes to total number of classes. CHARF is the probability of availability risk of complete class hierarchy. The CHARF for class hierarchy is defined as follows:

$$\text{Class Hierarchy Availability Risk Factor (CHARF)} = \frac{\text{count}(RC)}{\text{count}(TC)}$$

The RC is set of risky classes in class hierarchy. TC is used to count total classes.

4.4.2 Security Pattern Availability Risk Factor (SPARF)

The metric SPARF measures probability of availability risk of security pattern. SPARF is defined as the ratio of total number of risky classes to total number of sensitive classes. The SPARF for class hierarchy is defined as follows:

$$\text{Security Pattern Availability Risk Factor (SPARF)} = \frac{\text{count}(RC)}{\text{count}(SC)}$$

The RC is set of risky classes present in class hierarchy. SC is used to count total number of sensitive classes.

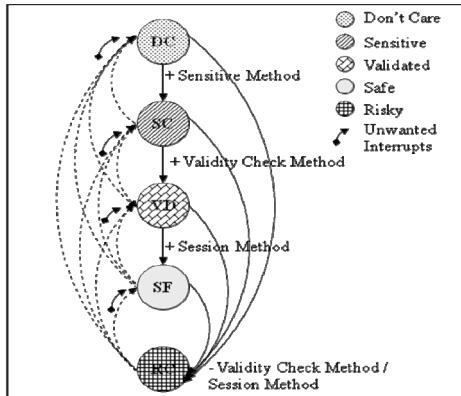


Fig. 2. Availability State Transition Model (ASTM)

5 Experimental Validation

For experimental validation of metrics, some prerequisite requirements need to be fulfilled i.e. set of non-sensitive classes, sensitive classes, safe classes, risky classes. For identification of such classes, some mechanisms already exist such as UMLsec etc. The focus of the research is quantifying security. For identification of such methods, some suggestive checks have been introduced in section 3.

In this section, we design an experiment, which carried out for empirical validation of developed metric suites. One of the important aspects of the experiment is to present a clear picture of what we believe and what we observe [18]. This leads us to formulate hypothesis. Empirical validation of developed metric suite has been performed in four steps.

5.1 Hypothesis Formulation

Null Hypothesis H₁₀: Availability risk factor of hierarchy will not increase as the number of risky classes increases.

Alternate Hypothesis H₁₁: Availability risk factor of hierarchy will increase as the number of risky classes increases.

Null Hypothesis H₂₀: Availability risk factor of hierarchy will not decrease as the number of safe classes increases.

Alternate Hypothesis H₂₁: Availability risk factor of hierarchy will decrease as the number of safe classes increases.

Null Hypothesis H₃₀: Availability risk factor of hierarchy will not decrease as the number of don't care classes increases.

Alternate Hypothesis H₃₁: Availability risk factor of hierarchy will decrease as the number of don't care classes increases.

Null Hypothesis H₄₀: Availability risk factor of hierarchy will not increase as the number of sensitive classes increases.

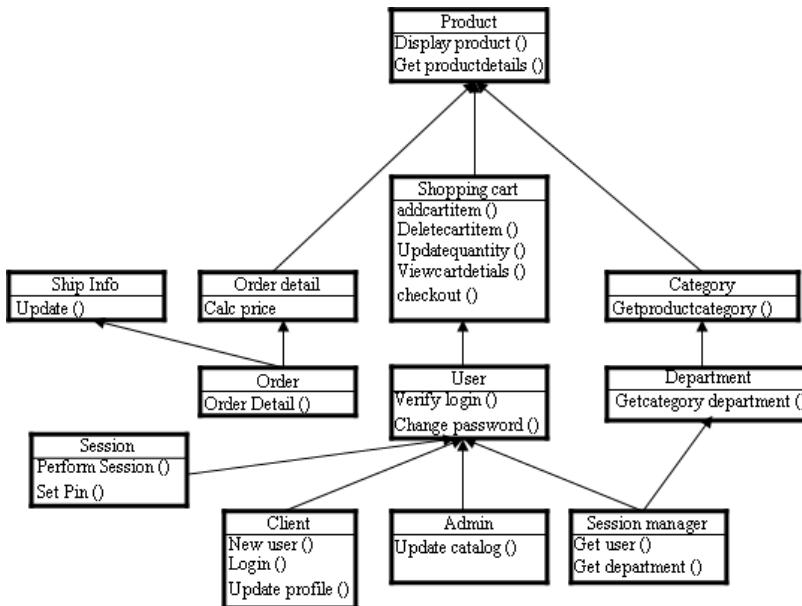
Alternate Hypothesis H₄₁: Availability risk factor of hierarchy will increase as the number of sensitive classes increases.

Null Hypothesis H₅₀: Sensitivity of class hierarchy does not affect integrity risk.

Alternate Hypothesis H₅₁: Sensitivity of class hierarchy affects availability.

5.2 Collection of Data

For collection of data, it is required to implement developed methodologies on UML designs. The methodology has been implemented on realistic projects. These implementations have been presented in two forms. First is algorithm implementation and second is availability states modeling. Implementation results of the methodology have been given. Due to organization's policy, it is not feasible to give detail implementation of all versions of all projects. Only relevant data has been presented as an outcome of the algorithm and state transition model. Projects have been developed in iterative versions. In order to investigate robustness of metrics, these methodologies have been implemented on all versions of software and data have been collected.

**Fig. 3.** Class Hierarchy (Availability)

5.2.1 Algorithm Implementation

Implementation of the algorithm has been shown in fig 4. First version of the class hierarchy is shown in fig 3. The classes **Session**, **User**, **Client**, **Admin**, **User Session Manager**, **Shopping Cart**, **Orders**, **Shipping Info**, **Product**, **Category**, **Department**, and **Order detail** is denoted as $C_0, C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}$, and C_{11} respectively.

Ist Implementation (class hierarchy)

TC= { $C_0, C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}, C_{11}$ }

SC= { $C_0, C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}, C_{11}$ }

DC= {}

For class C_0

$SMC_0 = \{SM(C_0) \cup SM(SP(C_0))\} = \{SM(C_0) \cup SM(C_1)\} = \{SM_{02}, SM_{12}\}$

$VMC_0 = \{VM(C_0) \cup VM(SP(C_0))\} = \{VM(C_0) \cup VM(C_1)\} = \{VM_{01}, VM_{11}\}$, ARCP=T

Fig. 4. AC²H Algorithm Implementation.

Total number of classes in version 1 is 12. First classes have been categorized into sensitive and don't care. As a result we get sensitive classes, $SC = C_0, C_1, C_2, C_3, C_4, C_5, C_6, C_8, C_9, C_{10}, C_{11}$ and don't care classes $DC = \emptyset$. For every sensitive method of sensitive class C_1 we go through ARCP. At last *cflag* is **null** which states that there is no availability risk for class C_1 (**ARCP** is **true** for every sensitive method of class C_1). Implementation of the AC²H algorithm on the class C_1 has been shown in fig 4. Repeat the procedure for all sensitive classes. In this class hierarchy, total 12 classes

have been used. In this class hierarchy, all classes have sensitive methods, displaying, altering, or processing sensitive information. Therefore, number of don't care classes are 0. All classes have sensitive methods, so, number of sensitive classes are 12. After implementation of algorithm, it has been found that in this class hierarchy, total classes 7 are at risky state and 5 classes are at safe state. These all safe classes have its corresponding validity check method and session method both. As an outcome of the implementation, input details for further security quantification are TC=12, DC=0, SC=12, SF=5, and RC=7. Same algorithm has been applied on 9 samples of class hierarchy. Results of implementation on all 9 samples are given in table 3.

5.2.2 Availability States Modeling

ASTM is simple and easy to implement on class hierarchies. Same procedure has been followed on all nine hierarchies. States of classes gets change in the same manner as already discussed in section 4. Outcome of availability state transitions and metric results calculated and represented in table 3.

5.3 Metric Results

This section shows how proposed security metric is able to assess availability risk of class hierarchy at design stage. Table 3 shows the results of applying security metrics to 9 hierarchies. To make it easy to understand, results have been represented in graphical form in fig 5. Figure shows the variations of availability risk factors CHARF and SPARF corresponding to DC, SC, SF, VD and RC classes. Based on results the graph has been plotted. Graph reflects the relationship of effect of increment and decrement between class states and risk factor.

5.4 Hypothesis Testing

For hypothesis testing, security attributes risk factors and their states have been calculated. To analyze the effect of variations on risk factors their correlation coefficients have been calculated. Correlation coefficient (CC) interpretations are a). .00 to .20 (negligible), b). .20 to .40 (low), c). .40 to .60 (moderate), d). .60 to .80 (substantial) and e). .80 to 1.00 (high). In order to accept or reject hypothesis correlation coefficient and their corresponding p-values has been considered.

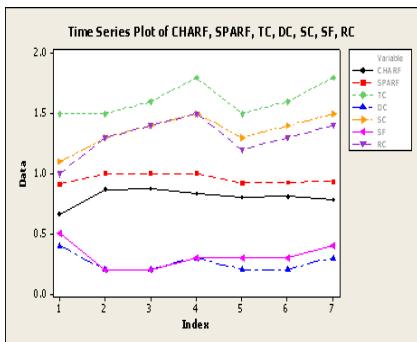
The results of the experiment are given in table 4. Availability states affect availability risk factor. Risk factors vary according to its states as shown in fig 5.

To test Hypothesis H_{10} , we computed the CC between availability risk factors and risky classes. CHARF and SPARF have strong positive correlation to risky classes. It concludes that when the number of risky classes increases availability risk factors will also increase. The results does not support null hypothesis H_{10} . Hence, we reject H_{10} hypothesis and accept alternative hypothesis H_{11} .

To test Hypothesis H_{20} , we computed the CC between availability risk factor and safe classes. CHARF and SPARF have a negative correlation to safe classes. It concludes that when number availability risk factors will decrease as the number of safe classes increase. The results does not support null hypothesis H_{20} . Hence, we reject the H_{20} hypothesis and accept alternative hypothesis H_{21} .

Table 3. Availability State Modeling on 9 Class Hierarchies and Metric Results

Class Hierarchy	TC	DC	SC	SF	RC	CHARF	SPARF
1 st	15	3	12	8	7	0.466667	0.583333
2 nd	16	3	13	8	8	0.500000	0.61538
3 rd	18	3	15	8	10	0.444444	0.57143
4 th	15	3	12	7	8	0.533333	0.66667
5 th	16	3	13	7	9	0.562500	0.69231
6 th	18	3	15	7	11	0.500000	0.64286
7 th	15	1	14	1	14	0.933333	1.00000
8 th	16	1	15	1	15	0.937500	1.00000
9 th	18	2	16	2	16	0.888889	1.00000

**Fig. 5.** Graphical Representation of Availability States and Risk Factor.

To test Hypothesis H_{30} , we computed the CC between availability risk factors and don't care classes. CHARF and SPARF have a negative correlation to don't care classes. It concludes that integrity risk factors will decrease as the number of don't care classes increase. The results does not support hypothesis H_{30} . Hence, we reject H_{30} hypothesis and accept alternative hypothesis H_{31} .

To test Hypothesis H_{40} , we computed the CC between availability risk factors and sensitive classes. CHARF and SPARF have a positive correlation to sensitive classes. It concludes that chances to increase availability risk factors increase as the number of sensitive classes increase. The results does not support hypothesis H_{40} . Hence, we reject H_{40} hypothesis and accept alternative hypothesis H_{41} .

To test Hypothesis H_{50} , we computed the CC between availability risk factors and sensitivity of the class hierarchy. CHARF and SPARF have a strong positive correlation to sensitive classes. This is because of the high correlation of availability risk factor with sensitive classes and risky classes. Sensitivity of the class hierarchy is completely dependent on both. The relationships between sensitive classes, risky classes and total classes have also been shown in fig 5. The sensitivity level will increase only if the sensitive classes increase because don't care classes will also increase the total number of classes. Availability risk will increase if sensitive classes and risky classes increase in parallel. Results conclude that when the sensitiveness level increases, the chance to increase availability risk factor also increase. The results does not support hypothesis H_{50} . Hence, we reject hypothesis H_{50} and accept alternative hypothesis H_{51} .

Table 4. Correlation Analysis Summary between Availability States and Risk Factors.

Metrics	TC	DC	SC	SF	RC	Sensitivity
CHARF	-0.084	-0.928	0.700	-0.997	0.969	0.976
(p-value)	(0.830)	(0.000)	(0.036)	(0.000)	(0.000)	(0.000)
SPARF	-0.039	-0.902	0.723	-0.992	0.979	0.956
(p-value)	(0.920)	(0.001)	(0.028)	(0.000)	(0.000)	(0.000)

6 Set Regression Line

Based on relationship between constructs and risk factors, the relative significance of individual constructs that influence risk factor is weighted proportionally. Multiple linear regression investigates and models the linear relationship between dependent variables (risk factors) and independent variables (constructs). The regression equation is an algebraic representation of the regression line. The regression equation takes the form of:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \dots + \alpha_i x_i$$

Where, y is the dependent (response) variable. Constant α_0 is the value of the dependent variable when the independent variable (s) is zero. Coefficients ($\alpha_0, \alpha_1 \dots \alpha_k$) represents the estimated change in mean response for each unit change in the dependent value.

The security attributes risk factors of class hierarchy depend upon one or more constructs. Component weightage (CW) of individual components have been calculated in terms of regression coefficients α_i . 9 medium size projects were used to fit the regression line of real data from commercial projects of software Company. These projects include the number of classes (11-18). The projects were independently completed over a period of six months. Using these data, the CW coefficient calculated for TC, DC, and SF.

After regression equation establishment, regression analyses have been presented. The Coefficient table lists the estimated coefficients for the independent variables (predictors) as shown in table 5. Linear regression examines the relationship between dependent and independent variables. In order to determine whether or not the observed relationship between the dependent and independent variables is statistically significant, coefficient p-values have been identified and comparison between coefficient p-values to commonly used α -level (0.05) has been done. Here s is measured in the units of the response variable and represents the standard distance data values fall from the regression line. For this study, the better the equation predicts the response, the lower S is. R^2 (R-Sq) describes the amount of variation in the observed response values that is explained by the independent variables. Adjusted R^2 is a modified R^2 that has been adjusted for the number of terms in the model. If unnecessary terms will be included, R^2 can be artificially high. Adjusted R^2 is used to compare models with different numbers of predictors.

Table 5. Estimated Coefficients for Availability Risk Factors

Estimated Coefficients for CHARF _{cal}					Estimated Coefficients for SPARF _{cal}				
Predi.	Coef	SE Coef	T	P	Pred.	Coef	SE Coef	T	P
Const.	0.76405	0.08145	9.38	0.001	Const	0.84603	0.09392	9.01	0.001
TC	0.01552	0.01394	1.11	0.328	TC	-0.02634	0.01608	-1.64	0.177
DC	-0.02672	0.02096	-1.27	0.271	DC	0.04781	0.02417	1.98	0.119
SF	-0.05546	0.01429	-3.88	0.018	SF	-0.03246	0.01648	-1.97	0.120
RC	0.00084	0.01422	0.06	0.956	RC	0.03763	0.01639	2.30	0.083
$S = 0.00980016$ R-Sq = 99.9%					$S = 0.0113000$ R-Sq = 99.8%				
R-Sq(adj) = 99.8%					R-Sq(adj) = 99.6%				

Regression Equation for Availability Risk Factor: During the establishment of multiple linear regression model, it has been identified that Sensitive Classes (SC) are highly correlated with other availability constructs. Therefore, SC has been removed from the equations. The Regression equations are:

$$\text{CHARF}_{\text{cal}} = 0.764 + 0.0155 \text{ TC} - 0.0267 \text{ DC} - 0.0555 \text{ SF} + 0.0008 \text{ RC} \quad (1)$$

$$\text{SPARF}_{\text{cal}} = 0.846 - 0.0263 \text{ TC} + 0.0478 \text{ DC} - 0.0325 \text{ SF} + 0.0376 \text{ RC} \quad (2)$$

For $\text{CHARF}_{\text{cal}}$ estimated coefficients are shown in table 5. The independent variables (TC, DC, SF, and RC) explain 99.9% of variation in the rate observations. The adjusted R is 99.8%, which is a decrease of 0.1%.

For $\text{SPARF}_{\text{cal}}$, estimated coefficients are shown in table 5. The independent variables (TC, DC, SF, and RC) explain 99.8% of variation in the rate observations. The adjusted R is 99.6%, which is a decrease of 0.2%.

Table 6. Calculated and Tabulated Values of Availability Risk Factors.

TC	DC	SC	SF	RC	Using Metrics		Using Equation	
					$\text{CHARF}_{\text{tab}}$	$\text{SPARF}_{\text{tab}}$	$\text{CHARF}_{\text{cal}}$	$\text{SPARF}_{\text{cal}}$
15	2	13	2	13	0.867	1	0.8425	0.9709
16	2	14	2	14	0.875	1	0.8588	0.9822
18	3	15	3	15	0.833	1	0.8084	0.9825
15	2	13	3	12	0.8	0.923	0.7862	0.9008
16	2	14	3	13	0.813	0.929	0.8025	0.9121
18	3	15	4	14	0.778	0.933	0.7521	0.9124

7 Tryout

The six projects of the validation suite have been evaluated using the proposed security metrics. Results of the metric indicated as tabulated values. In this section, only outcome of the metrics as tabulated values have been given named as $\text{CHARF}_{\text{tab}}$ and $\text{SPARF}_{\text{tab}}$. Calculated values of risk factors have been computed using regression equations (1) and (2) named as $\text{CHARF}_{\text{cal}}$ and $\text{SPARF}_{\text{cal}}$ are shown in table 6.

8 Statistical Analysis

Students-t test has been used to test the significance of the correlation between observed and expected security attribute risk factors. Students-t test is used to perform a hypothesis test and compute confidence interval of the difference between two sample means.

i). For $\text{CHARF}_{\text{tab}}$ (using metrics) versus $\text{CHARF}_{\text{cal}}$ (using the regression equation) two-tailed Students-t test: **H₀ (the null hypothesis):** There is no significant difference between $\text{CHARF}_{\text{tab}}$ and $\text{CHARF}_{\text{cal}}$. **H₁ (the alternative hypothesis):** There is a significant difference between $\text{CHARF}_{\text{tab}}$ and $\text{CHARF}_{\text{cal}}$.

$H_0: \mu_1 - \mu_2 = \delta_0$ versus $H_1: \mu_1 - \mu_2 \neq \delta_0$, where μ_1 and μ_2 are the sample means and δ_0 is the hypothesized difference (zero) between the two sample mean.

Mean, StDev and SEMean for six samples have been calculated and results shown in table 7. For the given sample, a 95% confidence interval is (-0.0309, 0.0692) which includes zero. The reference value of 0 is within the confidence interval, so we can accept H_0 with 95% confidence and conclude that the means are same. The test statistic is 0.86, with p-value of 0.410, and 9 degrees of freedom. Because the p-value is greater than α -level, then we reject H_1 and conclude that $\mu_1 - \mu_2$ is equal to the reference value.

Table 7. Student-t test for Availability Risk Factors.

Students-t test data for CHARF					Students-t test data for SPARF				
	N	Mean	StDev	SEMean		N	Mean	StDev	SEMean
CHARF _{tab}	6	0.8275	0.0382	0.016	SPARF _{tab}	6	0.9642	0.0394	0.016
CHARF _{cal}	6	0.8084	0.0385	0.016	SPARF _{cal}	6	0.9435	0.0388	0.016
Difference = $\mu (\text{CHARF}_{\text{tab}}) - \mu (\text{CHARF}_{\text{cal}})$ Estimate for difference: 0.0191 95% CI for difference: (-0.0309, 0.0692) T-Test of difference = 0 (vs not =): T-Value = 0.86 P-Value = 0.410 DF = 9					Difference = $\mu (\text{SPARF}_{\text{tab}}) - \mu (\text{SPARF}_{\text{cal}})$ Estimate for difference: 0.0207 95% CI for difference: (-0.0304, 0.0718) T-Test of difference = 0 (vs not =): T-Value = 0.92 P-Value = 0.384 DF = 9				

ii). For SPARF_{tab} (using metrics) versus SPARF_{cal} (using the regression equation) two-tailed Students-t test: **H_0 (the null hypothesis):** There is no significant difference between SPARF_{tab} and SPARF_{cal}. **H_1 (the alternative hypothesis):** There is a significant difference between SPARF_{tab} and SPARF_{cal}.

$H_0: \mu_1 - \mu_2 = \delta_0$ versus $H_1: \mu_1 - \mu_2 \neq \delta_0$, where μ_1 and μ_2 are the sample means and δ_0 is the hypothesized difference (zero) between the two sample means. Two-sample T test SPARF_{tab} vs SPARF_{cal}

Mean, StDev and SEMean have been calculated and results are shown in table 7. For the given sample, a 95% confidence interval is: (-0.0304, 0.0718) which includes zero. The reference value of 0 is within the confidence interval, so we can accept H_0 with 95% confidence and conclude that the means are same. The test statistic is 0.92, with p-value of 0.384, and 9 degrees of freedom. Because the p-value is greater than α -level, then we reject H_1 and conclude that $\mu_1 - \mu_2$ is equal to the reference value.

9 Significance

Development of the methodology is based on the suggestions of a report of the best practices and metrics teams [12]. Some suggestions have been extracted: first to assess the risk to which the information may be exposed with respect to security requirements confidentiality, integrity, availability, and privacy. Second establish threshold for those risks and third to identify implement security strategies, policies and controls to mitigate known risks and maintain it at acceptable levels. Developed methodology can be used to implement the suggestions perfectly. Based on availability risk factor it is easy to decide acceptable level of risk. Further, dynamic behavior of classes may help to derive root causes, which may lead to software failure or make software insecure.

The focus of the methodology is quantification of availability risk in design stage. With the help of quantitative results, it becomes easy to find out most risky class hierarchy and most secure class design. As availability risk factor decreases, application becomes more secure. In these versions of class hierarchies, second-class hierarchy is found to be most risky as compare to other versions of class hierarchy. Identified set of sensitive classes and risky classes may use for further analysis or for improvements during design stage. There is possibility of improve methodology. The methodology, it is cable enough to quantify availability risk of class design including algorithm, state transitions, and metrics.

10 Conclusions

Availability is one of the important security requirements. It becomes essential in real-time systems. Degree of quality of applications such as e-commerce, e-banking etc. is highly affected by availability of services for these time critical systems. In this paper, a methodology has been proposed to assess the availability risk at design level using methods and classes. Session methods play important role to maintain availability of services. Numerical results shown in the work supports the claim of acceptability of the proposed methodology to assess availability risk of class hierarchy. In the next phase of the research, we will implement the methodology on large scale of industrial data. Using methodology class movement graph may be developed, and the graph will depict the dynamic behavior of class.

References

- [1] RAP and ComponentBee, <http://www.vtt.fi/proj/cosi/index.jsp>
- [2] Evesti, A., Niemela, E., Henttonen, K., Palviainen, M.: A Tool Chain for Quality-driven Software Architecting. In: IEEE International Software Product Line Conference (2008)
- [3] Madan, B.B., Goševa-Popstojanova, K., Vaidyanathan, K., Trivedi, K.S.: A method for modeling and quantifying the security attributes of intrusion tolerant systems. An International Journal of Performance Evaluation 56, 167–186 (2004)
- [4] Chandra, S., Khan, R.A.: Availability State Transition Model. ACM SIGSOFT Software Engineering Notes 36(3), 1–3 (2011)
- [5] Leangsuksun, C., Shen, L., Liu, T., Song, H., Scott, S.L.: Dependability Prediction of High Availability OSCAR Cluster Server. In: The 2003 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2003), Las Vegas, Nevada, USA, June 23-26 (2003)
- [6] Muppala, J., Ciardo, G., Trivedi, K.S.: Stochastic Reward Nets for Reliability Prediction. Communications in Reliability, Maintainability and Serviceability: An International Journal published by SAE International 1(2), 9–20 (1994)
- [7] Jansen, W.: Directions in Security Metrics Research. National Institute of standards and technology, NISTR 7564 (March 2009)
- [8] Deng, Y., Wang, J., Tsai, J.J.P.: Formal Analysis of Software Security System Architectures. In: Proceedings of International Symposium on Autonomous Decentralized Systems, Dallas, TX, USA, March 26-28, pp. 426–434 (2001)

- [9] Chandra, S., Khan, R.A., Agrawal, A.: Software Security Factors in Design Phase. In: Prasad, S.K., Routray, S., Khurana, R., Sahni, S. (eds.) ICISTM 2009. CCIS, vol. 31, pp. 339–340. Springer, Heidelberg (2009)
- [10] Chandra, S., Khan, R.A.: Software Security Metric Identification Framework (SSM). In: Proceedings of International Conference on Advances in Computing, Communication and Control (ICAC3 2009), Mumbai, Maharashtra, India, January 23-24, pp. 725–731. ACM (2009)
- [11] Sabelfeld, A., Myers, A.C.: Language-Based Information-Flow Security. IEEE Journal on Selected Areas in Communications, special issue on Formal Methods for Security 21(1), 5–19 (2003)
- [12] Corporate Information Security Working Group, Report of the Best Practices and Metrics Teams Subcommittee on Technology, Information Policy, Intergovernmental Relations and the Census Government Reform Committee, United States House of Representative, November 17 (2004) (Revised January 10, 2005)
- [13] Mustafa, K., Khan, R.A.: Quality Metric Development Framework(qMDF). Journal of Computer Science 1(3), 437–444 (2005)
- [14] Bansiya, J., Davis, C.G.: A Hierarchical Model for Object-Oriented Design Quality Assessment. IEEE Transaction on Software Engineering 28(1), 4–17 (2002)
- [15] More Never Again IV, The availability digest (February 2010),
http://www.availabilitydigest.com/public_articles/0502/more_never_agains_4.pdf
- [16] Chandra, S., Khan, R.A.: Confidentiality Checking an Object-Oriented Class Hierarchy. Network Security 2010(3), 16–20 (2010)
- [17] Chandra, S., Khan, R.A.: A Methodology to Check Integrity of a Class Hierarchy. International Journal of Recent Trends in Engineering, Academy 2(4), 83–85 (2009)
- [18] Cardoso: Process control-flow complexity metric: An empirical validation. In: IEEE International Conference on Services Computing, IEEE SCC 2006, 18-22 September, pp. 167–173. IEEE Computer Society (2006)

Performance Analysis of Fast DOA Estimation Using Wavelet Denoising over Rayleigh Fading Channel on MIMO System

A.V. Meenakshi*, R. Kayalvizhi, and S. Asha

ECE, Periyar Maniammai University, Thanjavur
meenu_gow@yahoo.com, {kaya12007,ashasugumar}@gmail.com

Abstract. This paper presents a tool for the analysis, and simulation of direction-of-arrival estimation in wireless mobile communication systems over the Rayleigh fading channel. It reviews three subspace based methods of Direction of arrival estimation algorithms. The standard Multiple Signal Classification (MUSIC) can be obtained from the subspace based methods. In improved MUSIC procedure called Cyclic MUSIC, it can automatically classify the signals as desired and undesired based on the known spectral correlation property and estimate only the desired signal's DOA. The next method is an extension of the Cyclic MUSIC algorithm called Extended Cyclic MUSIC by using an extended array data vector. By exploiting cyclostationarity, the signal's DOA estimation can be significantly improved. In this paper, the DOA estimation algorithm using the de-noising pre-processing based on time-frequency conversion analysis is proposed, and the performances are analyzed. This is focused on the improvement of DOA estimation at a lower SNR and interference environment. This paper provides a fairly complete image of the performance and statistical efficiency of each of above three methods with QPSK signal model for coherent system.

Keywords: MUSIC, QPSK, DOA, MIMO.

1 Introduction

The goal of direction-of-arrival (DOA) estimation is to use the data received on the downlink at the base-station sensor array to estimate the directions of the signals from the desired mobile users as well as the directions of interference signals. The results of DOA estimation are then used by to adjust the weights of the adaptive beam former. So that the radiated power is maximized towards the desired users, and radiation nulls are placed in the directions of interference signals. Hence, a successful design of an adaptive array depends highly on the choice of the DOA Estimation algorithm which should be highly accurate and robust. Array signal processing has found important applications in diverse fields such as Radar, Sonar, Communications and Seismic explorations. The problem of estimating the DOA of narrow band signals using antenna arrays has been analyzed intensively over fast few years.[1]-[9].

* Corresponding author.

The wavelet denoising is a useful tool for various applications of image processing and acoustic signal processing for noise reduction. There are some trials for DOA estimation by applying the wavelet transform method into several sub bands MUSIC and CYCLIC MUSIC scenarios [6{8]. But they do not consider larger noise bandwidth with interference signal included in processing samples. In this paper, the DOA estimation algorithm using a time-frequency conversion pre-processing method with a signal OBW (Occupied Bandwidth) analysis was proposed for CYCLIC MUSIC and the effectiveness was verified through the simulation. This is focused on the improvement of DOA estimation performance at lower SNR and interference environment. This is in compliance with the radio usage trends of lower power and widening signal bandwidth especially.

This paper is organized as follows. Section 1 presents the narrow band signal model with QPSK signal for coherent system. In section 2 the above mentioned data model is extended to multi path Rayleigh fading channel. Here we describe two-channel models namely coherent and non-coherent frequency selective slow fading channels. Section 3. a ,b and c briefly describes the algorithms we have used. Section 4 deal with MUSIC and Cyclic MUSIC algorithms with proposed extended Cyclic MUSIC method. MUSIC procedures are computationally much simpler than the MLM but they provide less accurate estimate [2]. The popular methods of Direction finding such as MUSIC suffer from various drawbacks such as 1.The total number of signals impinges on the antenna array is less than the total number of receiving antenna array. 2. Inability to resolve the closely spaced signals 3. Need for the knowledge of the existing characteristics such as noise characteristics. Cyclic MUSIC algorithm overcomes the above drawbacks. It exhibits cyclostationarity, which improves the DOA estimation. Extended Cyclic MUSIC shows dramatic improvements than the Cyclic MUSIC of its extended data vector. Finally Section 5 describes the simulation results and performance comparison of all three methods namely MUSIC, Cyclic MUSIC and Extended Cyclic MUSIC. Section 6 concludes the paper.

2 Narrow Band Signal Model

The algorithm starts by constructing a real-life signal model. Consider a number of plane waves from M narrow-band sources impinging from different angles θ_i , $i = 1, 2, \dots, M$, impinging into a uniform linear array (ULA) of N equi-spaced sensors, as shown in Figure 1.

In narrowband array processing, when n signals arrive at an m -element array, the linear data model

$$y(t)=A(\Phi)x(t)+n(t) \quad (1)$$

is commonly used, where the $m*n$ spatial matrix $A=[a_1, a_2, \dots, a_n]$ represents the mixing matrix or the steering matrix. In direction finding problems, we require A to have a known structure, and each column of A corresponds to a single arrival and

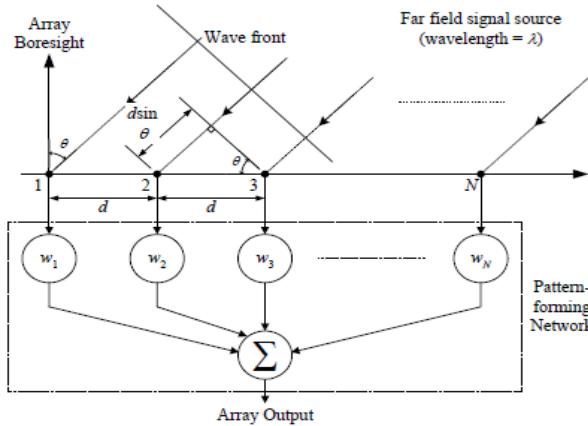


Fig. 1. Uniform linear array antenna

carries a clear bearing. $\mathbf{a}(\Phi)$ is an $N \times 1$ vector referred to as the array response to that source or array steering vector for that direction. It is given by:

$$\mathbf{a}(\Phi) = [1 \ e^{-j\varphi} \dots \ e^{-j(N-1)\varphi}]^T \quad (2)$$

where T is the transposition operator, and φ represents the electrical phase shift from element to element along the array. This can be defined by:

$$\varphi = (2\pi/\lambda)d \cos\theta \quad (3)$$

where d is the element spacing and λ is the wavelength of the received signal.

Due to the mixture of the signals at each antenna, the elements of the $m \times 1$ data vector $\mathbf{y}(t)$ are multicomponent signals. Whereas each source signal $\mathbf{x}(t)$ of the $n \times 1$ signal vector, $\mathbf{x}(t)$ is often a monocomponent signal. $\mathbf{n}(t)$ is an additive noise vector whose elements are modeled as stationary, spatially and temporally white, zero mean complex random processes that are independent of the source signals. That is

$$\begin{aligned} E[\mathbf{n}(t+\Gamma) \mathbf{n}^H(t)] &= \sigma \delta(\tau) \mathbf{I} \\ E[\mathbf{n}(t+\Gamma) \mathbf{n}^T(t)] &= 0, \quad \text{for any } \tau \end{aligned} \quad (4)$$

Where $\delta(\tau)$ is the delta function, \mathbf{I} denotes the identity matrix, σ is the noise power at each antenna element, superscripts H and T , respectively, denote conjugate transpose and transpose and $E(\cdot)$ is the statistical expectation operator.

In (1), it is assumed that the number of receiving antenna element is larger than the number of sources, i.e., $m > n$. Further, matrix \mathbf{A} is full column rank, which implies that the steering vectors corresponding to n different angles of arrival are linearly independent. We further assume that the correlation matrix

$$\mathbf{R}_{yy} = E[\mathbf{y}(t) \mathbf{y}^H(t)] \quad (5)$$

is nonsingular and that the observation period consists of N snapshots with $N > m$.

Under the above assumptions, the correlation matrix is given by

$$R_{yy} = E[(y(t)y^H(t))] = AR_{xx}A^H + \sigma I \quad (6)$$

Where $R_{xx} = E[(x(t)x^H(t))]$ is the source correlation matrix.

Let $\lambda_1 > \lambda_2 > \lambda_3, \dots, \lambda_n = \lambda_{n+1} = \dots = \lambda_m = \sigma$ denote the eigen values of R_{yy} . It is assumed that $\Lambda_i, i=1, 2, 3, \dots, n$ are distinct. The unit norm Eigen vectors associated with the columns of matrix $S = [s_1 s_2 \dots s_n]$, and those corresponding to $\lambda_{n+1} \dots \lambda_m$ make up matrix $G = [g_1 \dots g_{m-n}]$. Since the columns of matrix A and S span the same subspace, then $A^H G = 0$;

In practice R_{yy} is unknown and, therefore, should be estimated from the available data samples $y(i), i=1, 2, 3, \dots, N$. The estimated correlation matrix is given by

$$R_{yy} = 1/N \sum_{n=1}^N (y(t)y^H(t)) \quad (7)$$

Let $\{s_1, s_2, \dots, s_n, \dots, g_{m-n}\}$ denote the unit norm eigen vectors of R_{yy} that are arranged in descending order of the associated eigen values respectively. The statistical properties of the eigen vectors of the sample covariance matrix R_{yy} for the signals modeled as independent processes with additive white Gaussian noise are given in [9].

3 Mimo Signal Model

The MIMO received signal data model is given by

$$y_1(t) = \sum_{k=1}^K \alpha_1(k) x_{mk}(t) + n_1(t) \quad (8)$$

Where $\alpha_1(k) = \alpha(k)a_k(\Phi)$; $a_k(\Phi)$ is the antenna response vector. Where $x_{mk}(t)$ is the signal transmitted by k^{th} user of m^{th} signal, $\alpha_1(k)$ is the fading coefficient for the path connecting user k to the 1^{st} antenna, $n_1(t)$ is circularly symmetric complex Gaussian noise. Here we examine two basic channel models [4]. In the first case, fading process for each user is assumed to be constant across the face of the antenna array and we can associate a DOA to the signal. This is called coherent wave front fading. In coherent wave front fading channel the fading parameters for each user is modeled as $\alpha_1(k) = \alpha(k)a_k(\Phi)$, where $\alpha(k)$ is a constant complex fading parameter across the array, Φ_k is the DOA of the k^{th} user's signal relative to the array geometry, and $a_k(\Phi)$ is the response of the 1^{st} antenna element to a narrow band signal arriving from Φ_k . The signal model is represented in vector form as

$$y_1 = \sum_{k=1}^K \alpha_1(k) g_{mk}(k) + n_1 \quad (9)$$

Here g_{mk} is a vector containing the k^{th} user's m_k^{th} signal.

The second model we consider is non-coherent element- to- element fading channel on which each antenna receives a copy of the transmitted signal with a different fading parameter. In this case, the dependency of the array response on the DOA for each user cannot be separated from the fading process, so that no DOA can be exploited for conventional beam forming.

4 Algorithms Used

A. MUSIC

MUSIC is a method for estimating the individual frequencies of multiple times – harmonic signals. MUSIC is now applied to estimate the arrival angle of the particular user [1],[2].

The structure of the exact covariance matrix with the spatial white noise assumption implies that its spectral decomposition is expressed as

$$R = APA^H = U_S AU^H s + \sigma^2 U_n U^H n \quad (10)$$

Where assuming APA^H to be the full rank, the diagonal matrix U_S contains the M largest Eigen values. Since the Eigen vectors in U_n (the noise Eigen vectors) are orthogonal to A .

$$U_n a(\phi) = 0, \text{ where } \phi \in \{\phi_1, \phi_2, \dots, \phi_m\} \quad (11)$$

To allow for unique DOA estimates, the array is usually assumed to be unambiguous; that is, any collection of N steering vectors corresponding to distinct DOAs Φ_m forms a linearly independent set $\{a_{\phi_1}, \dots, a_{\phi_m}\}$. If $a(\cdot)$ satisfies these conditions and P has full rank, then APA^H is also full rank. The above equation is very helpful to locate the DOAs in accurate manner.

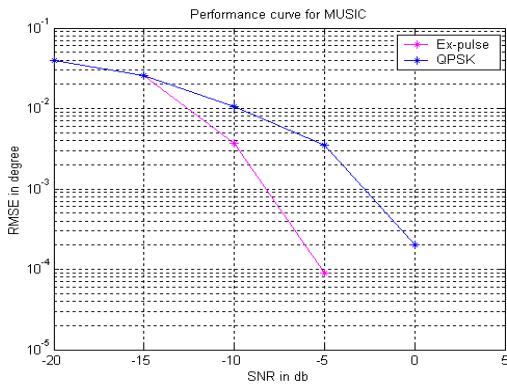
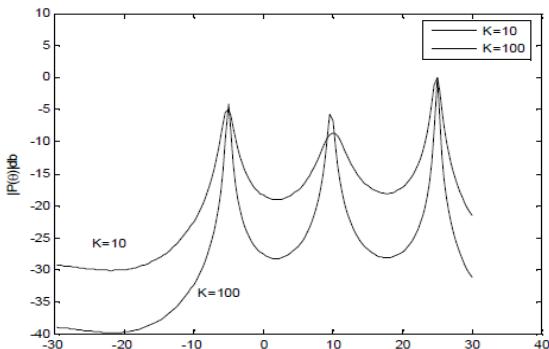
Let $\{s_1, \dots, s_n, g_1, \dots, g_{m-n}\}$ denote a unit norm eigenvectors of R , arranged in the descending order of the associated Eigen values, and let \hat{S} and \hat{G} denote the matrices S and G made of $\{s_i\}$ and $\{g_i\}$ respectively. The Eigen vectors are separated in to the signal and noise Eigen vectors. The orthogonal projector onto the noise subspace is estimated. And the MUSIC ‘spatial spectrum’ is then defined as

$$f(\phi) = \left[a^*(\phi) \hat{G} \hat{G}^* a(\phi) \right] \quad (12)$$

$$f(\phi) = \left[a^*(\phi) [I - S \hat{S}^*] a(\phi) \right] \quad (13)$$

The MUSIC estimates of $\{\Phi_i\}$ are obtained by picking the n values of Φ for which $f(\Phi)$ is minimized.

To conclude, for uncorrelated signals, MUSIC estimator has an excellent performance for reasonably large values of N , m and SNR. If the signals are highly correlated, then the MUSIC estimator may be very inefficient even for large values of N , m , and SNR.

**Fig. 2.** Performance comparison of MUSIC**Fig. 3.** Spectrum of MUSIC for two snapshots

B. Cyclic MUSIC

We assume that m_a sources emit cyclostationary signals with cycle frequency α ($m_a \leq m$). In the following, we consider that $x(t)$ contains only the m_a signals that exhibit cycle frequency α , and all of the remaining $m-m_a$ signals that have not the cycle frequency α .

Cyclic autocorrelation matrix and cyclic conjugate autocorrelation matrix at cycle frequency α for some lag parameter τ are then nonzero and can be estimated by

$$R_{yy\alpha}(\tau) = \sum_{n=1}^N y(t_n + \tau/2) y^H(t_n - \tau/2) e^{-j2\pi\alpha t_n} \quad (14)$$

$$R_{yy\alpha}^*(\tau) = \sum_{n=1}^N y(t_n + \tau/2) y^T(t_n - \tau/2) e^{-j2\pi\alpha t_n} \quad (15)$$

where N is the number of samples.

Contrary to the covariance matrix exploited by the MUSIC algorithm [1], the Cyclic MUSIC method [8] is generally not hermitian. Then, instead of using the Eigen Value decomposition (EVD), Cyclic MUSIC uses the Singular value decomposition (SVD) of the cyclic correlation matrix. For finite number of time samples, the algorithm can be implemented as follows:

- Estimate the matrix $R_{yy}^a(\tau)$ by using (15) or $R_{yy}^{a*}(\tau)$ by using (16).
- Compute SVD

$$[\mathbf{U}_s \quad \mathbf{U}_n] \begin{bmatrix} \Sigma_s & \mathbf{0} \\ \mathbf{0} & \Sigma_n \end{bmatrix} [\mathbf{V}_s \quad \mathbf{V}_n]^H \quad (16)$$

Where $[\mathbf{U}_s \quad \mathbf{U}_n]$ and $[\mathbf{V}_s \quad \mathbf{V}_n]$ are unitary, and the diagonal elements of the diagonal matrices Σ_s and Σ_n are arranged in the decreasing order. Σ_n tends to zero as the number of time samples becomes large.

- Find the minima of $\|\mathbf{U}_n^H \mathbf{a}(\Phi)\|^2$ or the max of $\|\mathbf{U}_s^H \mathbf{a}(\Phi)\|^2$

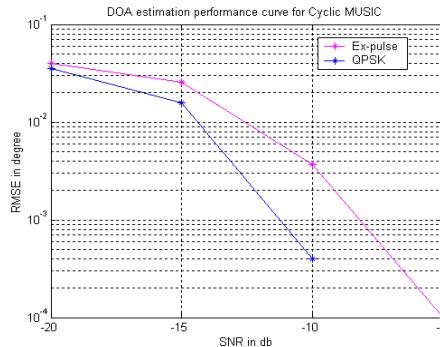


Fig. 4. Performance comparison of Cyclic MUSIC

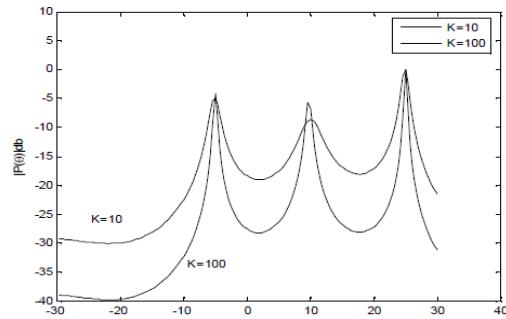


Fig. 5. Spectrum of Cyclic MUSIC for two snapshots

C. Extended Cyclic MUSIC

Here we give an extension of the conventional model in order to exploit the cyclostationarity of the incoming signals. We form the following extended data vector

$$\mathbf{Y}_{ce}(t) = [y(t); y^*(t)] \quad (17)$$

We can estimate the cyclic correlation matrix for the extended data model as

$$\mathbf{R}_{ce} = \sum_{n=1}^N I^\alpha_{2m}(t_n) \mathbf{Y}_{ce} (tn + \tau/2) \mathbf{Y}_{ce}^H (tn - \tau/2) \quad (18)$$

where the time dependent matrix

$$\mathbf{I}_{2m}(t) = \begin{bmatrix} I_M e^{-j2\pi t} & 0 \\ 0 & I_M e^{+j2\pi t} \end{bmatrix}$$

I_M is the M -dimensional identity matrix.

By computing the SVD of \mathbf{R}_{ce} similarly to the Cyclic MUSIC algorithm, the spatial spectrum of the Extended Cyclic MUSIC method is given by

$$p(\phi) = \frac{1}{a^H(\phi) U_n a(\phi) - \|a^T(\phi) U_n a(\phi)\|} \quad (19)$$

But this method is dedicated to cyclostationary signals that have no particular limitation.

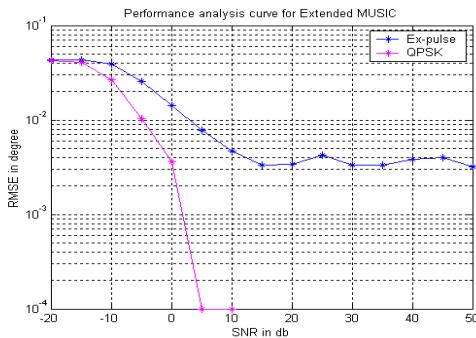


Fig. 6. Performance comparison of Extended Cyclic MUSIC

5 New Approach of DOA Estimation

A signal subspace based DOA estimation performance is affected by the two factors of an accurate array manifold modeling and a spatial covariance matrix of a received array signal. A higher SNR signal for a target source is required for an accurate estimation from finite received signal samples. But the DOA estimation performance is limited by the lower SNR from interference signals and environmental noise. For the

performance improvement of DOA estimation, this paper proposed a pre-processing technique of time-frequency conversion methodology for signal filtering. This method includes a time-frequency conversion technique with a signal OBW (Occupied Bandwidth) measurement based on wavelet de-noising method as shown in Fig. 7. This is a DOA method for SNR improvement based on time-frequency conversion approach. The improvement of a DOA estimation performance was verified by the simulation.

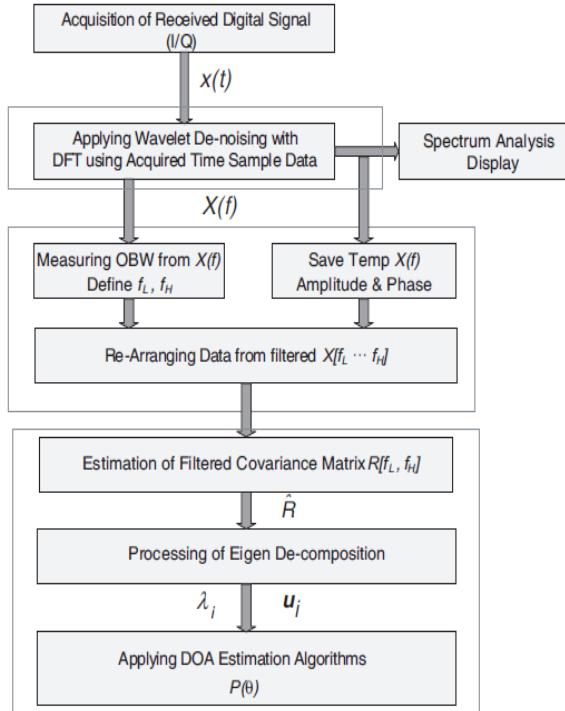


Fig. 7. Proposed Model for MUSIC

This is proposed to overcome the limitation of existing DOA estimation techniques based on only time domain analysis. The more effective estimation is expected by the improvement of SNR from the proposed pre-processing techniques of frequency domain analysis. The proposed method collects a time sampled signal $y(t)$ from an array antenna as shown in Figure 8. The upper and lower 99% - OBW limits f_L and f_H of a signal are determined from $y(f)$, which is the DFT result of a received signal $y(t)$. The filtered covariance matrix $R[f_L : : : f_H]$ can be obtained from the estimated signal energy, $y[f_L : : : f_H]$ with an improved SNR. And the more exact OBW measurement is expected through the proposed wavelet de-noising method based on time-frequency analysis. The proposed OBW limits are defined as following measurement concepts of Equations (9) ~ (12). This process can effectively eliminate small interference noises from the target signal streams by the frequency domain analysis [15, 16].

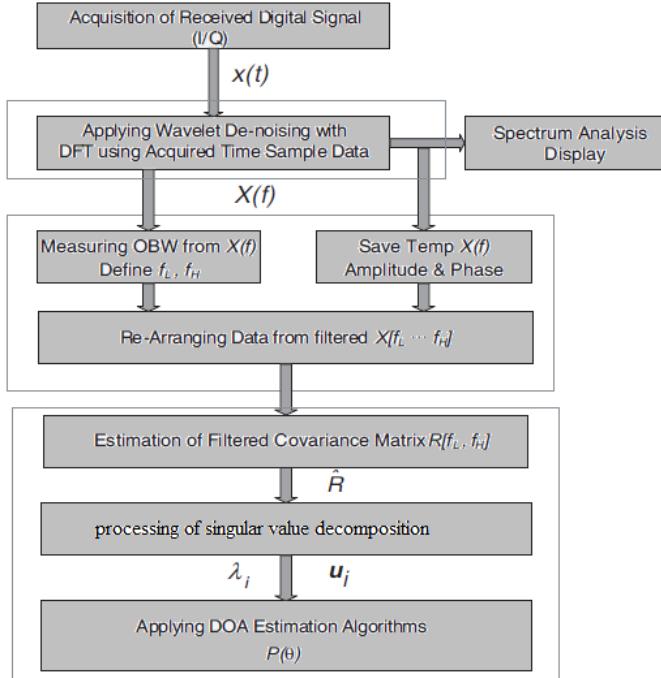


Fig. 8. Proposed Model for CYCLIC MUSIC and Extended Cyclic MUSIC

Where P_x is a power of each spectrum frequency elements $\{f_1, \dots, f_N\}$. The 99% OBW is calculated from the upper limit f_H and the lower limit f_L of 0.5% OBW point from each spectrum boundary.

$$P_{rel} = \sum_{i=f_1}^{f_N} P_y(i) \quad (20)$$

$$\Delta P_{\beta/2} = P_{rel} \times \beta / 2 [\%] \quad (21)$$

($\beta = 1$ for 99% OBW analysis)

$$f_L = \arg \min_{f_L} \left\| \sum_{f_L=f_1}^{f_N} P_y(f_L) - \Delta P_{\beta/2} \right\| \quad (22)$$

$$f_H = \arg \min_{f_H} \left\| \sum_{f_H=f_1}^{f_N} P_y(f_H) - \Delta P_{\beta/2} \right\| \quad (23)$$

An improved DOA estimation is expected from the filtered covariance matrix and Eigen-decomposition processing at particularly low SNR signal conditions. By the proposed pre-processing, it can effectively reject adjacent interferences at low SNR conditions. Moreover, it can acquire the signal spectrum with an improved DOA estimation spectrum simultaneously without additional computation. This improved signal spectrum is important results for radio surveillance procedure. The signal denoising is achieved by the discrete wavelet transform-based thresholding to the resulting coefficients, and suppressing those coefficients smaller than certain amplitude. An appropriate transform can project a signal to the domain where the signal energy is concentrated in a small number of coefficients. The proposed Wavelet de-noising process get a de-noised version of input signal obtained by thresholding the wavelet coefficients. In this paper, the wavelet procedure applied the heuristic soft thresholding of wavelet decomposition at level one. This de-noising processing model is depicted as following simple model.

$$S(n) = f(n) + \sigma e(n), \quad n=0, \dots, N-1 \quad (24)$$

In this simplest model, $e(n)$ is a Gaussian white noise of independent normal random variable $N(0; 1)$ and the noise level is supposed to be equal to 1. Using this model, it follows the objectives of noise removal by reconstruct the original signal f . It can be assumed that the higher coefficients are result from the signal and the lower coefficients are result from the noise. The noise eliminated signal is obtained by transforming back into the original time domain using these wavelet coefficients.

6 Simulation and Performance Comparison

Data Specification

Signal specification:

Data Model: QPSK and FM signal

Input bit duration T	= 0.5μsec
Sampling interval t	= T/10;

Antenna Array Model:

Type: Uniform Linear array antenna

No. of array Elements	N	= 16
Free space velocity	c	= 3×10^8
Centre frequency	fc	= 2.4GHz
Wavelength	λ	= c / fc
Inter element Spacing	d	= $\lambda/2$
Angle of arrival in degrees	θ	= -5 to 20

Channel model: Rayleigh fading

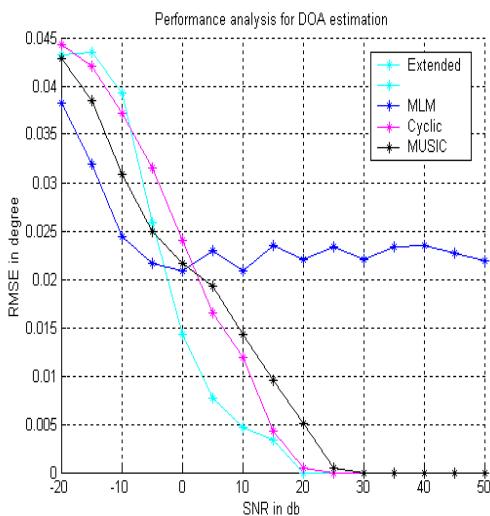


Fig. 9. Performance comparison Of all

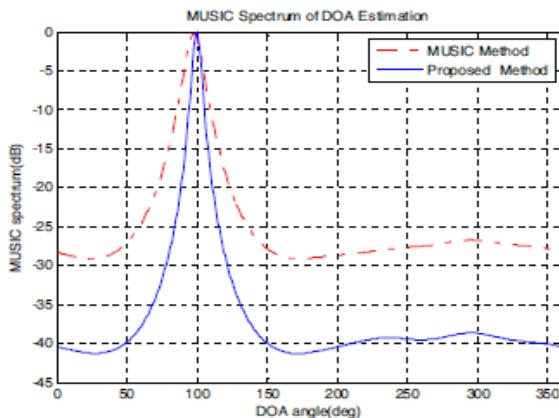


Fig. 10. DOA estimation spectrum Methods with fading channel

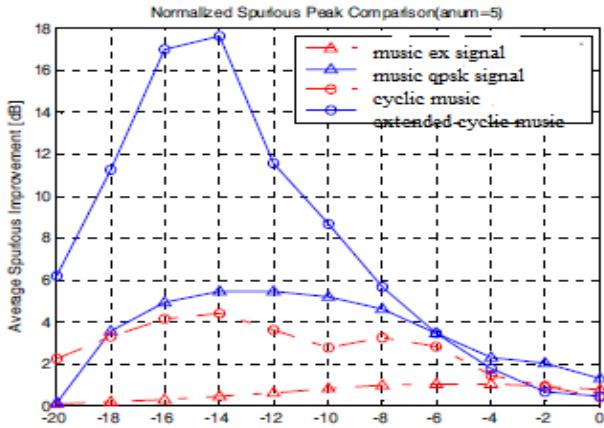


Fig. 11. Comparison of spurious peak

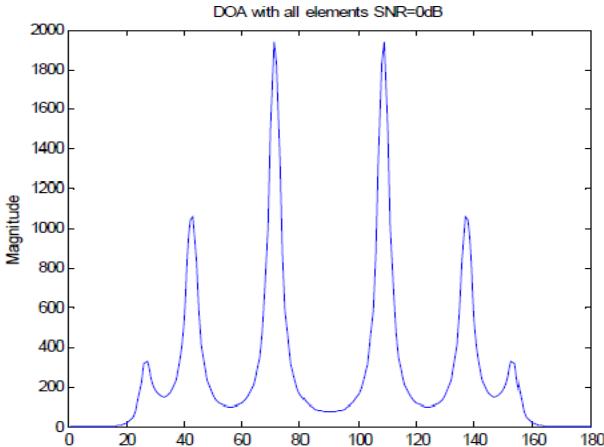


Fig. 12. Direction of Arrival

In this section, we present some simulation results, to show the behavior of the three methods and to compare the performance with the analytically obtained Root Mean Squared Error (RMSE). We consider here a linear uniformly spaced array with 16-antenna elements spaced $\lambda/2$ apart. Incoming QPSK signals are generated with additive white Gaussian noise with signal to noise ratio 10dB, the bit rate of the QPSK signal of interest is 2Mb/s and other QPSK modulated signals with data rate 1Mb/s are considered as interference. MUSIC is simulated using the specified parameters. The Cyclic MUSIC algorithm is also simulated with some cyclic frequency of 4MHz and some lag parameter of 2. One QPSK signals arrived at 20 degree and an interferer at 5 degree DOA. RMSE Vs SNR plots and there corresponding spectrum for the three methods as shown in figure 2, 3,4, 5 and 6. This section presents Mont Carlo computer simulation results to illustrate the performance of the proposed algorithms

for synchronous system. Each Monte Carlo experiment involved a total of 1000 runs, and each estimation algorithm is presented with exactly the same data intern. It is interesting to note that QPSK signal performs better than FM signal. So that bandwidth requirement is as low as possible for QPSK signals as FM signals. It is interesting to note that the conventional MUSIC would require more data samples than Cyclic MUSIC and extended cyclic MUSIC to achieve the same RMSE. Form the simulation results, the proposed method improves the DOA estimation performance of accuracy and spurious peak of spatial spectrum especially for lower SNR signals. Figure 9, 10, 11 and 12 show the comparison results of DOA estimation performance for low SNR signals with interference and noise. The DOA estimation performance was compared by the spurious peak of DOA estimation spectrums which increase a measurement ambiguity and probability of successful DOA estimation. The proposed method improves the spurious peak characteristic more than 10 dB at less than -10 dB SNR signal condition by applying MUSIC and cyclic MUSIC estimation.

7 Conclusion

Unlike MUSIC, Cyclic and Extended Cyclic MUSIC does not suffer from the drawback of requiring a higher number of antenna elements than sources. Good signal selective capability and high resolution is achieved in Extended Cyclic MUSIC than Cyclic MUSIC. This algorithm exploits cyclostationarity, which improves the signal Direction of Arrival estimation over the Rayleigh fading channel in MIMO wireless systems. Extended Cyclic MUSIC shows the dramatic improvements than the Cyclic MUSIC. Extended Cyclic MUSIC method allows perfect selection of the Signal of Interests and ignores the interference signal. Therefore the proposed method shows an improved ability of DOA resolution and estimation error at the noise and interference conditions. These are the measurement limits at on-air environment.

References

- [1] Lee Swindlehurst, A., member, IEEE, Stoica, P., Fellow, IEEE: Maximum likelihood methods in radar signal processing (February 1998)
- [2] Krim, A., Viberg, M.: Two decades of array signal processing Research. IEEE Signal Processing Magazine (July 1996)
- [3] McCloud, M.L., Varanasi, K.: Beamforming, Diversity, and Interference Rejection for Multiuser communication over fading channels with a receive antenna array. IEEE Trans. on Comm. 51 (January 2003)
- [4] Kumaresan, R., Tufts, D.W.: Estimating the angles arrival of multiple plane waves. IEEE Trans. Aerosp. Electron. Syst. AES-19 (January 1983)
- [5] Sharman, K.C., Durrani, T.S.: Maximum Likelihood parameter estimation by simulated annealing. In: Proc. IEEE Int. Conf. Acoust. Speech Processing (April 1988)
- [6] Miller, M., Fuhrmann, D.: Maximum Likelihood Direction of Arrival Estimation for multiple narrow band signals in noise. In: Proc. 1987 Conf. Inform. Sciences, Syst., pp. 710–712 (March 1987)
- [7] Schell, S.V., member, IEEE: Performance analysis of the Cyclic MUSIC method of Direction Estimation for Cyclostationary Signals. Trans. (November 1994)

- [8] Stoica, P., Sharman, K.C.: A novel eigenanalysis method for direction estimation. In: Proc. Inst. Elec. Eng., pt. (February 1990)
- [9] Schmidt, R.O.: Multiple emitter location and signal (August 2000)
- [10] Pesavento, M., Gershman, A.B., Wong, K.M.: Direction of arrival estimation in partly calibrated time-varying sensor arrays. In: Proc ICASSP, Salt Lake City, UT, pp. 3005–3008 (May 2001)
- [11] Pesavento, M., Gershman, A.B., Wong, K.M.: Direction finding in partly-calibrated sensor arrays composed of multiple subarrays. IEEE Trans. Signal Processing 50, 2103–2115 (2002)
- [12] See, C.M.S., Gershman, A.B.: Subspace-based direction finding in partly calibrated arrays of arbitrary geometry. In: Proc. ICASSP, pp. 3013–3016 (April 2002)
- [13] Pesavento, M., Gershman, A.B., Wong, K.M.: On uniqueness of direction of arrival estimates using rank reduction estimator (RARE). In: Proc. ICASSP, Orlando, FL, pp. 3021–3024 (April 2002)
- [14] Pesavento, M., Gershman, A.B., Wong, K.M., Böhme, J.F.: Direction finding in partly calibrated arrays composed of nonidentical subarrays: A computationally efficient algorithm for the RARE estimator. In: Proc. IEEE Statist. Signal Process. Workshop, Singapore
- [15] Boubaker, N., Letief, K.B., Much, R.D.: Performance of BLAST over frequency-selective wireless Communication channels. IEEE Trans. on Communications 50(2), 196–199 (2002)
- [16] Choi, J.: Beamforming for the multiuser detection with decorrelator in synchronous CDMA systems: Approaches and performance analysis. IEEE Signal Processing 60, 195–211 (1997)
- [17] Sathish, R., Anand, G.V.: Spatial wavelet packet denoising for improved DOA estimation. In: Proceedings of the 14th IEEE Signal Processing Society Workshop on Machine Learning for Signal Process., pp. 745–754 (October 2004)
- [18] ITU-R SM.1794, Wideband Instantaneous Bandwidth Spectrum Monitoring Systems, International Telecommunication Union (January 2007)

DAGITIZER – A Tool to Generate Directed Acyclic Graph through Randomizer to Model Scheduling in Grid Computing

D.I. George Amalarethinam¹ and P. Muthulakshmi²

¹ Department of Computer Science, Jamal Mohamed College, Trichirappalli, Tamil Nadu, India
di_george@jmc.edu

² Department of Computer Science, SRM University, Tamil Nadu, India
rmlakshmi2004@yahoo.co.in

Abstract. Scheduling is absolutely the resource management. A group of interdependent jobs/tasks forms the workflow application and scheduling is to map the jobs/tasks on to the collection of heterogeneous resources available in a massive geographic spread. Most complicated applications consist of interdependent jobs that coordinate to solve a problem. The completion of a particular job is the criterion function essentially to meet in order to start the execution of those jobs that depend upon it [1]. This kind of workflow application may be represented in the form of a Directed Acyclic Graph (DAG). Grid Workflow is such an application and is modeled by DAG. This paper proposes a tool that generates Directed Acyclic Graph through Randomizer, which helps in solving the scheduling problem among the dependent tasks by considering the parameters, computation cost (COMPCost) of the nodes and the communication cost (COMMCost) between the nodes. This tool is developed in Java, considering it as a platform independent and web authoring application developer. The task dependencies are made random, the computation cost and communication cost are also randomly allocated by the randomizer. The output generated by the tool includes (i) a visual component of an actual DAG,(ii) a table with complete information on task, its predecessors, COMPCost, COMMCost and (iii) detailed description about the number of levels, number of tasks at each level, identification of a tasks in a level and relationship between the nodes.

Keywords: Grid Workflow, Scheduling, Directed Acyclic Graph, Randomizer, Communication cost, Computation cost.

1 Introduction

The evolution of distributed computing results in a new technology called Grid computing. Grid Computing Environment is a large scale heterogeneous environment and is deployed by applications from various fields such as medicine, bioinformatics, image processing, biotechnology, etc. The aim of grid computing is to achieve high performance through effective and efficient utilization of resources available in the grid infrastructure. Grid is a collection of resources, includes applications, computing, data storage, machines, and networks. Furthermore, grid computing enables careful selection, sharing, coordination and integration among the resources. The aim of such computing technology will be met only when the resources are effectively allocated with appropriate jobs. High performance will be achieved only when the scheduling

of group of dependent tasks is intensively done. In grids, users may face hundreds of thousands of computers to utilize. It is impossible for anyone to manually assign jobs to computing resources in grids. An effective scheduling aims at minimum turnaround time. To show such scheduling through an illustration, the best ever known possibility is Directed Acyclic Graph. Grid Workflow scheduling is replicated through Directed Acyclic Graph, and the paper focuses on the generation of DAG, which is made in a much automated way through randomizer based on the number of tasks involved.

A directed acyclic graph is a directed graph with no directed cycles and is formed by a collection of vertices and directed edges, each edge connecting one vertex to another, such that there is no way to start at some vertex v and follow a sequence of edges that eventually loops back to v again[2][3][4].

A directed acyclic graph (DAG) can be used to model a parallel program where the tasks (nodes) represent the vertices and the edges (lines connecting the vertices) represent the dependencies between the tasks. DAG workflow modeling is to describe the workflow as a series of subtasks. The dependency between the subtasks generates the directed acyclic graph represented as $G=(V,E,W)$, in which $V(v_1,v_2,v_3,\dots,v_n)$ represents the set of all nodes (all tasks in the workflow), $E(e_1,e_2,e_3,\dots,e_n)$ represents the set of all edges that connect the nodes in V (shows the dependency between nodes in V) and $W(w_1,w_2,\dots,w_n)$ represents the weights of all the nodes (the computation cost of all the nodes, usually integer). Each edge e_i ($1 \leq i \leq n$) is noted by (v_i, v_j) , where $i \neq j$ and associated with each edge is an integer called the communication cost (between task and its successor). Associated with each edge (v_i, v_j) , there can be a value $d_{i,j}$ that represents the amount of data to be transmitted from task v_i to task v_j [5] [6]. Any node that has no predecessor is an entry node (root node). Similarly, the node which has no successor is called the exit node (leaf node).

Scheduling decisions in dynamic scheduling algorithms are made at run time [7]. The objective of dynamic scheduling algorithms includes not only creating high quality task schedules, but also minimizing the run time scheduling overheads [8] [9]. The proposed tool generates arbitrary DAGs with a required number of tasks, used to test the task scheduling algorithms developed by researchers.

2 Applications of DAG

A directed graph may be used to represent a network of processing elements called as nodes; the data transmission is done through edges; data enters a node through the incoming edges and goes out of the node through the outgoing edges. DAG finds part in (i) electronic circuit design; (ii) Bayesian network that represents a system of probabilistic events as nodes in a directed acyclic graph, where the chance of an event may be calculated from the likelihoods of its predecessors in the DAG; (iii) Dataflow, with which programming languages describe structures of values that are related to each other by a directed acyclic graph; (iv) Compilers, particularly in linear execution codes, DAGs are used to describe the inputs and outputs of each of the arithmetic operations performed; (v) Software designing; (vi) Information flow modeling in networks; (vii) Numerical methods like Gaussian Elimination or Fast Fourier Transformations where the iterations of the loop can be represented by an node in DAG [10]; (viii) to find the order in which files should be compiled; (ix) DAG is used to represent the dependency between the cells of a spreadsheet and more than the above said, algorithms represented through DAG are simpler rather doing it with ordinary graphs.

3 Related Works

The idea of developing a tool that could generate a DAG had been derived by (i) PYRROS, a tool developed by Yang and Gerasoulis [11] is a compile time scheduling and a code generator, which is consisted with a task graph language with an interface to C language. The tool could use only a particular algorithm and is not exclusively a DAG generator; (ii) The application specification tool in PARSA (software developed for automatic scheduling and partitioning of sequential user programs) is accepted with a sequential program written in the SISAL functional language and converted into a DAG and is represented in textual form by an acyclic graphical language called IF1 (Intermediate Form 1) [12]; (iii) The idea proposed by Y. K. Kwok and I. Ahmad [10] under scheduling arbitrary DAGs without communication stated that nodes in the DAG can be assigned priorities randomly; (iv) The node and edge weights are usually obtained by estimation at compile time.[13]; (v) Hu's Algorithm [14] for Tree Structured DAGs, where in-tree structured DAGs had been proposed with unit computations and without communications and the number of processors is assumed to be limited.

4 Architecture and Operations of the Tool

The proposed tool efficiently generates a DAG for a Random Workflow. In random workflow, dependency and number of successors of a task are generated randomly.

The operational blocks that involved in building the tool very effective are shown in Figure 1 and are listed as,

1. Resource Portal
2. Fragmentor
3. Mapper
4. Filter
5. Ascriber
6. Evaluator
7. Finalizer

The randomization involved in almost all the functional blocks and is done by using Java's Random class. The Random class is a generator of pseudo random numbers and they are uniformly distributed sequences. The way of setting different seed values will reduce the possibility of getting repeated sequences [15].

In Statistics, a type of probability distribution in which all outcomes are equally likely is called uniform distribution.

It is a family of probability distributions such that for each member of the family, all intervals of the same length on the distribution's support are equally probable. The support is defined by the two parameters, a and b , which are its minimum and maximum values respectively. The distribution is often abbreviated $U(a,b)$. The uniform distribution on the interval $[a,b]$ is the maximum entropy distribution among all continuous distributions which are supported in the interval $[a, b]$ (which means that the probability density is 0 outside of the interval). The Probability Density Function (pdf) of Uniform distribution function is given by,

$$P(x) = 1/(b-a); a \leq x \leq b \quad (1)$$

where, $P(x)$ is the pdf of the random variable x and a, b are maximum and minimum values of the interval respectively.

4.1 Tool Architecture

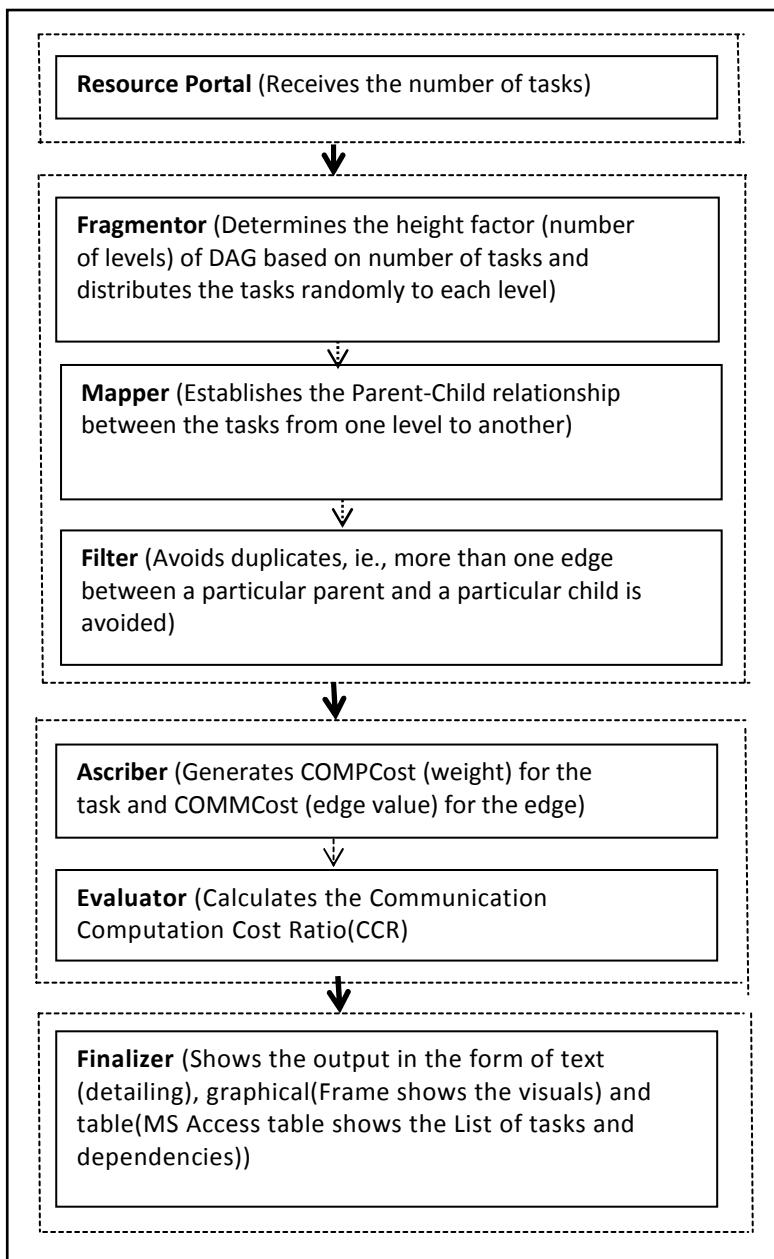


Fig. 1. Architecture of DAGITIZER

4.1.1 Resource Portal

The tool starts its functioning in this block, where the expected inputs are (i) number of tasks (nv) involved in constructing the DAG, (ii) the maximum COMPCost(MCOMPCost), which is the upper seed for the randomizer to generate the COMPCost for individual node (between 1 and upper seed), (iii) the maximum COMMCost(MCOMMCost), which is the upper seed for generating the COMMCost of each edge through randomization.

4.1.2 Fragmentor

The Height Factor (HF) is generated by this block, the number of levels that the DAG has to distribute the tasks (nodes) and the maximum nodes that a level can hold (**Maximum Nodes Per Level by Tool**).

Also, there is a possibility of specifying the maximum nodes (**Maximum Nodes Per Level by User**) that a level should hold, if this is the case as that of the one generated by the tool (MNPLT) itself, then distribution of nodes will be done taking MNPLT as upper seed; otherwise the nodes will be distributed among the levels based on the user specification (MNPLU) taking MNPLU as upper seed.

Now, the nodes in each level is distributed between 1 and upper seed by the randomizer. And the HF is revised according to MNPLU, when MNPLT and MNPLU are not equal. The randomizer then generates the number of nodes to be shared by each level and this will not be uniform for all the levels.

4.1.3 Mapper

This block takes the responsibility of random assignment of number of successors of each node. In which, each node is identified in a particular level as v_{ij} , where 'v' is the vertex at level 'i' and 'j' is the position of the node in the particular level; $1 \leq j \leq (\text{number of nodes in level } i)$. A node can choose its successors randomly from the nodes in the successive levels, i.e., symbolically the parent – child relationship is established randomly for count of successors generated by the randomizer. The tool is carefully devised in such a way that no node from the last level of the DAG should have a successor.

4.1.4 Filter

This block is responsible for avoiding duplicates. As the mapping of child task to a parent task is done randomly, there might be a possibility of duplicates by generating the same link between the tasks more than once on different levels. This block eliminates such duplicates and maintains only one link between a pair of nodes/tasks.

4.1.5 Ascriber

The process of randomization is applied in assigning the COMPCost for each node in the DAG, $1 \leq \text{COMPCost} \leq \text{MCOMPCost}$ and COMMCost for edges, $1 \leq \text{COMMCost} \leq \text{MCOMMCost}$. Thereby, the summation all the COMPCost (Total COMPCost) and the summation of all the COMMCost (Total COMMCost) is calculated.

4.1.6 Evaluator

The Communication - Computation Ratio (**CCR**) is calculated in this block. CCR is defined as the ratio between the average communication cost of all the edges and the average computation cost of the nodes.

This can be shown as,

$$CCR = \text{Average}(\sum COMMCost(e_i)) / \text{Average}(\sum COMPCost(v_j)) ; 1 \leq i \leq nv, 1 \leq j \leq ne \quad (2)$$

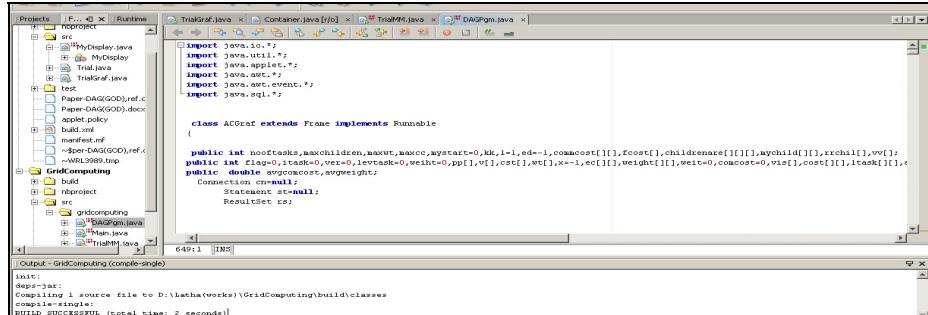
where, $\text{Average}(\sum \text{COMMCost}(e_i))$ is the average of all the edge values, $\text{Average}(\sum \text{COMPCost}(v_j))$ is the average weights of all tasks.

4.1.7 Finalizer

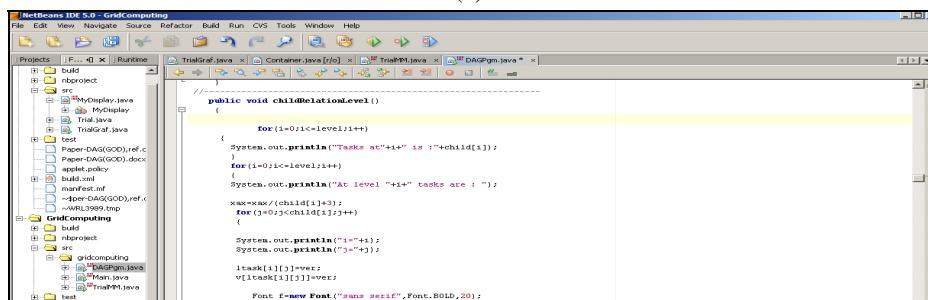
The collection of various forms of output is generated by this block. Outputs are shown in the form of (i) text description, (ii) pictorial visuals and (iii) table (database). NetBeans IDE (Integrated Development Environment) is used to develop the tool in Java.

The text description of the output generated by the tool is shown in the output container (Figure 3).

The graphical view of the output is presented in the Frame [15], which will show the task identification number v_i ; $1 \leq i \leq nv$ in a TextField [15] (a box to hold characters) depicted to be the node, where the Editable property is set to false, the COMPCost of each node is shown on the top of the TextField. The TextFields (nodes) are connected to their respective child(ren)/tasks (also TextFileds containing the task identification number in the successive levels). The connections are shown through lines of various colors having the COMMCost to the middle of the line length and ensured that the values should not overlap each other by adjusting the display to either a bit top or bottom if more than one value shares the same place of displaying the COMMCost.



(a)



(b)

Fig. 2. Parts of Tool's development in source code level

MS-Access is used to show the tabulated results in its table containing information as each record consists of four attributes (i) the task identification number, (ii) child/task, (iii) CommunicationCost(COMMCost between task (representing parent/task) and the child/task), (iv) weight(COMMCost) of the task.

5 Implementation and Results

The simulation results of the program execution are shown below; Figure 2 shows the program (source code) of the tool, Figure 3 shows the text description of the sample output of the program, Figure 4 shows the pictorial view of the description in the Frame [15], and Figure 5 shows the stored table view of MS-Access.

<pre> int: dshape.jar; compile-single; run-single; GOD IS GREAT Enter the number of tasks: 15 Height Factor(approximated to)4 Tasks per level(approximated to)3 Enter the maximum number of children: 3 Enter the maximum Weight of the tasks: 5 Enter the maximum communication cost of the edges weight: 7 Tasks at 0 is :3 Tasks at 1 is :2 Tasks at 2 is :3 Tasks at 3 is :2 Tasks at 4 is :3 Tasks at 5 is :2 At level 0 Tasks are : =>0 =>0</pre>	<p>Weight Generator</p> <pre> weight of task 0 is 2 weight of task 1 is 1 weight of task 2 is 1 weight of task 3 is 5 weight of task 4 is 1 weight of task 5 is 1 weight of task 6 is 5 weight of task 7 is 4 weight of task 8 is 3 weight of task 9 is 1 weight of task 10 is 3 weight of task 11 is 2 weight of task 12 is 3 weight of task 13 is 1 weight of task 14 is 3</pre>
(a)	(b)
<pre> *****After avoiding duplicates***** children of 0 is 7 task:v[ed] is 0 child: pp[ed] is7 weight: wt[v[ed]] is2 Edge No:0 and the cost is 4 children of 0 is 6 task:v[ed] is 0 child: pp[ed] is6 weight: wt[v[ed]] is2 Edge No:1 and the cost is 3 children of 0 is 4 task:v[ed] is 0 child: pp[ed] is4 weight: wt[v[ed]] is2 Edge No:2 and the cost is 6 children of 0 is 3 task:v[ed] is 0 child: pp[ed] is3 weight: wt[v[ed]] is2 Edge No:3 and the cost is 1 children of 0 is 9 task:v[ed] is 0 child: pp[ed] is9</pre>	<p>Summary</p> <pre> ----- Number of tasks =15 Number of edges=54 Total of all WEIGHTS =36 Total of all COMMUNICATION COSTS =211 Average Task Weight =2.0 Average Communication cost =3.0 CCR RATIO:1.5</pre>
(c)	(d)

Fig. 3. Sample output in a text view

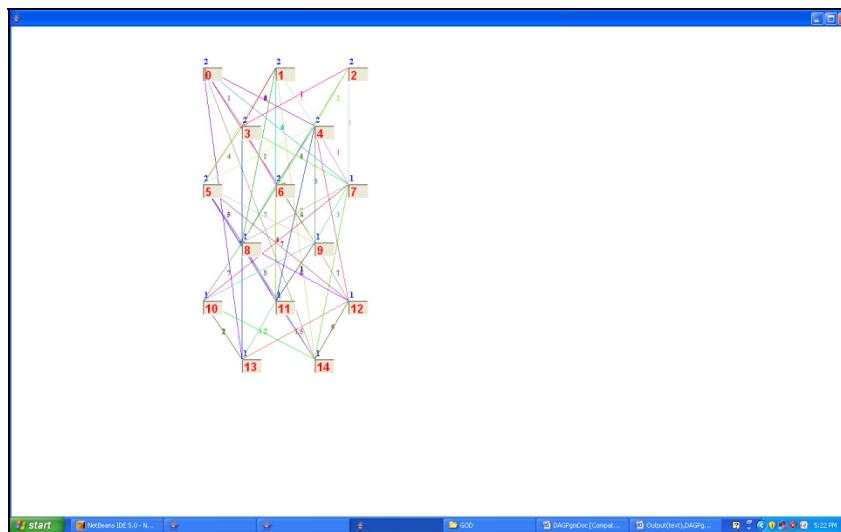


Fig. 4. Pictorial View, showing a Frame with all the nodes and edges along with COMPCost and COMMCost

A screenshot of Microsoft Access showing a table named 'DAObjTask'. The table has columns: Task, Child, Communicate, and Weight. The data is as follows:

	Task	Child	Communicate	Weight
Table1	0	7	4	2
Table2	0	6	3	2
Table2 - Table	0	4	6	2
Table4	0	0	3	1
Table4 - Table	0	5	4	2
Table5	0	13	5	2
Table5 - Table	0	14	7	2
Table5 - Table	1	11	2	1
Table5 - Table	1	5	6	1
DAObjTask	1	3	2	1
DAObjTask	1	4	1	1
DAObjTask	1	14	4	1
DAObjTask	1	8	2	1
DAObjTask	1	6	6	1
DAObjTask	1	2	5	1
DAObjTask	2	8	3	1
DAObjTask	2	4	2	1
DAObjTask	3	3	1	1
DAObjTask	3	7	6	5
DAObjTask	3	13	4	5
DAObjTask	3	5	4	5
DAObjTask	3	6	3	5
DAObjTask	4	8	7	1
DAObjTask	4	9	5	1
DAObjTask	4	11	4	1
DAObjTask	4	12	3	1
DAObjTask	5	4	1	

Fig. 5. View of the table showing task list

6 Time Complexity of the Algorithm

This tool has utilized Random Polynomial Time; An algorithm is said to be of polynomial time if its running time is upper bounded by a polynomial expression in the size of the input for the algorithm, and the time complexity is found to be, $T(n) = O(n^k)$ for some constant k, where $T(n)$ is the time to run the algorithm with n inputs, $O(n^k)$ is the order of upper bound on the value of $T(n)$ [16][17]. Random Polynomial

Time of an algorithm is the polynomial time in the input size and thereof the decisions on the size of outputs of the randomizer.

7 Conclusion

There are large volumes of data flow computation problems and many numerical algorithms where indeterminism is found, and are modeled through DAGs; that scheduling techniques can be applied. The inspiration of such idea made us to devise a tool that could help in efficient and proper scheduling of tasks through the generation of DAG. Grid Computing is a technology where arbitrary participation of resources, tasks are happening and the resource mapping is found to be very challenging. Resource mapping or scheduling is simulated through DAG and in this paper, the goal is achieved for arbitrary participants and fine list of results has been observed by simulating with various inputs. The tool shown in the paper is potentially developed in such a way that it would be giving the expected results for users and makes the users very comfortable with using and analyzing the randomness through the pictorial view rather in the form of text. The tool will be very much helpful to the researchers who are developing task scheduling algorithms for multiprocessor systems and for Grid computing environment.

References

- [1] Lopez, M.M., Heymann, E., Senar, M.A.: Analysis of Dynamic Heuristics for Workflow Scheduling on Grid Systems. In: IEEE Proceedings of The Fifth International Symposium on Parallel and Distributed Computing (2006)
- [2] Christofides, N.: Graph theory: an algorithmic approach, pp. 170–174. Academic Press (1975)
- [3] Thulasiraman, K., Swamy, M.N.S.: Acyclic Directed Graphs. In: Graphs: Theory and Algorithms. John Wiley and Son (1992) ISBN 9780471513568
- [4] Bang-Jensen, J.: 2.1 Acyclic Digraphs, Digraphs: Theory, Algorithms and Applications, 2nd edn. Springer Monographs in Mathematics, pp. 32–34. Springer (2008)
- [5] Hwang, K.: Advanced Computer Architecture: Parallelism, Scalability, Programmability. McGraw-Hill, Inc., New York (1993)
- [6] Topcuoglu, H., Hariri, S., Wu, M.Y.: Performance-Effective and Low Complexity Task Scheduling for Heterogeneous Computing. *IEEE Trans. Parallel Distributed Systems* 13(3), 260–274 (2005)
- [7] Ilavarasan, E., Thambidurai, P., Mahilmannan, R.: Performance Effective Task Scheduling Algorithm For Heterogeneous Computing System. In: Proceedings of the Fourth International Symposium on Parallel and Distributed Computing, France, pp. 28–38 (2005)
- [8] Sih, G.C., Lee, E.A.: A Compile-Time Scheduling Heuristic For Interconnection-Constrained Heterogeneous Processor Architectures. *IEEE Trans. Parallel Distributed Systems* 4(2), 175–187 (1993)
- [9] Kim, J., Rho, J., Lee, J.-O., Ko, M.-C.: CPOC: “Effective Static Task Scheduling For Grid Computing”. In: Proceedings of the 2005 International Conference on High Performance Computing and Communications, Italy, pp. 477–486 (2005)

- [10] Kwok, Y.-K., Ahmad, I.: Static Algorithms for Allocating Directed Task Graphs to Multiprocessors. *ACM Computing Surveys* 31(4) (December 1999)
- [11] Yang, T., Gerasoulis, A.: PYROSS: Static Task Scheduling and Code Generation for Message Passing Multiprocessors. In: Kennedy, K., Polychronopoulos, C.D. (eds.) *Proceedings of 1992 International Conference on Super Computing (ICS 1992)*, Washington DC, July 19-23, pp. 428–437. ACM press, New York (1992)
- [12] Shirazi, B., Kavi, K., Hurson, A.R., Biswas, P.: PARSA: A Parallel Program Scheduling and Assessment Environment. In: *Proceedings of the International Conference on Parallel Processing*, pp. 68–72. CRC Press Inc., Boca Raton (1993)
- [13] Chu, W.W., Lan, M.T., Hellerstein, J.: Estimation of Intermodule Communication (IMC) and its Applications in Distributed Processing Systems. *IEEE Transactions and Computing* C-33, 691–699 (1984)
- [14] Hu, T.C.: Parallel Sequencing and Assembly Line Problems. *Operational Research* 19, 841–848 (1961)
- [15] Schildt, H.: *The Complete Reference Java2*, 5th edn. Tata McGraw Hill Publishing Company (2002)
- [16] Papadimitriou, C.H.: Computational complexity. Addison-Wesley, Reading (1994) ISBN 0-201-53082-1
- [17] Sipser, M.: *Introduction to the Theory of Computation*. Course Technology Inc. (2006) ISBN 0-619-21764-2

A New Fault Tolerant Routing Algorithm for Advance Irregular Alpha Multistage Interconnection Network

Ved Prakash Bhardwaj and Nitin

Jaypee University of Information Technology,
P.O. Waknaghat, Solan-173234, Himachal Pradesh, India
ved.juit@gmail.com, delnitin@ieee.org

Abstract. Parallel processing system (PPS) provides high communication speed and utilization of resources to the information processing system. In order to attain an effective PPS, Multistage Interconnection Network (MIN) is used. Designing a fault tolerant and cost effective MIN is a major key issue. This paper presents a new fault tolerant irregular MIN named as Advance Irregular Alpha Multistage Interconnection Network (AIAMIN). The proposed MIN is the advance form of modified alpha network (ALN). It has been analyzed that the proposed AIAMIN yields better fault tolerance by providing more alternate path between any pair of source and destination address, as compare to the existing modified ALN.

1 Introduction and Motivation

Interconnection Network (IN) plays a key role in providing communication amid processors, amid processors and memory modules [1-10]. MIN is a cost effective IN and therefore, it is used in PPS [3-10]. INs have uniform or non-uniform connection pattern and it classify the MINs to be regular or irregular respectively [11-20]. Generally, a MIN has $n = \log_2 N$ stages and each stage has $N/2$ switching elements (SEs), where N is the size of network. Basically, 2×2 SEs are used in MIN [21-25]. MINs have multipath nature and it yields a good fault tolerant capability in the network. The basic idea of having fault tolerant is to establish multiple path between a source and destination pair and the alternate path can be used in case of faults in the primary path [4-18]. Many authors have proposed various network model with their routing algorithm, in order to increase the fault-tolerance [4-21], however a unique solution has not yet obtained. In the present paper, we have proposed a new interconnection network named as Advance Irregular Alpha Multistage Interconnection Network and its routing algorithm. The basic architecture of AIAMIN is based on previously proposed modified ALN [25]. It has been observed that the AIAMIN engender better fault tolerance in the network as compare to modified ALN. It provides 2-switch fault tolerant capability at the same cost.

The rest of the paper is as follows: Section 2 discusses the Structure of previously proposed Modified ALN. In Section 3, we have proposed our network model and its routing algorithm. Section 4 discusses the issue of path availability of proposed network. In Section 5, the cost of the proposed network is analyzed and section 6 is followed by conclusion and references.

2 Structure of Modified Alpha Network

The modified ALN has N sources and N destinations [25]. There are $n = \log_2 N$ stages in this network and hence it has 4 stages. This network is divided in two subgroups and each one has $N/2$ sources and $N/2$ destinations [25]. In This network, source and destination addresses are connected through multiplexers and demultiplexers respectively. The auxiliary links connects the switches of first, second and third stages [25]. Modified ALN is a single switch fault tolerant network with limited number of alternate paths between every pair of source and destinations [25].

3 Proposed Network Model

3.1 Advance Irregular Alpha Multistage Interconnection Network

The AIAMIN is an irregular MIN and it has N sources and N destinations and here the value of N is 16. This network has $n = \log_2 N$ stages [18-25]. In the figure (1), the S represents the source address, Mux represent the Multiplexer, Demux represent the demultiplexer and D represent the destination address. In first stage, it has 8 switches of size 2×3 .The first switch is named as A, second switch named as B, and in the same way we have given names to all the switches of first stage. Each switch of first stage is connected with a 2×1 multiplexer for each input link. In first stage each source address is connected with three switches through Mux e.g. the source address 0 is connected with A, E and F. In second stage, this network has 3 switches of size 8×2 and the first switch of this stage is named as I, second switch named as J, and third switch named as K. The switches of first stage are connected with the switches of second stage through their output links e.g. A is connected with I, J and K through its output link.

In third stage, this network has two switches of size 3×8 and the first switch of this stage is named as L and second switch named as M. The switches of second stage are connected with the switches of third stage through their output links e.g. I is connected with L and M through its output link. In fourth stage, this network has 8 switches of size 2×2 and the first switch of this stage is named as N, second switch named as O, and equally we have given names to all the switches. Each switch of fourth stage is connected with a 1×2 demultiplexer for each output link. In fourth stage each destination address is connected with three switches through Demux e.g. the destination address 0 is connected with N, R and S. The switches of third stage are connected with the switches of fourth stage through their output links e.g. L is connected with N, O, P, Q, R, S, T and U through its output link. This network shows the 2-switch fault tolerance capability in second stage and 1-switch fault tolerance capability in third stage and therefore, the switches of second stage are shown by red colour, the switches of third stage are shown by green colour in figure (1).

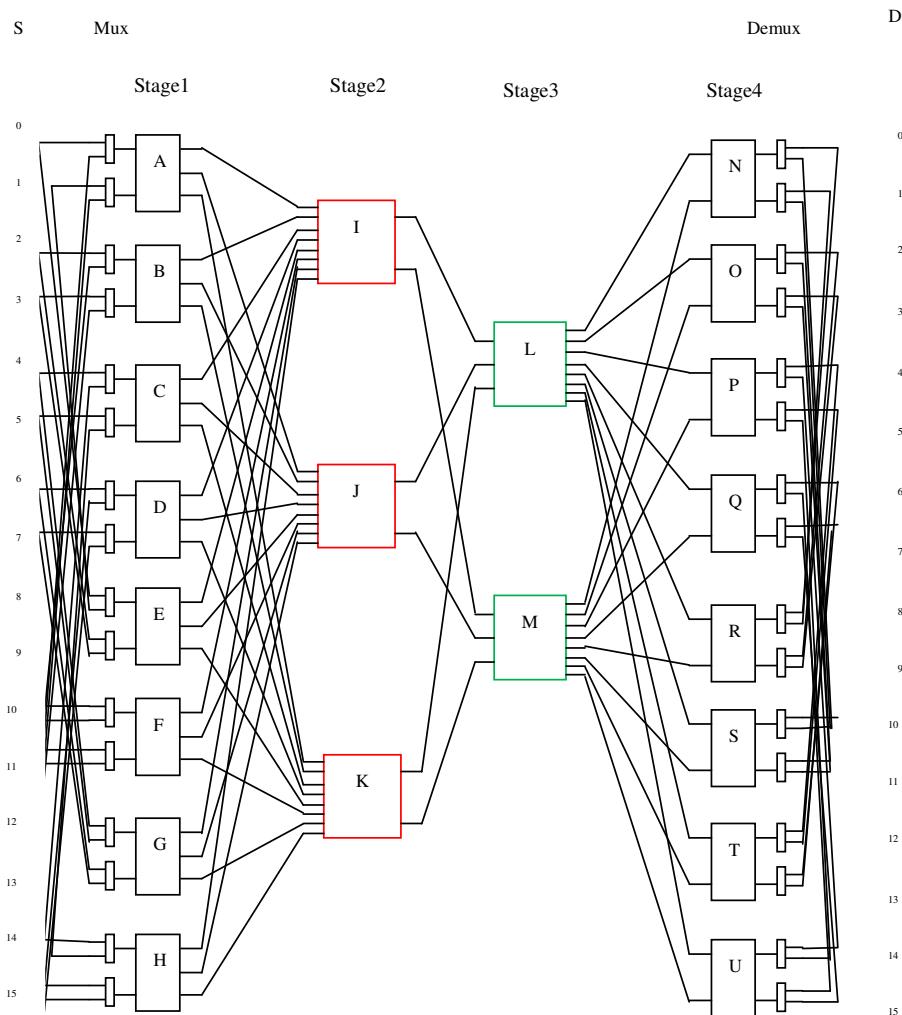


Fig. 1. 16×16 Advance Irregular Alpha Multistage Interconnection Network.

3.2 Routing Algorithm of AIAMIN

In the proposed algorithm, any source can send data to any destination in presence of multiple faults. In first step, the source and its destination address is selected. Now send the request to the suitable SE of first stage, if it is busy or faulty then follow step 7 otherwise go to step 4. In the next step, we have to send the request from SE of first stage to the SE of second stage. If the SEs of second stage are busy or faulty then we will follow step 8 otherwise go to step5. Now we have to send the request from second stage to third stage, if the SEs of third stage are busy or faulty then we will follow step 9 otherwise go to step6. In the next step, we have to send the request from

SE of third stage to the SE of fourth stage and therefore, we will send the request to the appropriate SE of fourth stage. In this way, we can send data from any source to any destination.

1. Begin
2. Get the source and its destination address.
3. Send the request to the appropriate SE of stage1. If SE of stage1 is busy or faulty then go to step7, otherwise go to step4.
4. Now send the request from SE of first stage to the first SE i.e. I of stage 2. If I of stage 2 is busy or faulty then go to step8, otherwise go to step5.
5. Send the request from SE of second stage to the SE L of stage 3. If L is busy or faulty then go to step9, otherwise go to step6.
6. Receive the request on SE L or M and go to step 10.
7. Now send the request to the first auxiliary SE of stage 1 and go to step 4, if it is busy or faulty then send the request to the second auxiliary SE of stage1 and go to step 4, if this SE is also busy or faulty then drop the request and go to step11.
8. Now send the request to J and go to step 5, if J is busy or faulty then send the request to K and go to step 5, if K is also busy or faulty then drop the request and go to step11.
9. Now send the request to M and go to step 6, if M is also busy or faulty then drop the request and go to step11, otherwise go to step 6.
10. Send the request from L or M to the appropriate SE i.e. SE that contain the destination address.
11. End.

Example1. Let the source and destination addresses are 0 and 2 respectively. In this example we will consider only the worst case, i.e. when proposed network has more than one fault in first, second and fourth stage and one fault in third stage.

Algorithmic Step1: Begin.

Algorithmic Step2: Get the source and its destination address; here we have the source address 0 and its destination address 2.

Algorithmic Step3: It is assumed that A and E are busy or faulty therefore, we will follow step 7 of algorithm and send the request to second auxiliary SE i.e. F. Now go to step 4 of algorithm. The auxiliary links are shown by red colour in figure (2).

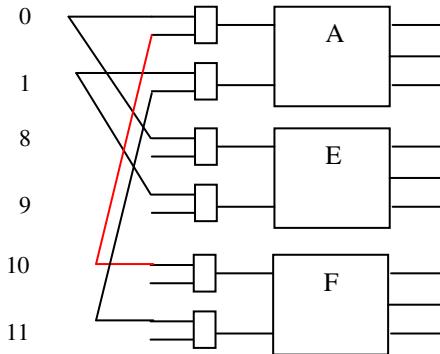


Fig. 2. Sending Request from source 0 to F through Mux

Algorithmic Step4: It is assumed that I and J are busy or faulty therefore step 8 will be followed. In this step, the request will be sent from F to K and then go to step 5 as shown in figure (3). In the next step, the request will be sent from K to M and go to step 6 of algorithm as shown in figure (4). In the next step, SE M will receive the request and now follow step 10 of algorithm.

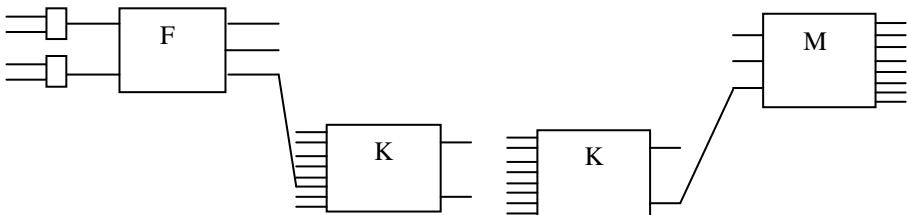


Fig. 3. Sending Request from F to K.

Fig. 4. Sending Request from K to M.

Algorithmic Step5: Now we will send the request from M to the appropriate SE i.e. at SE O and this SE will forward this request to its destination address i.e. at 2 through demultiplexer.

Algorithmic Step6: End.

In this way, we have sent the data from source address 0 to 2. This is the complete explanation of the proposed routing algorithm of the proposed network model.

4 Path Availability

Path Availability explains that how many paths are available between two nodes i.e. between source node and destination node. If we have more number of paths in the network then we can send data from any source to any destination in case of multiple faults. The theorem1 and 2 shows that the proposed network model has more number of paths as compare to the existing modified ALN.

4.1 Theorem 1

There are six alternate paths between every pair of source and destination in AIAMIN.

Proof:

Let us suppose the source address is 2 and destination address is 12 then the possible number of alternate paths between SE B and T are as follows:

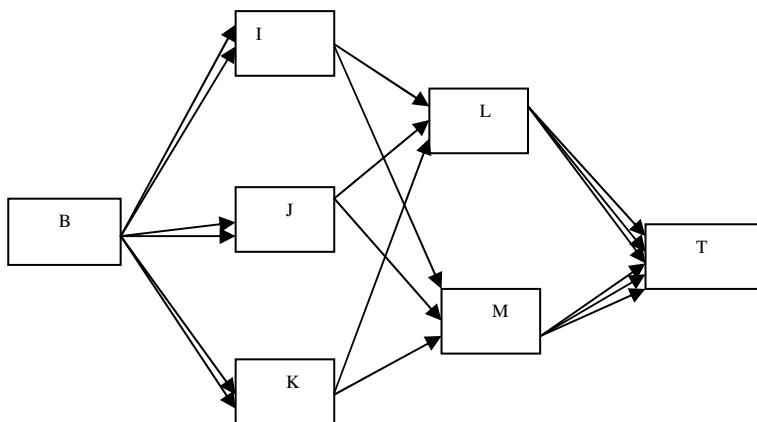


Fig. 5. Available Paths between B and T.

From figure (5), it is clear that there are six alternate paths between B and T and therefore, the above shown paths prove the theorem.

4.2 Theorem 2

If the SEs of stage3 is connected through the auxiliary links then there are twelve alternate paths between every pair of source and destination in AIAMIN.

Proof:

In order to get the more alternate path the SEs of third stage should be connected through the auxiliary links. These links are shown by red colour in figure (6).

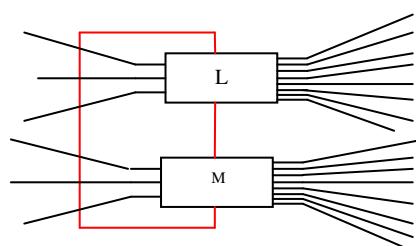


Fig. 6. AIAMIN With Auxiliary Links.

Now let us suppose the source address is 2 and destination address is 12 then the six alternate paths will be same as explained in theorem1 and the other six alternate paths are shown in figure (7).

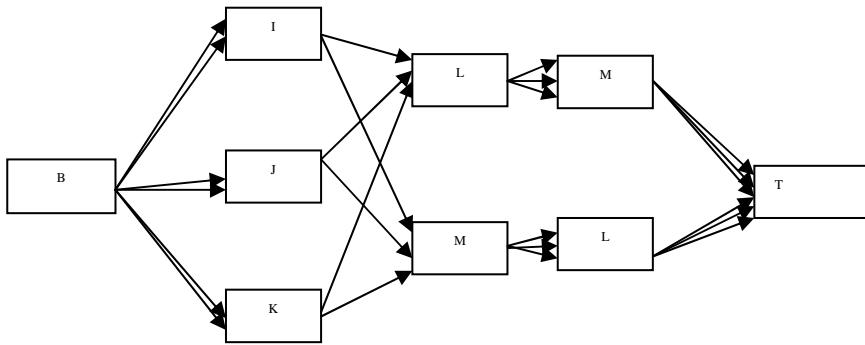


Fig. 7. Path Availability between B and T

The above shown paths prove the theorem.

5 Cost Analysis

Basically cost of the network depends on its component's complexity for e.g. the cost of a switch is proportional to the number of gates counts within a switch. Similarly we can calculate the cost of multiplexers and demultiplexers. The cost of AIAMIN is calculated in both conditions i.e. when it does not have the auxiliary links:

Total number of 2×3 SEs = 8, cost = 48

Total number of 8×2 SEs = 3, cost = 48

Total number of 3×8 SEs = 2, cost = 48

Total number of 2×2 SEs = 8, cost = 32

Total number of 2:1 multiplexers = 16, cost = 32

Total number of 1:2 demultiplexers = 16, cost = 32

Total cost of the AIAMIN is 240 units. If it has auxiliary links then this network will have 2 SEs of 4×9 size, instead of 3×8 size as shown in figure (6) and therefore, the cost of this network will be 264 units. The total cost of modified ALN is 240 units and it is a single switch fault tolerant network and our proposed network is 2-switch fault tolerant with more alternate paths in both cases i.e. with auxiliary links or without auxiliary links. If this network will use the auxiliary links then it will be little costly as compare to the modified ALN and it will provide more alternate paths as explained in theorem2. If this network will not use the auxiliary links then the cost of proposed network and modified ALN will be same.

6 Conclusion

In this paper, we have proposed a new fault tolerant network model named as advance irregular alpha multistage interconnection network (AIAMIN). A brief explanation of section 3 and theorems of section 4 shows that the AIAMIN has better fault tolerance capacity with more alternate path as compare to the previously proposed modified ALN. Further, the design idea of the proposed network can be applied to the other interconnection networks to obtain a good fault tolerant network at low cost.

References

1. Bataineh, S.M., Allosl, B.Y.: Fault-Tolerant Multistage Interconnection Network. *Telecommunication Systems* 17(4), 455–472 (2001)
2. Cheema, K.K., Aggarwal, R.: Design Scheme and Performance Evaluation of a new Fault-tolerant Multistage Interconnection Network. *International Journal of Computer Science and Network Security* 9(9) (September 2009)
3. Mun, Y.: Performance Analysis of Banyan-Type Multistage Interconnection Network Under Non Uniform Traffic Pattern. *The Journal of Supercomputing* 33, 33–52 (2005)
4. Chen, C.-W.: Design schemes of dynamic rerouting networks with destination tag routing for tolerating faults and preventing collisions. *J. Supercomput.* 38, 307–326 (2006)
5. Chen, C.-W., Chung, C.-P.: Designing A Disjoint Path Interconnection Network with Fault Tolerance and Collision Solving. *The Journal of Supercomputing* 34, 63–80 (2005)
6. Dou, W.-Q., Yao, E.-Y.: On 1-rate and 2-rate multicast 3-stage Clos networks, *Appl. Math. J. Chinese Univ.* 24(2), 151–156 (2009)
7. Hao, D., Shen, X.: Rearrangeability of 7-stage 16×16 shuffle exchange networks. *Front. Electr. Electron. Eng. China* 3(4), 440–458 (2008)
8. Mahajan, R., Vig, R.: Performance and Reliability Analysis of New Fault Tolerant Advance Omega Network. *WSEAS Transactions on Computers* 7(8) (August 2008) ISSN: 1109-2750
9. Gupta, A., Bansal, P.K.: Proposed Fault Tolerant New Irregular Augmented Shuffle Network. *Malaysian Journal of Computer Science* 24(1) (2011)
10. Garofalakis, J., Stergiou, E.: An Approximate Analytical Performance Model for Multistage Interconnection Networks with Backpressure Blocking Mechanism. *Journal of Communication* 5(3) (March 2010)
11. Bataineh, S., Qanzu'A, G.E.: Reliable Omega Interconnected Network for Large-Scale Multiprocessor Systems. *The Computer Journal* 46(5) (2003)
12. Veselovsky, G.: An Approach for Exploring Combinatorial Properties of R-path Omega Interconnection Networks. *AU J.T.* 13(3), 143–150 (2010)
13. Rastogi, R., Nitin, Chauhan, D.S., Govil, M.C.: On Stability Problems of Omega and 3-Disjoint Paths Omega Multi-stage Interconnection Networks. *International Journal of Computer Science Issues* 8(4(2)) (July 2011)
14. Tutsch, D., Hommel, G.: MLMIN: A multicore processor and parallel computer network topology for multicast. *Computers & Operations Research* 353807-3821 (2008)
15. ALqerem, A.H.: Congestion Control By Using A Buffered Omega Network. In: *International Conference on Applied Computing* (2005)
16. Ghai, M.: A Routing Scheme for a New Irregular Baseline Multistage Interconnection Network. *International Journal of Advanced Computer Science and Applications* 2(5) (2011)

17. Nitin, Subramanian, A.: Efficient Algorithms to Solve Dynamic MINs Stability Problems using Stable Matching with Complete TIES. *Journal of Discrete Algorithms* 6(3), 353–380 (2008)
18. Sharma, S., Bansal, P.K., Kahlon, K.S.: On a Class of Multistage Interconnection Network in parallel processing. *International Journal of Computer Science and Network Security* 8(5) (2008)
19. Sharma, S., Kahlon, K.S., Bansal, P.K.: Reliability and Path length Analysis of Irregular Fault tolerant Multistage Interconnection Network. *ACM SIGARCH Computer Architecture News* 37(5) (2009)
20. Nitin: Component Level Reliability analysis of Fault-tolerant Hybrid MINs. *WSEAS Transactions on Computers* 5(9), 1851–1859 (2006) ISSN 1109–2750
21. Rastogi, R., Nitin, Chauhan, D.S.: 3–Disjoint Paths Fault-tolerant Omega Multi-stage Interconnection Network with Reachable Sets and Coloring Scheme. In: Proceedings of the 13th IEEE International conference on Computer Modeling and Simulation (IEEE UK-Sim), UK (2011)
22. Rastogi, R., Nitin: Fast Interconnections: A case tool for Developing Fault-tolerant Multi-stage Interconnection Networks. *International Journal of Advancements in Computing Technology* 2(5), 13–24 (2010)
23. Ghai, M., Chopra, V., Cheema, K.K.: Performance Analysis of Fault-Tolerant Irregular Baseline Multistage Interconnection Network. *International Journal on Computer Science and Engineering* 02(09), 3079–3084 (2010)
24. Fan, C.C., Bruck, J.: Tolerating Multiple Faults in Multistage Interconnection Networks with Minimal Extra Stages. *IEEE Transactions on Computers* 49(9) (September 2000)
25. Gupta, A., Bansal, P.K.: Fault Tolerant Irregular Modified Alpha Network and Evaluation of Performance Parameters. *International Journal of Computer Applications* (0975–8887) 4(1) (July 2010)

Comparing and Analyzing the Energy Efficiency of Cloud Database and Parallel Database*

Jie Song, Tiantian Li, Xuebing Liu, and Zhiliang Zhu

Software College, Northeastern University, Shenyang, P.R. China

{songjie, zzl}@mail.neu.edu.cn, {litiantian_neu, neu_lxb}@163.com

Abstract. To study the Energy Efficiency (*EE*) of cloud database so as to achieve green computing, the measurement model and approach of *EE* should be defined, the *EE* characteristics of cloud database should be investigated, and the *EE* of cloud database should be compared with that of parallel database. In this paper, the measurement model of *EE* and its mathematical expression are proposed; the test cases including data loading, querying and analyzing are defined; the measurement approach of cloud database's *EE* is described; the *EE* characteristics of HBase (a cloud database) when executing loading, retrieving, querying, aggregation and join operations are analyzed and compared with that of GridSQL (a parallel database). Plenty of experiments show that, despite that cloud database is an application of "green cloud computing", the *EE* of HBase remains to be further optimized.

Keywords: Cloud Database, Parallel Database, Energy Efficiency.

1 Introduction

Currently, the definition of cloud database is not clearly given [1]. It is generally accepted that cloud database is a new merged database based on CAP theory [2, 3], BASE theory [4] and database Sharding [5] technique. Although considered as a "green database", cloud database still lacks mature solutions to evaluate and reduce the energy consumption; approaches for optimizing its Energy Efficiency (*EE*) are on demand. For the optimization, the measurement model and benchmarking approach of *EE* should be defined, and the *EE* characteristics of cloud database should be analyzed and compared with that of traditional database. Current researches on evaluating cloud database mostly focus on performance issues; few of them evaluate and optimize the *EE*, introduce the measurement model, measurement instruments, test cases and regularities of *EE*. In addition, as far as we know, there is no report on comparison of *EE* between cloud database and parallel database. Therefore, our researches in this paper focus on the following aspects:

- Firstly, a measurement model for *EE* is defined. Current performance models mainly focus on "speed" and "storage", not consider energy issues, while the proposed *EE* model focuses on "performed computation" and "energy consumption".

* Supported by the Fundamental Research Funds for the Central Universities of China (N110417002), the Natural Science Foundation of Liaoning Province (200102059), the National Natural Science Foundation of China (61173028).

- Secondly, a set of test cases evaluating the *EE* of cloud database are designed. TPC-H, a typical test case for traditional database, is too complex to be implemented in cloud database for its specificity and immaturity. The test cases in [6], which have already been used in the benchmarking of MapReduce based data analysis [7, 8], are adopted and modified to suit for the cloud database.
- Finally, the *EE* of cloud database is compared with that of parallel database under the test cases of loading, reading, querying, aggravation, and join [9]. Comparison results show that, despite that HBase is an application of “green cloud computing”, it does not inherit the characteristics of high *EE*; its *EE* remains to be further optimized.

Currently, there exist some works evaluating the cloud database. For example, [10] evaluated the performance of database management system which doesn’t support structured queries, such as Google BigTable. The YCSB (Yahoo Cloud Serving Benchmark) [11] framework developed by Yahoo provides a set of test cases combined by insert, read, update and scan operations, which could be adopted to measure the performance of cloud database; but these test cases only involve the basic data operations which are too simple for database. [6] implemented MapReduce based inserting, querying, aggregation and join operations on Hadoop HDFS, and compared their performance with that of parallel database. Most researches only proposed simple energy measurement models under no constraints; they did not provide specific measurement approach. The problem of energy efficiency hasn’t been paid enough attention. Therefore, we propose an *EE* model under performance constraint based on the existing researches in this paper. It includes the explanation of mathematical definition and test cases; it measures the performed computation per energy unit; it considers the dynamic power, different kinds of workload and various execution conditions.

2 Energy Efficiency Model

There are three factors contributing to the energy consumption: computers, network equipments and other accessorial infrastructure (such as air conditioner). In this paper, we only consider the first one. Literally, Energy Efficiency (*EE*) is the combination of “energy” and “efficiency”. And FLOPS (Floating Point Operations Per Second) or MIPS (Million Instructions Per Second) is often adopted to evaluate the efficiency of a computer, while power (watt) is adopted to measure the energy consumed or generated by the electrical equipments in a second. But it is difficult to calculate how many floating-point-operations a task contains, so a new measurement for workload needs to be proposed. Actually, workload unit can be defined according to the algorithm. Taking sorting algorithm as an example, we can define sorting 1000 records as one workload unit, but such definition is related to the algorithm’s complexity, and lacks generality. To simplify the *EE* model, we define workload unit as follows.

Definition 1 Workload Unit: Workload unit (U) is used to measure the size of the workload. We define the performed computation by a 1GHz CPU per second as one workload unit, denoted as $1U$.

Definition 2 Energy Efficiency: Let $L(T)$ be the workload (unit is U), and $E(T)$ be the consumed energy of target system during time T , then the energy efficiency during time T is defined as:

$$\eta(T) = \frac{L(T)}{E(T)} \quad E(T) \neq 0 \quad (1)$$

EE unit is denoted as η , $\eta = U/Joule$, $1m\eta = 10^{-3} \eta$.

For a cloud database, let N be the number of nodes, c_i ($1 \leq i \leq N$) be a node, $f_i(t)$ (GHz) and $\omega_i(t)$ be the CPU frequency and usage of c_i , and $p_i(t)$ (watt) be the power at time t , then:

$$L_i(t) = f_i(t)\omega_i(t) \quad E_i(t) = p_i(t) \quad (2)$$

$$L(T) = \sum_{i=1}^N \int_0^T f_i(t)\omega_i(t)dt \quad E(T) = \sum_{i=1}^N \int_0^T p_i(t)dt \quad (3)$$

η is calculated as:

$$\eta_i(t) = \frac{L_i(t)}{E_i(t)} = \frac{f_i(t)\omega_i(t)}{p_i(t)} \quad \eta(T) = \frac{L(T)}{E(T)} = \frac{\sum_{i=1}^N \int_0^T f_i(t)\omega_i(t)dt}{\sum_{i=1}^N \int_0^T p_i(t)dt} \quad (4)$$

Definition 3 Energy Consumption: Energy consumption is the consumed energy per request case under a certain experimental condition, denoted as $E(case, nodes, amount, concurrency)$.

3 Measurement Approach

The table schemas we adopt are listed in table 1. Details are abbreviated (refer to [6]). In this section, we explain how *Loading*, *Grep*, *Query*, *Aggregation* and *Join* cases are implemented in HBase and GridSQL, which are different from that in [6].

Loading: The data is loaded into *Grep*, *Rankings* and *UserVisits* tables under fixed concurrent requests.

Grep: The *Grep* case performs the function as the following SQL described. The pattern is matched once in every 10,000 records.

```
SELECT * FROM Grep WHERE field LIKE '%XYZ%'
```

Selection: *Selection* case performs the function as the following SQL described. It is a lightweight filter to find the *pageURLs* in *Rankings* table with a *pageRank* above a user-defined threshold. In our experiment, we set this threshold to 10, and the system yields 90% of the total records.

```
SELECT pageURL FROM Rankings WHERE pageRank > 10
```

Table 1. Table schemas and data generation approach

Table	Column	Type	Generation Approach
<i>Grep</i>	<i>key (PK)</i>	Long	Auto-increased column in GridSQL, omitted in HBase
	<i>field</i>	Varchar (90)	Random value from ASCII 32 to 126 (94 charters)
<i>Rankings</i>	<i>pageURL(PK)</i>	Varchar (100)	Randomly selected from 10,000 real URL addresses
	<i>pageRank</i>	Double	Random value between 0 and 100.0
	<i>avgDuration</i>	Int	Random value between 0 and 1000
<i>UserVisits</i>	<i>sourceIP</i>	Varchar (16)	Formatted as “202.118.X.Y”, X and Y are values between 0 and 255 (65536 IPs totally)
	<i>destURL</i>	Varchar (100)	Randomly selected from 10,000 real URL addresses
	<i>visitDate</i>	DateTime	Random value between 1900-01-01 and 2000-12-31
	<i>adRevenue</i>	Double	Random value between 0.0 and 1.0
	<i>userAgent</i>	Varchar (64)	Random string
	<i>countryCode</i>	Char (3)	Random value between 0 and 100, converted to string
	<i>languageCode</i>	Char (3)	Random value between 0 and 100, converted to string
	<i>searchWord</i>	Varchar (32)	Random words
	<i>duration</i>	Int	Random value between 0 and 1000

Aggregation: Aggregation case performs the function as the following SQL described. It calculates the total *adRevenue* in *UserVisits* table grouped by *sourceIP*. It measures the performance of parallel analytics on a single table. There are 65,536 kinds of *sourceIP* (shown in Table 1.), thus this case produces 65,536 groups.

```
SELECT sourceIP, SUM(adRevenue) FROM UserVisits
    GROUP BY sourceIP
```

Join: Join case performs the function as the following SQL described, which is a common operation in data management system. It consists of 4 tasks on two data sets. The first task is querying records by a given range of *visitDate*, the query rate is about 10%. In the second task, *Rankings* and *UserVisits* tables are joined by their *pageURL* (*destURL*). The third task is calculating the total *adRevenue* and average *pageRank* on joined results grouped by *sourceIP*. In the last task, the aggregated results are written into *Results* table.

```
SELECT INTO Results sourceIP, AVG(pageRank) as avgPageRank,
        SUM(adRevenue) as totalRevenue
    FROM Rankings AS R, UserVisits AS UV
    WHERE R.pageURL = UV.destURL
    AND UV.visitDate BETWEEN Date('2005-1-1') AND
        Date('2006-1-1')
    GROUP BY UV.sourceIP;
```

The critical task of *Join* case is the join task. Currently, HBase does not provide explicit command for joining two or more disparate data sets, and they don't support for creating foreign keys on sparse table. We have to implement the join operation by programming. There are lots of MapReduce based join algorithms, such as *Map Side Join*, *Reduce Side Join*, *Semi Join*, *Distributed Hash Join* and *Bloom Filter based Join*. Among these algorithms, *Reduce Side Join*, shown as Algorithm 1, is the most conventional join strategy in the MapReduce framework.

Algorithm 1. MapReduce implementation of join task

```

1. Map( $Key_M^{In}$   $rowId$ ,  $Value_M^{In}$   $row$ ){
2.   If ( $row$  is Ranking)
3.     Write( $row.pageURL$  as  $Key_M^{Out}$ ,  $row.pageRank$  as  $Value_M^{Out}$ );
4.   End if
5.   If ( $row$  is UserVisits)
6.     If ( $row.visitDate$  between begin_date and end_date)
7.       Pair  $pair = \text{new Pair}(row.sourceIP, row.adRevenue)$ 
8.       Write( $row.destURL$  as  $Key_M^{Out}$ ,  $pair$  as  $Value_M^{Out}$ );
9.     End if
10.   End if
11. }
12. Reduce( $Key_R^{In}$   $ip$ ,  $Value_R^{In}$   $arrays$ ){
13.   Set  $ipSet$ ;
14.   Set  $pageRankSet$ ;
15.   Foreach  $e$  in  $arrays$ 
16.     If ( $e$  is UserVisits)
17.        $ipSet.add(e.sourceIP);$ 
18.       Write( $e.sourceIP$  as  $Key_R^{Out}$ ,  $e.adRevenue$  as  $Value_R^{Out}$ );
19.     End if
20.     Else if ( $e$  is Ranking)
21.        $pageRankSet.add(e$  as pageRank)
22.     End else
23.   End for
24.   Foreach  $ip$  in  $ipSet$ 
25.     Foreach  $pageRank$  in  $pageRankSet$ 
26.       Write( $ip$  as  $Key_R^{Out}$ ,  $pageRank$  as  $Value_R^{Out}$ );
27.     End for
28.   End for
29. }
```

4 Comparison between HBase and GridSQL

In this section we compare different aspects of HBase, such as energy consumption and EE, with that of GridSQL. Details of the testbed are shown in Table 2.

Table 2. Description of the experiment environment

Items	Descriptions
Node	1 admin node, 10 data nodes, homogeneous computers. Inter Core i5-2300 2.80GHz, 8GB memory, 1TB hard disk, onboard video, audio and network card. The energy consumption of each node excluding that of monitor, keyboards and mouse.
Operation System	CentOS 5.6, Linux 2.6.18 Kernel
Databases	Hadoop HDFS 0.20.2, HBase 0.90.3, ZooKeeper 3.3.2
CPU Working Mode	GridSQL 2.0, edb-jdbc 1.4 (JDBC for GridSQL)
Program Language	Conservative model
Power Instrument	Java 6
Time interval of ω and f	PowerBay power-meter (http://www.northmeter.com/index-en.html), power precision $\pm 0.01\sim 0.1W$, maximum 2200W, measurement frequency 1.5-3 second.
Time interval of ω and f	Acquire CPU frequency and usage in every second ($\Delta t = 1s$), when CPU is busy, the request on measuring ω cannot be responded in time, Δt is larger than 1 second.

Table 2. (continued)

Measurement Units	Computing Efficiency: U as defined in definition 1; Energy Consumption: Joule;
	Power: Experiment adopts real power, unit is watt, the real power of experimental computer shifts between 50 watt and 100 watt;
	Energy Efficiency: unit is η and $m\eta$, $\eta=U/Joule$, $1m\eta=10^{-3}\eta$, the maximum energy efficiency of experimental computer is about 28 $m\eta$
	CPU frequency: GHz
	CPU usage: rate ω ($0 \leq \omega \leq 1$)
Concurrent Clients	5 computers as clients; each client send 1-20 requests to the database concurrently. Can simulate maximum 100 concurrent requests.
Data Amount	Data amount is based on <i>Grep</i> and <i>UserVisits</i> tables; data amount of <i>UserVisits</i> table is 10 times larger than that of <i>Rankings</i> table. there are three kinds of data amounts: 0.5 million, 5 million and 50 million; data is averagely distributed on each data node.

Firstly, the *Loading* case is studied. We load 0.5 million records to *Grep* table under different concurrent threads. The comparison result between the two databases is shown in Fig.1.

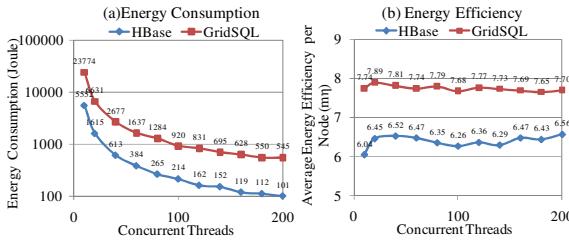
**Fig. 1.** Comparison of energy consumption of loading case on HBase and GridSQL

Fig.1-a (logarithmic Y axis) shows that the energy consumption of HBase is less than that of GridSQL on *loading* case; such superiority is more obvious when the number of concurrent threads increases. In GridSQL, there are lots of operations on checking constraints and converting data format; In HBase, the table schema and data format are quite simple, and the data is written to the distributed file system directly once the timestamp is added. *Loading* in HBase is simpler than that in GridSQL, so the former consumes less energy. Fig.1-b shows that the Energy Efficiency (EE) of both two databases is as low as about $7m\eta$; this is because *loading* involves mainly I/O operations, therefore, CPU is relatively idle. This result further proves that *loading* algorithm of GridSQL is much more complex than that of HBase.

Fig.2 compares the energy consumption, performance and energy efficiency of *Grep*, *Query*, *Aggregation* and *Join* cases running on HBase and GridSQL with various number of concurrent requests and 50 million data. It can be easily concluded that the observed energy consumption of GridSQL is strikingly better than that of HBase. The former is 10 times less than the later in *query* cases (Fig.2 a-1,-2), and 20 times less in *analysis* cases (Fig.2 a-3,a-4).

Fig.2-a could be explained from the performance aspect. Fig.2-b shows the execution time of each case corresponding to Fig.2-a, and GridSQL outperforms

HBase greatly. Basically, the longer the execution time is, the larger the energy consumption is. Since the power of experimental computer shifts from 50w to 100w, that is to say, two-second-computation in 50 watt consumes as much energy as one-second-computation in 100 watt; but in this comparison, execution time takes the most responsibility of energy consumption.

The performance of MapReduce has been well discussed in [6-7]. So, we won't discuss it anymore; we will explain Fig.2-a from another aspect. Under the same testbed and the same test case, there are only two possible reasons causing HBase to consume more energy than GridSQL:

- Algorithm complexity: For a same case, HBase performs more operations than GridSQL. The algorithm of HBase is more complex than that of GridSQL;
- Energy efficiency: For a same amount of energy, HBase performs less operation than GridSQL. *EE* of HBase is lower than that of GridSQL.

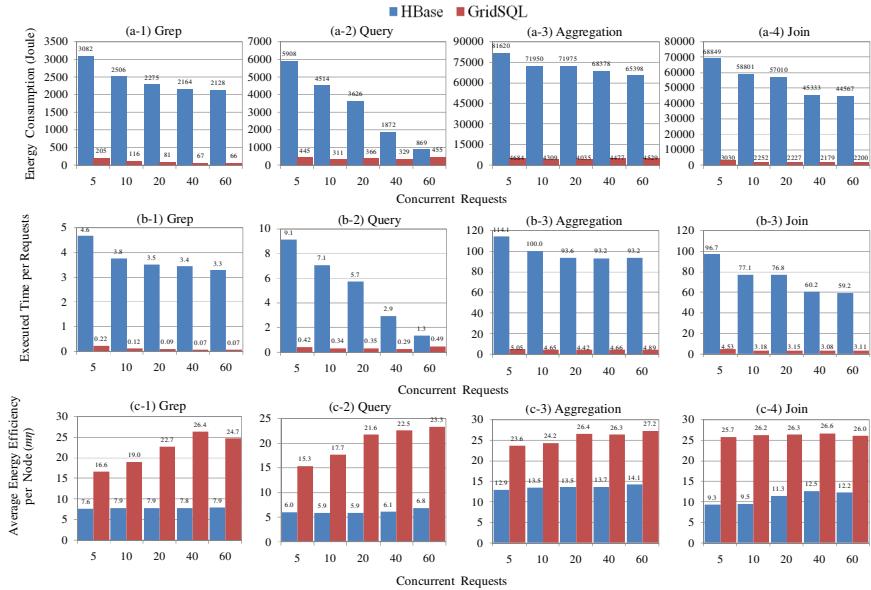


Fig. 2. Comparison of energy consumption, performance and energy efficiency on HBase and GridSQL

Fig.2-c shows the *EE* of cases corresponding to Fig.2-a. It proves the two reasons above. The maximum *EE* of the experimental computer is about $28mJ$ ($f=2.8GHz$, $\omega=1$, $E\approx100watt$). In GridSQL, *EE* is close to the maximum value. When concurrency increases, GridSQL isn't fully loaded, nodes are idle while waiting new tasks, and *EE* is lower; when concurrency increases, the workload is sufficient, and *EE* is close to the maximum. In HBase, *EE* is quite lower. *EEs* of *Grep* and *Query* cases (Fig.2, c-1,-2) are about a quarter of the maximum, while *EEs* of *Aggregation* and *Join* cases are about a half of the maximum. *EE* becomes stable after the concurrency increases to a certain value, which means that the lower *EE* is caused by other reasons

(such reasons will be studied in next section), rather than the insufficient workload. On the other hand, take *Query* case as an example, the energy consumption of HBase is 10 times more than that of GridSQL (Fig.2, a-2). If the algorithm complexity of HBase and GridSQL are the same, *EE* of HBase should be 10 time less than that of GridSQL rather than 3 times less (Fig.2, c-2). Now, that the *query* case performed by HBase contains more computation than by GridSQL has been proved.

Statistically, we analyze the CPU frequency $f(t)$ and usage $\omega(t)$ of data node and admin node of HBase and GridSQL during execution time. The value of $f(t)$ could be one of the 13 values between 1.6GHz and 2.8GHz with a interval of 0.1; the value of $\omega(t)$ ($0 \leq \omega \leq 1$) is normalized to the 10 equal-width intervals between 0.1 and 1. The frequency-ratio, which is the ratio between interval count and total count, is shown in Fig.3.

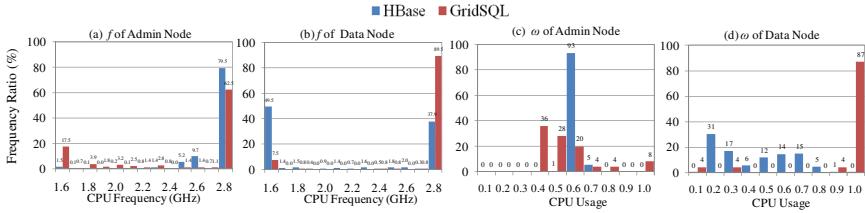


Fig. 3. Histogram of CPU frequency and usage of admin node and data node of HBase and GridSQL

For the admin node, Fig.3-a,-c shows that CPUs of GridSQL and HBase almost work on the highest frequency, while CPU usages are lower, the maximum value of $\omega(t)$ seldom occurred. We increase the concurrency, but the statistic of ω does not change much. It can be seen that the workload of admin node is light in our experiment.

For the data node, Fig.3-b,-d shows that CPU of GridSQL mostly works on the highest frequency and usage, while CPU frequency of HBase works on either maximum (busy) or minimum (idle) and CPU usage is as lower as about 0.5. f is a instant value, and ω is an average value in a short period. Fig.3-b,-d proves that the data nodes of HBase are idle in half of the execution time. The idleness is not caused by insufficient workload but caused by algorithm bottleneck; CPUs are waiting for other resources, such as schedules from admin node, I/O operations of hard disk and network communication. Compared with the *EE* regularities of GridSQL, the *EE* of HBase could be improved by a large space.

5 Conclusions and Future Works

All the experiments in section 4 show the energy consumption and Energy Efficiency (*EE*) characteristics of HBase as a typical cloud database by comparison with that of GridSQL as a typical parallel database, and some conclusions could be drawn here. **Firstly**, there are two reasons causing higher energy consumption of HBase (except *loading* case): higher algorithm complexity and lower energy efficiency. For the later, *EE* can be improved by reducing CPU idleness. **Secondly**, the maximum *EE* of experimental computer occurs when CPU frequency and usage are maximums.

Thirdly, there are two reasons causing lower CPU frequency and usage: insufficient workload and bottleneck of algorithm. The former is minor and could be improved by increasing concurrency; the latter is the main responsible reason which causes CPU waiting for I/O operations, network communication and schedules.

According to the features of HBase, we deduce the possible reasons leading to the high complexity and bottleneck of the algorithm as follows: ① HBase does not provide index mechanism, thus the data has to be queried by full-scanning. The algorithm complexity is higher. ② Since the data is stored as text, we have to parse it first, which is an additional computing work increasing the query computation. ③ I/O operations may be a node's bottleneck. MapReduce is independent from storage system, but in our experiment, it is implemented on HDFS, a distributed file system. The performance of HDFS's I/O interface should be optimized. ④ The intermediate results of MapReduce system are stored as files in HDFS, which brings more additional I/O operations. ⑤ Reduce functions require data from other nodes; CPU has to wait when the data is transferred over the network; it could be a bottleneck if there is no optimization applied in "shuffle" stage. ⑥ The schedule algorithm may contains defects, which may cause the data nodes idle for waiting the schedule. The above deductions are based on the experimental results and features of HBase and MapReduce. Further validation of these deductions is our future works.

References

- [1] Daniel, A., Michael, J.C., Surajit, C., Hector, G., Jignesh, M.P., Raghu, R.: Cloud Databases: What's New? Proc. of the VLDB Endowment 3(2), 1657 (2010)
- [2] Brewer, E.A.: Towards Robust Distributed Systems. In: Proc. of PODC 2000, p. 7 (2000)
- [3] Seth, G., Nancy, A.L.: Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-tolerant Web Services. SIGACT News 33(2), 51–59 (2002)
- [4] Dan, P.: BASE: An Acid Alternative. ACM Magazine Queue 6(3) (2008)
- [5] Database Sharding, <http://www.dbshards.com/articles/database-sharding-whitepapers/>
- [6] Andrew, P., Erik, P., Alexander, R., Daniel, J.A., David, J.D., Samuel, M., Michael, S.: A comparison of approaches to large-scale data analysis. In: Proc. of SIGMOD, pp. 165–178 (2009)
- [7] Jiang, D.W., Ooi, B.C., Shi, L., Wu, S.: The Performance of MapReduce: An In-depth Study. Proc. of the VLDB Endowment 3(1), 472–483 (2010)
- [8] Shi, Y.J., Meng, X.F., Zhao, J., Hu, X.M., Liu, B.B., Wang, H.P.: Benchmarking Cloud-based Data Management Systems. In: Proc. of CloudDB 2010, pp. 47–54 (2010)
- [9] Michael, S., Daniel, J.A., David, J.D., Samuel, M., Erik, P., Andrew, P., Alexander, R.: MapReduce and Parallel DBMSs: Friends or Foes? Commun. ACM (CACM) 53(1), 64–71 (2010)
- [10] Chang, F., Dean, J., Ghemawat, S., Hsieh, W.C., Wallach, D.A., Burrows, M., Chandra, T., Fikes, A., Gruber, R.: Bigtable: A Distributed Storage System for Structured Data. In: Proc. of the 7th Conference on USENIX Symposium on Operating Systems Design and Implementation, pp. 205–218 (2006)
- [11] Cooper, B., Silberstein, A., Tam, E., Ramakrishnan, R., Sears, R.: Benchmarking Cloud Serving Systems with YCSB. In: Proc. of ACM Symposium on Cloud Computing 2010, pp. 143–154 (2010)

A Grid Fabrication of Traffic Maintenance System

Avula Anitha, Rajeev Wankar, and C. Raghavendra Rao

Department of Computer and Information Sciences

University of Hyderabad

anitha_aavula@yahoo.com, {wankarcs,karcs}@uohyd.ernet.in

Abstract. This Paper describes Traffic Maintenance System as a Grid System and proposes an approach for reducing congestion on road traffic. It introduces terminology, notations and evolves congestion controlling strategy. The steps used in this strategy such as clustering, scheduling and communication are described with an illustrative example.

Keywords: Traffic Maintenance, Congestion, Grid System, Controlling Strategy, Clustering, Scheduling, communication.

1 Introduction

A Traffic Management System [2] as discussed is a complex temporal graph theory problem. It has junctions and connecting nodes giving rise to topology with their capacities as weights. Vehicles start from a location/node with an intention to reach a destination/another node in a stipulated time with desired travelling conditions which is dynamic in nature and makes transport as temporal. Vehicles from different sources with different destinations create different loads on roads as well as junctions which lead to the spatial complexity. Thus the problem is associated with traffic or spatiotemporal nature [5]. Enhancing throughput and utilization factors of a traffic management system can be achieved by reducing delays as well as enhancing edge utilization by adopting apt routing strategies.

Advance Information Systems will aid the vehicles to choose alternative path or a diverted path to avoid waiting or discomfort in the proposed route because of congestion, choking or overloading. A web service solution will definitely address information publishing aspects whereas developing routing strategies or alternative routing strategies are computationally intensive and needs High Performance Computing Facilities. Thus a Grid Fabrication for Traffic Maintenance System (GFTMS) is thought as a better solution which provides computational support as well as stateful web service.

In this paper, the following section 2 gives a brief account of Grid System and Web Services. Section 3 provides Traffic Maintenance analogy in Grid System. Section 4 presents a proposed congestion controlling strategy by utilizing local information system. Section 5 provides illustrative explanation of Grid Fabrication for Traffic Maintenance System (GFTMS) Section 6 specifies the tools used to work with road network. Conclusion and assumptions made in this paper are dealt in Section 7.

2 Grid System

The supercomputing held in San Diego in 1995 has lead to the emergence of Grid Computing platform. The Global Grid Forum (GGF) also known as Open Grid Forum (OGF) developed standard interfaces, behaviour, core semantics etc. for grid applications based upon web services [7].

We can classify service state management in web services into two forms. 1. Interaction aware state, in which a client may interact with a server for a long period of time. These interactions are correlated with some information passed from the client along with the message such as session-id, cookies etc. In this the server will not create a specific instance for client and it won't manage any client state information. 2. Application aware state, in which the services are aware of its clients and create a specific instance for the specific client and pass the instance information back to the client for interaction. Therefore the client is holding a reference to the specific instance of the service/application and hence can interact with the service instance without passing any correlation information. These services are referred to as stateful services.

Grid services are the stateful web services with a well defined interfaces and behaviour for interactions.

In our problem, at any moment any junction can behave as an information requestor or information responder, to achieve this Service Oriented Approach (SOA) approach is required which can be done using either Web Services or Grid Services. As Web Services are stateless and the information gathered should be valid and existing for a long time, the statefulness should be provided. Therefore in our approach Grid Technology is used.

Grid Resource Access Manager is the module in Grid that provides the remote execution and status management of the execution. When a job is submitted by a client, the request is sent to the remote host and handled by the gatekeeper daemon located in the remote host. The gatekeeper creates a job manager to start and monitor the job. When the job is finished, the job manager sends the status information back to the client and terminates. Job manager is created by the gatekeeper daemon as part of the job requesting process using Resource Specification Language (RSL) by the clients. All job submissions requests are described in RSL including the executable file and condition on which it must be executed. Job manager provides the interface that controls the allocation of each local resource manager such as a job scheduler like PBS or load-leveler. The functions are parse RSL, allocate job requests to the local resource manager, send call-backs to clients, receive status, cancel requests from clients and send output results to clients using Global Access to Secondary Storage if requested (GASS) [8].

3 Traffic Maintenance Analogy in Grid System

In the following snapshot of road network, the junctions are considered as Grid Nodes. Road is a connection between two nodes. These are the services or resources available to nodes. These resources will have the capacity. The roads are utilized by the vehicles which will occupy the road capacity. Therefore the current available capacity of the road will be varying which leads to dynamic nature. The vehicles are the users and mobiles in the road network. These will form the load and utilizes the capacity of resources.

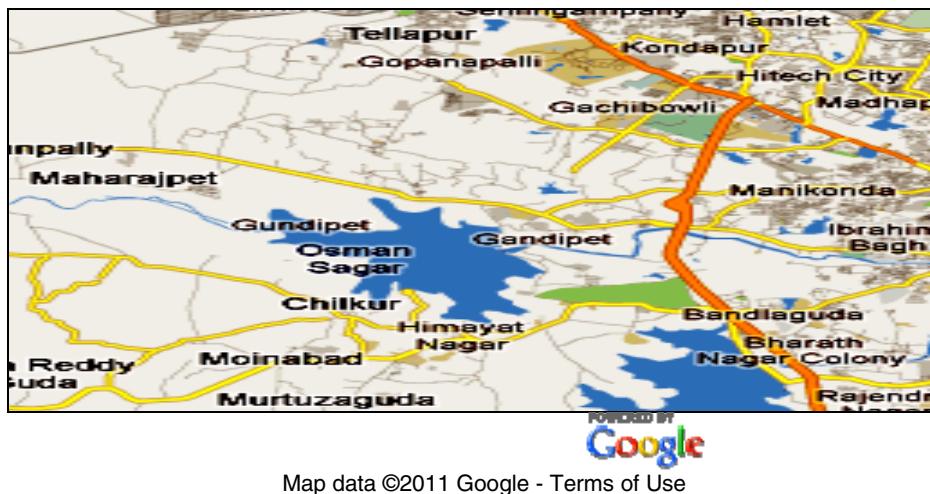


Fig. 1. Snapshots of Road Maps

The Grid nodes will maintain the information about inflow from each resource connected to it and outflow to each direction i.e. load information, the destination of each vehicle reaching to the node, the next hop which the vehicle is going to take to reach the destination region. The information present in the Grid nodes is published to GRAM of Grid Server. The GRAM pushes the information available with it to all the nodes present in the Grid System. This helps the nodes to decide about the best hop and alternate hop to be taken by each vehicle accurately to avoid the formation of congestion at other nodes. Therefore GRAM will help the Grid nodes to take the decision appropriately by taking into consideration about the current load at different resources.

Each vehicle has to maintain the data about vehicle no., source region, intermediate hop (dynamic, which will be changing depending on best hop or alternate next best hop taken to reach the destination), and destination region.

The following table shows the details at different instances about the source region from which the vehicle has started, intermediate hop it has taken at different junctions, the destination region.

Table 1. Intermediate hops of a vehicle at different instances

Time	Source	Intermediate Hop	Destination
T1	JS	JS	JD
T2	JS	JA	JD
T3	JS	JB	JD
T4	JS	JC	JD
T5	JS	JD	JD (Destination reached)

4 Problem Analysis and Proposed Approach

4.1 Problem Analysis

In this problem a network of crisscross roads with junctions as Grid nodes to which different roads or resources connected considered. The different vehicles will have different destination. Here instead of a specific destination, destination region is considered. There may be many situations which lead to congestion. The situations may be accident, rallies, peak hours etc.. In such cases, the congested road deters the vehicles from travelling at their desired speed. Therefore, the distribution of vehicles should be done in an optimal way, where for a vehicle even though, the alternate best hop road is allotted, it can able to reach the destination with desired speed in time, rather than going through the best hop at less speed which may take more time to reach the destination when it is congested.

4.2 Proposed Approach

The Grid node will analyze the data with the help of information maintained at it and the load information pushed by GRAM to the Grid node. It does this by following steps.

- a. Clustering
- b. Scheduling
- c. Communication/updating the information to nearby nodes and GRAM.

The data collection about the resources or links connected to a Grid node can be done using GPS system where vehicles are equipped by GPS terminals or from multiple distributed data sources such as traffic lights, sensors, induction loops and video cameras or through the sensor network model proposed by Feilong Tang, Minyi Guo, Minglu Li and Cho-Li Wang where homogenous vehicles, buses are taken into consideration. It is assumed that sensors are deployed in each bus [1,3].

Clustering

When the vehicles are accumulated at nodes from different resources, the process at the node will cluster the vehicles, depending on the criteria of destination region of the vehicles.

Scheduling

The GRAM will publish the information present at it to the Grid nodes i.e. about the load information of different resources connected at each Grid node and the road

network graphs. This information is used by the Grid nodes to find the best hop. The best hop is determined by using the criteria about the load information and then the distance. For example, let JA and JB are the two junctions through which destination region may be reached by a cluster of vehicles with destination region as JD. If JA is used the distance, time to reach it and load information of the resource used to reach it is 7 km, 15 minutes and semi-occupied and if JB is used it is 8 km, 18 minutes and free, then JB is selected as the best hop and the vehicle is scheduled to go to JB. The priority taken here is in the order of load, time and then distance. If the information publishing about JA and JB to the current Grid node is not present, then it would have selected JA as the best hop which may result to congestion at JA very soon in future.

If more than one node is scheduling the cluster of vehicles present at them to JB simultaneously as the JB is free, it may lead to amalgamation of clusters which is shown in figure 2. Therefore, the requirement to split into sub-clusters arises. For this ST-DBSCAN or ST-GRID can be adapted [5]. After splitting rerouting of one of the sub clusters to alternate best hop will take place by analyzing the nodes using the criteria of load information, time to reach and distance by adapting the Split Multipath Routing protocol [6] as shown in Figure 3.

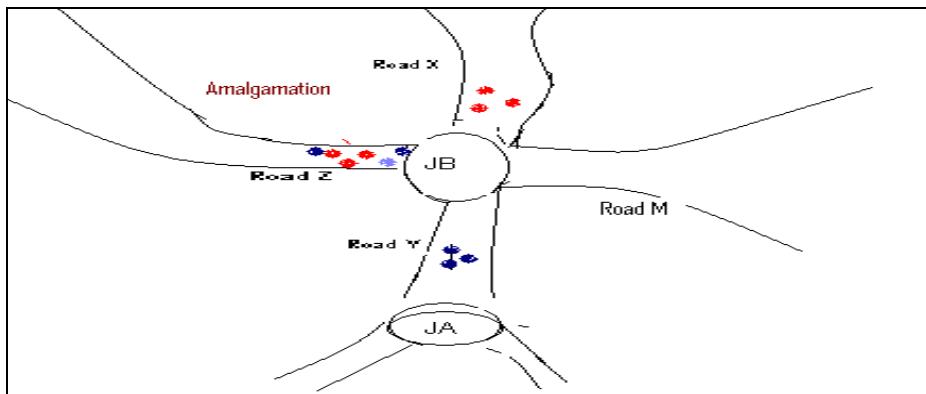


Fig. 2. Amalgamation of Clusters

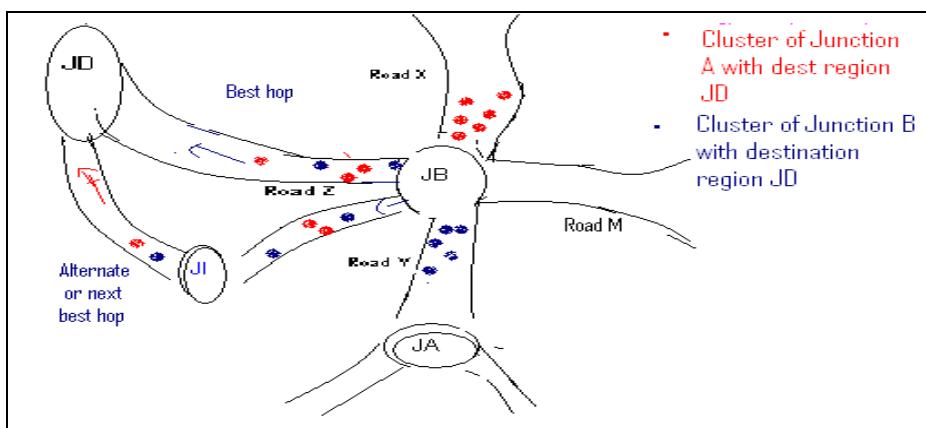


Fig. 3. Splitting the Clusters into Sub-Clusters

Communication/Updation of information

After scheduling of the cluster of vehicles, the current load information at each Grid node may change and this should be communicated to other Grid nodes and published to GRAM.

5 GFTMS-An Illustrative Example

In the following illustration the vertices A, B, C, D and E are the junctions which are treated as Grid Nodes. Let B is considered as the Grid Server and the other nodes are treated as Grid Clients. Let Source node or current node is 'E' where the vehicles from different resources are reaching to E. The Grid Node E will maintain the information about vehicles, the capacity of the roads connected to it. The Grid node will now group the vehicles according to the criteria of destination region. Let 10 vehicles has to reach C and 50 vehicles has to reach B. The 2 clusters let CLC and CLB are created using the destination region C & B. The information present initially in the CLB cluster of vehicles is

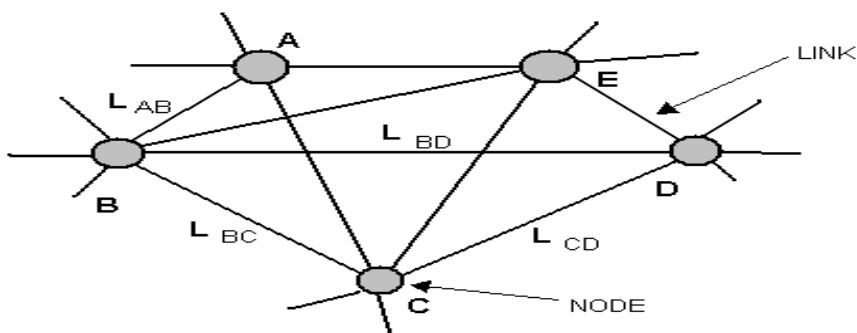


Fig. 4. Nodes and their Links

Table 2. Initial Status of Cluster B Vehicles

Time	Source	Intermediate Hop	Destination
T1	EJ	-----	BJ

And the information present in the CLC cluster of vehicles is

Table 3. Initial Status of Cluster C Vehicles

Time	Source	Intermediate Hop	Destination
T1	EJ	-----	CJ

The Grid node E will have the information about current load of different resources connected to it, their maximum capacities, threshold value, distance, and time to reach. The current load of any resource can be calculated using induction loops [3]. If LA is 5 veh/km, 10veh/km, 8veh/km, 7km, 10min, LB is 10veh/km, 15veh/km, 12veh/km, 15km, 30min, LC is 8veh/km, 20veh/km, 15veh/km, 20 km, 40min and LD is 3veh/km, 15veh/km, 12veh/km, 7km, 10min, then for CLC cluster of vehicles the links or resources LB and LD can be used and for CLB vehicles the links or resources LB and LA can be used. After clustering the current resource information is published with the Grid Server. For Eg. Grid node E publishes the information present at it to Grid Server B is as shown in the following table.

Table 4. Information published by Grid Node E to B

Links	Curr. Load	Max. Capacity	Threshold	Distance	Time to reach(if not congested)
LA	5	10	8	7	10
LB	10	15	12	15	30
LC	8	20	15	20	40
LD	3	15	12	7	10

In the same way all other Grid nodes also publish their information to Grid Server. The Grid Server now will help the Grid node(s) in taking the decision of scheduling to allocate the best resources to CLC and CLB cluster of vehicles by publishing the information present with it to Grid client E. Now Grid client will use the load information of the other Grid Nodes received from the Grid Server to take the decision which will not again result in future congestion in the following hops. For Eg. let Grid Server has the load information from A is LA to LB is 6veh/km and LA to LE is 5veh/km, from C is LC to LB is 10veh/km, LC to LE is 8veh/km.

Current Load Information for different links present at Grid Server B is shown in the following Table 5.

Table 5. Current load information at different links

Links	LA	LB	LC	LD	LE
LA	-	6	12	-	5
LB	6	-	10	5	10
LC	12	10	-	4	8
LD	-	5	4	-	3
LE	5	10	8	3	-

The above information is published to Grid Client E, it will now decide about the best hop for CLC and CLB cluster of vehicles.

The path for CLC vehicles can be EDC or EC. In this illustration only two paths are considered i.e. EDC or EC even though other paths are available from E to reach C to reduce the complexity. For EDC, E-D and D-C links had the 3veh/km,

15veh/km, 12veh/km, 7km, 10min and 4veh/km, 15veh/km, 12veh/km, 2km, 8min respectively in the order of load, maximum capacity of the resource, threshold value, distance, and time to reach.

For EC path it is 8veh/km, 20veh/km, 15veh/km, 20km, 40min in the order of load, maximum capacity of the resource, threshold value, distance, and time to reach.

While scheduling the best hop the priority is taken in the order of load, distance, time to reach the next hop and also it is seen that it will not exceed the threshold value of the resource otherwise even though the selected hop may be the best in terms of load, distance and time it further may leads to congestion in future. The assumption used here is the allowable current load can be little greater than the threshold value which can be given by a constant. And also it is assumed that the vehicles are travelling at constant speed.

$$\text{Curr-load} \leq \text{threshold-value} + C$$

C is a constant whose value should be very less. If it is more, the Curr-load may reach closer to the maximum capacity which may lead to congestion very soon.

Table 6. Updated Path Information with C=2 (If Selected)

Paths	Curr-load	Distance	Time	Threshold
EDC	3+4	7+2	10+8	12+12
EC	8	20	40	15

As 10 vehicles has to reach the destination region CJ the current load of EDC path increases by 10 and the load will be $3+4+10=17$ and the threshold value accepted is 24. The current updated load i.e.17 is not exceeding 24 therefore DJ node can be selected as the best hop. It will reject the EC path as the new updated load will be $8+10=18$ and the threshold value it accepts is only 15.

As the next best hop is through the link LD the next best hop can be DJ. At the Grid Node DJ, current load information for the link LD is updated as $3+10=13$ veh/km. This information is further communicated to Grid Server and to other Grid Clients for supporting the future decisions taken by them at their nodes. The current status of CLC cluster of vehicles is

Table 7. Status of Cluster C Vehicles at T2 instance

Time	Source	Intermediate Hop	Destination
T2	EJ	DJ	CJ

In the same way for CLB cluster of vehicles also the next best hop is decided.

The same procedure is followed in grouping, scheduling and communicating at the other Grid Nodes for the vehicles which use the resources.

While doing the scheduling at different Grid nodes there is a chance of amalgamation of clusters. Then, at that time the cluster of clusters can be splitted and instead of sending all the clusters through the best hop which may lead to congestion. Some sub

clusters are sent through the best hop and other sub clusters may be sent by finding the alternate best hop. For this the Split Multi Path Routing Protocol may be adapted [6].

6 Tools

Tools which can be useful while working with this problem are AIMSUN2 is a Advanced Interactive Microscopic Simulator for Urban and Nonurban Network for modeling the traffic scenarios, GETRAM includes animated simulation display which shows vehicles moving through the network, OpenStreetMap provides detailed information about the road network and the zones of activity in the country, VANET Simulators for simulating Vehicular Adhoc Networks [3, 4].

7 Conclusion

Proposed paper is a theoretical approach which can solve the problem of congestion by taking the decision to distribute vehicles optimally and acts like an admission control. Realization and implementation of this approach definitely be beneficial to reduce congestion and to control and release the traffic according to the roads capacities. The assumptions made here are by considering unidirectional and single channel roads with some capacity. It is assumed that all the vehicles are travelling at the same constant speed once the resource through which they have to travel is identified and scheduled. The approach can be extended to multichannel and bidirectional roads with heterogeneous vehicles and a mathematical model can be developed.

Acknowledgements. I would like to express my gratitude to my Supervisors and DRC members Prof. Arun Agarwal, Department of Computers and Information Science, University of Hyderabad and Dr. Atul Negi, Department of Computers and Information Science, University of Hyderabad for their valuable suggestions and comments which helped to think further and write this paper. I sincerely want to extend my thanks for my supervisors for spending their precious time and for encouraging me to proceed further with this idea presented in this paper. I would like to wholeheartedly thank the management of Keshav Memorial Institute of Technology for providing the facilities to pursue my Ph.D. at University of Hyderabad.

References

- [1] Tang, F., Guo, M., Li, M., Wang, C.-L.: Implementation of an Intelligent Urban Traffic Management System Based on a City Grid Infrastructure. *Journal of Information Science and Engineering* 24, 1821–1836 (2008)
- [2] Mukherjee, S., Pan, I., Dey, K.N.: Traffic Organization by utilization of resources through Grid Computing Concept. In: 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC 2009). IEEE (2009)
- [3] Pigne, Y., Danoy, G., Bouvry, P.: A Vehicular mobility model based on real traffic counting data. In: Proceedings of the Third International Conference on Communication Technologies for Vehicles. Springer, Heidelberg (2011)

- [4] Hughes, J.T.: AIMSUN2 Simulataion of a congested Auckland Freeway. In: 6th EURO Working Group on Transporation, Goteberg, Sweden
- [5] Kisilevich, S., Mansmann, F., Nanni, M., Rinizivillo, S.: Spatio-Temporal clustering: a Survey. Technical Report, ISTI-CNR, Italy, Submitted to Data Minig and Knowledge Discovery Handbook, vol. 6, pp. 855–874. Springer (2010)
- [6] Lee, S.-J., Gerla, M.: Split multipath routing with maximally disjoint paths in Ad hoc networks. *IEEE Communications*, 3201–3205 (2001)
- [7] Kunszt, P.Z., Guy, L.P.: The open grid services architecture and data grids. In: Berman, F., Fox, G.C., Hey, A.J.G. (eds.) *Grid Computing: Making the Global Infrastructure a Reality*. Wiley series in Communications, Networking and Distributed Systems, pp. 9–50. Wiley, Chichester (2003)
- [8] Cameron, D.G., Carvajal-Schiaffino, R., Millar, A.P., Nicholson, C., Stockinger, K., Zini, F.: Evaluating scheduling and replica optimizsation strategies. In: Proceedings of the Fourth International Workshop on Grid Computing (GRID 2003), pp. 65–77 (2003)

Authors



V. Anitha, Pursuing PhD at University of Hyderabad, Gachibowli. Areas of Interest are Grid Computing and Intelligent Transportation Systems. Presently working as an Assoc. Professor in Keshav Memorial Institute Technology, Narayanguda, Hyderabad.



Dr. Rajeev Wankar, Reader, DCIS, University of Hyderabad. Gachibowli. Areas of Interest are Parallel and Grid Computing and Analysis of Algorithms.



Dr. C. Raghavendra Rao, Professor, DCIS, University of Hyderabad. Gachibowli. Areas of Interest are Simulation and Modelling and Knowledge Discovery.

Intrusion Detection and QoS Security Architecture for Service Grid Computing Environment

Raghavendra Prabhu*, Basappa B. Kodada, and K.M. Shivakumar

Dept. of Computer Science and Engineering

Canara Engineering College Benjanapadavu

{rghprabhu7, basappabk, shivakumar2555}@gmail.com

Abstract. Grid Computing is information technology which used to share resources across the global to solve the large scale problem. It is based on networks to enable large scale aggregation and sharing of computational, data, sensors and other resources across global. Grid Computing Environment provides the services like Job Executing Environment and web services as well. So Grid Computing Environment should be secured from the outside and inside intruder. Grid Computing is a Global Infrastructure on the internet has led to a security attacks on the Computing Infrastructure. The wide varieties of IDS (Intrusion Detection System) are available which are designed to handle the specific types of attacks. No technique can give QoS along with IDS. So this paper proposes a Mobile Agent-based Intrusion Detection System (MA-IDS) architecture, is a secured architecture to provide the security, maximizing the user's benefits and Quality of Service (QoS). The Most Benefit Travelling Salesman Problem (MBTSP) is introduced to describe how the Agent acts in this model by using optimized routing algorithm.

Keywords: Service Gird, Intrusion Detection System, MA-IDS, OGSA, QoS.

1 Introduction

Grid service is one of the key technologies to reuse of resources and achieve high efficiency in service grid. A service grid, in service architecture, combines the Grid service and the Web service together. In a service grid, all resources including calculation, communication and memory, are in forms of services. With the growing use of internet, attackers have become more and more active in identifying the flaw of the application or Operating system connected to the network protocols are able to make the attacks on the network resources to make the damage on the network system or Application running in the system. In this paper we discuss about the Intrusion Detection System (IDS) and Resource management and scheduling in a grid environment and propose Mobile Agent-based Intrusion Detection System (MA-IDS) to detect threats and to maximize the user's benefit as well as assure QoS requirements.

Grid infrastructure can be divided into three layers: the computational grid, the data grid and the service grid. Agents can provide useful abstraction at each of the

* Correspondent author.

three grid levels. Mobile agents have the features of autonomy, mobility, intelligence, cooperation and security. Applying Mobile Agents (MA) to the service grid can make it possible to avoid the shortcoming of inflexibility and need to be controlled caused by the process migration mechanism, which is traditionally used in heterogeneous computing systems, and enhances the adaptability of grid application to a complex and changeable grid environment.

Intrusion Detection Systems have a very important role in the Grid Computing Environment For the execution of large scale application or in service grid there is clearly need to detect the known or unknown intrusion and any other kind of dangerous events. At the same time it should provide the Quality of Services to maximize the users benefits and grid service should not be denied as well.

Rest of the paper organized as follows. Section 2 of this paper contains the related work of IDS in the Grid Computing Environment. Section 3 explains about the Proposed Architecture of Mobile Agent Intrusion Detection System (IDS). Section 4 of this paper proposes the Implementation of Security Architecture of service Grid-IDS and Section 5 explains the different QoS algorithms that are used in providing better QoS to the user and finally Section 6 concludes this paper and presents the future work.

2 Related Work

Grid Computing has many security mechanisms by integrating into Grid Security Infrastructure (GSI) which offers basic authentication and secure communication based on the X.509 certificates for authentication. [1] provides an agent method to IDS respectively, but they could not resolve the problems caused by the heterogeneity and dynamic of the Grid; [2] provides distributed IDS based on data fusion method, but it lack the ability of automatic reorganization. [11] Has proposed the Intrusion Detection System at the node level but it does not provide the QoS.

Leu et al. [3] developed a performance-based grid intrusion detection system (PGIDS) which exploited grid's abundant computing resources to detect logical, DoS and DDoS attacks real-time so that the drawbacks that traditional IDSs suffer were then eliminated. However, PGIDS is performed on a static environment. Its detection flexibility is limited. Park and Lee in 2002 [4] raised a route-based packet filtering (RPF) approach checking whether each packet comes from a correct link and source. Moreover, many IDS prototypes have been developed in recent years, such as distributed attack detection (DAD) [8], Multicast Intrusion Detection and Alerting System (MIDAS) [6] and Distribution Intrusion Detection System (DIDS) [7].

Zhan Gao et. al [10] said that Grid computing is a kind of important information technology, which enables resource sharing in a wide area and system integration. Resource management and scheduling in a grid environment is a challenging problem worth researching. A Mobile Agent-based service scheduling model (MASSM) in the service grid is presented, which aims at maximizing the user's benefit. The Most Benefit Travelling Salesman Problem MBTSP) is introduced to describe how the Agent acts in this model and a routing algorithm is also proposed. ASSM is adaptable to the dynamic environment in service grid and able to maximize the user's benefit as well as assure different QoS requirements.

Jin et al. [5] proposed a fault tolerance mechanism to handle three types of aspects a grid faults, node crash and agent crash based on java exception handling, java threads state capturing technique and mobile agent technology. If grid nodes failed, necessary software components were transferred to other location through mobile agent or other mechanisms. Applications were then reconfigured to provide uninterrupted services. Part of the system-level components that failed was also corrected by this mechanism. Finally grid node that crashed was dynamically detected. User tasks specified with fault-tolerant sign would be processed correctly.

In a Grid level intrusion detection system is presented but it offers a very complicated system, without an emphasis on performance or the reduction of the number of sent messages. It describes the problem and the need for a higher level of intrusion detection but doesn't express the importance of having a real-time image of the applications running on the Grid to be able to detect the more advanced types of attack. Another important related work is the one presented which offers a good solution for Denial-of-Service and Distributed Denial-of-Service attacks. Also, it presents a type of Grid Intrusion Detection System but it does not provide a solution to other types of attack. Hence this paper proposes the security architecture to prevent the all type of known attacks and future attacks as well.

3 MA-IDS for Grid Computing

Mobile Agent based Intrusion Detection System (MA-IDA) combines the modules of service grid – Intrusion Detection System [9] and the mobile agent model to give better performance [10]. The MA – IDS contains the Grid-IDS and Mobile agent architecture at the node level. The Grid-IDS will detect the all type of attacks at grid and node level as well. It will provide a secure and high performance and better Quality of Service depending upon the user needs in service grid computing.

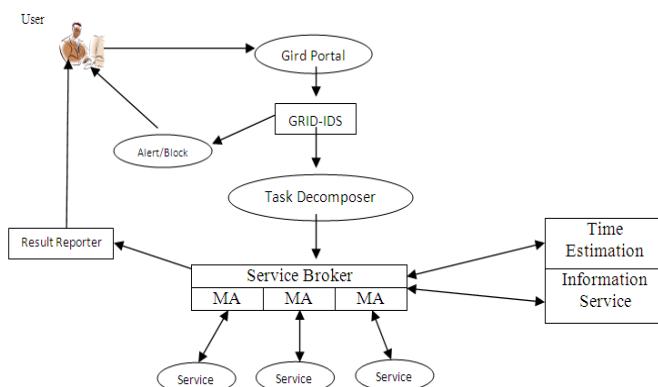


Fig. 1. MA-IDS (Mobile Agent – Intrusion Detection System) with QoS

In the figure 1 grid user submits his application through the grid portal along with his requirement of the application's deadline, the least acceptable benefit, and the time-sensibility of the application. The application processes through Grid-IDS to detect all types of malicious attacks which could be known attack or unknown attacks. So the intrusion detection system (IDS) for Service Grid should be a system which could dynamically integrate with node and detect the resources of a Grid Computing to execute application by ensuring the Resource security of Grid Computing. If any type of attack is detected by the Grid-IDS then the alert message will be generated to the node as well as Attacker. Once the Request is attack free application will move to the Task Decomposer is responsible for the analyzing the application and decomposition of it into different requirements of services, which are provided by different sites in the service grid. The required services will be filled into a service list, which will be delivered to Task Broker.

Task Broker consists of Mobile Agents (MAs), each of which has a service list. After the arrival of an application, Task Broker will allocate an available MA (one not serving other applications) for it, whose service list can be obtained from Task Decomposer. Then it will work on behalf of the corresponding user. In order to generate a schedule meeting the user's QoS requirement as well as maximizing the benefit, the MA needs the help of Information Service and Time Estimator. Information Service is composed of Service State, Service Price and Service Capability. Service State is responsible for providing information about a service, for example the availability, queue length and expected wait time. The upto date information about a service's function and capability is stored in Service Capability and price in Service Price, respectively. If there is any change in a service's state, price or capability, the current active MAs as well as Time Estimator will be notified. So the MAs can adapt their schedules to such changes in time. Time Estimator consists of Execution Estimator and Transition Estimator, by which the time needed for a MA to migrate between different sites and the time needed for a site to complete a service is estimated respectively. Up to now there has been a lot of researches into the problem of grid transition time estimation and execution time estimation and corresponding software's have been developed [12], [13].

Result Reporter provides a means by which a grid user can monitor his applications. A user can use the commands or APIs provided by Result Reporter to query or control the execution of his applications. And after a MA has accomplished its task, it can notify the user through Result Reporter, which includes two cases. On one hand if a MA succeeds, it will inform the user about the total cost and time of its predetermined route. Then the user will evaluate the result and he can modify his beneficial function if not satisfied. On the other hand if a MA fail (not meeting the user's QoS requirement), it will notify the user of this failure and the user can decrease his QoS requirement. With the help of Information Service and Time Estimator a MA uses a routing method to determine the most beneficial route for a certain user and our routing method will be presented in the following section.

4 GRID-IDS

Figure 2 summarizes the architecture of the Grid – Intrusion Detection System (Grid-IDS). The Sampler randomly/heuristically picks up sample packet windows (series of

contiguous packets) and sends them to the Network Packet Analyzer component. The Analyzer and the preprocessing engine analyze the packets and convert them into a standard XML format by stripping the network and DLL headers. This metadata is sent for processing to the next component i.e. the “Rules Engine” which can be an OGSA (Open Grid service Architecture) component. The Rules engine is a OGSA enabled component of the application that facilitates the XML packet to be checked for anomalies against suspicious activities and predefined business rules. This component should be able to detect packets from invalid/entrusted IPs and domains. DoS attacks, Filtering, Screening, Authentication, Trust, etc. related issues can be addressed at this component. The Rules engine should be enabled to allow the organization to implement and customize the rules based on the location of the IDS on the network. Rules must be classified as preemptive/non-preemptive.

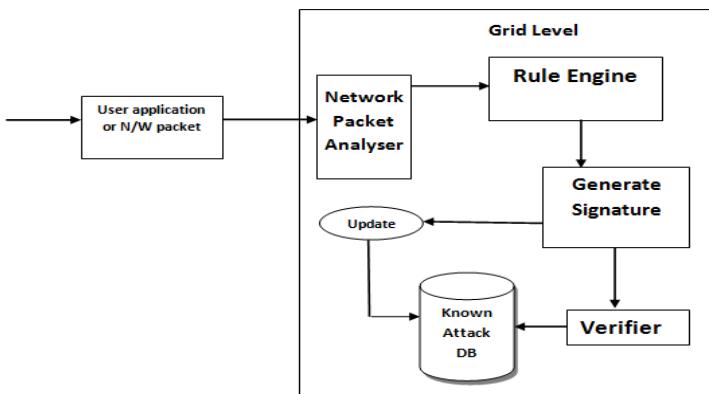


Fig. 2. Grid – Intrusion Detection System (Grid – IDS)

The rules engine upon detecting anomaly will automatically forward to alert agent component or manual intervention component. If directed to alert agent component then the alerts are audited, logged, and mailed to concerned authorities. If all packets in a sample packet window are cleared by Rules Engine, then the packets go for a check of known attack signatures to the verifier. The Verifier component checks the packets against attacks picked from a local signature known attack database. This DB is pre-populated from external and publicly known signatures and other IDS instance detected signatures. If the verifier detects the known attack signature then it directed to alert agent component then the alerts are audited, logged, and mailed to concerned authorities otherwise the packet will be accepted to continue. If the rule engine detects the anonymous attack, then Gen-sig will generate the signature and the updater then picks up these XMLs and their packet payloads and digests them using fast and compressive hashing algorithms that compact this information and store it in the local signature DB for the known attacks.

5 QoS for Grid Computing Environment

Two algorithms are used to provide QoS for Grid Computing Environment. The Most Benefit Travelling Salesman Problem (MBTSP) & Most Benefit Routing Algorithm (MBRA) which are described next.

5.1 Most Benefit Travelling Salesman Problem (MBTSP)

The traditional Travelling Salesman Problem (TSP) is defined as follows: Given n cities with different distances between each other, salesmen will traversal all the cities and return to the city he starts from. TSP arises from finding the shortest route for the salesman. In MASSM the MA routing problem is not the same as TSP and has its own features:

- There may be many sites providing the same service, varying in service capability and service price.
- A MA do not need to traversal all sites and it only needs finish its task.
- Different MAs may have different beneficial functions.
- For every site the status and capability of its service may change.
- In the dynamical grid environment, the load of network may change and any service provider may join or quit at any time.
- A MA's routing target is the smallest total costs not the shortest length.

In order to describe the MA's routing features in MA–IDA, we now define MBTSP as follows: Given n cities with different distances between each other, every city needs an amount of certain kinds of goods and places a tax on them. A salesman departs from one city to visit others unpeated, carrying different kinds of goods. And travelling between different cities will cost him different fares. MBTSP arises from finding a travelling route for the salesman to sell out all his goods and obtain the most profit.

5.2 Most Benefit Routing Algorithm (MBRA)

In MA–IDA, the routing problem for the MA is similar to MBTSP if we take the MA as the salesman and the services as the goods. Due to different distances and bandwidths between different sites, it takes a MA different time to migrate from one site to another. To maximize the user's benefit the MA is supposed to find a route which costs as little time as possible. Except for the factor of time, the MA should also take the service price into account. Because the shortest route may pass sites with the highest prices and it cannot be the best choice to maximize the user's benefit.

In MBRA, MAs migrate under the direction of evaluating different paths. In the beginning the path between every two sites has an initialized selection value. After leaving one site, the MA will delete from its service list the service provided by the site and select the next site with a certain probability. When the service list is empty, the first loop comes to the end. The MA will evaluate all the paths using an evaluating function and change their selection values. Then the MA recovers its service list and the second loop begins. In the second loop the MA uses the selection values produced in the previous one to decide which path to select. After the second loop, the MA

evaluates all the paths again and changes their selection values. After recovering its service list, the MA goes on for the third loop. When the whole migration route selected by the MA doesn't change or the iteration time has reached a predefined number, we believe that the algorithm has converged at an optimal route for the MA.

6 Conclusion

The proposed architecture can detect all types of attacks which could be known attacks or unknown attacks efficiently and also provide better quality of service which aims at maximizing the users' benefit and guaranteeing different QoS. It allows new computing resources and services to be added dynamically and also previous unknown attacks will become known attacks by updating the new attack signature to known attack database. We plan future work in three major areas. First, we will address problems of service advanced reservation and co-allocation in MA-IDS and this is necessary when the MA needs several services at the same time, which is not involved in current MA-IDS. Future work is to implement for the Performance evaluation to interface this architecture into the grid computing environment.

Acknowledement. We are very thankful our Beloved Principal, Department and Faculties Ramesh Nayak, Suresha D., Hemalatha V., Namratha Padiyar, Alok Ranjan for their valuable help, suggestions, guidance and encouragement.

References

- [1] Li, J., Zhang, G.-y., Gu, G.-c.: A Multi-agent Based Architecture for Network Attack Resistant System. In: Li, M., Sun, X.-H., Deng, Q.-n., Ni, J. (eds.) *GCC 2003, Part I*. LNCS, vol. 3032, pp. 980–983. Springer, Heidelberg (2004)
- [2] Wang, Y., Yang, H., Wang, X., Zhang, R.: Distributed Intrusion Detection System Based on Data Fusion Method. In: Proceedings of the 6th World Congress on Intelligent Control and Automation (2004)
- [3] Leu, F.Y., Lin, J.C., Li, M.C., Yang, C.T.: A Performance-Based Grid Intrusion Detection System. In: Proc. of IEEE Annual International Computer Software and Applications Conf., pp. 525–530 (July 2005)
- [4] Park, K., Lee, H.: On the Effectiveness of Route-Based Packet Filtering for Distributed DoS Attack Prevention in Power-Law Internets. In: Proc. ACM SIGCOMM, pp. 15–26 (August 2001)
- [5] Jin, L., Tong, W.Q., Tang, J.Q., Wang, B.: A Fault-Tolerant Mechanism in Grid. In: Proc. of IEEE International Conference on Industrial Informatics, pp. 457–461 (August 2003)
- [6] Sebring, M.M., Shellhouse, E., Hanna, M.E., Whitehurst, R.A.: Expert Systems in Intrusion Detection: A Case Study. In: Proceedings of the Eleventh National Computer Security Conference, Washington, D.C. (October 1988)
- [7] Snapp, S.R., et al.: DIDS (Distributed Intrusion Detection System) – Motivation, Architecture, and An Early Prototype. In: Proceedings of the Fifteenth National Computer Security Conference, Baltimore, MD (October 1992)
- [8] Wan, K.K., Chang, R.: Engineering of a Global Defense Infrastructure for DDoS Attacks. In: Proc. IEEE Int'l. Conf. Net. (August 2002)

- [9] Kodada, B.B., Prasad, M.: Security Architecture for Building IDS in the Service Grid Environment. International Journal of Computer Science issues (accepted, 2012)
- [10] Gao, Z., Luo, S., Ding, D.: A Service Scheduling Model in the Service Grid Environment. In: The Sixth International Conference on Grid and Cooperative Computing (GCC 2007). IEEE Computer Society (2007) 0-7695-2871-6/07
- [11] Kodada, B.B., Nayak, R., et al.: Intrusion Detection System – Inside Grid Computing Environment (IDS-IGCE). (IJGCA) International Journal of Grid Computing and Application (2011)
- [12] Vazhkudai, S., Schopf, J.M.: Predicting sporadic grid data transfers. In: Proceedings of the 11th IEEE International Symposium on High Performance Distributed Computing (HPDC 2002) (2002)
- [13] Nudd, G.R., Kerbyson, D.J., Papaefstathiou, E., Perry, J.S.C., Wilcox, D.V.: Pace — a toolset for the performance prediction of parallel and distributed systems. Int. J. High Performance Computing Applications, Special Issues on Performance Modelling 14(3), 228–251 (2000)

Service Composition Design Pattern for Autonomic Computing Systems Using Association Rule Based Learning

Mohammed A.R. Quadri¹, Vishnuvardhan Mannava¹, and T. Ramesh²

¹ Department of Computer Science and Engineering,
K.L. University, Vaddeswaram, 522502, A.P., India

mohammad.ataulla@gmail.com, vishnu@kluniversity.in

² Department of Computer Science and Engineering,
National Institute of Technology, Warangal, 506004, A.P., India
rimesht@nitw.ac.in

Abstract. The adaptability in software is the main fascinating concern for which most of the software architects today are really interested in providing the Autonomic computing. In order to provide remedy for the service failures that occurs at the servers of the respective service providers, there is a need to introduce the self-reconfiguration planes to be applied astronomically without the interruption of the administrator to solve the problem manually. Different programming models have been introduced for providing the dynamic behavior of the services being provided. Few among them are the Aspect Oriented Programming (AOP) and Feature Oriented Programming (FOP) both of them having the ability to modularize the crosscutting concerns, where the former is dependent on aspects, advice and lateral one on the collaboration design and refinements. In this paper we will use the design patterns which will satisfy the properties of autonomic computing system: for the Decision-Making phase we will introduce Case-Based Reasoning design pattern, and for Reconfiguration phase we will introduce Reactor design pattern. The most important proposal in our design pattern is that we will use the Association Rule Learning method of Data Mining to learn about new services that can be added along with the requested service to make the service as a dynamic composition of two or more services. Then we will include the new service as an aspectual feature module code without interrupting the user. The pattern is described using a java-like notation for the classes and interfaces. A simple UML and Sequence diagram are depicted.

Keywords: Autonomic System, Design Patterns, Aspect-Oriented Design Patterns, Feature-Oriented Programming (FOP), Aspect-Oriented Programming (AOP), Data Mining.

1 Introduction

The most widely focused elements of the autonomic computing systems are self-* properties. So for a system to be self-manageable they should be self-configuring, self-healing, self-optimizing, self-protecting and they have to exhibit self-awareness,

self-situation and self-monitoring properties [1]. For providing dynamic behavior in currently developed systems some of the newly introduced programming language features are required. The most popular and interesting research area in providing dynamic adaptability in today's programming world is with Aspect-Oriented Programming (AOP) [2][3] and Feature-Oriented Programming (FOP) [4]. Design patterns are most often used in developing the software system to implement variable and reusable software with object oriented programming (OOP) [2]. Most of the design patterns have been successfully developed in the OOPs, but at the same time developers have faced some problems like as said in [6] they found the lack of modularity, composability and reusability in respective object oriented designs [7]. The cause of problem in Applying OOPs in developing design patterns was found due to the crosscutting concerns. Crosscutting concerns are the problems that result in code tangling, scattering, and replication of code when software is decomposed along one dimension [8]. So in order to overcome this problem some advanced modularizing techniques have been introduced like AOP and FOP. With the help of Aspect-Oriented Programming we can separate the crosscutting concerns from the main functional logic. We can handle these crosscutting concerns in separate Aspect and pointcut like modules. On the other hand the Feature-Oriented Programming has made the way for research in software-product lines with the inclusion of new features in already developed software as a separate Feature Module by making the life of developers easier by not develop again from scratch. In this paper we will propose a design pattern that will develop an autonomic system that will use Case-Based Reasoning Design Pattern [9] for Decision-Making, and for Reconfiguration phase we will use Reactor Design Pattern [10]. With the help of the Aspects Feature Module based technique proposed in [3] we can include the new service into the existing system as a new Feature. Here we have concentrated on the Learning process with the help of Data Mining methods. The Association Rule Learning in data mining is the main concept that we have used to create a Dynamic composition services from the requested service. To understand this let's look at an example, if a customer visits a site and requests for a service say he buys a computer and then after that he will also buy the UPS like this different customers who buy the computer will also buy an UPS and also printer. So with our pattern these transactions are stored in the Learning data base which will be used by the Association rule method of data mining to determine a rule like {computer, UPS} => {printer} means the customer who buys a computer and at the same time UPS, he will also buy a printer, with the help of Association rule data mining instead of requesting the three items individually we can dynamically compose service which is a composition of three services as a single service. So when a customer who comes to buy the three items at a single request can just access this composed service and no need to request three items separately. Then we will just include this new composed service a Feature into the Service Repository so that the plan to take decisions regarding like this services will be easy. By providing the Learning process with association learning rule of data mining we can dynamically reconfigure the system to be adaptive to the new services. Refer Fig 1 for the Autonomic computing system in proposed in [9].

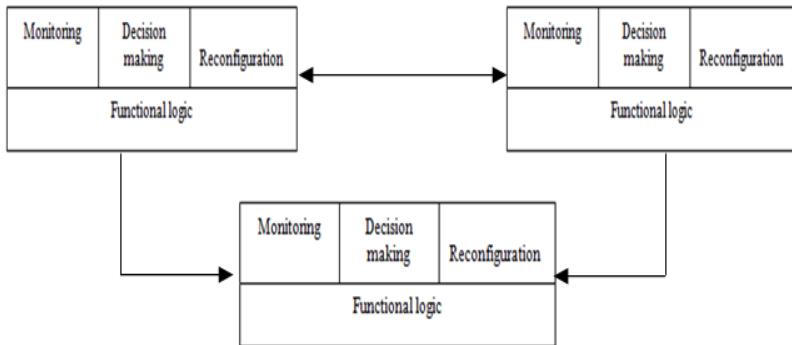


Fig .1. Autonomic Computing

2 Related Work

In this section we present some works that deal with different autonomic systems design. There are number of publications reporting the adaptive nature of the systems where changes occur depending upon the environments in which they are deployed. They provide the ability to monitor, to make decisions and to reconfigure at run-time.

In M.Vishnuvardhan and T.Ramesh paper [11] discuss applying the Adaptive Monitoring Compliance Design Pattern for autonomic systems. The authors of the paper uses adaptive design pattern called adaptive sensor factory have been proposed to make the monitoring infrastructure of the adaptive system more dynamic by fusing the sensor factory pattern, observer and strategy patterns. This pattern will determine the type of sensor that suits best for monitoring the client.

In Sven Apel, Thomas Leich, and Gunter Saake [3] they proposed the symbiosis of FOP and AOP and aspectual feature modules (AFMs), a programming technique that integrates feature modules and aspects. they provide a set of tools that support implementing AFMs on top of Java and C++.

In Olivier Aubert, Antoine Beugnard [12] they proposed an Adaptive Strategy Design Pattern that can be used to analyze or design self-adaptive systems. It makes the significant components usually involved in a self-adaptive system explicit, and studies their interactions. They show how the components participate in the adaptation process, and characterize some of their properties.

3 Proposed Autonomic Design Pattern

One of the objectives of this paper is to apply the Association rule based Learning method of Data Mining to determine the new service and include it as a Aspectual Feature-oriented code into the existing system. Initially the customer will request for a service then the Trigger class will trigger an event to the server. With the help of the Case-Based Reasoning Design Pattern [9] we can take a decision that will decide which plan should be applied to provide the service to the user. Here the Decision is taken based on the predefined rules in the Fixed Rules class then these rules are given as input to the Decision class by the Inference class. After a service is planned the data related to this transaction will be stored in the trigger repository for future use and also in the Learning Repository. Once the Service providing plan is selected then it

will be given as input to the Reactor Design Pattern [10] which intern provides the service to the customer with a Service Handling Mechanism in Reactor Pattern.

On the other hand by performing the Association Rule based learning of data mining on the data stored in Learning Repository the server will generate new service which is composition of two or more services. So the Association Rule learning helps to include a new service with the help of Aspectual Feature module code in to the already existing Service Repository. So that more efficient access of services can be provided to the customers and all this process is done at run time with the inclusion a developer manual, means self-reconfiguration is provided at run-time.

“A design pattern is a particular form of recording information about a design such that the same pattern can be applied in future, if same situation is repeated to solve problem”. So the design patterns are accepted in wide range of object-oriented designs. Collections of design patterns can be found in numerous publications. Some of the design patterns that are employed to in the autonomic system are described below.

4 Design Pattern Template

To facilitate the organization, understanding, and application of the adaptation design patterns, this paper uses a template similar in style to that used by Ramirez et al.[9].

4.1 Pattern Name

Service Composition Design Pattern

4.2 Classification

Structural-Monitoring

4.3 Intent

Systematically applies the Design Patterns to an Autonomic Computing System and insertion of a new dynamically composed service interns of a Aspectual Feature Module (AFM) [3] with the help of Association rule based Learning method of data mining.

4.4 Context

Our design pattern may be used when:

- a) Self-learning type of autonomic computing property has to be achieved.
- b) For dynamically composing web services and to generate a new service which will perform the complex tasks of the customers.
- c) To include the newly discovered services as a Feature Module into the currently running system and satisfy the self-reconfiguration property.

4.5 Proposed Pattern Structure

A UML class diagram for the proposed design Pattern can be found in Fig 2.

4.6 Participants

- a) **ClientAPI:** The client will supply the service he wants to access and the CleintAPI class will generate a Event to the Trigger class to serve the Requested service.

- b) **Trigger:** The trigger class will notify the Inference Engine that an event has been generated and it has to handle a requested service.
- c) **Inference Engine:** This class will just takes the event generated by trigger class and it will check whether a rule that can select a correct plan exists in the Decision class.
- d) **Fixed Rules:** This class will check for any rules that satisfy the given service request and then it will respond with a rule that satisfies most.
- e) **Decision:** After a rule is selected by the Inference Engine it will select the Correct Plan that can fulfill the service requested by the customer with Decision class.
- f) **TriggerRepository:** Here this class will store the details of the event generated and the cause of the event and also timestamp like that it store some information.
- g) **Learner:** This class will store all the service transactions that have taken place in the server and it will be used for future purpose to derive new services and for service composition. When new service is composed in the Association Rule Learning it will also add the new rule of that respective Service.
- h) **Association Rule Learning:** This is the important class where it will perform the Association Rule based learning method of the Data Mining. It will access the information stored in the Learner class and perform the Association rule based learning on that information and then it will generate the new services which are composited of two or more service.
- i) **Log:** it will store the trigger, Rule, Decision Related information for the record storing purpose.
- j) **Initiation Dispatcher:** It defines an interface for registering, removing, and dispatching the event handlers.
- k) **Synchronous Event Demultiplexer:** The synchronous Event Demultiplexer is responsible for waiting until new events occur. When it detects new events, it will inform the Initiation Dispatcher to call back application-specific event handler.
- l) **Event Handler:** Specifies an interface consisting of a hook method [10] [13] that abstractly represents the dispatching operation for service-specific events. This method must be implemented by application-specific services.
- m) **Concrete Event Handler:** Implements the hook method [10], as well as the methods to process these events in an application-specific manner. Applications register Concrete Event Handlers with the Initiation Dispatcher to process certain types of events. When these events arrive, the Initiation Dispatcher calls back the hook method of the appropriate Concrete Event Handler.
- n) **Handle Event Aspect:** This is the aspect-oriented implementation module that will be viewed into the concrete event handling class, so that only a particular requested service method get viewed into the code at run-time.
- o) **Refines Class Event Handler:** This will add a new event handler for a new service and also the composed service in such a way that the insertion will be done as a new feature with the help of FOP.

4.7 Consequences

- a) This design pattern will eliminate the number of service requests that should be sent to the server.
- b) With the help of Association Rule based Learning we can easily achieve the Dynamic Service Composition Techniques.

- c) We can handle the complex service Requests of the customers with the help of composition of two or more services as a new single service.
- d) Also the Reconfiguration of the system takes place at the run-time.
- e) Without interrupting the current running system we can easily insert a new composed service as a Feature Module.
- f) With the help of the case-based learning design pattern we can choose perfect decisions about the service plan to be accomplished to fulfill the Requested service.

4.8 Related Design Patterns

- a) **Strategy Design Pattern [5]:** This pattern can be used to define a family of algorithms, encapsulate each one, and make them interchangeable. Strategy lets the algorithm vary independently from the clients that use it. In our proposed design pattern we use this pattern to choose or make a decision about the plan to be selected to fulfill the customer request or can be used as a decision-making pattern.
- b) **Adaptation Detector Design Pattern [9]:** This design pattern can be used to interpret monitoring data and determine when an adaptation is required. With the help of observer design pattern it can monitor the system/application for any changes in the environment. When ever any changes are detected then it will generate the Event as a Trigger which is then handled by Case-Based Reasoning pattern.
- c) **Architecture-Based Design Pattern [9]:** This design pattern provides an architectural of selecting reconfiguration plans.

4.9 Roles of Our Design Pattern in Autonomic System

- a) **Reactor Design Pattern:** Reactor Design Pattern in [10] handles service requests that are received concurrently from more than one client. In our proposed pattern we use this design pattern for efficiently handling the requested services. Each service in an application may consist of and is represented by a separate event-handler that dispatches the service-specific request. So this task of dispatching is done by initiation dispatcher, which itself manages the registered event handlers.
- b) **Case-Based Reasoning Design Pattern:** The Case-Based Reasoning Design Pattern in [9] will apply the rule based decision making mechanism to determine a correct reconfiguration plan. This design pattern will separate the decision-making logic from the functional logic of the Application. In our proposed pattern we will use this pattern to provide the perfect suitable plan to implement the customer requested service.

5 Feature Based Service insertion into Server Repository with Association Rule Based Learning

The main concept in our proposed pattern is that with the help of already stored information in the Learning repository, we can use this information and then use the Association Rule based Data Mining method to perform data mining operation on this information. With this type of operation we can derive the new services from the already existing

services. But the newly derived services are a composition of two or more services. So with the help of this association rule based data mining we can include the new services into the Service Repository as a Aspectual Feature Module [3]. With this we can provide composition of services to fulfill the complex service requests of the customers.

The view of our proposed design pattern for the unstructured peer-to-peer computing System can be seen in the form of a class diagram see Fig 2.

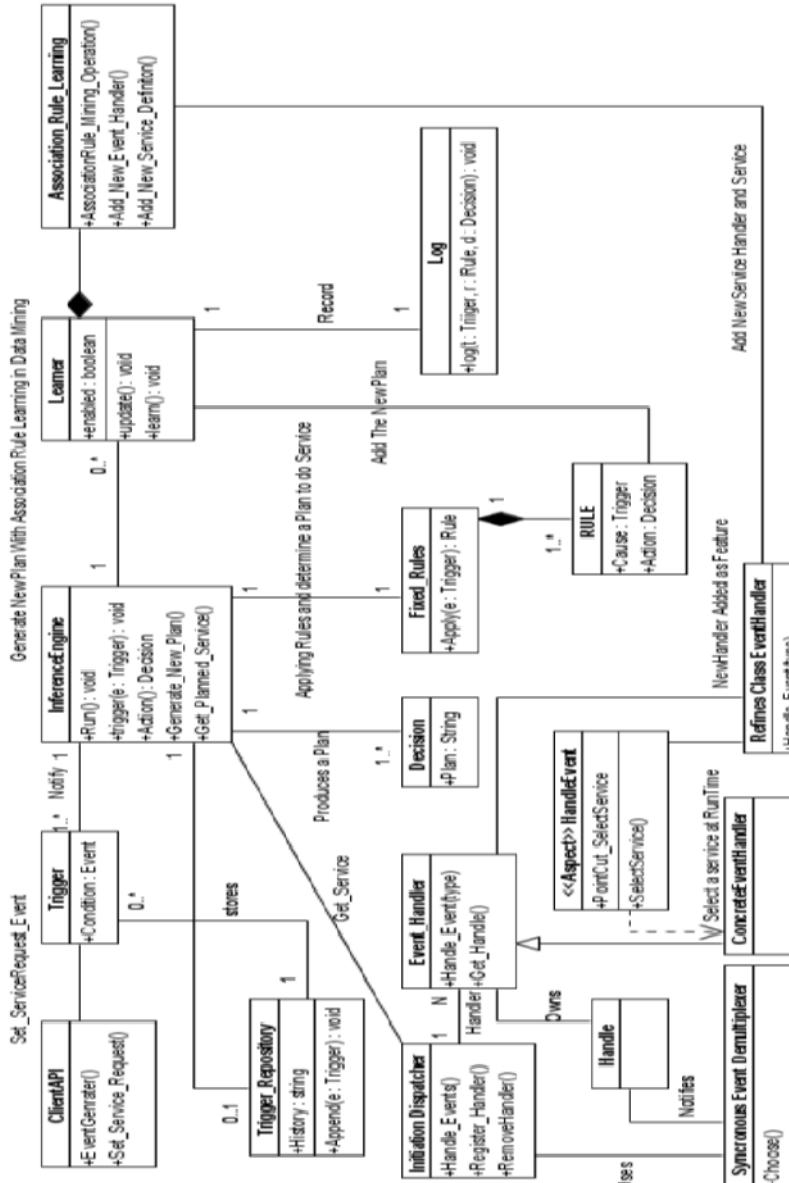


Fig. 2. Applying Design Pattern for the Autonomic System

6 Profiling Results

We are presenting the profiling results taken for ten runs without applying this pattern and after applying this pattern using the profiling facility available in the Netbeans IDE. The graph is plotted taking the time of execution in milliseconds on Y-axis and the run count on the X-axis. The graph has shown good results while executing the code with patterns and is shown in Fig 3. This can confirm the efficiency of the proposed pattern.

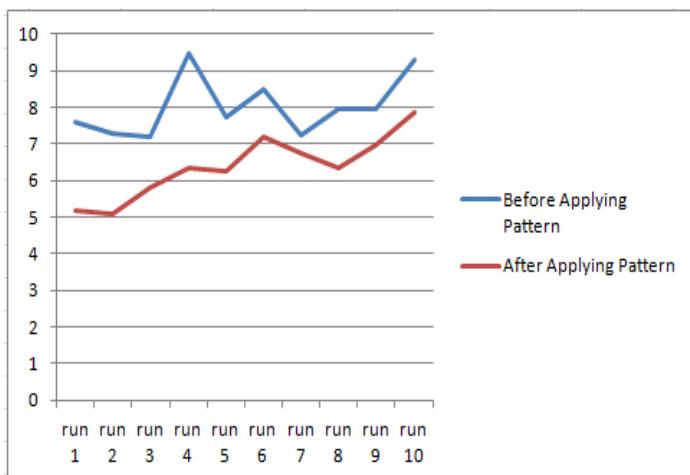


Fig. 3. Profiling data after and before applying design pattern

7 Conclusion and Future Work

In this paper we have proposed a pattern to facilitate Dynamic Service Composition. Here we have shown how the Association rule based Learning method of Data Mining can be used to determine the new services which are the composition of two or more services. Then we have also studied how the new services can be inserted into the service repository as a Aspectual Feature Module. Several future directions of work are possible. We are examining how these design patterns can be applied in the Software Product Lines (SPL). Also focusing upon the efficient use of the Data Mining methods for providing the flexible use of the design patterns for the solving most recurring problems.

References

1. Dobson, S., Sterritt, R., Nixon, P., Hinckley, M.: Fulfilling the Vision of Autonomic Computing, vol. 43, pp. 35–41. IEEE Computer Society (2010), doi:10.1109/MC.2010.14
2. Laddad, R.: AspectJ in Action, 2nd edn., ch. 12. Manning (2010)
3. Apel, S., Leich, T., Saake, G.: Aspectual Feature Modules. IEEE Transactions on Software Engineering 34(2) (2008), doi:10.1109/TSE.2007.70770

4. Batory, D., Sarvela, J.N., Rauschmayer, A.: Scaling Step-Wise Refinement. *IEEE Transactions on Software Engineering* 30(6), 187–197 (2003), doi:10.1109/ICSE.2003.1201199
5. Gamma, E., Helm, R., Johnson, R., Vlissides, J. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley (1995)
6. Kuhlemann, M., Rosenmüller, M., Apel, S., Leich, T.: On the Duality of Aspect-Oriented and Feature-Oriented Design Patterns. In: *Proceedings of the 6th Workshop on Aspects, Components, and Patterns for Infrastructure Software*. ACM, New York (2007), doi:10.1145/1233901.1233906
7. Hannemann, J., Kiczales, G.: Design Pattern Implementation in Java and AspectJ. In: *Proceedings of the International Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA)*, pp. 16–17 (2002), doi:10.1145/583854.582436
8. Tarr, P., Ossher, H., Harrison, W., Sutton, J.S.M.: N Degrees of Separation: Multi-Dimensional Separation of Concerns. In: *Proceedings of the International Conference on Software Engineering (ICSE)*, pp. 107–119 (1999), doi:10.1145/302405.302457
9. Ramirez, A.J., Betty, H.C., Cheng: Design Patterns for Developing Dynamically Adaptive Systems. In: *Proceedings of the 2010 ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems*. ACM, New York (2010), doi:10.1145/1808984.1808990
10. Schmid, D.C.: Reactor: An Object Behavioral Pattern for Demultiplexing and Dispatching Handles for Synchronous Events. Addison-Wesley (1995)
11. Mannava, V., Ramesh, T.: A Novel Adaptive Monitoring Compliance Design Pattern for Autonomic Computing Systems. In: Abraham, A., Lloret Mauri, J., Buford, J.F., Suzuki, J., Thampi, S.M. (eds.) *ACC 2011, Part III. CCIS*, vol. 190, pp. 250–259. Springer, Heidelberg (2011), doi:10.1007/978-3-642-22709-7_26
12. Aubert, O., Beugnard, A.: Adaptive Strategy Design Pattern, Laboratoire d’Informatique des Télécommunications, ENST Bretagne, France, June 25 (2001), doi: 10.1.1.100.4541

An Enhancement to AODV Protocol for Efficient Routing in VANET – A Cluster Based Approach

M.C. Aswathy¹ and C. Tripti²

¹ Department of Computer Science & Engineering

² Department of Computer Science & Engineering

Rajagiri School of Engineering & Technology, Rajagiri valley, Cochin, India

aswathymc@gmail.com, triptic@rajagiritech.ac.in

Abstract. Vehicular Ad-hoc Networks (VANET) are a special kind of Mobile Ad-hoc network (MANET), in which vehicles on the road forms the nodes of the networks. Now a days, VANETs find several applications as an Intelligent Transportation System. Dynamic network architectures and node movement characteristics differentiates VANETs from other kinds of ad hoc networks. Since VANETs have a dynamic network topology, routing in VANETs are complicated. Ad-hoc On-Demand Distance Vector (AODV) routing protocol is the most commonly used topology based routing protocol for VANET. During the route discovery process AODV broadcasts route request message (RREQ). It creates many unused routes between a source and a destination node. This paper aims at improving the performance of AODV by enhancing the existing protocol by creating stable clusters and performing routing by Cluster Heads and Gateway nodes.

Keywords: VANET, AODV, Cluster.

1 Introduction

The highly dynamic network topology and node movement characteristics differentiate VANETs from other kinds of ad hoc networks. Therefore, the design of an efficient routing protocol for VANETs is very crucial. Ad-hoc On-Demand Distance Vector (AODV) routing protocol is the most commonly used topology based routing protocol for VANET. AODV is a reactive kind of protocol where the route from a source to a destination is created only when it is needed. In AODV the route discovery is based on query, and the reply to this query is used to take a routing decision. All intermediate nodes stores routing table, which contains the route information. During the routing process, AODV uses some control packets. A node broadcasts routing request message (RREQ) to find a route to another node. The intermediate node which has routing information replies with a routing reply message (RREP). A node uses route error message (RERR) to notify other nodes about the loss of link. Every node uses HELLO message to detect and monitor links to its neighbours. During the route discovery process AODV floods the entire network with large number of control packets, and hence it finds many unused routes between the source and destination[3]. This becomes a major drawback to AODV since this causes routing overhead, consuming bandwidth and node power.

AODV can be optimized in many different ways. Channel Availability [10] can be used to enhance AODV for vehicle safety applications. Mobility parameters can be used while route discovery. Another method is to cluster the nodes of the network and managing routing of packets by cluster heads. The main idea behind clustering is to divide nodes of the network into multiple separate groups called clusters and forming a cluster structure. By clustering, the routing process can be focused to only a subset of nodes of the network, thus the routing task can be simplified.

In a clustered network, instead of broadcasting the RREQ message, it can be sent to Cluster Heads. The Cluster Heads can find the routing information among the cluster members. If the route is available, the node sends a RREP message otherwise the RREQ is forwarded to other Cluster Heads via Gateway nodes. Thus the number of RREQ message for route discovery can be reduced considerably. This reduction in control messages can reduce the congestion in the network. Also overhead of the network to manage large number of packets can be reduced. Thus the performance of AODV can be improved.

The remainder of the paper is structured as follows. Section 2 discusses the related works done on clustering and improvements in AODV. Section 3 introduces the proposed enhancement to AODV using Clusters. Section 4 describes the experimental setup for the implementation. Section 5 concludes the paper.

2 Related Works

The AODV-Clustering [1] uses two route discovery mechanisms; Quick Route Discovery mechanism and the Traditional AODV Route Discovery mechanism. The protocol first uses the Quick Route Discovery mechanism. If a suitable route cannot be found, then it uses traditional AODV route discovery mechanism. When the algorithm uses the traditional AODV route discovery mechanism, it will flood the network with many control packets. In C-AODV[2], Clusters are formed based on the distance between the nodes and its cluster head. This is not suitable for VANETs since the distance between the nodes(vehicles) changes rapidly in VANETs. Enhanced AODV for directional flooding using Coordinate System [3] uses the concept of polar coordinate system. It limits the route discovery process to a limited region using GPS data like position, speed and track angle of source and destination node. Mobility-based Clustering in VANETs using Affinity Propagation [4] uses the idea of Affinity Propagation from a communications perspective, in a distributed manner. In the algorithm, each node in the network transmits the responsibility and availability messages to its neighbours, and then makes a decision on clustering independently. In Toward Strongly Connected Clustering Structure in Vehicular Ad hoc Networks [5], vehicles are assumed to use control channel to exchange periodic messages and gather information about their neighbourhood, and use one service channel to form the clusters and perform all intra-cluster communication tasks. Here the slowest vehicle among non-clustered neighbours initiates the cluster formation process. The algorithm avoids grouping vehicles whose relative velocity is greater than the threshold in one cluster.

3 Proposed Routing Using Clusters

The dynamic nature of VANET makes the node management and routing a difficult process. Clustering in VANET is an approach which can help to resolve these issues to some extent. Figure 1 shows the cluster formation in a VANET. The proposed algorithm for clustering in VANET can be done using three basic steps.

- i. Cluster Formation
- ii. Cluster Maintenance
- iii. Clustered Routing

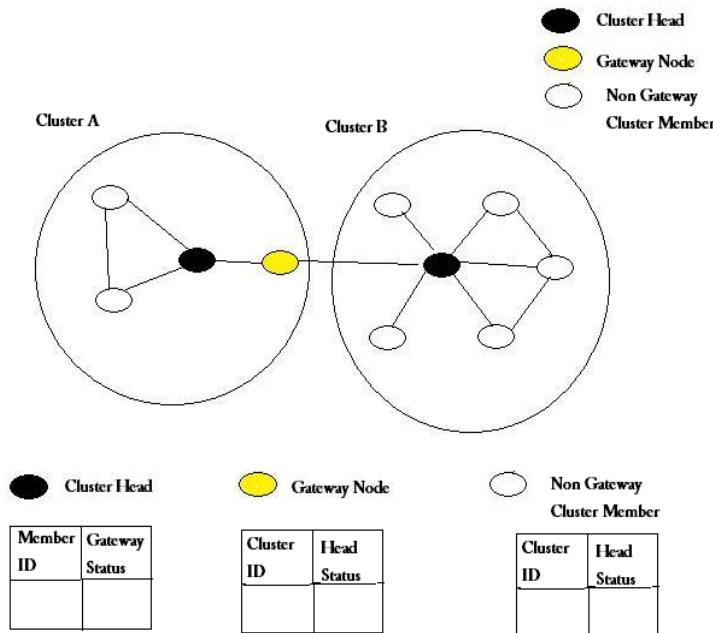


Fig. 1. VANET Clusters with 10 Nodes

3.1 Cluster Formation

A cluster is characterised by a leader node called the Cluster Head node and a set of member nodes called the Cluster Members, having a single-hop link from the Cluster Head. A cluster has only one Cluster Head. The Cluster members may or may not be Gateways. Cluster Heads and Gateways are vital elements in routing.

In the proposed algorithm, to reduce the network overhead, cluster size is limited to single hop. Every node has a NODE-ID and a CLUSTER-ID. The CLUSTER-ID is the identified of the cluster of which the node is a member. Every node maintains a GATEWAY-TABLE which contains the CLUSTER-IDs of Cluster Heads, it can reach by single hop. A cluster head maintains a MEMBERS-TABLE to keep the list of members joined to this cluster. The number of cluster members is called its DEGREE.

Every node exchanges HELLO message to inform about their existence to its neighbour nodes and waits for CLUSTER_STATUS message from cluster head for a time period. The neighbouring Cluster Head will reply with its CLUSTER_STATUS message, which contains the Degree of the cluster. If the node gets more than one CLUSTER_STATUS message, it will update this information in the GATEWAY-TABLE and it will choose the cluster with highest degree to join. Then the node can send a request message JOIN_REQUEST to join the cluster, indicating whether it is a Gateway node or not. The Cluster Head will check its DEGREE. The node is allowed to join the cluster only if the DEGREE of the cluster is within the limits. This limits the number of Cluster Members and controls overhead of managing large number of Cluster Members by the Cluster Head. The cluster head sends a JOIN_ACCEPT message to the node if it can accept the request and updates the CLUSTER-MEMBERS table and the DEGREE. The requesting node sets the Head Status field in the GATEWAY-TABLE. If the node doesn't receive a CLUSTER_STATUS message within the time period, the node will elect itself as Cluster Head and updated its CLUSTER-ID with its NODE-ID. It will set its DEGREE as zero.

3.2 Cluster Maintenance

If the cluster head doesn't receive HELLO message from its member within a time period, it will assume that the node is dead. The cluster head will delete the entry of the node from the members table. It also reduced its Degree by 1. If a node doesn't receive CLUSTER_STATUS message from its Cluster Head within a time period, it will do the Cluster Join Procedure.

3.3 Clustered Routing

In Clustered AODV a source node seeking to send a data packet to a destination node checks its route table to see if it has a valid route to the destination node. If a route exists, it simply forwards the packets to the next hop along the way to the destination. On the other hand, if there is no route in the table, the source node begins a route discovery process. It sends a route request (RREQ) packet to its Cluster Head. The Cluster Head checks to see whether its member node has a route to the destination or the destination node itself. If it has a route, it will reply with a route reply (RREP) packet. If not, the Cluster Head will forward the RREQ packet to its Gateway Members. The Gateway members will forward the packet to Cluster Heads in its GATEWAY-TABLE. This process continues until the request reaches either an intermediate node with a route to the destination or the destination node itself. This route request packet contains the IP address of the source node, current sequence number, the IP address of the destination node, and the sequence number known last. An intermediate node will reply to the route request packet only if they have a destination sequence number that is greater than or equal to the number contained in the route request packet header. When an intermediate node forwards route request packet, it will record in its route table the address of the neighbour from which the first copy of the packet has come from. This recorded information is later used to construct the reverse path for the route reply (RREP) packet. When the route reply packet arrives from the destination or the

intermediate node, the nodes forward it along the established reverse path and store the forward route entry in their route table. The clustered routing can be summarised as:

- Source node sends RREQ to its Cluster Head.
- The Cluster Head checks whether Destination node is in its MEMBERS-TABLE.
 - If so, it will forward the packet to the member
 - Otherwise, it will forward packet to Gateway nodes in its MEMBERS-TABLE.
- When an intermediate node receives a RREQ packet, it checks whether it is the Destination node or is there any path to destination. If so, it will reply with an RREP packet
- When a Gateway node receives a RREQ packet, it will forward the packet to Cluster Heads in its GATEWAY-TABLE

4 Experimental Setup

The proposed algorithm can be implemented in NS2.34. The mobility model for the VANET can be created using VanetMobiSim-1.1. The trace file generated from VanetMobiSim serves as the input to NS2. To evaluate the performance of the network, the trace file generated from NS2 can be analysed and various performance parameters such as number of packets dropped, packet delivery ratio and end to end delay can be computed by using an AWK script.

5 Conclusion

In the proposed method, the VANET is made into small clusters with long Cluster Head duration. The AODV protocol is optimized by replacing broadcasting of RREQ packets with forwarding of RREQ packets to Cluster Heads and thereby managing routing by Cluster Heads and Gateway Nodes. This effectively reduces the total number of control packets generated during the route discovery process. Thus the overhead of network in routing packets can be reduced and the efficiency of the protocol can be improved.

References

1. Zheng, K., Wang, N., Liu, A.-F.: A new AODV based clustering routing protocol. In: International Conference on Wireless Communications, Networking and Mobile Computing. IEEE (2005)
2. Thirumurugan, S.: C-AODV: Routing Protocol for Tunnel's Network. International Journal of Computer Science and Technology, IJCST 2(1) (March 2011)
3. Reno Robert, R.: Enhanced AODV for directional flooding using Coordinate System. In: 2010 International Conference on Networking and Information Technology. IEEE (2010)

4. Shea, C., Hassanabadi, B., Valaee, S.: Mobility-based Clustering in VANETs using Affinity Propagation. In: Global Telecommunications Conference, GLOBECOM 2009. IEEE (2009)
5. Rawshdeh, Z.Y., Mahmud, S.M.: Toward Strongly Connected Clustering Structure in Vehicular Ad hoc Networks. In: Proceedings of the 2009 IEEE 70th Vehicular Technology Conference: VTC 2009-Fall, Anchorage, Alaska, USA, September 20-23 (2009)
6. Bononi, L., Di Felice, M.: A Cross Layered MAC and Clustering Scheme for Efficient Broadcast in VANETs. In: International Conference on Mobile Adhoc and Sensor Systems, MASS 2007. IEEE (2007)
7. Luo, Y., Zhang, W., Hu, Y.: A New Cluster Based Routing Protocol for VANET. In: 2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing. IEEE (2010)
8. Maslekar, N., Boussedjra, M., Mouzna, J., Labiod, H.: A Stable Clustering Algorithm for Efficiency Applications in VANETs. In: Wireless Communications and Mobile Computing Conference (IWCMC). IEEE (2011)
9. Venkata Manoj, D., Manohara Pai, M.M., Pai, R.M., Mouzna, J.: Traffic Monitoring and Routing in VANETs – A Cluster Based Approach. In: 2011 11th International Conference on ITS Telecommunications. IEEE (2011)
10. Yawan, N., Keeratiwintakorn, P.: Efficiency Improvement of AODV for Vehicular Networks with Channel Availability Estimation. In: The 8th Electrical Engineeringl Electronics, Computer, Telecommunications and Information Technology (ECTI) Association of Thailand - Conference 2011 (2011)
11. Fan, P., Haran, J.G., Dillenburg, J., Nelson, P.C.: Cluster-Based Framework in Vehicular Ad-Hoc Networks. In: Syrotiuk, V.R., Chávez, E. (eds.) ADHOC-NOW 2005. LNCS, vol. 3738, pp. 32–42. Springer, Heidelberg (2005)
12. Akbari, A., Soruri, M., Jalali, S.V.: Survey of Stable Clustering for Mobile Adhoc Networks. In: 2009 Second International Conference on Machine Vision. IEEE (2010)

Human Emotion Recognition and Classification from Digital Colour Images Using Fuzzy and PCA Approach

Shikha Tayal¹ and Sandip Vijay²

¹Department of Electronics Engineering, College of Engineering Roorkee, India
shikha.tayal@gmail.com

²Department of Electronics Engineering, DIT, Dehradun, India
vijaysandip@gmail.com

Abstract. In this paper, we proposed a new model for recognizing various emotions of humans with different age groups and gender. Fuzzy is used for extracting more accurate region of interest, i.e., face. The dimensionality of face image is reduced by the Principal Component Analysis (PCA) [12] and finally emotion is recognized and classified using Euclidean Distance. Database is prepared and some performance metrics like recognition-rate v/s Eigen-range has been calculated. The proposed method was also tested on FACES Collection database [13]. The experiment results demonstrate that the emotion recognition system has been successful with average recognition rate of 96.66% (with both experiment databases) when approximately or more than 60% eigenfaces used. It is also shown that database can be easily expanded to classify faces and non faces images.

Keywords: Emotion Recognition, Fuzzy logic, Principal component analysis (PCA), Euclidean Distance, Eigen Range.

1 Introduction

Day by day, we are more depending upon highly intelligent machines but still there are some situations in which task performed or any decision taken by machines must depend upon the state of mind in which human is presently going through. Emotion detection problem originates when we are able to develop highly intelligent humanoid that is not able to interact properly with people due to lack of emotions. Our target is to introduce some more intelligence, regarding human emotions in present humanoid. As emotions are more reflected from facial features as compared to voice, our contribution is limited to detecting emotions from facial features.

This paper presents a holistic approach using fuzzy and PCA approach for efficient and robust emotion recognition. Face plays a major role in conveying distinctiveness and sentiments. Thus, for emotion identification the starting step involves extraction of human face from whole image by the Fuzzy Face Detector And Segmentor

(FFDAS). Then PCA technique is used to simplify a dataset into lower dimension while retaining the characteristics of dataset.

2 Related Work

A wide range of classification algorithms have been applied to the emotion recognition problem (e.g. Support Vector Machine (SVM), Neural network (NN), PCA, LDA, Learning Fuzzy with Genetic Algorithm [2]. Above emotion recognition methods fall into two categories: Feature based and Holistic approach. In feature-based method, emotion recognition relies on localization and detection of facial features such as eyes, nose, mouth, chin and eyebrows and their geometrical relationships, with the help of deformable templates and extensive mathematics. Yullie and Cohen [11] used deformable templates in contour extraction of face images. In holistic approach, whole image is preset into a point on high dimensional space. In this method, information that best describes a face is derived from the entire face image. Based on the Karhunen-Loeve expansion in pattern recognition, Kirby and Sirovich [5], [8] have shown that any particular face can be represented in terms of a best coordinate system termed as "eigenfaces". These are the eigen functions of the average covariance of the ensemble of faces. Later, Turk and Pentland [10] proposed a face recognition method based on the eigenfaces approach. In research paper [3], author proposed PCA for classification of emotions using Singular Value Decomposition. In [4], Facial features are extracted using DCT and these extracted features are further used to classify emotions using three different Neural Network models: MLP, PCA, GFFNN. Kishore and Varma [6], proposed new facial emotion classifier based on wavelet fusion, which combines the features extracted by Gabor wavelet and Discrete Cosine Transform (DCT). Author, concludes that the effectiveness of extraction expression feature is completely dependent on the effectiveness of pre -processing of the raw image. Facial expressions are extracted from the detailed analysis of eye region images is given in [9]. Another method of classification of facial features using Linear Discriminant Analysis (LDA) is explained in [1]. In this, gabor features are extracted using gabor filter banks and compressed by two stage PCA method. Kernel Eigen Space method based on class features for expression analysis is explained in [7].

3 Proposed Technique

An unsupervised pattern recognition system is proposed in this paper which is independent of extreme geometry and computation. Here, Fuzzy logic is applied for face detection and segmentation. PCA for emotion recognition and it is based on the information theory approach in which the relevant information in a face is extracted as efficiently as possible. Further Euclidean Distance, used for emotion classification, provides the measure of similarity between training images and image under test. The emotion recognition system is shown in fig.1:

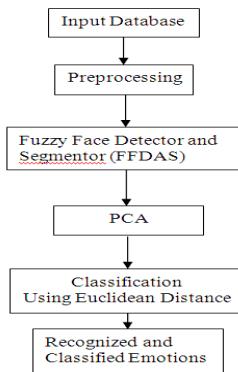


Fig. 1. Flow Diagram of Proposed Emotion Recognition and Classification Approach

3.1 Database Preparation

Face images with different emotions: neutral, happy, sad, anger, disgust and surprised as well as non-face images are stored in a library in the system known as database. The database is further divided into two sets – training dataset (40 images) and testing dataset (12 images). Images are in jpg format. Another set of database is taken from FACES Collection database affiliated by Max Planck Institute for Human Development [13]. This database consists of different individual in various emotions with different age groups and gender. Images are in bmp format.

3.2 Preprocessing

Both training and testing images undergoes following preprocessing steps:

1. Image size normalization: This module automatically reduce every face image to 205*214 pixels
2. Noise Removal
3. Background Minimization
4. Illumination control

3.3 FFDAS

After preprocessing images are fed to FFDAS to reduce the work space, by locating and cropping facial region more accurately from the image. RGB images are first converted into YCbCr Color space. Then Cb and Cr values are used to separate face skin color from the full image. Sugeno-Fuzzy inference system is designed on the basis of set of rules defined by Cb and Cr values of YCbCr color model as input to the fuzzy system and skin_pixel as output.

The fuzzy logic rules applied for skin detection are the following:

1. IF Cb is Light and Cr is Light THEN the skin_pixel =0
2. IF Cb is Light and Cr is Medium THEN the skin_pixel =0
3. IF Cb is Light and Cr is Dark THEN the skin_pixel =0

4. IF Cb is Medium and Cr is Light THEN the skin_pixel =0
5. IF Cb is Medium and Cr is Medium THEN the skin_pixel =1
6. IF Cb is Medium and Cr is Dark THEN the skin_pixel =1
7. IF Cb is Dark and Cr is Light THEN the skin_pixel =0
8. IF Cb is Dark and Cr is Medium THEN the skin_pixel =1
9. IF Cb is Dark and Cr is Dark THEN the skin_pixel =0

3.4 PCA for Feature Extraction

PCA is applied to extract unique facial characteristics important in distinction between different human emotions. Applied PCA algorithm is as follows:

1. Prepare a data set of M facial human images (each of NxN pixels) that is needed to extract the feature, i.e., $I = \{I_1, I_2, I_3 \dots, I_M\}$
2. Convert each image matrix $N \times N$ to vector $N^2 \times 1$.
3. Create the database matrix I that adds all images in one matrix $N^2 \times M$. Here each column of matrix I represents an image.
4. Calculate mean for each image dimension (each column of matrix I). Resulting matrix I_{MEAN} will be a row vector of dimension $M \times 1$.
5. Subtract the mean of each image dimension from image vector. Resultant matrix is known as mean adjusted matrix (say Y).
6. Calculate the covariance matrix C of mean adjusted matrix Y obtained in step 5.
7. Calculate the eigen vector matrix λ_k and eigen values u_k of covariance matrix such that $C = \lambda_k u_k$
8. Feature vector is prepared by collecting all eigenvectors in one matrix.
9. Multiply feature vector by mean adjusted data to calculate image feature.
10. Face space F is created by collecting image features of all images.
11. Feature spaces F are used for classification of different emotions.

3.5 Computing Euclidean Distances

This basic idea, ‘dimensionality reduction followed by distance calculation in a subspace’, is one of the primary tools for managing complexity and for finding the patterns hidden within massive amounts of real world data. In the original eigen face paper [12], distance between two face images is the Euclidean distance between their projected points in a PCA subspace, rather than the distance in the original $M \times N$ dimensional image space. By computing distance between face images, we've replaced $M \times N$ differences between pixel values with a single value. In eigen face, computing the distance between faces in this lower dimensional subspace is the technique that eigen face uses to improve the signal-to-noise ratio.

3.6 Emotion Recognition

The Euclidean distance between two feature vectors of images i and j provides a measure of similarity between the corresponding images i and j . Euclidean distances between test images with other training images are calculated. Then the expression of training image, which results in minimum Euclidean distance, is said to be final recognized emotion associated with that test image.

4 Experimental Results and Analysis

Algorithm for facial expression recognition classifies the given image into seven basic facial expression categories (happiness, sadness, fear, surprise, anger, disgust and neutral).

The proposed method is tested on FACE Collection database consists of images of different individuals with different age groups and genders. It is also tested on our own prepared database. Two different training and testing database has been prepared to check the performance of the system. Emotion associated with each image in training database is also fed to the system. Testing data set contains 12 different test images of same individuals as shown in training data set. Here emotion associated with each image is unknown to the system. And this is what we expect from our system to correlate suitable and best emotion with these images.

Image Processing Toolbox in Matlab is used as simulation tool. And some of the simulation results are shown below:

First, FFDAS is used for face extraction from both training and testing data set. And simulation result of testing dataset is shown in fig.2

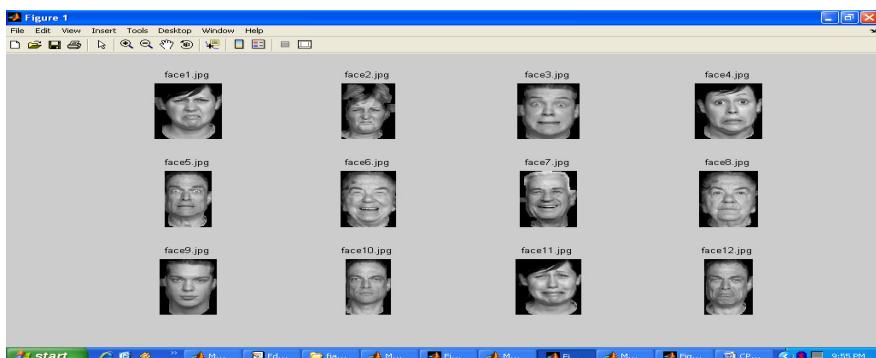


Fig. 2. Cropped face region of testing dataset 1 using FFDAS

Then, PCA is applied for dimensionality reduction in input data to retain those characteristics of the data set that contribute most to its variance, by keeping lower-order principal components which contain “most important” aspect of the data and ignoring higher-order ones. The extracted feature vectors in the reduced space are used to project test image on face space.

Euclidean Distance of eigen face of each test image is calculated with training dataset using equation and the one with minimum Euclidean distance between eigen face of test and train image is considered as best match and expression associated with corresponding train image is supposed to be the emotion of test image. Simulation results of this stage are shown in fig.3. Here test images are displayed with the resultant emotion.

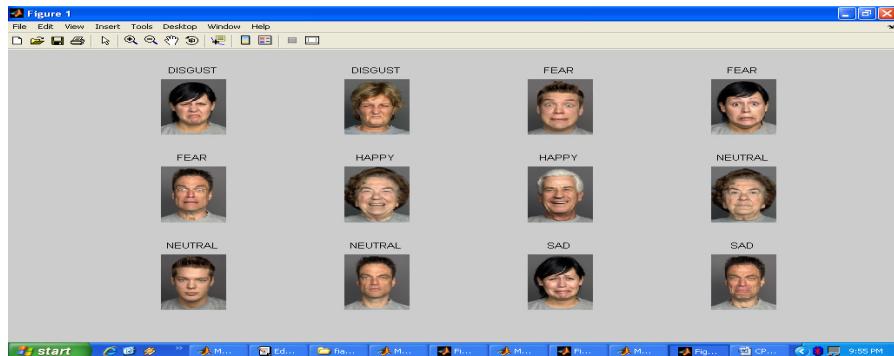


Fig. 3. Recognized emotions for testing dataset 1

Dataset 2 also contains some non-face images. Simulation results shown in fig:4, shows that our proposed system is capable to differentiate face and non-face images.



Fig. 4. Recognized emotions for testing dataset 2

The proposed technique is also analyzed by varying the number of eigen faces used for feature extraction. The recognition performance is shown in Figure 5.

The experiment is based on:

$$\begin{aligned} \text{Number of Train images} &= 40 \\ \text{Number of Test images} &= 12 \end{aligned}$$

The result of proposed algorithm is:

Maximum recognition rate = 100% (when approximately half number of Eigen faces used)

Minimum recognition rate = 41.66%

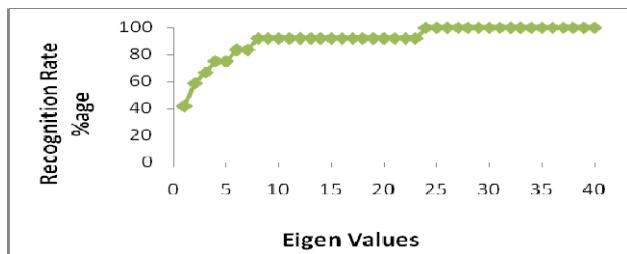


Fig. 5. Graph of Recognition Rate v/s Eigen Values for Database 2

5 Conclusion

This paper presents an emotion recognition system using fuzzy and PCA techniques. Experiment results with average recognition rate of approximately 96.66% when testing seven emotions on standard image data set, show that Fuzzy and PCA approach is successful in recognizing emotion using facial expressions. This research could help in future works, like capturing non-static images in real time and simultaneously analyzing these images according to affective computing techniques. By making these analyses some of the user's emotional states could be seen like joy, fear, angry, and with these probable results, assistants and computer optimizers could help users in the most different applications.

References

- [1] Deng, H.-B., Jin, L.-W., Zhen, L.-X., Huang, I.-C.: A New Facial Expression Recognition Method Based on Local Gabor Filter Bank and PCA plus LDA. International Journal of Information Technology 11(11), 86–96 (2005)
- [2] Amir, J.: A Learning Fuzzy Model for Emotion Recognition. European Journal of Scientific Research 57(2), 206–211 (2011)
- [3] Kaur, M., Vashisht, R., Nirvair, N.: Recognition of Facial Expressions with Principal Component Analysis and Singular Value Decomposition. International Journal of Computer Applications 9(12), 36–40 (2010)
- [4] Kharat, G.U., Dudul, S.V.: Emotion Recognition from Facial Expression Using Neural Networks. In: HIS, Krakow, Poland, May 25–27. IEEE (2008)
- [5] Kirby, M., Sirovich, L.: Application of the Karhunen-Loeve procedure for the characterization of human faces. IEEE PAMI 12(1), 103–108 (1990)
- [6] Kishore, K.V.K., Varma, G.P.S.: Efficient Facial Emotion Classification with Wavelet Fusion of Multi Features. IJCSNS International Journal of Computer Science and Network Security 11(8) (2011)
- [7] Kosaka, Y., Kotani, K.: Facial Expression Analysis by Kernel Eigen Space Method based on Class Features (KEMC) Using Non-Linear Basis For Separation of Expression Classes. In: International Conference on Image Processing, ICIP (2004)
- [8] Sirovich, L., Kirby, M.: Low-dimensional procedure for the characterization of human faces. Journal of the Optical Society of America A Optics and Image Science 4(3), 519–524 (1987)

- [9] Moriyama, T., Kanade, T., Xiao, J., Cohn, J.F.: Meticulously Detailed Eye region Model and It's Application to Analysis of Facial Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(5) (2006)
- [10] Turk, M., Pentland, A.: Eigenfaces for recognition. *Journal of Cognitive Neuroscience* 3(1), 71–86 (1991)
- [11] Yuille, A.L., Cohen, D.S., Hallinan, P.W.: Feature extraction from faces using deformable templates. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Proceedings CVPR 1989*, June 4-8, pp. 104–109 (1989)
- [12] Dimitri, P.: Eigenface-based facial recognition (February 2003)
- [13] <http://faces.mpg.de/album/escidoc:57488> for downloading the FACE Collection database

A New Process Placement Algorithm in Multi-core Clusters Aimed to Reducing Network Interface Contention

Ghobad Zarrinchian, Mohsen Soryani, and Morteza Analoui

Iran University of Science and Technology, Tehran, Iran
{Zarrinchian, soryani, analoui}@comp.iust.ac.ir

Abstract. The number of processing cores within computing nodes which are used in current clustered systems, are growing up rapidly. Despite this trend, the number of available network interfaces in such nodes almost has been remained unchanged. This issue can lead to high usage of network interface in many workloads, especially in workloads which have high inter-process communications. As a result, network interface would become a performance bottleneck and can degrade the performance drastically. The goal of this paper is to introduce a new process mapping algorithm in multi-core clusters aimed to reducing network interface contention and improving the performance of running parallel applications. Comparison of the new algorithm with other well-known methods in synthetic and real workloads indicates that the new strategy can gain 5% to 90% performance improvement in heavy communicating workloads.

1 Introduction

Parallel processing is one of the basic approaches to obtain high processing power. This power is necessary to run many scientific and economic applications which are known as Grand Challenge Applications (GCA). In this regard, various architectures have been introduced. Of these architectures, cluster computing has gained more popularity such that based on last published issues in 2011 [1], up to 82% of 500 top supercomputers in the world, used this architecture. Besides, recent advancements in multi-core processor technology have made these processors an excellent choice to use in clustered nodes. Magny-cours series of AMD Opteron and Westmere series of Intel Xeon which have 12 and 10 cores per chip respectively are some examples of multi-core processors which are becoming common in recent computing nodes.

Although multi-core processors, can improve computational capability, but they raise some challenges. The main challenge in this regard, is the contention of various cores for using shared resources like memory and buses. In the presence of such contention, shared resources can be performance bottleneck and can degrade the performance of running parallel applications drastically. Consequently, efficient execution of parallel applications in such systems needs more deliberations of these systems. In doing so, there are lots of studies including [2-6] which provide insights into the conditions in which efficient performance of clustered systems can be gained.

When multi-core computing nodes are used individually, memories and buses are the main shared resources that contention on them can adversely affect the performance. But when these nodes are connected together to form a clustered system, network interfaces are raised as another important resources. This is because various

processes of a parallel job (when placed on different nodes) use these interfaces for communications and synchronizations. In spite of considerable growth in the number of processing cores within computing nodes, the number of available network interfaces almost has been remained unchanged and this number is 1 or 2 for most systems. This issue can lead to high usage of network interfaces in many workloads, especially in workloads which have high inter-process communications. Since network interface port can service just one request at a time, other communication requests received from different cores must be queued to service later. The more cores in a node, the more requests for network interface. As a result, waiting time of messages at interface queue will be increased. This issue can finally prolong the execution time of parallel programs. According to these notes, if we could distribute parallel processes in available computing nodes in a way that requests arriving in each network interface be decreased, queuing time will also be decreased and we can expect performance improvement. In doing so, our goal in this paper, is to present a solution for mapping parallel processes to multi-core clusters in order to reduce network interface contention. After presenting our proposed mapping strategy, we will compare it with some well-known methods and it is shown that the new mapping method can obtain 5% to 90% performance improvement based on used scenarios.

2 Related Works

Various methods have been proposed for mapping parallel processes to processing cores. Of these methods, Blocked and Cyclic are two common approaches which are already investigated in [7-8]. In Blocked method, the mapping is started by selecting a computing node and assigning parallel processes to its free cores one-by-one. When there is no free core, another node will be used and this procedure is repeated until the end of assignment. In Cyclic method, parallel processes are distributed among computing nodes as Round Robin. As a result, maximum number of nodes and minimum number of cores in each node is used in this method (in contrast to Blocked method which uses minimum number of nodes and maximum number of cores in each node).

Although Blocked and Cyclic methods are used in many situations as a default method, but these approaches have little intelligence and do not consider the volume of communications between processes. Because of this issue, other techniques have been proposed which are more intelligent than Blocked and Cyclic. Some of these methods are [9-13]. Proposed mapping algorithm in these studies is based on graph partitioning techniques. The main idea in these techniques is to find processes that communicate to each other frequently and to map them near each other (e.g. place them in the same node). By this way, those processes can benefit from higher bandwidth of memory compared to network interface bandwidth. In order to do this, Application Graph (AG) and Cluster Topology Graph (CTG) are established and then, it is tried to find an efficient mapping from AG to CTG. In AG, vertices represent parallel processes and edges represent communications between processes. In CTG, vertices and edges represent processing cores and available bandwidth between them respectively. Since graph mapping problem is known as NP problem, some heuristics have been introduced which are based on graph partitioning approaches. Dual recursive bipartitioning (DRB) and K-way graph partitioning are two common heuristics. In DRB, AG is divided into two subgroups such that processes which communicate to each other frequently will be grouped in the same subgroup, but processes which communicate to each other infrequently, will be placed in different subgroups. By ‘frequently’ we

mean the total volume of data exchanged between each pair of processes. The CTG is also divided into two subgroups in the same way as done with AG. Then, each subgroup of AG is assigned to the peer subgroup of CTG. This operation is repeated on each subgroup recursively until one process in AG or one core in CTG remains. K-way graph partitioning is the same with DRB except that instead of two groups, graphs are divided into K groups.

Although graph partitioning techniques try to improve performance by mapping frequently communicating processes near each other, but when we try to put such processes near themselves, some shared resources can become performance bottlenecks and these methods are oblivious to this issue. Studies that propose a mapping approach to mitigate contention problem are very limited. Of these studies, we can point to [14-16]. [14] Introduces a mapping algorithm to avoid congestion on Torus interconnection networks. But this study does not consider congestion problem on network interface. In [15] the problem of contention on network interface is investigated. This study tries to put a combination of parallel jobs which have high inter-node communications and low inter-node communications in one node. By this way, network interface contention is alleviated while maximum number of processing cores is used in an efficient way. However this study does not provide a systematic algorithm to use in all scenarios and under every condition. [16] uses a scheduling method to mitigate contention and does not benefit from an intelligent mapping.

3 Proposed Mapping Algorithm

In order to reduce network interface contention, the conditions in which contention is raised, must be recognized. By determining such conditions, we can present the solution. If we could accommodate all processes of a parallel job in just one computing node, there will be no usage of network interface and hence, there is no contention. But when the number of processes is high, or the number of free cores in computing nodes is low, parallel processes must be placed in more than one node inevitably. In this case, high volume of inter-process communications can raise the contention on network interface and hence, degradation of performance will be occurred. To tackle this problem, we should determine a threshold on the number of processes which reside in a node and have high inter-node communication demands. This means that we should distribute processes among available nodes in order to reduce network requests arriving to each interface. Consequently, waiting time at interface queue for inter-node messages will be decreased. In this paper, we tried to determine an appropriate value for threshold using the number of adjacent processes (for each process) and the number of available free cores in computing nodes. Fig. 1 shows our mapping algorithm pseudocode. The first step is to separate parallel jobs based on the length of messages they send. Since larger messages need more service time, processes which send larger messages should use intra-node communications to benefit from high bandwidth of memory. We categorized messages into 3 groups: large messages (1MB or higher), medium messages (2KB to 1MB), and small messages (2KB or less). Based on these categories, we separate parallel jobs. First we select parallel jobs which send large messages (step 1), and then, it is the time to select and map jobs which send medium and small messages respectively (steps 4,6). If processes of a job send messages with different lengths, largest message length is considered for action. After partitioning jobs, parallel jobs in each group are sorted (step 2) based on average number of adjacent processes for each process (Adj_{avg}). Jobs which have more

average adjacency are mapped earlier. This is because these jobs may need to distribute between the nodes to have efficient performance. As a result, these jobs should be mapped before other jobs to use available capacity of computing nodes. After choosing a job to map, processes of this job are sorted based on their communication demands and processes which have more communications, are mapped earlier. In

proposed algorithm, communication demand for process i is calculated by: $\sum_{j=1, j \neq i}^P L_{ij} \lambda_{ij}$

in which, L_{ij} is the length of messages sent from process i to process j (largest length when having different lengths), λ_{ij} is the rate of sending messages from i to j , and P represents number of parallel processes for current job. After determining process with most communication demand (given process ‘A’), this process is assigned to a node with most free cores. Then, adjacent processes of ‘A’ are sorted based on the communication demand between ‘A’ and them, and it is tried to map adjacent processes of ‘A’ in the same node as ‘A’.

Now, it must be noted that if the number of adjacent processes is high, or the number of available free cores in current node is low, some adjacent processes must be mapped to other computing nodes. In such situations, as mentioned earlier, high

```

New_Mapping_Algorithm( )
Input: Workload graph, Cluster architecture
Output: Mapping information
{
    1. job_pool = select_jobs ( high_length );
    2. sort_jobs ( job_pool );
    3. while ( job_pool is not empty )
    {
        crnt_job = select_job ( job_pool );
        If ( Adjavg <= FreeCoresavg -1 )
            No threshold is determined;
        Else
            Threshold = 
$$\left\lceil \frac{\sum_{i=1}^P \frac{Adj_{pi}}{Adj_{max}}}{num\_of\_nodes} \right\rceil ;$$

        sort_process ( crnt_job );
        crnt_process = select_process ( crnt_job );
        crnt_node = selec_node ( cluster_arch );
        crnt_socket = select_socket ( cluster_arch );
        map_process ( crnt_process, crnt_node, crnt_socket );
        sort_adj ( crnt_process );
        map_adj_processes ( threshold );
    }
    4. job_pool = select_jobs ( medium_length );
    5. repeat steps 2,3;
    6. job_pool = select_jobs ( small_length );
    7. repeat steps 2,3;
}

```

Fig. 1. Pseudocode of the proposed mapping algorithm

volume of inter-process communications can lead to severe contention on network interface and degrade the performance. So before mapping processes of current job, we should determine a threshold on the number of processes which reside in a node and use network interface for their inter-node communications. To determine the threshold, we act as follow: If average adjacency for processes is less than or equal to average number of free cores ($\text{FreeCores}_{\text{avg}}$) in computing nodes (except one processing core which is used to place process ‘A’), approximately, we can say that ‘A’ and its adjacent processes can reside in just one node and there is no significant inter-node communications, probably. In such case, there is no need to determine a threshold. In contrast, if average adjacency is higher than the average free cores, some processes must be placed out of current node. In this case, threshold is determined by:

$$\text{Threshold} = \left\lfloor \frac{\sum_{i=1}^p \frac{\text{Adj}_{pi}}{\text{Adj}_{\max}}}{\text{num_of_nodes}} \right\rfloor \quad (1)$$

In eq. 1, a weight ($\frac{\text{Adj}_{pi}}{\text{Adj}_{\max}}$) is assigned to each process. In this weighted value, Adj_{pi} represents number of adjacent processes for process pi and Adj_{\max} represents maximum adjacency between processes. The reason for choosing a weighted threshold is because high amount of adjacency makes us determine a threshold. Consequently, processes which have more adjacency should have more impact (or weight, as a result) on selected threshold than others. The weighted value is then divided by the number of nodes (num_of_nodes) to distribute processes between all computing nodes. It is to be mentioned that although distributing processes between all cluster nodes, does not always lead to optimum results, but our experiments show that in many scenarios, it can result in efficient performance. An important note about eq. 1 is that if number of computing nodes is more than parallel processes, the threshold will be equal to 0 which is meaningless. In this case, we set the threshold value to 1.

4 Evaluation of the New Mapping Algorithm

4.1 Simulation Testbed

In this paper, we used Omnet++ v4.1 simulator to perform our experiments. The system which we considered for simulation, is a multi-core cluster containing 16 computing nodes which are connected through an intermediate switch. Each computing node has 4 sockets and each socket is a 4-core processor, so each node contains 16 processing cores. The architecture of each node is based on the NUMA¹ architecture. This means that each socket can access to its local memory (although it can also access to remote memories but with more latency). In each node, we used a network interface with InfiniBand technology. InfiniBand, is one of the most advanced technologies which is used to establish high performance clusters. Table 1 lists the parameters we used in our simulations.

¹ Non Uniform Memory Access.

Table 1. Simulation parameters

Parameter	Value
Main memory bandwidth	4GB/s
Remote memory access latency	10% more than local memory access latency
Cache bandwidth (for intra-chip communications)	Corresponds to AMD Opteron 2352 chip
Maximum length of common buffer in cache	1MB
Network interface bandwidth	1GB/s (corresponds to InfiniHost MT23108 4x)
Switching latency at intermediate switch	100ns (independent of message length)

4.2 Experimental Results

To evaluate the new mapping method, we used synthetic and real workloads. In synthetic workloads, messages which had different lengths and rates were generated. In these traffics, we used four different communication patterns between parallel processes. These patterns which are based on communication patterns in message passing libraries are: Bcast/Scatter, Gather/Reduce, All-to-All and Linear. In Bcast/Scatter, one process as the root process broadcasts its messages to other processes and other processes are just receiver. In Gather/Reduce, one process as the root process, receives messages from other processes and other processes are just senders. In All-to-All, each process sends messages to all other processes. In Linear, each process receives messages from a previous process and sends its messages to a next process (there is a linear communication pattern between processes). Tables 2 to 5 show the definition of 4 synthetic workloads which each, contains a number of parallel jobs with different communication patterns. Real workloads were extracted from communication behavior of NPB² benchmarks. Tables 6 to 9 show the definition of 4 real workloads which each, contains some benchmarks with different number of processes and different benchmark classes.

For performance evaluation, we used sum of the waiting times of messages at server queues (network interface and memory) as our main metric. We compared our results with the results obtained from Blocked, Cyclic and DRB methods. Fig. 2 shows

Table 2. Synt_workload_1

Job	No. of Processes	Pattern	Length	Rate	Message Count
0	64	All-to-All	64KB	100m/s	2000
1	64	Bcast/Scatter	64KB	100m/s	2000
2	64	Gather/Reduce	64KB	100m/s	2000
3	64	Linear	64KB	100m/s	2000

Table 3. Synt_workload_2

Job	No. of Processes	Pattern	Length	Rate	Message Count
0	64	All-to-All	2MB	10m/s	2000
1	64	Bcast/Scatter	2MB	10m/s	2000
2	64	Gather/Reduce	2MB	10m/s	2000
3	64	Linear	2MB	10m/s	2000

² NAS Parallel Benchmarks.

Table 4. Synt_workload_3

Job	No. of Processes	Pattern	Length	Rate	Message Count
0	32	All-to-All	2MB	10m/s	2000
1	32	Bcast/Scatter	2MB	10m/s	2000
2	32	Gather/Reduce	2MB	10m/s	2000
3	32	Linear	2MB	10m/s	2000
4	32	All-to-All	64KB	10m/s	2000
5	32	Bcast/Scatter	64KB	10m/s	2000
6	32	Gather/Reduce	64KB	10m/s	2000
7	32	Linear	64KB	10m/s	2000

Table 5. Synt_workload_4

Job	No. of Processes	Pattern	Length	Rate	Message Count
0	24	All-to-All	2MB	10m/s	2000
1	24	Bcast/Scatter	2MB	10m/s	2000
2	24	Gather/Reduce	2MB	10m/s	2000
3	24	Linear	2MB	10m/s	2000
4	24	All-to-All	64KB	10m/s	2000
5	24	Bcast/Scatter	64KB	10m/s	2000
6	24	Gather/Reduce	64KB	10m/s	2000
7	24	Linear	64KB	10m/s	2000

Table 6. Real_workload_1

Job	No. of Processes	Benchmark	Class
0	25	SP	C
1	32	IS	C
2	32	FT	B
3	16	FT	B
4	16	IS	C
5	32	CG	C
6	8	IS	B
7	25	BT	C
8	16	CG	B

Table 7. Real_workload_2

Job	No. of Processes	Benchmark	Class
0	8	IS	B
1	32	FT	B
2	32	IS	C
3	32	MG	C
4	32	CG	C
5	32	IS	B
6	32	MG	B
7	32	CG	B
8	16	BT	C

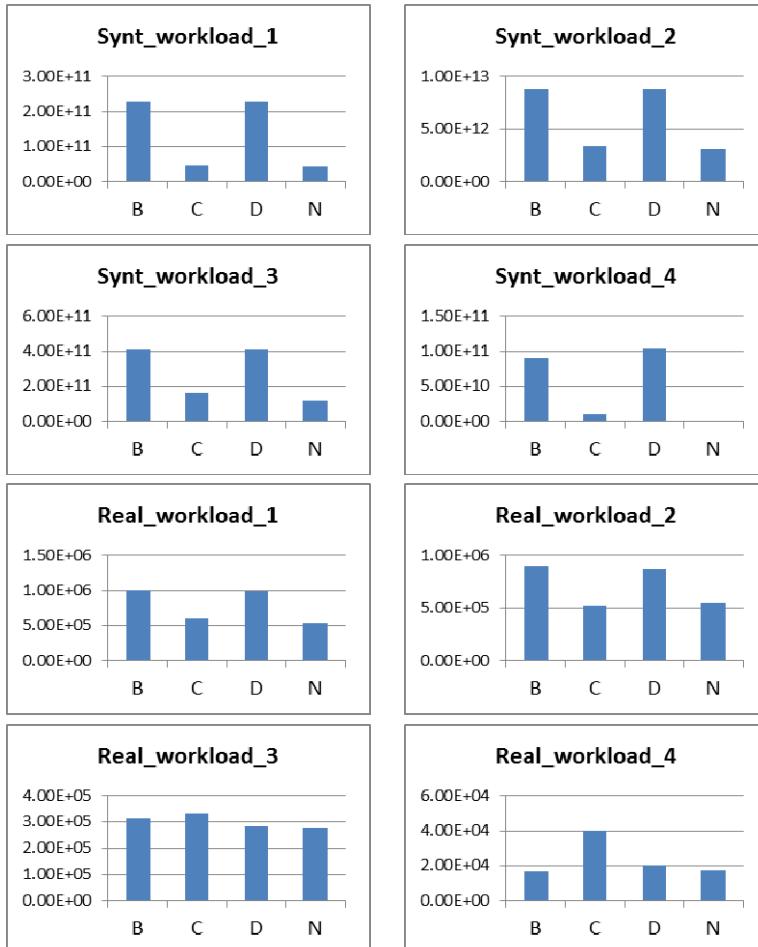
Table 8. Real_workload_3

Job	No. of Processes	Benchmark	Class
0	25	BT	B
1	32	CG	B
2	32	EP	B
3	32	FT	B
4	32	IS	B
5	25	LU	B
6	32	MG	B
7	25	SP	B

Table 9. Real_workload_4

Job	No. of Processes	Benchmark	Class
0	25	SP	C
1	32	CG	C
2	32	EP	C
3	32	MG	C

the performance results for 4 synthetic workloads and 4 real workloads. In this figure, ‘B’ indicates Blocked, ‘C’ indicates Cyclic, ‘D’ indicates DRB, and ‘N’ indicates our new mapping algorithm.

**Fig. 2.** Waiting time of messages for synthetic and real workloads (in mili-seconds)

According to Fig.2 it can be seen that the new mapping strategy, has produced better results compared to other methods. In synthetic workloads, the number of processes in parallel jobs is more than the number of processing cores within a node. Besides, significant part of communications is due to jobs which have All-to-All patterns. These factors cause

synthetic workloads to be heavy communicating workloads. In such workloads, the Blocked technique which tries to accommodate parallel processes in minimum number of nodes, has led to severe contention on network interface and has unacceptable performance, consequently. In contrast, Cyclic has gained better performance by distributing processes among computing nodes. Since in DRB method, processes which are communicating frequently, are mapped near each other, process mapping is done as Blocked and the results are not efficient. The reason that the new method has performed more efficient than Cyclic is that in the new algorithm, efficient mapping conditions is determined for each parallel job independent of other jobs. In other words, if the amount of adjacency and communications between processes, is high, the new method will distribute the processes, otherwise it acts like Blocked. Based on performance results, the new mapping technique has gained performance improvements up to 5%, 8%, 29% and 91% for Synt_workload_1 to Synt_workload_4 respectively (performance gain is calculated compared to the best result from other methods, i.e. Cyclic in here). In Real_workload_1 and Real_workload_2 scenarios, IS and FT benchmarks were used more than other benchmarks. These benchmarks have high communications and their communication pattern is All-to-All entirely. As a result, the above mentioned workloads are heavy and as can be seen in Fig. 2, the Cyclic method has performed better than the Blocked and DRB methods. In these workloads, the new approach has acted as efficient as Cyclic and even better (in Real_workload_1 scenario, 11% performance improvement is observed). In order to show that our approach can perform efficiently not only in heavy workloads, but also in non-heavy workloads, Real_workload_3 and Real_workload_4 were used. Real_workload_3 is a medium workload in term of communications and as can be seen in Fig. 2, there is no significant difference between performance results of different methods for this scenario. Despite this, the new mapping technique has performed a little bit better than others. Real_workload_4 is a scenario which has light communications and as we can expect, Blocked and DRB methods have better results than Cyclic. Performance results for this scenario show that the new mapping method has performed as well as Blocked which indicates that the new approach can have efficient results even in light communicating workloads.

5 Conclusion

In this paper, we proposed a new process mapping algorithm to assign parallel processes to multi-core clusters aimed to reducing network interface contention. Since the number of processing cores within recent computing nodes is growing up rapidly, contention on shared resources is posing itself as a serious challenge and should be considered for optimizing performance. Here, we tackled this problem and proposed a process placement algorithm to alleviate contention on network interface as one of the main shared resources. We compared our technique with other well-known methods and observed that improved performance was gained (5% to 90%) in experimental workloads. Our mapping algorithm is easy to implement and its efficiency makes it usable in recent high performance multi-core clusters.

Acknowledgments. We thankfully appreciate Research Institute for ICT (ITRC) of Iran for supporting us in this project. We hope this work be a valuable research to extend the technological knowledge of this institute.

References

1. <http://www.top500.org>
2. Hood, R., Jin, H., Mehrotra, P., Chang, J., Djomehri, J., Gavali, S., Jespersen, D., Taylor, K., Biswas, R.: Performance Impact of Resource Contention in Multicore Systems. In: IEEE International Symposium on Parallel and Distributed Processing, Atlanta (2010)
3. Chai, L., Gao, Q., Panda, D.K.: Understanding the Impact of Multi-Core Architecture in Cluster Computing: A Case Study with Intel Dual-Core System. In: 7th IEEE International Symposium on Cluster Computing and the Grid, Rio De Janeiro, Brazil (2007)
4. Jokanovic, A., Rodriguez, G., Sancho, J.C., Labarta, J.: Impact of Inter-Application Contention in Current and Future HPC Systems. In: IEEE Annual Symposium on High-Performance Interconnects, Mountain View, U.S.A (2010)
5. Kayi, A., El-Ghazawi, T., Newby, G.B.: Performance issues in emerging homogeneous multi-core architectures. Elsevier Journal of Simulation Modeling Practice and Theory 17(9) (2009)
6. Narayanaswamy, G., Balaji, P., Feng, W.: Impact of Network Sharing in Multi-core Architectures. In: 17th IEEE International Conference on Computer Communications and Networks, Virgin Islands, U.S.A (2008)
7. Dummler, J., Rauber, T., Rungger, G.: Mapping Algorithms for Multiprocessor Tasks on Multi-core Clusters. In: 37th IEEE International Conference on Parallel Processing, Portland, U.S.A (2008)
8. Ichikawa, S., Takagi, S.: Estimating the Optimal Configuration of a Multi-Core Cluster: A Preliminary Study. In: IEEE International Conference on Complex, Intelligent and Software Intensive Systems, Fukuoka, Japan (2009)
9. Chen, H., Chen, W., Huang, J., Robert, B., Kuhn, H.: MPIPP: An Automatic Profile-guided Parallel Process Placement Toolset for SMP Clusters and Multiclusters. In: 20th Annual International Conference on Supercomputing, New York, U.S.A (2006)
10. Mercier, G., Clet-Ortega, J.: Towards an Efficient Process Placement Policy for MPI Applications in Multicore Environments. In: 16th European PVM/MPI Users' Group Meeting on Recent Advances in Parallel Virtual Machine and Message Passing Interface, Berlin, Germany (2009)
11. Rodrigues, E.R., Madruga, F.L., Navaux, P.O.A., Panetta, J.: Multi-core Aware Process Mapping and Its Impact on Communication Overhead of Parallel Applications. In: IEEE Symposium on Computers and Communications, Sousse, Tunisia (2009)
12. Khoroshevsky, V.G., Kurnosov, M.G.: Mapping Parallel Programs into Hierarchical Distributed Computer Systems. In: 4th International Conference on Software and Data Technologies, Sofia, Bulgaria (2009)
13. Jeannot, E., Mercier, G.: Near-Optimal Placement of MPI Processes on Hierarchical NUMA Architectures. In: D'Ambra, P., Guerracino, M., Talia, D. (eds.) Euro-Par 2010, Part II. LNCS, vol. 6272, pp. 199–210. Springer, Heidelberg (2010)
14. Agrawal, T., Sharma, A., Kale, L.V.: Topology-Aware Task Mapping for Reducing Communication Contention on Large Parallel Machines. In: 20th IEEE International Symposium on Parallel and Distributed Processing, Rhodes Island, Greece (2006)
15. Koop, M.J., Luo, M., Panda, D.K.: Reducing Network Contention with Mixed Workloads on Modern Multicore Clusters. In: IEEE International Conference on Cluster Computing and Workshops, New Orleans, U.S.A (2009)
16. Koukis, E., Koziris, N.: Memory and Network Bandwidth Aware Scheduling of Multiprogrammed Workloads on Clusters of SMPs. In: 12th International Conference on Parallel and Distributed Systems, Minneapolis, U.S.A (2006)

Resource Based Optimized Decentralized Grid Scheduling Algorithm

Piyush Chauhan and Nitin

Department of CSE and ICT
Jaypee University of Information Technology,
P.O. Waknaghat, Solan-173234, Himachal Pradesh, India
shbichauhan@gmail.com, delnitin@ieee.org

Abstract. Peer to peer (P2P) grid system has good potential for decentralized grid scheduling. The existing P2P grid resource management algorithms allow detection of resources after generation of task. Recently, few new P2P grid resource management schemes proposed discovery of resources before task is generated. These schemes simply shortlist best possible subset of grid resources on basis of one or many overlays. In this paper, we suggest that these algorithms have potential to include the step of organizing shortlisted resources in non-increasing order on basis of optimization criteria. Addition of this step will reduce time to find out schedule of DAG's interdependent tasks on fully decentralized P2P grid. An optimization criterion is based on computation capability and communication cost of grid resources.

1 Introduction

Scenario of utility computing is shifting very rapidly. Newly achieved fame of cluster, cloud and grid computing is changing the utility computing for better results in fewer expenses. In grid computing owning powerful super computers is not feasible for every organization and therefore, researchers are expecting next generation grids consist of hundreds and thousands of minute size clusters or individual nodes.

Essential building block for any grid computing systems [1] is scheduling [2] various subtasks among resources of grid [3] in optimized fashion. To distribute jobs with help of central scheduler is impossible because of scalability issue. Therefore, meta scheduling came into existence. Many grid computing platforms use meta-scheduling approach. In this approach, each cluster posses one local scheduler and job is given to the cluster, which has the capability of executing it.

However, next generation grids are expected to be of petite size clusters and single nodes. Minuscule size of cluster makes impossible the scheduling of huge task on single cluster. This is the reason of shifting from traditional meta-scheduler based client-server model to decentralized peer to peer grid [4], [5] system.

P2P [6] based grid systems are based on cooperation among peers. Peers share work and obtain results. P2P based grid systems achieve results, by splitting work among participating nodes; instead of depending on expensive centralized server. In P2P based grid computing algorithms each computer node be accountable for setting up its own task implementation agenda, and intentionally overlooks structure of grid as pool of clusters.

Nodes of P2P grid computing environment communicate in a gossip based pattern. By gossip [7] based pattern they shortlist subset of grid nodes that can best accommodate requirements of all submitted task. Here task can generate at any node of grid. Gossip based unstructured P2P grid [8] computing systems show self healing behavior in case of highly dynamic environments. High dynamic environment symbolizes that rate of joining and leaving of grid nodes is very high.

To handle dynamicity, one important property of P2P is that nodes/peers of grid jointly keep an overlay network [9]. Structures of these overlay networks depend upon logical relationships between peers, not just on physical existence and connectivity of nodes. Hence, overlay management is vital in P2P based grid computing systems. Gossip based overlays come in category of unstructured P2P systems. Set of two layered overlay gossip protocol is used in this paper. Bottom overlay called CYCLONE [10]; its work is to feed top overlay with nodes randomly selected from grid. Second overlay does not select nodes randomly; it keeps the best nodes on basis of overlay metric. Second layer is called VICINITY [11] layer. VICINITY layer takes as input set of nodes shortlisted by CYCLON and this set is further shortlisted on basis of VICINITY overlay metric.

In this paper, we propose an algorithm which concentrate on the output of top gossip based overlay. Top gossip based overlay gives in output a set of shortlisted nodes. These nodes are as close as possible on basis of metric of VICINITY overlay. We arrange these shortlisted nodes in non increasing order on basis of optimization criteria. Optimization criteria depend on computation capability and cost of communicating with node. This hierarchical arrangement is done before task is generated on grid node. This proactive step of arranging shortlisted nodes before task is generated, prove to be stitch in time for reducing computations and comparisons, which arise in finding optimal schedule. Optimal schedule here is obtained for DAG's interdependent tasks on fully decentralized P2P grid upon generation of task on any grid node.

The rest of the paper is structured as follows: Section 2 contains motivation and background work. Section 3 reveals the proposed algorithm in detail. In section 4 we discuss about simulation results, finally section 5 gives conclusion and future scope of work.

2 Motivation and Background

Many researchers have accepted need for decentralized grid scheduling algorithm. However, there is no good algorithm, which emphasize on shortlisted grid nodes usage scheduling [12]. Most of existing work is on resource detection. Hence, no approach deals with scheduling of grid nodes and scalability issues concurrently.

Zorilla [13] is one specimen of scalable scheduling scheme. In Zorilla when job request come up then it hunts for vacant nodes. Zorilla performs this step with the help of dissemination algorithm. Shortcoming of Zorilla is that when job request come up there should be an adequate number of unused nodes. However, nodes shortlisted by Zorilla show geographical locality. Another approach is LHC [14] computing grid project at CERN which gives applications the right to choose resources for their use. This approach gives results because application specific resources are chosen, but it only depends upon resource discovery. We need advance planning for resource utilization.

One example of system, which shortlist nodes before task is generated is CYCLON. Cyclon is gossip based approach for unstructured peer to peer systems. Cyclon approach is very simple each node recognizes a small constantly changing set

of its neighboring nodes. Occasionally, node into consideration changes its caches with neighbor whose information was the most primitive one to be injected in the network. Overlay of Cyclon show property of random graph, and is robust enough in case of enormous node failures. In place of basic shuffling Cyclon use enhanced shuffling. In basic shuffling node take any randomly chosen neighbor and exchange cache with that neighbor, whereas in enhanced shuffling neighbor is not chosen randomly. Motive behind this enhancement is to halt pointers to dead nodes from roaming around forever. This all become possible by adding one extra field in cache. This new field is called age. Age field denote coarsely age of entry since the instant it came into existence. Hence, Cyclon gives a framework for extremely huge P2P [15] overlays of scalable and inexpensive membership management. Not only it show property of random graph, Cyclon also helps in lowering of contamination of caches from out-of-date references to previous members. Shortcoming of Cyclon is that machines are selected in random; hence, even finest found set will scarcely satisfy user expectation.

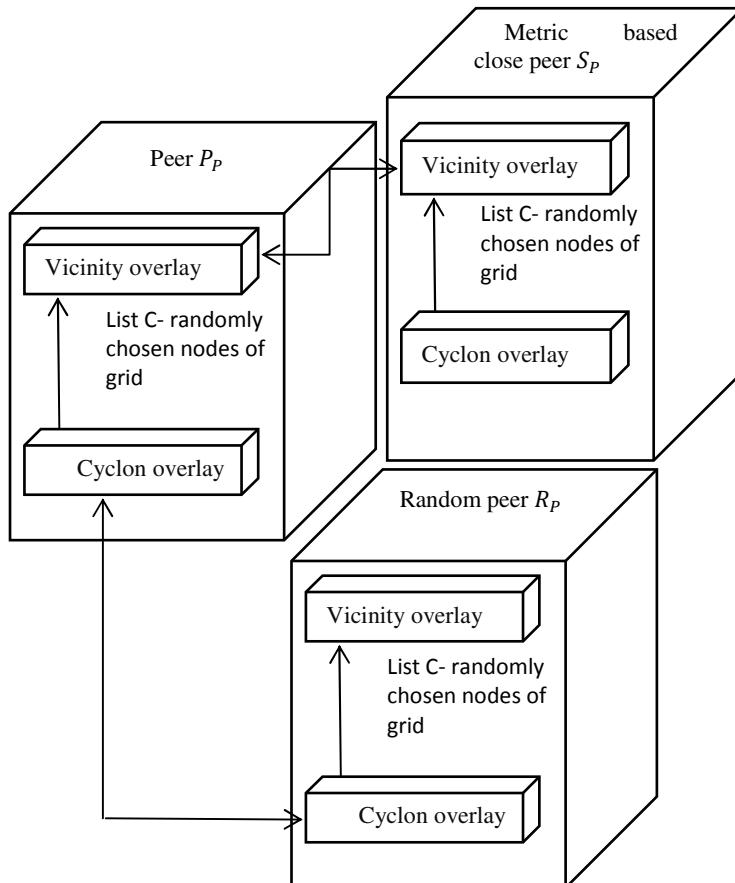


Fig. 1. Cyclon and Vicinity based model.

Whereas, Cyclon uses single overlay to shortlist resources before task is generated, there is other scheme that uses multiple overlays in a layer wise structure. Second overlay layer is called Vicinity [11], [16], [17]. Vicinity comes above Cyclon layer. Cyclon's output is shortlisted list of nodes which is used as initial input for Vicinity. Vicinity further shrinks this shortlisted nodes list based on overlay metric. Here Genetic algorithm [18], [19] is used for selection process to achieve ideal results in polynomial time. Hence, it is possible with Vicinity to shape an overlay connecting nodes with largely coinciding first available time slot. Here Vicinity act as helper overlay which connects nodes sharing certain characteristics. Vicinity and Cyclon and their relationship are visible in figure 1. We can add more overlays over Vicinity, also we can use overlays based on other metric in place of Vicinity.

Work in [4] uses Cyclon and Vicinity as two layered framework present at each peer of grid for decentralized grid scheduling. First layer Cyclon selects fresh randomly selected set of nodes from grid. Next layer Vicinity takes output of Cyclon layer as input and yield set of nodes having some similar characteristic. This second layer act as lightweight helper overlay. We manipulate list of nodes that are output of second layer and hierarchically arrange these nodes on basis of optimization criteria. This proactive step proves very fruitful in grid scheduling algorithm.

3 Proposed Decentralized Grid Scheduling Algorithm

In this section we have proposed a model with fully decentralized schedule for grid computing system. We start with first part of our approach that is modified version of [4], in second portion of our model we have used modifications of first part to speed up scheduling algorithm; and yield good quality schedules for grid.

Our approach is divided into two sections. In first section each node achieve hierarchical organized resource list before task is generated. Second section involves algorithmic steps to efficiently schedule sequentially arranged mutually dependent parts of task, represented by nodes of DAG [20]. Second section depends upon final output of first section to produce good and fast decentralized scheduling algorithm for grid system.

First section of our model consists of three steps: First step uses Cyclon gossip protocol. Each node keeps a small list of arbitrary links to other nodes in a grid. Every node from time to time selects one node C from this list, and exchanges links of C node with its links having highest age field value. In this way, every node is randomly supplied with a refreshed set of links to other arbitrarily chosen nodes. Unavailable nodes are immediately removed from list of other nodes. Sturdy overlay is created which can withstand crash of huge number of nodes in grid.

Second list C and nodes in list C into consideration. These nodes are randomly chosen with help of Cyclon gossip protocol in step one. This list C act as input for second step which uses Vicinity gossip layer to short list number of nodes from list C , such that new list contain only semantically related nodes. This means Vicinity overlay short list nodes based on optimization function. Various metrics are available on basis of which optimization function shortlist nodes in list C . One such metric is connecting nodes with largely overlapping first accessible time slot, by utilizing size of this overlap as optimization function. This step yield further shortened List V of nodes from list C .

Third step includes hierarchical arrangements of the nodes listed in V on bases of their computation cost and communication cost in list HRL . Hence, third step uses optimization criteria, which depend on computation capabilities of node and cost of communication involved with this node. We arrange nodes of list V in non increasing order on basis of above mentioned metric in list HRL . This hierarchical arrangement of resources present in list V helps in reducing time required to calculate the best schedule for tasks generated in various nodes of grid. First proactive portion of our algorithm yields hierarchically arranged shortlisted nodes. Tri-layered structured model mentioned above is pictured in figure 2.

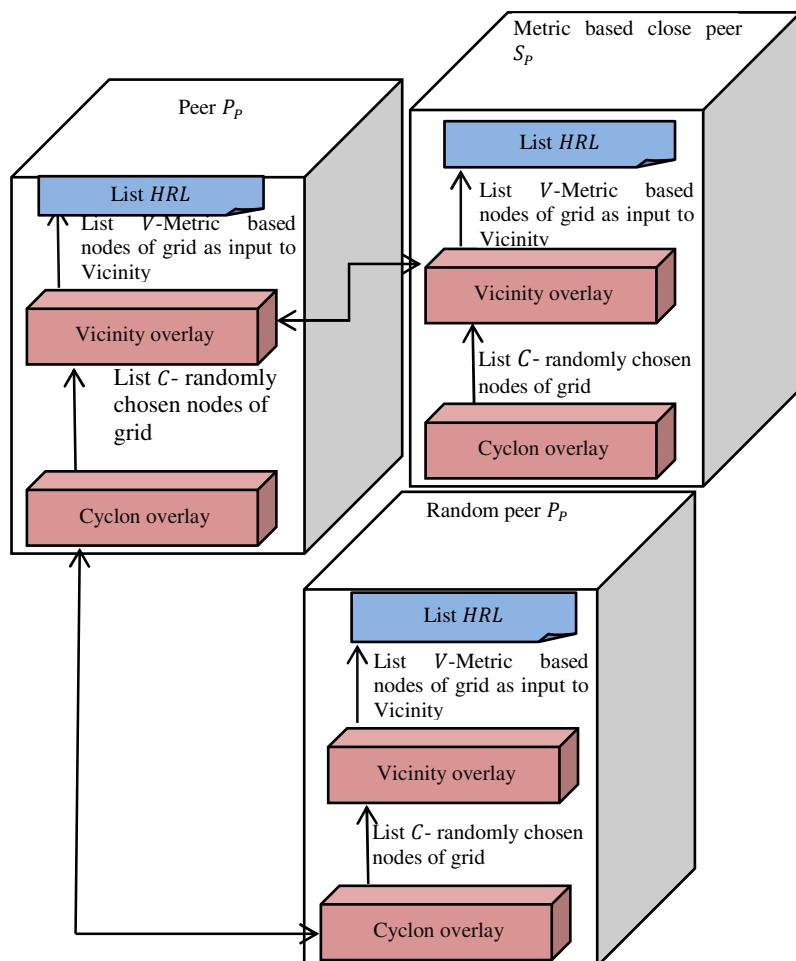


Fig. 2. New model with List HRL.

3.1 Proactive Hierarchical Arrangements of Resource

Proposed algorithm for resources optimization before task is generated in decentralized grid requires that each grid resource contain fixed size cache of Z links. Ordering and shuffling of resources is carried out when initiating peer C performs following steps:

1. C select arbitrary subset of A neighbors from set of neighbors of C .
2. Age of all links in List A is incremented by one unit.
3. Choose Neighbor O having maximum age in all neighbors marked in list A .
4. Replace O 's links with link of age zero and having C 's address.
5. Send modified subset to O node's list.
6. Receive from O a subset of no more than S of its own links.
7. Remove entries pointing towards C and entries already present in C 's list of links.
8. Update C 's list of links to incorporate all entries, initially using vacant cache slots; if not available replacing entries among the ones sent to O .
9. Using metric that shortlist resources from C 's list with overlapping first accessible time slot generate refined subset list V (Size of overlap is unit for metric).
10. List V 's entries are arranged hierarchically in decreasing order on basis of Computation ability and communication cost (This list is called hierarchical resource list (HRL)).

With help of above mentioned steps we will obtain a list of resources available to any grid node for processing its sub task. $\{R_1, R_2, R_3 \dots R_i \dots, R_{(X-1)}, R_X\}$ is resources sequence acquired by executing above mentioned steps.

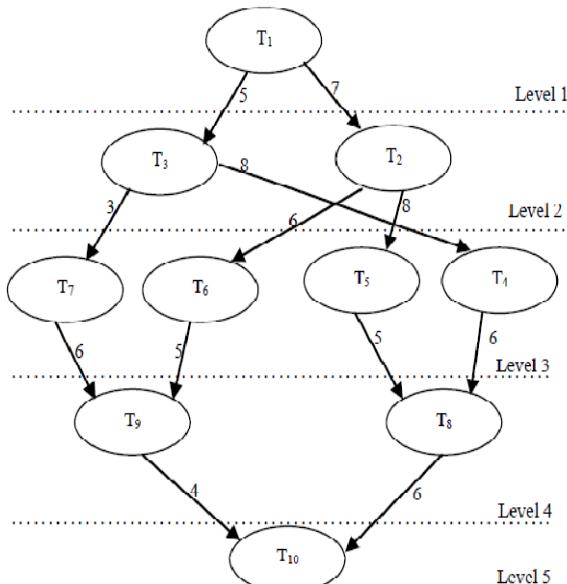


Fig. 3. Random DAG.

3.2 Resource Based Optimized Decentralized Grid Scheduling Algorithm

DAG [21] denotes task generated, which is further subdivided into fine grain subtask, in figure 3. This DAG is divided into various precedence levels. At each precedence level of DAG, tasks are decreasing in size from right to left. Until we have executed all tasks in one precedence level we will not jump to next precedence level. This is because tasks in next precedence level will depend upon output of parent tasks present in previous precedence level. Sample DAG is shown in figure 3. As per task based proactive algorithm (pseudo code shown in figure 4.) we will execute task T_1 on first resource R_1 present in μ list. (Initially μ list will be vacant and later on empty resources will be added to μ list from *HRL* list). New vacant resource will be added to μ list because all existing resources are occupied by other sub task. Next task T_2 , which is on next precedence level, will get executed on same resource R_1 because it is right most child task of parent task T_1 . Now add task T_3 to resource in list μ , which is capable of finishing it first. This way we will assign all tasks of Sample DAG to best fit grid resources.

Algorithm utilizing output of hierarchical resource list (*HRL*) is given in figure 4 below:

RESOURCE BASED OPTIMIZED DECENTRALIZED GRID SCHEDULING ALGORITHM

```

1. Generate priority based task sequence  $\alpha$ .
2. Generate empty list  $\mu$  of processors utilized.
3. do{
4.     Select initial unscheduled task in sequence
 $\alpha$ ;
5.     if( no vacant resource in  $\mu$  list )
6.     {
7.         Add first resource from HRL to  $\mu$  list;
8.     }
9.     if(task is right most child of parent
node)
10.    {
11.        Assign it to resource executing parent
task;
12.    }
13.    else
14.    {
15.        for(all processors in  $\mu$  list)
16.            :execute task on resource in
list  $\mu$  giving minimum finish time for
task;
17.    }
18. }while(unscheduled task present in task
sequence  $\alpha$ );

```

Fig. 4. Pseudo code of Task based proactive algorithm.

This way we will assign task to resources best fit for it. Here we will get good schedule in less comparison between resources and tasks. This reduction in comparison between resources and tasks proves to be very fruitful for fine grain DAG's with huge sizes.

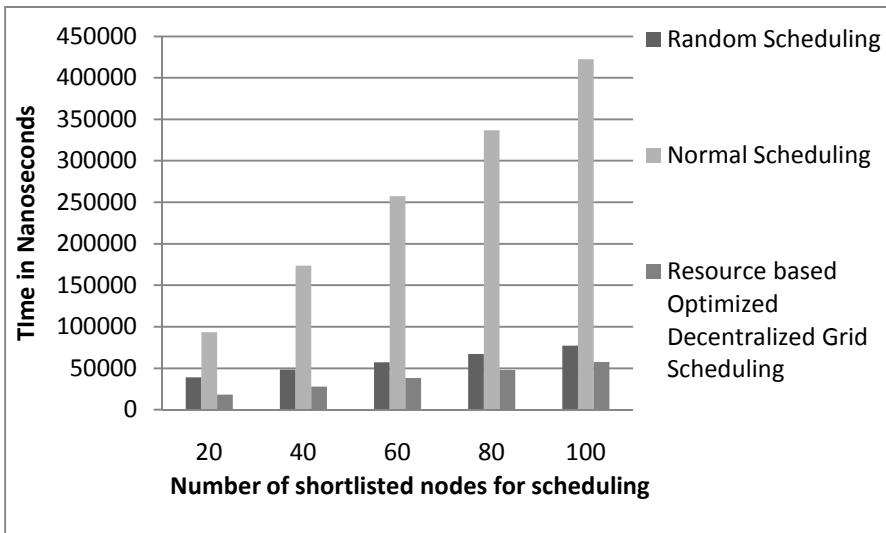


Fig. 5. Comparison in terms of time of random, normal and resource based optimized decentralized grid scheduling algorithms.

4 Experimental Results

Experimental results for Resource based optimized decentralized grid scheduling algorithm for grid is shown in figures 5 in the simulated grid environment. The simulated set of experiments compare the performance of the Resource based optimized decentralized grid scheduling algorithm in terms of time utilized to find the schedule with respect to various grid sizes.

The total time cost in finding optimum schedule of task graph for various grid sizes is average of 10 simulation runs. Java version Net Beans IDE 6.9.1 was used on computer AMD Athlon(tm) 64x2 dual core processor 4600+2.41GHz with 1gb DDR2 RAM as test bed to run java based simulation. Simulation was run 10 times for set of different numbers of shortlisted grid resources for scheduling tasks of DAG. The five sizes of shortlisted resources 20, 40, 60, 80 and 100 were taken as set of shortlisted nodes.

In figures 5 experimental results is shown for random allocation of Grid resources for interdependent tasks of DAG. Time required to obtain random schedule is very less when compared with time to obtain normal scheduling as shown in figure 5. It shows the time cost to obtain schedule for all task of DAG. This schedule is outcome of comparing all tasks of DAG for all shortlisted nodes and assigning task to grid node which yield most suitable value of optimization criteria. Schedule obtained by

this method is much better than random scheduling but time to calculate schedule also increases manifolds. We remark it as normal scheduling.

Using Resource based optimized decentralized grid scheduling algorithm we can reduce time to calculate schedule as shown in figures 5. In addition, quality of schedule remains same as normal scheduling. Hence good schedule in less time are obtained for Resource based optimized decentralized grid scheduling algorithm. Also, gap between time taken by normal scheduling algorithm and Resource based optimized decentralized grid scheduling algorithm tends to increase with increase in size of shortlisted nodes of grid.

Hence, resource based optimized decentralized grid scheduling algorithm will continue to produce schedule results even faster than random scheduling as size of DAG and Grid increases.

5 Conclusion and Future Scope of Work

This paper proposes a new lightweight decentralized grid scheduling algorithm. Unlike existing grid scheduling algorithm, this approach does not have centralized scheduler and hence avoids single point of failure. In addition, absences of central scheduler eliminate performance bottlenecks caused by it. Our approach depends upon gossip based P2P overlay. Our approach is not only decentralized in nature but it gives schedule for task sequence in less time and comparisons. This was possible due to less comparison involved to assign tasks to best fit resources. Huge DAGs require large set of resources. It is not intelligent to compare all these resources for every task. This becomes time consuming with bulky list of shortlisted resources. This algorithm reduces the cost of comparison of task, with resources of grid by taking into consideration only relevant resources from *HRL*. As for the future consideration, we can include more overlay layers to get more refined set of resources of grid for computation of grid tasks.

References

1. Dong, F., Akl, S.G.: Scheduling Algorithms for Grid Computing: State of the Art and Open Problems. Technical Report No. 2006–504, School of Computing, Queen’s University, Kingston, Ontario (2006), doi: 10.1017/s030574101000024x
2. You, S.Y., Kim, H.Y., Hwang, D.H., Kim, S.C.: Task Scheduling Algorithm in GRID Considering Heterogeneous Environment. In: Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, pp. 240–245 (2004)
3. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
4. Fiscato, M., Costa, P., Pierre, G.: On the feasibility of Decentralized Grid Scheduling. In: Proceedings of the 2nd IEEE Conference on Self-Adaptive and Self-Organizing Systems Workshop, pp. 225–229 (2008)
5. Beaumont, O., Carter, L., Ferrante, J., Legrand, A., Marchal, L., Robert, Y.: Centralized versus distributed schedulers for multiple bag-of-tasks applications. IEEE Transaction on Parallel Distributed System 19, 698–709 (2008)

6. Risson, J., Moors, T.: Survey of Research towards Robust Peer-to-Peer Networks: Search Methods. *The International Journal of Computer and Telecommunications Networking* 50, 3485–3521 (2004), doi:10.1016/j.comnet.2006.02.001
7. Jelasity, M., Guerraoui, R., Kermarrec, A.-M., van Steen, M.: The Peer Sampling Service: Experimental Evaluation of Unstructured Gossip-Based Implementations. In: Jacobsen, H.-A. (ed.) *Middleware 2004*. LNCS, vol. 3231, pp. 79–98. Springer, Heidelberg (2004)
8. Voulgaris, S., Jelasity, M., Steen, M.V.: A robust and scalable peer-to-peer gossiping protocol. In: 2nd International Workshop on Agents and Peer-to-Peer Computing, pp. 47–58 (2003)
9. Jelasity, M., Babaoglu, O.: T-Man: Fast gossip-based construction of large-scale overlay topologies. Technical Report UBLCS-2004-7, University of Bologna, Department of Computer Science, Bologna, Italy, vol. 7, pp. 2004–2007 (2004)
10. Voulgaris, S., Gavidia, D., Steen, M.V.: Cyclon: Inexpensive Membership Management for Unstructured P2P Overlays. *Journal of Network and Systems Management* 13(2), 197–217 (2005)
11. Voulgaris, S., van Steen, M.: Epidemic-Style Management of Semantic Overlays for Content-Based Searching. In: Cunha, J.C., Medeiros, P.D. (eds.) *Euro-Par 2005*. LNCS, vol. 3648, pp. 1143–1152. Springer, Heidelberg (2005), doi:10.1007/11549468
12. Schopf, J.M.: Ten Action when scheduling. In: *Grid Resource Management-State of the Art and Future Trends*, ch. 2, pp. 15–23. Kluwer Academic Publishers (2003)
13. Drost, N., Van Nieuwpoort, R.V., Bal, H.E.: Simple Locality-Aware Co-allocation in Peer-to-Peer Supercomputing. In: *Sixth IEEE International Symposium on Cluster Computing and the Grid Workshops*, pp. 8–14 (2006), doi:10.1109/CCGRID.2006.1630909
14. Lamanna, M.: The LHC computing grid project at CERN. *Nuclear Instruments and Methods in Physics Research* 534(1-2), 1–6 (2004)
15. Ganesh, A.J., Kermarrec, A.M., Massoulie, L.: Peer-to-Peer Membership Management for Gossip-based Protocols. *IEEE Transactions on Computers* 52(2), 139–149 (2003), doi:10.1109/TC.2003.1176982
16. Voulgaris, S., van Steen, M.: An Epidemic Protocol for Managing Routing Tables in Very Large Peer-to-Peer Networks. In: Brunner, M., Keller, A. (eds.) *DSOM 2003*. LNCS, vol. 2867, pp. 41–54. Springer, Heidelberg (2003)
17. Voulgaris, S., Kermarrec, A., Massoulie, L., Steen, M.V.: Exploiting semantic proximity in peer-to-peer content searching. In: *10th International Workshop on Future Trends in Distributed Computing Systems*, Suzhu, China, vol. 7695, pp. 2118–2125 (2001)
18. Eiben, A., Smith, J.: *Introduction to Evolutionary Computing*. Natural Computing Series. Springer (2003)
19. Iordache, G., Boboila, M., Pop, F., Stratan, C., Cristea, V.: A Decentralized Strategy for Genetic Scheduling in Heterogeneous Environments. *Journal Multiagent and Grid Systems* 3(4), 355–367 (2007)
20. Forti, A.: DAG Scheduling for Grid Computing systems. Ph.D. Thesis, University of Udine – Italy (2006)
21. Cao, H., Jin, H., Wu, X., Wu, S., Shi, X.: DAGMap: Efficient Scheduling for DAG Grid Workflow Job. In: *9th IEEE Grid Computing Conference*, pp. 17–24 (2008), doi:10.1109/GRID.2008.4662778

Web-Based GIS and Desktop Open Source GIS Software: An Emerging Innovative Approach for Water Resources Management

Sangeeta Verma, Ravindra Kumar Verma*, Anju Singh, and Neelima S. Naik

Centre of Environmental Studies, National Institute of Industrial Engineering (NITIE),
Vihar Lake, Mumbai- 400087, India
ravindraraj2008@gmail.com

Abstract. This paper introduces an overview of Web-based GIS and its applications and some of the easily assessable Desktop Open Source GIS Software with easy-to-follow guidance that will help water resource decision-maker and interested stakeholder. Web-based GIS is a prospective application in GIS and represents an important advancement over the traditional desktop GIS. Its application eliminates duplication and inconsistency (which is often possible between GIS professionals) and makes location information conveniently and intuitively accessible across organization/s, at a lower cost per user. Internet provides a medium for processing geo-related information and spatial information to users at an amount larger than traditional GIS. XML and Java have been developed to facilitate the utilization of the internet as well as to provide a coding standard in the software industry. Therefore, in the last few years, there has been a significant development in the area of free and open source GIS software. This paradigm shift from stand-alone GIS to open access Web GIS services provide greater opportunities for sustainable solutions in water resource management and planning.

Keywords: Geographic Information System, Web-based GIS, Water Resource Management, Open Source GIS Software.

1 Introduction

Water is a scarce natural resource. Therefore, managing water resource is extremely important for sustainable development in many parts of the world. The fact that the world faces water crises has become increasingly clear in recent years. These challenges will intensify unless effective and concerted actions are taken [1]. These challenges call for innovative approaches {like Modelling, Remote sensing (RS), Geographic information systems (GIS), Interoperability, Data models, Web services, Web-based GIS, Mobile GIS} because of the dramatic change in water resource management that has occurred during the past few years. Each of these approaches will help water professionals with their work and provide a foundation for continued success.

* Corresponding author.

With respect to water resource management domain, the immediate challenges are (1) to identify the ways in which GIS can facilitate more effective and/or more efficient water resource management (such as development of a GIS module specific for modelling applications that includes multidimensional and time series display capabilities, development of water resource model code for GIS insertion) [2], (2) to develop WebGIS-based Spatial Decision Support Systems (SDSS) that will address specific water resource challenges and problems, and (3) to train the next generation of water resource scientists, engineers, and policy analysts to sustain the continued evolution and appropriate use of Web-based GIS water resource applications [3].

GIS is a system of hardware, software and procedure that capture, store, edit, manipulate, manage, analyse, share and display georeferenced data. In order to take advantage of GIS to improve water resources planning and management, it is an attractive idea to integrate GIS with traditional hydrological models (like HEC-1, HEC-2, MODFLOW, SHE, SWAT, BASIN etc.) more efficiently and to include at least some level of spatial effects by partitioning entire watersheds into smaller sub-watersheds and further predict surface, ground water and rainfall-runoff flows. As a result, the interpretation of spatial data becomes easy and increasingly simple to understand.

Since 1980s, various researchers have adopted different approaches, such as loosely coupling, tight coupling and system embedding, to integrate GIS and hydrological models [4]. However, this integration research approach has several important drawbacks. The most important drawback is that the users must install the expensive GIS software and hydrological models in order to use the integrated system. Meanwhile, the integrated system usually has unfriendly user interface and relies on the specific operating system, such as Windows, Linux and Unix. In embedded coupling, another drawback is that the mathematical model is limited by power of the user interface [5]. Consequently, this drawback brings several serious problems, such as the high cost of system deployment and system maintenance, the difficulties in using the integrated system, and the difficulties in collaborating between different users [6, 7] and also increase in system complexity and becomes a challenge for the developers.

Nowadays, GISs have revolutionized many aspect, especially with the advent of the Internet and Web. In WebGIS, the internet technologies are connected with GIS in order to take advantage of their special characteristics, such as easy usability, use the GIS data such as input, adjustment, manipulation, analysis, output of geographical information and to bring out related service on the Internet. Whereas previous stand-alone GISs had restricted application capability on network, the Internet GIS makes it possible to retrieve and analyse spatial data through the web.

Internet also provides a medium for processing geo-related information with no location restrictions. As such, internet reshapes all functions of GIS including: gathering, storing, retrieving, analysing and visualizing data [9]. This shows that internet enables the popularity of the results of GIS analysis and spatial information to users at an amount larger than traditional GIS. Furthermore, respective user may set up an online browse or raising online analysis of spatial subjects free of cost and without any investment for any GIS Software. Extensible Markup Language (XML) and Java have been used to facilitate the development of the internet, to provide a coding standard in the software industry [9] and to make the development of the GIS software much easier. Thus, WebGIS is useful tool for water resource management as

it gives users better understanding of the overall picture i.e. locations of rivers/basins, topography of the flooded/drought areas, linkages of geographical factors and natural disasters occurred, water demand and supply thus gives users the ability to find best solution for each area and manage water resources in a sustainable manner.

The Open Geospatial Consortium (OGC) is an organization for geospatial data and web services. In 2000, it has initiated the Open Web Service (OWS) program based on service-oriented architectures and web service, and has proposed several geospatial specifications to support geospatial data sharing and interoperation, such as Web Map Service (WMS), Web Feature Service (WFS), and Web Processing Service (WPS) [9]. WMS has its ability to produce maps rather than its ability to access specific data holdings, and generates spatially referenced maps dynamically. It contains a HTTP interface for requesting geo registered maps from different distributed databases. [10]. WFS defines the interfaces for the access and manipulation of geographical features and elements through Geography Markup Language (GML) [11]. WPS provides standardized interfaces to facilitate publishing, discovering and binding geospatial services that enable spatial processing functions across a network [12]. Therefore, in past few years, there has been a significant development in the area of Free and Open Source geospatial Software (FOSS) in the GIS community. GIS software fulfilling the specific requirements have been distributed with licenses that grant more freedom of use and that support openness, such as licenses used by FOSS GIS projects. For instance, in the last two years, 20 entries have been added to the list of software projects on the website FreeGIS.org (now containing 330 entries). This shift from traditional planning to open access Web GIS planning services is related to the recognition that structured cooperation provides greater opportunities for sustainable solutions and that cost effective access to baseline data is needed for effective management and planning [13].

This paper focuses on need to adopt Web-based GIS tool and easily accessible FOSS in decision making related to water resource management mainly in developing countries that have limited financial resources. The rest of the paper is structured as follows; Section 2 introduces advantages and architecture of Web-based GIS system. While, Sub-section 2, provides easy-to-follow guidance (web address and sources) from where water professions can collect data and download softwares free of cost for WebGIS and SDSS or both, and Section 3 presents a short review of WebGIS applications in water resource management. Finally, Section 4 concludes there view.

2 The Web GIS

WebGIS is a relatively inexpensive way of disseminating spatial data and basic GIS functionality. A good portion of the basic functionality of desktop GIS is now available to users interacting with GIS databases via the World Wide Web (WWW) or an Intranet. The tool has following advantages over conventional GIS (modified from Tripathy (2002) [14]):

- It can be used to disseminate information to large number of local and international users.
- It provides opportunities for public feedback, participation and collaboration in managerial decision making process.

- Users do not have to purchase commercial GIS software.
- Users typically do not need extensive training
- The application will lead to reduction in cost, since only a GIS server software is required to be deployed at a centralized location for distributed users.
- The GIS data can be accessed any-time any-where, since it requires only Internet access.
- Transfer of copyrighted GIS data need not take place as transfer may be allowed in image format only.
- Data can be stored as centralized data warehouse in an organization to allow easy maintenance of the data.
- Data can be dynamically accessed, so changes made to the data are available immediately to the distributed users.
- Data can be concurrently accessed by many distributed users.
- The system can act as full GIS system with input, output, storage, editing, manipulation, analysis and query function.
- Customized application can be developed so that specific information can be served at specific scale and distributed to the users.
- It is also possible to deploy the Internet GIS organization wide using Intranet capability.

2.1 An Architecture of Typical WebGIS System

The architecture of general WebGIS system is similar to the client/server typical three-tier architecture (Figure 1). In this system, computers are connected in a world wide network and communicate with each other. The communication between various nodes takes place through a protocol, namely TCP/IP. Web browser and Web server are important constituents of Internet. Web browser software (e.g. Mozilla, Internet Explorer) resides on client machines e.g. desktop or workstation computers. A computer where server software is installed may act as a Web server. It consists of Web server, Web GIS softwares and Database.

The Internet uses a client server approach for distributing information. Information is stored and processed on servers and then sent to a client when it is requested. The request contains information regarding location of Web resource and additional information. HTML is the language that is used to display web pages content on a client's computer. HTML code is sent between the client and server using the Hypertext Transfer Protocol (HTTP). In web based application scripts are embedded in HTML code. Client side scripting are executed at client-side by web browsers whereas in server side scripting the scripts runs on server side or application servers. The popular server side scripts are PHP, ASP and JSP. JavaScript is a client side object oriented scripting language popular in developing client side application [15]. Common Gateway Interface (CGI) is a standard for writing applications to create Web pages dynamically. CGI standards define how Web servers interact with these software. The applications can be written in a programming language, e.g. C, C++ etc. or a scripting language. The variables are passed through query string to these programs. The applications are called CGI scripts [16].

2.2 Components of Typical Web GIS Systems Include

2.2.1 Hardware

- ✚ Central server computer
- ✚ Client computers
- ✚ Connection through the Internet or, for intranet sites, through a LAN or WAN.

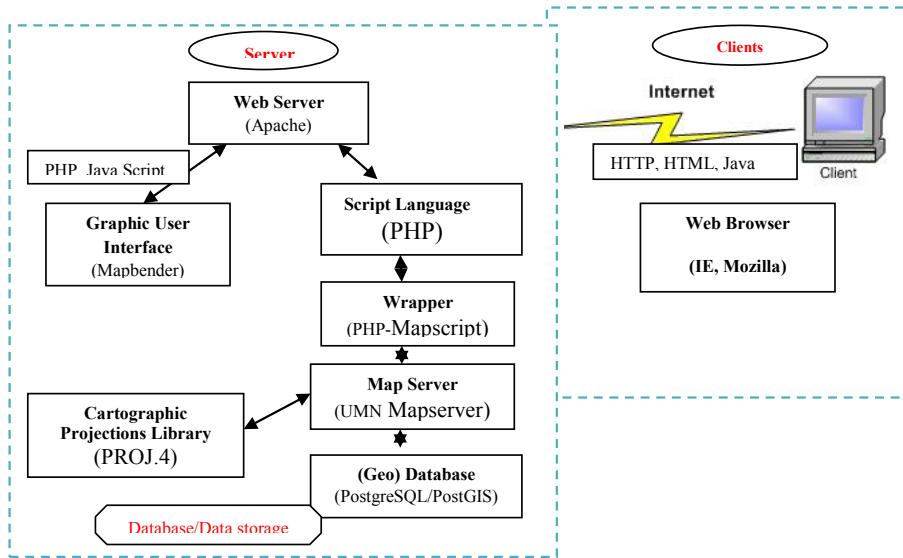


Fig. 1. In, General Web-GIS Architecture

2.2.2 Database

Data is available free of cost on Internet and Web related to GIS and RS. Only registration is mandatory. Links (as given below) can be used to search more information for water resource management and development.

Several important sectoral sources of information exist on the Internet. The following are the examples of sectoral data:

- ✚ FAO's Aquastat (FAO 2006) [17] provides data on water resources, irrigation and land use in the following main categories: Land use and population; Climate and water resources; Water use, by sector and by source; Irrigation and drainage development and Environment and health.
- ✚ World Water and Climate Atlas, provides data on the world's meteorology and water uses. This is considered as part of the tools introduced by International Water Management Institute (IWMI 2008) [18].
- ✚ FRIEND was founded by UNESCO in 1985 (FRIEND 2007) [19], it covers a diverse range of topics including low flows, floods, variability of regimes, rainfall/runoff modelling, processes of stream flow generation, sediment transport, snow and glacier melt, climate change and land use impacts.

- ✚ USGS (the U.S. Geological Survey) [20] provide access to water-resources data such as quantity, quality, distribution, and movement of surface and underground waters around the united states and disseminates the data to the public, State and local governments, public and private utilities, and other Federal agencies involved with managing our water resources. The USGS has also developed and continues to develop extensive regional datasets on changes in groundwater levels and the status of the nation's major aquifer systems [21].
- ✚ StreamStats [URL 22] allows users to easily obtain stream flow statistics, drainage-basin characteristics, and other information for user-selected sites on streams. Stream Status users can choose locations of interest from an interactive map and obtain information for these locations.
- ✚ The National Oceanic and Atmospheric Administration (NOAA) monitors nationwide precipitation input- e.g., through the NEXRAD program [23] and individual precipitation stations [24].
- ✚ India Water Portal [25] provides district wise monthly precipitation data, available on Internet for water resource management.

Other important sources of information are the RS and GIS sites which are available on the World Wide Web sites (adapted from Awad,et al. 2009) [26]:

- ✚ Global Land Cover Facility (GLCF) and GLOVIS. GLCF located at University of Maryland provides satellite data of various resolutions and thematic maps, in particular for land use and cover applications at local and global scale [URL 27]. Data available from the site are ETM+, TM, MSS and ASTER. The data are geo-referenced in UTM coordinate system and may be utilized for earth sciences applications. Data may be stacked to create BIL format data. Since 2008, Landsat archives are in public domain and thus data can be downloaded from Internet free of cost [URL 28]. The registration is mandatory. The data which are not processed can be ordered free of cost. High resolution data are also available from the site.
- ✚ In India, RS data may be purchased on slashed rate from National Remote Sensing Centre (NRSC)-Hyderabad, India [URL 29], Department of Space.
- ✚ India-WRIS (Water Resources Information System) (2010) [URL 30], a Web-GIS provide a comprehensive, credible, and contextual view of India's water resources data along with allied natural resources data and information. It also allows users to search, access, visualize, understand, analyze, look into context and study spatial patterns. It is a 'Single Window' solution of all water resources and related data in a standardized GIS format in a national framework for water resource assessment and monitoring.
- ✚ NASA (the USA space research agency), Hyderabad (2008), provides worldwide satellite images, climate data ...etc.
- ✚ NOAA AVHRR [URL 31], imagery is one of the most stable sources of information available freely from the Internet. It is also compatible with NASA's Earth Observing System Data and Information System, enabling global change researchers to more readily gain a greater understanding of planet Earth.

- ✚ MapServer (2007) [32] GIS search engine is an open source development environment for building spatially-enabled internet applications. MapServer is not a full featured GISsystem, nor does it aspire to be. Instead, MapServer excels at rendering spatial data (maps, images and vector data) for the web. Beyond browsing GIS data, MapServer allows to create “geographic image maps”, that is, maps that can direct users to content. For example, the Minnesota Department of Natural Resource (MNDNR) recreation compass provides users with more than 10,000 web pages, reports and maps via a single application. The same application serves as a “map engine” for other portions of the site, providing spatial context where needed.
- ✚ ESRI's World Map 1:1,000,000 (ESRI 2004) [33] includes Country Boundaries, Coastlines, Administrative Boundaries, Water Bodies, Perennial Rivers, Intermittent Rivers, Populated Places, Major Roads, Major Trails, Capital Cities, and Major Cities.
- ✚ ESRI launched the Geography Network [URL 34], global network of spatial data users and providers. It provides the Internet infrastructure to support sharing of geographic information among data providers, service providers, and users around the world. Through the Geography Network, a user can access many types of geographic content, including dynamic maps, downloadable data, and more advanced Web services.
- ✚ GlobeXplorer is a satellite images portal (GlobeXplorer 2008) [35], commercial satellite images provider, and it is free to browse, search and examine all kinds of satellite images around the world.
- ✚ Alexandria Digital Library (ADL 2005) [36] is part of the University of California Santa Barbara Libraries. This portal provides search engine to find any type of aerial and satellite images of the world by specifying the geographic extent and the date of the image, these images are free to download.
- ✚ Another source of an up-to-date satellite images with high resolution is Google Earth (Google, 2008) [37]. This portal provides an easy method to surf the world and search for any spatial feature (river migration, hydro-network, river boundaries etc...)

2.2.3 GIS Software

Desktop GIS softwares are needed for processing the satellite data and extracting thematic information. Web GIS software may be used to disseminate spatial data over Web. Several open source and commercial off-the-shelf (COTS) software are available for GIS and Web GIS application. Open source desktop GIS softwares are ILWIS, Quantum GIS, Mapwindow etc. COTS GIS softwares are ArcGIS, MapInfo etc. Open source Web GIS softwares are Mapserver, Geoserver, MapGuide etc. COTS Web GIS software are ArcGIS server, ERDAS Apollo, MapExtreme, Manifold Internet Map Server (IMS) etc. The softwares are available for various platforms, e.g. Windows, Linux, Mac OS and Unix both at 32 bit and 64 bit. COTS softwares vary in their cost. For Web GIS software either the data may be straight away published, e.g. in ArcGIS server, ERDAS Apollo, or require writing codes/HTML pages e.g. in Mapserver, MapGuide, MapExtreme etc. Mapserver is a CGI program. Software e.g.

MapGuide, MapExtreme works in modern programming environment. Mapserver software was developed in the mid 1990's by the University of Minnesota with the assistance of NASA and the Minnesota Department of Natural Resources (MNDNR) [38]. Mapserver is currently managed by the Open Source Geospatial Foundation (OSGeo). The server can be run on several operating systems, e.g. Windows, Linux, Mac OS X etc. It is a CGI program written in C language. It supports shape file format. Map elements, namely thematic map, legend, scalebar, reference etc. can be created. The program uses a Map file, which is a text file. The Map file contains various parameters and details needed for creation of an image file. The image file is embedded in the Web page and depicts the spatial data. HTML documents are often Template HTML files. Mapserver replaces variables, objects etc. in the template file with their values (obtained from query strings, Map files or both). Mapserver tutorial contains HTML files, Java scripts, Map file and data [39].

2.3 Open Source Desktop GIS Software

Open source software is becoming popular and increasingly more reliable than in the past particularly in developing countries for use towards sustainable water resource management [9]. It is commonly used in GIS applications along with their Web address for a further reference (Table 1). Software is considered open source once it complies with the following characteristics [40]:

-  The source code of the product must be made available.
-  The license allows unlimited redistribution of the product.
-  The license permits the creation of license free derived works.
-  The license does not limit how, where or by whom the product can be used.

Information of open source GIS software exists on the *Internet*. Few very useful links are:

-  Open source GIS [URL 41]: This site attempts to build a complete index of Open source/free GIS related software projects. The definition of GIS has been kept loose to encompass a broad range of projects which deal with spatial data. This site stands because of other projects, most notably OSRS,
-  Freegis [URL 42]: This site organizes software, geodata projects according to the OS, language features. More than 330 relevant softwares are available.
-  SourceForge [URL 43]: This is a site for all sorts of open source software. It helps to maintain and provide codes for downloads.
-  OSGeo [URL 44]: This is the website of The Open Source Geospatial Foundation that has been created to support and build the highest-quality open source geospatial software. A list of very popular software packages are listed as OSGeo projects in the OSGeo website.
-  Wikipedia provides a list of open source GIS software packages [URL 45] and comparisons of some packages were given at [URL 46]. It is noted that both commercial and open source softwares have been included.

3 Addressing Water Resources Management through Web-Based GIS and FOSS Approach

Water resource management is combination of (i) rainfall estimation, (ii) watershed management (iii) irrigation water management and identification of potential irrigable lands, (iv) reservoir or lake sedimentation (v) river basin management (vi) dam and drought management (vii) water quality and quantity assessment, (viii) ground water assessment and prospecting (ix) flood forecast and monitoring, (x) low flow estimation, (xi) climate change and Hydrological cycle and (xii) Impact of climate change on available water resources.

Since 1990s, various researchers have adopted WebGIS and Free Open Source Software (FOSS) approach for natural resource management i.e water resource, gained popularity in recent years. Fletcher, et al. [47] developed GIS for stream water management in West Virginia. Choi and Engel [48] have shown that geoprocessing via the web is possible, through creation of a web-based watershed delineation system. Their tool uses the University of Minnesota's MapServer (University of Minnesota, 2006) as the backend engine, using this map system to obtain an outlet point from the user to begin the delineation process. Their implementation uses a "double-seed array-replacement algorithm to obtain a watershed boundary form point coordinates. Another example of successful web-based geoprocessing is given by Anselin et al. [49] in their Java-based geoprocessing extensions. Their "specific focus is on methods to identify and visualize outliers in maps for rates or proportions. Rathore et al. [39] used an open source internet GIS software (Mapserver) to create an application for dams and drought information and functionalities of a modern internet GIS application for water resource in India. It will provide information about dam storage, hydropower, dam location, rainfall map for different Standardized Precipitation Index (SPI) various time scale etc over the district in India.

Choi et al. [50] examined the relationship between GIS technology and watershed management SDSS by utilizing the prototype conceptual framework SDSS [51] and form a Web-based SDSS, combined integration of two models in SDSS named Long-Term Hydrological Impact Assessment (L-THIA) and Web-based Hydrologic GIS (WHYGIS). This Web-GIS approach is capable to real-time watershed delineation, hydrologic data extraction/preparation functionality, generate watershed boundaries and prepare real time hydrologic data for straightforward operation of hydrologic models via the Internet.

L-THIA, Web DSS structure is comprised of a modeling system, a database system and a graphical user interface, and includes special features for users with limited hydrology knowledge. It is intended to support decision makers who need information regarding the hydrologic impacts of water quantity and quality resulting from land use change.

HYDRA5, Web-based GIS system for catchment management, designed for novice as well as expert users for planning and water quality. It provides a catalogue for search and retrieval of arbitrary data sets, a series of hyper maps with some GIS functionality and links to spatially referenced data, and an analysis and graphing tool for time-series data.

Table 1 List of Deckton GIS software and its applications

Name	Release	Developer	Homepage	Operating system	Programming Language	Application
Apache Batik	V 1.6	Apache	http://xmlgraphics.apache.org/batik/	Windows, Linux	Java 1.3	Use images in the Scalable Vector Graphic (SVG)
DIVA GIS	V 6.0.3	CIP (International Potato Center, Peru)	http://research.cip.cgiar.org/confluence/display/divas/Home	Windows only	Java with Eclipse	Make maps of species distribution data and their analysis
Deegree	V 2.1	latlon, Germany	http://www.deegree.org/	Windows only	Java 1.5, tomcat 5.5, C	Server/Client web applications and Desktop mapping
Fmmaps	V 0.0.2	Fmmaps team	http://sourceforge.net/projects/fmmaps/	Linux and Gnome	-	-
FWTools	V 2.0.6	Private	http://fwttools.maptools.org/	Windows, Linux	Java, Python	-
GeoOxygen	V 1.3	GeoOxygen Team	http://oxygene-project.sourceforge.net/	Independent	Java	-
GeoServer	V 2.1.1	Geoserver team	Welcome">http://geoserver.org/display/GEOS>Welcome	MacOS X, Unices, and Windows	WebGIS	Manipulating geographic and Cartesian data sets and producing Encapsulated PostScript File (EPS)
Generic Mapping Tools	V 4.3.1	University of Hawaii	http://gmt.soest.hawaii.edu/	Windows, Linux, Mac	C/C++	Manipulating geographic and Cartesian data sets and producing Encapsulated PostScript File (EPS)
GRASS GIS	V 6.2.3	GRASS Development Team	http://grass.itc.it/	Windows, Unix (Linux)	C	Spatial analysis and scientific visualization.
gvSIG	V 1.1.2	River, Generalitat Valencia, Universidad Jaume I, Prodevelop	http://www.gvsig.gvates.es	Windows, Linux, Mac OS X	Java	Vector data editing, easily digitises by snapping vertices to existing nodes and generate correct topology.
HidroSIG	V 3.1.1	University of Colombia, SedeMedellin	http://cancerbeta.unalm.edu.co/~hidrosig/index.php	Windows, Linux	Java	-
ILWIS	V 3-04-02	52° North	http://www.ilwiss.com.es/	Windows only	MS Visual 6	Image processing and analysis, digital mapping
KOSMO	V 1.2	SAIG S.L.	http://www.nividsolutions.com/its/jishome.htm	Windows, Mac OS X, Linux	Java	substitute Arcview for advanced user
ITS Topology Suite	V 1.8.0	VIVID Solutions	http://www.nividsolutions.com/its/jishome.htm	Windows, Linux	Java	-
Mapnik	V 0.5.1	BerliOS Developer Team	http://mapnik.org/	Windows, Linux, Mac OS	C++, Python	Toolkit for developing GIS
Map Window GIS	V 4.5 SR	Map Window open Source Team	http://www.mapwindow.org/	Windows only	NET (VB, C++, C#, .Net framework 2.0)	Providing core GIS and GIS functions, developing Decision Support System (DSS)
mezoGIS	V 0.1.5	Private, frozen now	http://www.mezogis.org/	Windows and Linux	python	-
monoGIS	V 0.7	MonGIS Team	http://www.monogis.org/	Windows and Linux	OGR/GDAL, (C++) Geotools	Editing and Vector Analysis
NRDB	V 2.3	Private	http://nrdb.co.uk/	Windows only	C++	Tool for displaying and editing of spatial data stored in shapefiles

Table 1. (continued)

							Geospatial mapping tool
Open JUMP	V 1.2	Open JUMP TEAM	http://www.openjump.org/	Windows, Linux, Mac OS X	Java	-	
OpenMap	V4.6.4	BBN Technologies	http://openmap.bbn.com/	Windows, Linux, Mac OS X	Java	-	
OSSIM	V1.7.9	OSSIM team	http://ossim.org/OSSIM/Home.html	Windows, Linux, Mac OS X	C++	RS, GIS, image processing, photogrammetry	
PostGIS	V 1.5.3	Refractions Research	http://postgis.refractions.net/	Windows, Linux, Mac OS and other Unices	C, Java	Spatial database for GIS	
Quantum GIS (QGIS)	V 1.7	QGIS Development Team	http://www.qgis.org/	MS Windows, Linux, Mac OS X, POSIX and other Unices	C++, Python, and C	View Editing, GRASS-Graphical user interface	
SAGA	V 2.0.3	Uni. of Goettingen	http://www.saga-gis.uni-goettingen.de/	Windows and Linux	C++	Analysis modelling, visualisation (raster), especially terrain hydrographic analysis	
SAMT	V 2.8.1	Institute of Landscape Systems Analysis (ZALF)	http://www.zalf.de/home/samt-lsa/	Unix	-	Integrating fuzzy-models and spatial simulation	
SavGIS	V2.1.5.0	IRD (Development Research French Institute)	http://savgis.org/	Windows only	-		
SharpMap	V 0.9	Sharpmap team	http://www.codeplex.com/SharpMap	Windows and Linux	C++	Renders GIS vector data for use in web and desktop applications	
TerraView	V 3.2.0	DPI of INPE	http://www.dpi.inpe.br/teraview/index.php	Windows, Linux	C++		
Thuban	V 1.2.2	Thuban Team	http://thuban.ineriaiion.org/	Windows and Mac OS, Linux and other Unices	python	Geographic data viewer	
uDig	V 1.2.1	Refractions Research	http://udig.refractions.net	Windows (not windows) 2000, Mac OS, Linux, and other Unices	Java with Eclipse	Viewing (OGC standards) application framework	
Whitebox	V 1.07	University of Guelph Centre for Hydrogeomatics	http://www.uoguelph.ca/~hydrogeo/Whitebox/index.html	Windows	Python, C#, .NET	Spatial analysis on raster data sets	
Kalypso	V 10.10	Jonny by Björnsen Consulting Engineers (BCE) and Hamburg University of Technology, Germany	http://kalypso.hjoransen.de/	Linux	Java	Numerical simulation in water management and ecology	
Capaware	V rc 2	Instituto Tecnológico de Canarias (ITC)	http://www.capaware.org/	Windows	C++	Geographic analysis and visualization	
FalconView	V 4.2.1	Georgia Tech Research Institute	http://www.falconview.org/	Window 2000, XP, Vista	C++	Mapping system	

(Source: modified after Chen, et al. 2010)

Wang et al. [52] developed Web-GIS based river simulation model named GIS-ROUT system for estimating environmental exposure and risk assessment after disposal of chemicals in surface water and their contributions to surface water quality throughout continental United States. It has a number of advantages over common modelling approach such as; (1) its components (WWW, GIS and ROUT) can be as integrated or independent from each other at the same time (2) GIS provides wide range of spatial analytical functions to prepare data for river modeling (3) user-friendly internet-based interface (4) capable of sharing data and simulation results (5) do not require powerful client computer system and (6) better reliability of spatial analysis of the modeling output.

Web-GIS can provide a user-friendly light-weight interface for the users to access geographical data and geographical services using web browser. However, the Web-GIS and hydrological model integrated systems need to manage and process various spatial data, such as, DEM, remote sensing images, land use data, topographic data and various temporal data, such as stream flow data, precipitation data and temperature data. While, FOSS permits users to use, change and improve the software codes within any programming language, and permit to rewrite in multiple languages for particular need. It is often developed in a collaborative manner in a public domain.

Rajani [53], who studied the relevance of FOSS for developing countries. Chen et al. [9] assessed number of Open Source Softwares and suggested 31 suitable Open Source GIS efficient software packages (Table 1), based on their potentiality to identify and shortlist the most suitable and user friendly, which could be used for analysis or designing a system applicable to water resources management mainly in developing countries. Some of related software and their link and application are summarized in Table 1.

4 Conclusions

This paper provides an overview of Web-based GIS and links & sources for RS and GIS data and web address for Desktop Open Source GIS softwares packages for water professionals of developing countries that have limited financial resources. Thus, information about the water resource management and development may be made available on Web using client, namely a Web browser. GIS software is not required at the user desktop. Also, user need not be familiar with GIS to use this data. This indicates great potential of Web GIS in information dissemination on water resource domain. Additionally, numbers of open source Web GIS software are available and thus applications in hydrology and water resources may be created with small budget. Which are freely available on internet and Web for water resource management. Web-based GIS system could be used by any user without installing traditional desktop GIS or COTS Softwares. The Web-based SDSS for watershed management is currently being validated and proved advantageous over traditional desktop applications. In addition, the Internet based approach increases the user base by reducing costs of access to users. This paradigm shift from stand-alone GIS software to Web-based GIS software and Open Source GIS softwares with COTS software will enable us to think above and beyond the technical drawbacks that have occupied us during the past 10 years. This Web-based GIS will also lead to a wide-range application with easy access for water resource management.

Acknowledgement. The authors would like to thank anonymous reviewer for the constructive comments and suggestions.

References

- [1] WWAP, Water for People, Water for Life, UN World Water Development Report. Prepared as a collaborative effort of 23 UN agencies and convention secretariats co-ordinated by the World Water Assessment Programme, UNESCO, Paris (2003), <http://www.unesco.org/wate/wwap/index.shtmlr>
- [2] Martin, P.H., LeBoeuf, E.J., Dobbins, J.P., Daniel, E.B., Abkowitz, M.D.: Interfacing GIS with water Resource models: A state-of-the-art review. Journal of the American Water Resource Association, 1471–1487 (2005)
- [3] http://dusk.geo.orst.edu/ucgis/web/apps_white/water.html
- [4] McDonnell, R.A.: Including the spatial dimension: Using geographical Information systems in hydrology. Progress in Physical Geography 20(2), 159–177 (1996)
- [5] Castrogiovanni, E.M., Loggia, G.L., Noto, L.V.: Design storm prediction and hydrologic modeling using a web-GIS approach on a free-software platform. Journal of Atmospheric Research 77, 367–377 (2005)
- [6] Sui, D.Z., Maggio, R.C.: Integrating GIS with hydrological modeing: practices, problems, and prospects. Journal of Computer, Environment and Urban Systems 23(1), 33–51 (1999)
- [7] Al-Sabhan, W., Mulligan, M., Blackburn, G.A.: A real-time hydrological model for flood prediction using GIS and the WWW. Journal of Computer, Environment and Urban Systems 27, 9–32 (2003)
- [8] Alesheikh, A.A., Helali, H., Behroz, H.A.: Web GIS: Technologies and Its Application. In: Symposium on Geospatial Theory, Processing and Application (2002), <http://www.isprs.org/proceedings/XXXIV/part4/pdfpapers/426.pdf> (Access on 10.12.2011)
- [9] Chen, D., Shams, S., Moreno, C.C., Leone, A.: Assessment of open source GIS software for water resources management in developing countries. Journal of Hydro-environment Research 4(3), 253–264 (2010)
- [10] OGC: WMS Implementation Specification, http://portal.opengeospatial.org/files/?artifact_id=1058
- [11] OGC: WFS Implementation Specification, http://portal.opengeospatial.org/files/?artifact_id=8339
- [12] OGC: OpenGIS Web Processing Service, http://portal.opengeospatial.org/files/?artifact_id=24151
- [13] Peng, Z.R.: An assessment framework for the development of Internet GIS. Environment and Planning B: Planning and Design 26(1), 117–132 (1999)
- [14] Tripathy, G.K.: Web-GIS based urban planning and information system for municipal corporations-A distributed and real-time system for public utility and town planning. Geospatial Application (2002), <http://www.gisdevelopment.net/application/urban>
- [15] <http://www.adobe.com/livedocs/coldfusion/6.1/htmldocs/introb5.htm>
- [16] http://en.wikipedia.org/wiki/Common_Gateway_Interface
- [17] FAO (2006), <http://www.fao.org/nr/water/aquastat/main/index.stm>
- [18] International Water Management Institute, IWMI (2008), <http://www.iwmi.cgiar.org/>

- [19] FRIEND (2007), <http://typo38.unesco.org/en/about-ihp/ihp-partner/friend.html>
- [20] <http://waterdata.usgs.gov/nwis>
- [21] <http://water.usgs.gov/cgi/rasabiblio>
- [22] <http://water.usgs.gov/osw/streamstats/index.html>
- [23] <http://www.roc.noaa.gov/>
- [24] <http://www.webscraping.noaa.gov/guide/sciences/atmo/precip.html>
- [25] <http://indiawaterportal.org/node/7160>
- [26] Awad, M., Khawlie, M., Darwich, T.: WebBased meta database and its role in improve water resource management in Mediterranean basin. *Journal of Water Resource Management* 23, 2669–2680 (2009)
- [27] <http://glcf.umiacs.umd.edu/index.shtml>
- [28] <http://glovis.usgs.gov/>
- [29] <http://www.nrsc.gov.in/>
- [30] <http://www.india-wris.nrsa.gov.in>
- [31] <http://www.saa.noaa.gov>
- [32] MapServer (2007), <http://mapserver.gis.umn.edu>
- [33] ESRI (2004), http://www.esri.com/data/free_data/index.html
- [34] <http://www.geographynetwork.com>
- [35] GlobalXplorer Satellite Images Portal (2008), <http://imageatlas.globalexplorer.com>
- [36] Alexandria Digitallibrary, ADLO, Satellite images and GIS Portal (2005), <http://webclient.alexandria.ucsb.edu/>
- [37] Google Earth GIS and Remote Sensing Databases (2008), <http://earth.google.com>
- [38] <http://mapserver.org/MapServer.pdf>
- [39] Rathore, D.S., Chalisgaonkar, D., Pandey, R.P., Ahmad, T., Singh, Y.: AWeb GIS Application for Dams and Drought in India. *Indian Society of Remote Sensing* 38(4), 670–673 (2010)
- [40] Open Source Initiative, <http://www.opensource.org/docs/definition.php>
- [41] Open Source GIS, <http://opensourcegis.org/>
- [42] Freegis, <http://freegis.org/>
- [43] SourceForge, <http://web.sourceforge.com/>
- [44] OSGeo, <http://www.osgeo.org/>
- [45] Wikipedia, http://en.wikipedia.org/wiki/List_of_GIS_software
- [46] http://en.wikipedia.org/wiki/Comparison_of_GIS_software
- [47] Fletcher, J.J., Sun, Q., Strager, M.P.: GIS application for stream water management in West Virginia. In: Proc. of International Conferences on Info-tech and Info-net (ICII), vol. 4, pp. 113–123 (2001)
- [48] Choi, J.Y., Engel, B.A.: Real-Time Watershed Delineation System Using Web-GIS. *Journal of Computing in Civil Engineering* 17(3), 189–196 (2003)
- [49] Anselin, L., Kim, Y.W., Syabri, I.: Web-based analytical tools for the exploration of spatial data. *Journal of Geographic Systems* 6, 197–218 (2004)
- [50] Choi, J.-Y., Engel, B.A., Farnsworth, R.L.: Web-based GIS and spatial decision support system for watershed management. *Journal of Hydroinformatics* 7(3), 165–174 (2005)
- [51] <http://pasture.ecn.psu.edu/~watergen/owls>
- [52] Wang, X., Homer, M., Dyer, S.D., White-Hull, C., Du, C.: A river water quality model integrated with a web-based geographic information system. *Journal of Environmental Management* 75, 219–228 (2005)
- [53] Rajani, N., Rekola, J., Mielonen, T.: Free as in education: significance of the free/libre and open source software for developing countries. *World Summit on the Information Society* (2003)

A Design Pattern for Service Injection and Composition of Web Services for Unstructured Peer-to-Peer Computing Systems with SOA

Vishnuvardhan Mannava¹, T. Ramesh², and Mohammed A.R. Quadri¹

¹ Department of Computer Science and Engineering,
K.L. University, Vaddeswaram, 522502, A.P., India

vishnu@kluniversity.in, mohammad.ataulla@gmail.com

² Department of Computer Science and Engineering,
National Institute of Technology, Warangal, 506004, A.P., India
rimesht@nitw.ac.in

Abstract. Adaptability in software is the main fascinating concern for which today's software architects are really interested in providing the autonomic computing. Different programming paradigms have been introduced for enhancing the dynamic behavior of the programs. Few among them are the Aspect oriented programming (AOP) and Feature oriented programming (FOP) with both of them having the ability to modularize the crosscutting concerns, where the former is dependent on aspects ,advice and lateral one on the collaboration design and refinements. In order to provide remedy for the service failures that occurs at the servers of the respective service providers, there is a need to introduce the self-reconfiguration planes to be applied autonomically without the interruption of the administrator to solve the problem manually. In this paper we will Propose an Service Injection Design Pattern for Unstructured Peer-to-Peer networks, which is designed with Aspect-oriented design patterns .We propose this pattern which is an amalgamation of the Strategy Design Pattern, Worker Object Aspect-Oriented Design Pattern, and Check-List Design Pattern these can be used to design the Self-Adaptive Systems. The main concept here is that when a client requests for a complex service then Service Composition should be down to fulfill the request. With the help of strategy pattern we can choose a service that can do the task of two or more services at a single click. With the Web Service Description Language (WSDL) which is an XML file that can be requested from all the Web Services providing peers (as HTTP GET). This file contains the all the description about the service, means the input and output parameters of that service and its location details. When a client requests for a service that is not loaded currently in the memory will be injected as Aspectual Feature Module code. Here we will be using the Service Oriented Architecture (SOA) with Web Services in Java to Implement the Design Pattern. The pattern is described using a java-like notation for the classes and interfaces. A simple UML class and Sequence diagrams are depicted.

Keywords: Autonomic system, Design Patterns, Aspect-Oriented Programming Design Pattern, Feature-Oriented Programming (FOP),Aspect-Oriented Programming (AOP),JXTA, Service Oriented Architecture (SOA), Web Services, Web Service Description Language (WSDL).

1 Introduction

The most widely focused elements of the autonomic computing systems are self-* properties. So for a system to be self-manageable they should be self-configuring, self-healing, self-optimizing, self-protecting and they have to exhibit self-awareness, self-situation and self-monitoring properties [1]. As the web continues to grow in terms of content and the number of connected devices, peer-to-peer computing is becoming increasingly prevalent. Some of the popular examples are file sharing, distributed computing, and instant messenger services. Each one of them provides different services, but shares the same mechanism like Discovery of peers, searching, file and data transfer. Currently developed peer-to-peer applications are inefficient with the developers solving the same problems and duplicating the similar infrastructure implementations [2]. Most of the applications are specific to a single platform and can't communicate and share data with different applications.

To overcome the current existing problems Sun Microsystems have introduced JXTA. JXTA is an open set, generalized peer-to-peer (P2P) protocols that allows any networked device –sensors, cell phones, PDA's, laptops, workstations, servers and supercomputers- to communicate and collaborate mutually as peers. The advantage of using the JXTA peer-to-peer programming is that it provides protocols that are programming language independent, multiple implementations, known as bindings, for different environments. The JXTA protocols are all fully interoperable. So with help of JXTA programming technology, we can write and deploy the peer-to-peer services and applications. JXTA protocols standardize the manner in which peers will discover each other, self-organize into peer groups, Advertise and discover network resources, communicate with each other, monitor other.

JXTA overcomes the many of the problems in current existing peer-to-peer systems, some of them are 1) Interoperability – enables the peers provisioning P2P services to locate and communicate with one another independent of network addressing and physical protocols. 2) Platform Independent - JXTA provides the developing code with independent form programming languages, network transport protocols, and deployment platforms. 3) Ubiquity – JXTA is designed to be accessed by any device not just the PC or a specific deployment platform. In this paper we propose a design pattern for providing the services to peer-clients in unstructured peer-to-peer network.

Design patterns are most often used in developing the software system to implement variable and reusable software with object oriented programming (OOP) [3]. Most of the design patterns in [3] have been successfully applied in OOPs, but at the same time developers have faced some problems like as said in [4] they observed the lack of modularity, composability and reusability in respective object oriented designs [5]. They traced this lack due to the presence of crosscutting concerns. Crosscutting concerns are the design and implementation problems that result in code tangling, scattering, and replication of code when software is decomposed along one dimension [6], e.g., the decomposition into classes and objects in OOP. To overcome this problem some advanced modularization techniques are introduced such as Aspect-oriented

programming (AOP) and Feature-oriented programming (FOP). In AOP the crosscutting concerns are handled in separate modules known as aspects, and FOP is used to provide the modularization in terms of feature refinements.

In our proposal of a design pattern for a peer-to-peer system, we use the Aspect-oriented design pattern called Worker Object pattern [7] and Checklisting Design Pattern [10]. When comes to the worker object pattern it is an instance of a class that encapsulates a method called a worker method. It will create and handle each client service request in separate thread by making the job of server easy from looking after each and every client until it completes serving its request. So the server can listen for new client requests if any to handle. The Checklisting Design Pattern is used to provide means for selecting a service plan that best suits the clients request and also that matches the WSDL information of a Service providing peer server. An application has to interact with considering some constraints. We come across this situation when a task needs to be accomplished by the collaboration of multiple peers. The strategy Design Pattern helps in making decision regarding which Checklist of items (items here are the services) have to be selected to solve the complex task requested by user with respect to the WSDL information available from each and every service providing peer servers.

2 Related Work

In this section we present some works that deal with unstructured peer-to-peer systems design. There are number of publications representing the design pattern oriented design of the peer-to-peer computing systems. The JXTA protocols standardization provides one of the autonomic computing system properties known as “self-organization” into peer groups. The self-organization is property that provides the autonomic capability in the peer-to-peer design of networks.

In V.S.Prasad Vasireddy, Vishnuvardhan Mannava, and T. Ramesh paper [8] discuss applying an Autonomic Design Pattern which is an amalgamation of chain of responsibility and visitor patterns that can be used to analyze or design self-adaptive systems. They harvested this pattern and applied it on unstructured peer to peer networks and Web services environments.

In Sven Apel, Thomas Leich, and Gunter Saake [9] they proposed the symbiosis of FOP and AOP and aspectual feature modules (AFMs), a programming technique that integrates feature modules and aspects. They provide a set of tools that support implementing AFMs on top of Java and C++.

In Alois Ferscha, Manfred Hechinger, Rene Mayrhofer, Ekaterina Chtcherbina, Marquart Franz, Marcos dos Santos Rocha, Andreas Zeidler [10] they proposed that The design principles of pervasive computing software architectures are widely driven by the need for opportunistic interaction among distributed, mobile and heterogeneous entities in the absence of global knowledge and naming conventions. Peer-to-Peer (P2P) frameworks have evolved, abstracting the access to shared, while distributed information. To bridge the architectural gap between P2P applications and P2P frameworks we propose patterns as an organizational schema for P2P based

software systems. Our Peer-it hardware platform is used to demonstrate an application in the domain of flexible manufacturing systems.

Because of the previous proposed works as described above, we got the inspiration to apply the aspect-oriented design patterns along with inclusion of the feature-oriented software development capability to peer-to-peer computing systems.

3 Proposed Autonomic Design Pattern

One of the objects of this paper is to apply the Aspect-oriented design patterns to the object-oriented code in the current existing application. So that a more efficient approach to maintain the system and providing reliable services to the client requests can be achieved.

The very important capability that our proposed design pattern provides in the peer-to-peer computing systems that, when a new service is to be added to the peer system in the network without disturbing the running server we can do this with the help of Aspectual Feature Module [9] oriented insertion of the new service into the peer-server code by using the feature refinements property of Feature-Oriented Programming (FOP). In our proposed Design Pattern for an Autonomic Computing System, initially all the peers in the network group will broadcast the advertisements that represents the respective services that are provided by them as the WSDL messages. The WSDL messages are the XML files that are used in web services. It defines the input and output parameters of a web service in terms of XML schema. With these sent advertisements all the peers will know the information about the services that are provided by different peer servers and whether they are Active or Deactivated at present. Here in each of the Web Service Description Language (WSDL) file it will include the status of the service whether it is active or deactivated currently.

The Client who requests for a Complex service that cannot be fulfilled with single service but has to take help of two or more services that are provided by different peers in order to satisfy the client's request. So with the help of the Checklist Design Pattern in [10] will help us in selecting the perfect combination of services that best suits the clients service request with respect to the WSDL messages send by the peer servers. Here the Strategy Design Pattern will select a CheckList Plan that better suits the Service Request to be fulfilled and also it will select the perfect WSDL of a peer server that can perfectly process the service request of a Client. For a clear idea of the pattern structure proposed see Figure 1 referred from [10]. Action means the Service definition, Check peer will check whether it can apply the input data retrieved to this service. P2P Infrastructure means any middleware technology like JXTA for establishment of Connections between clients and peers. Once the information of all the peers that are providing the web services is gathered as WSDLs from respective peer servers then it (Client) will invoke the respective peer's service that is responsible for the performing the task.

Here the Peer Web Service provider can take help of the services that are provided by the other peers in the network to fulfill the complex service request of a client. So this kind of service composition can be achieved at the server side with Service

Oriented Architecture (SOA) implementation with Web Services. With this invocation call the respective peers will check whether the requested service is currently Active and running. If this is not the case then it will loads the requested service into the memory as a new Feature with the help of the Aspectual Feature Module [9] code from the Service Repository without disturbing the already running services in that peer. Now once the service is loaded into the memory then the peer will invoke a call that will update the State of the service to Activated.

The Worker Object Pattern [7] is responsible for handling the different client requests in terms of separate Thread based requests. Means when more than one client request for same service or different one they are handled in different Threads.

4 Design Pattern Template

To facilitate the organization, understanding, and application of the proposed design patterns, this paper uses a template similar in style to that used in [11].

4.1 Pattern Name

Service injection design pattern along with composition of services with SOA.

4.2 Classification

Structural-Decision-Making

4.3 Intent

Systematically applies the Aspect-Oriented Design Patterns to an unstructured peer-to-peer Computing System and service injection with a Refinement class for providing new service in the peer in terms of a Feature Module.

4.4 Context

Our design pattern may be used when:

- a) The service requested is a complex service and need to be executed with the help of collaboration of peers in the network.
- b) To include the new service operations into peers as Aspectual Feature Modules [9].
- c) For providing the Distributed service request processing environment with the help of JXTA peer-to-peer Programming and Service Oriented Architecture (SOA) with Web Services.

4.5 Proposed Pattern Structure

A UML class diagram for the proposed design Pattern can be found in Fig 2. The Block diagram for our proposed design pattern is shown in Fig 1.

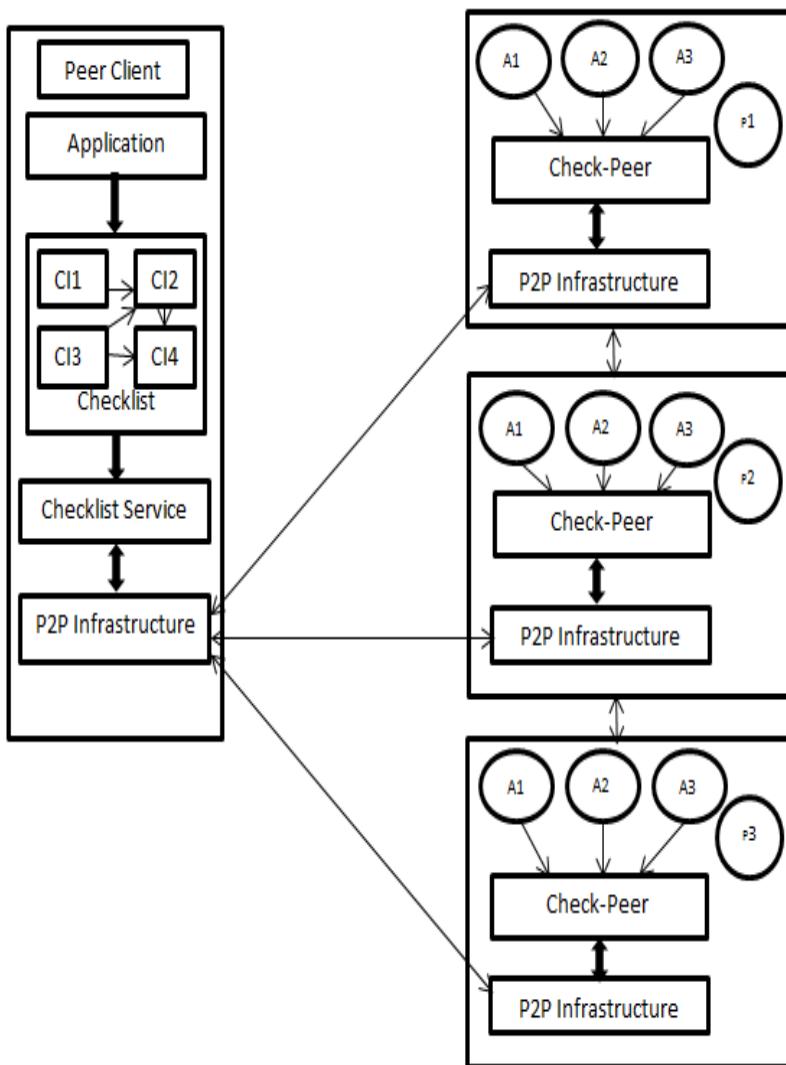
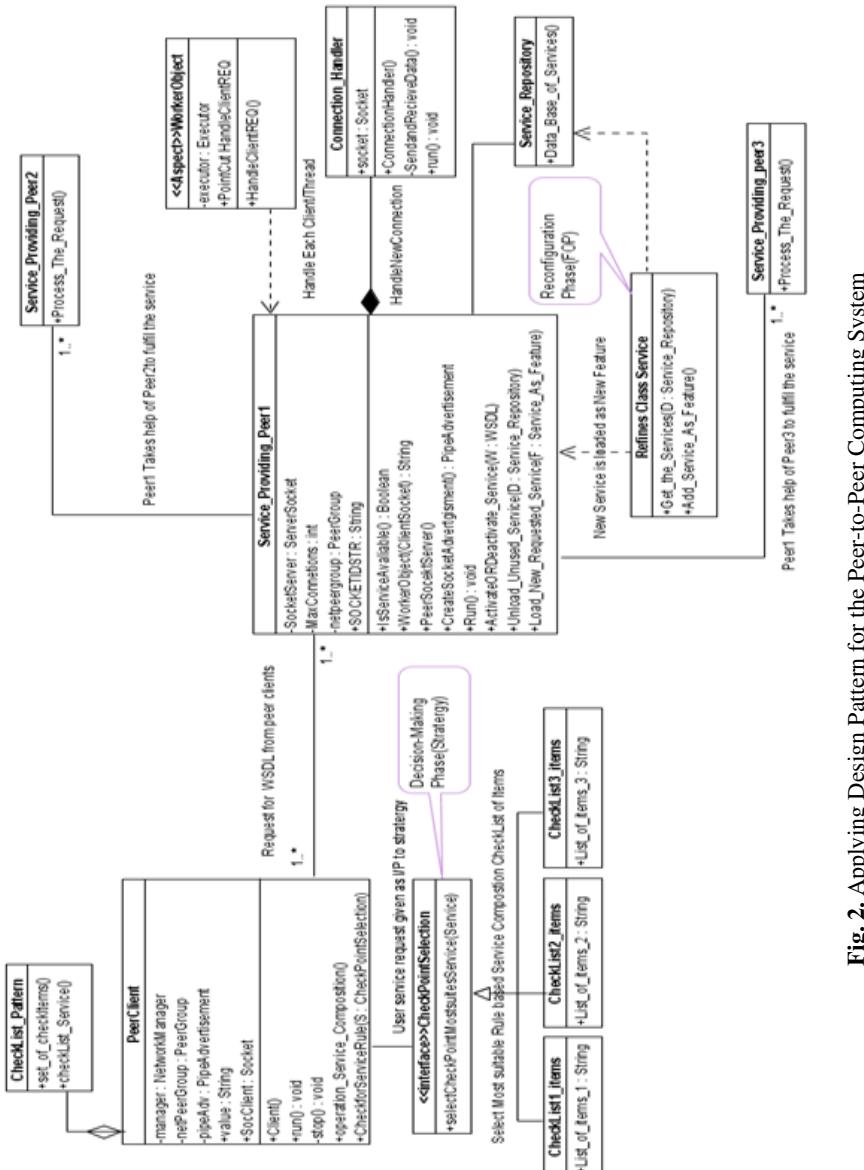


Fig. 1. The Proposed Design Pattern Structure

4.6 Participants

- Client:** This application creates JxtaSocket and attempts to connect to JxtaServerSocket. Peer Group will create a default net peer group and a socket is used to connect to JxtaServerSocket. After these steps the client will call the run method to establish a connection to receive and send data. The startJxta method is called in the client to create a configuration from a default configuration and then instantiates the JXTA platform and creates the default net peer group. Once the net peer group is created the client will send a Pipe Advertisement for requesting a service in the peer-to-peer network.



- b) **Checklist Pattern [10]:** In this class it will consists of Checklists that a set of Check items. Here it will provide the service execution plans for the client's requested service.
- c) **Checkpoint Selection:** Here the strategy pattern will help to select the perfect Checklist that matches for the execution of the requested complex service. The selection is based on some rule based selection statements and also that matches the input and output parameter details in WSDL of a particular Peer Server.
- d) **Service Providing PeerN (N=1 or 2 or 3):** First the default net peer group is created with Peer Group. Creates a JxtaServerSocket to accept connections, and then executes the run method to accept connections from clients and send receive data. The peer will start listening for the service requests. If a service that is requested by a client is not running in the memory then the peer will load the new service from the service repository into the memory space of the currently running services and at the same time it will invoke a functional call that will change the state of that service to Activated from deactivated state.
- e) **Connection Handler:** This will take care of the connections with multiple clients and sending and receiving the data between the clients and service providing peer.
- f) **Worker Object Aspect:** Once the connection is established between the peer and client, the worker object will create a separate thread for the connected client to run the requested service in a new thread for that it will call the Reactor Pattern method to handle the requested service execution.
- g) **Service Repository:** It will contain the services that are currently deactivated and that are not loaded into the memory for execution. The services not being requested by any of the clients are kept here.
- h) **Refines Class Service:** This will add a new service from the service repository that is requested by a client, in such a way that the insertion will be done as a new feature with the help of FOP [12].

4.7 Consequences

- a) With the use of this pattern we can get the benefits of worker object pattern, The application will handle each of the client requests in a separate thread by reducing the overhead on the main peer thread that is providing the service to get blocked until the first client request is served and making other client requests to get blocked.
- b) We use the Feature-Oriented Programming [12] to insert the new service into the memory from the Service Repository as a feature into the current executing services code.
- c) By the use of dynamic crosscutting concerns of the Aspect-Oriented Programming [7] the system will me executing fast as the decisions are made at run-time. With the help of the Web Service Description Language (WSDL) all the clients can get information about the services that are active and currently provided by the peer servers.

4.8 Roles of our Design patterns in Autonomic System

- a) **Worker Object Pattern [7]:** The worker object pattern is an instance of a class that encapsulates a worker method. A worker object can be passed around, stored, and invoked. The worker object pattern offers a new opportunity to deal with otherwise complex problems. It will provide the server with the facility to handle the service request from different clients in a separate per client connection.
- b) **Strategy Design Pattern [3]:** This pattern can be used to define a family of algorithms, encapsulate each one, and make them interchangeable. Strategy lets the algorithm vary independently from the clients that use it.
- c) **Checklisting Design Pattern [10]:** The Checklisting pattern provides means for maintaining a (usually ordered) list of services (usually running on different peers) an application has to interact with, considering arbitrary constraints. This problem typically arises in mobile scenarios where a task is accomplished by the collaboration of multiple peers.

5 Conclusion and Future Work

In this paper we have proposed a pattern to facilitate the ease of developing unstructured peer-to-peer computing systems. So with the help of our proposed design pattern named Service Injection Design pattern for unstructured peer-to-peer systems, provide services to clients with the help of Aspect-Oriented design patterns. So with this pattern we can handle the service-request of the clients and inject the new services into peer's code as feature modules. Several future directions of work are possible. We are examining how these design patterns can be inserted into a middleware technology , so that we can provide the Autonomic properties inside the middleware technologies like JXTA with the help of Features-Oriented and Aspect-Oriented programming methods.

The view of our proposed design pattern for the unstructured peer-to-peer computing System can be seen in the form of a class diagram see Fig 2.

References

1. Dobson, S., Sterritt, R., Nixon, P., Hinckley, M.: Fulfilling the Vision of Autonomic Computing, vol. 43, pp. 35–41. IEEE Computer Society (2010), doi:doi:10.1109/MC.2010.14
2. JXTA Java Standard Edition v2.5: Programmers Guide, © 2002-2007 Sun Microsystems, Inc. (September 10, 2007)
3. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley (1995)
4. Kuhleemann, M., Rosenmüller, M., Apel, S., Leich, T.: On the Duality of Aspect-Oriented and Feature-Oriented Design Patterns. In: Proceedings of the 6th Workshop on Aspects, Components, and Patterns for Infrastructure Software. ACM, New York (2007), doi:10.1145/1233901.1233906

5. Hannemann, J., Kiczales, G.: Design Pattern Implementation in Java and AspectJ. In: Proceedings of the International Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA), pp. 16–17 (2002), doi:10.1145/583854.582436
6. Tarr, P., Ossher, H., Harrison, W., Sutton, J.S.M.: N Degrees of Separation: Multi-Dimensional Separation of Concerns. In: Proceedings of the International Conference on Software Engineering (ICSE), pp. 107–119. 10.1145/302405.302457 (1999)
7. Laddad, R.: AspectJ in Action, 2nd edn., ch. 12. Manning (2010)
8. Prasad Vasireddy, V.S., Mannava, V., Ramesh, T.: A Novel Autonomic Design Pattern for Invocation of Services. In: Wyld, D.C., Wozniak, M., Chaki, N., Meghanathan, N., Nagamalai, D. (eds.) CNSA 2011. CCIS, vol. 196, pp. 545–551. Springer, Heidelberg (2011), doi:10.1007/978-3-642-22540-6_53
9. Apel, S., Leich, T., Saake, G.: Aspectual Feature Modules. IEEE Transactions on Software Engineering 34(2) (2008), doi:10.1109/TSE.2007.70770
10. Ferscha, A., Hechinger, M., Mayrhofer, R., Chtcherbina, E., Franz, M., dos Santos Rocha, M., Zeidler, A.: Bridging the gap with P2P patterns. In: Proceedings of the Workshop on Smart Object Systems, In conjunction with the Seventh International Conference on Ubiquitous Computing (2005)
11. Ramirez, A.J., Betty, H.C., Cheng: Design Patterns for Developing Dynamically Adaptive Systems. In: Proceedings of the 2010 ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems. ACM, New York (2010), doi:10.1145/1808984.1808990
12. Batory, D., Sarvela, J.N., Rauschmayer, A.: Scaling Step-Wise Refinement. IEEE Transactions on Software Engineering 30(6), 187–197 (2003), doi:10.1109/ICSE.2003.1201199

A Policy Driven Business Logic Change Management for Enterprise Web Services

M. Thirumaran¹, P. Dhavachelvan², and G. Naga Venkata Kiran¹

¹ Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry, India

² Department of Computer Science, School of Engineering and Technology, Pondicherry University, Puducherry, India

{thirumaran, nagavenkatakiran}@pec.edu, dhavachelvan.csc@pondiuni.edu.in

Abstract. Services might be moved, or relocated and may undergo changes during its life cycle but compromising the changes with respect to the business policy will be the key issue which has not been addressed by the existing business process change management framework. Hence there should be an effective framework for managing these changes without affecting the business functionality and also ensuring the associated business policy. This paper stresses mainly on handling these emergency changes to a business process which is capsuled as web services dynamically at the business analyst level. We use Finite State Machine for simulating the business logic set which includes the business policy constraints with its associated mapping function in order ensure the changes made are policy enabled during run time. It also facilitates Business Analyst directly to control and manage the business logic of the targeting web services dynamically and thereby eliminates the IT staff effort.

Keywords: Change Management, Policy Driven Approach, Finite State Machine, Business Logic, Business Process Automation.

1 Introduction

Web services provide a new approach for accessing systems in a loosely coupled, platform independent and standardized manner. However services might be moved, or relocated and may undergo changes during its life cycle. Hence there should be an effective framework for managing these changes without affecting the business functionality. Change management is the process responsible for controlling and managing the lifecycle of all changes in an IT environment. The goal of change management is to “ensure that standardized methods and procedures are used for efficient and prompt handling of changes, in order to minimize the impact of change-related incidents upon service quality and, consequently, improve the day-to-day operations of the organization. There are different approaches for managing changes in long term composed services at different levels. Changes can be propagated either at service developer’s side or at service provider’s site. If web service is exposed to changes the web service clients on the customer’s side is neither part of the upgraded

piece of software nor of the IT environment which is being upgraded. This probably results in erroneous behavior whose real cause is located somewhere in the system. Changes shouldn't attract attention to themselves in this way. Ideally they shouldn't be noticeable at all. If this is not possible consumers should be notified about upcoming changes in advance. Based on this notification they could request additional details of upcoming changes and prepare their Web service clients accordingly. The necessary measures often depend on the change that has been carried out on the Web service provider side. Changes can be mainly classified into two ways – Internal changes and External changes. Internal changes to a web service are propagated by the service developer. External changes are the changes made by third party service providers and here, service consumers have to manage changes should have the capability to understand service descriptions, discover the services dynamically and invoke them. We propose a framework for managing changes that are propagated during run time using FSM (Finite State Machine).

In our work we introduce a policy driven change management framework to manage the emergency changes by the business analyst. Usually Business Analyst deals with the business process in the administration level for any kind of modification or refinement of existing business process. Here Business Analyst can directly control and manage the business process at the development level through the business logic schema.

2 Framework for Policy Driven Web Service Change Management

A framework for satisfying dynamic run time changes in web service is proposed along with the evaluation of changes using Finite State Machine and to handle the critical and emergency changes that arise during runtime rather than the standard and normal changes [Fig. 1]. Also it provides a sophisticated change monitoring system through the change evaluation process. This approach incorporates the run time changes on the business logic more easily and also for evaluating the changes made at service business logic in terms of business workflow with respect to the change requirements, policy enforcement and the service performance. In our work we use finite state machine for run time evaluation and management of the business logic change activity. This measure of change evaluation ensures the confident level of business analyst who themselves feels that the business logics are always under their control. Hence the automated business logic evaluation model with dynamic change criterion analysis empower the task of service management tremendously by reducing the programming effort and development cost thus significantly enhancing the speed of business logic change adaption in the nascent business market. The business logic to be changed is the input to the change analyzer. This input is given as the change request to the change analyzer which analyzes the whether the changes can be validated or not. The service registry contains the list of services published. The business logic analyzer analyzes the changes and identifies the corresponding policies that might be affected partially from the policy set. The business logic set contains all the logics of the enterprise and they are retrieved as and when needed. The finite state machine simulator simulates the changes to be done automatically without the intervention of source developer. The state

transition storage table consists of transitions of functions from one state to the other. Then the changes affecting the other services or business logic is identified from the mapping function. The change evaluator evaluates the changes to be done and passes to the policy manager and the schema generator is used to generate the new schema for the change request. Finally the business logic schema is generated automatically and fruitful changes are done to the system.

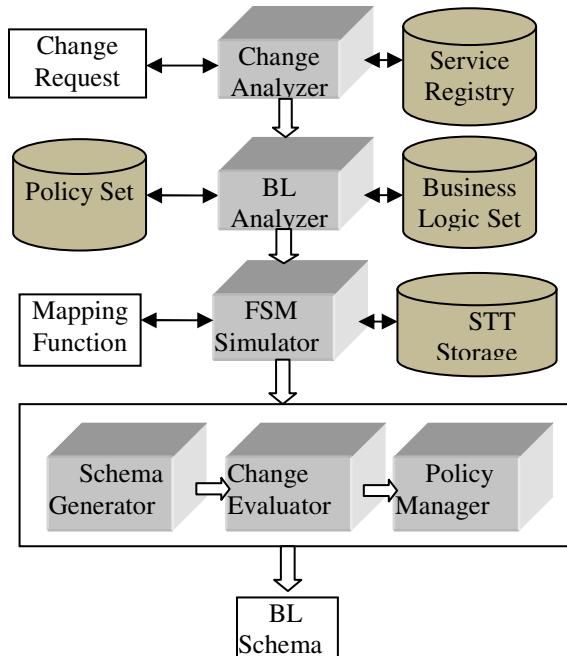


Fig. 1. Framework for Policy driven Web Service Change Management

3 Business Policy Driven Mapping Algorithm

The following algorithm builds the policy mapping existing between the business logic set L and the policy set. If any changes are to be done, i.e. after change requests are received, logic is extracted for the domain variable Dv. Then the business logic schema is generated based on the logic and policy mapping constraints. Then the changes to be enforced are measured to find the completeness of the set and a state transition table is build to analyze the states after changes are done.

Algorithm PolicyMapper (Business Logic, Policy Set)
 $\text{//L} \rightarrow (R, F, Pr, Py, D)$; where L is business Logic set
 $\text{//R} \rightarrow \text{set of rules}$
 $\text{//F} \rightarrow \text{set of functions}$

```

//Pr→set of parameters
//Py→set of policies
//D→function dependency relation
Begin
CR→GetChangeRequest
Domain→find Domain (CR)
//CR be the change request consisting of Domain Variable (DV)
L→extract logic (domain, DV)
L→(R, F, P, Py, D)
// STT be the state transition table
STT→Simulates (L)
BLSchema→Generate Schema (L, Py)
//“f” be mapping function
//policy mapping
P→gets Policy (Py)
State→start state
i=1
Do
Begin
Next→findTransition (L, P, STT)
State→next
f[i]→ next
i→ i +1
End
While (state==endstate && P==NULL)
Return f[i]
End

```

4 Case Study

We consider a travel agency scenario [Fig. 2] which outsources its functionality from the following services: (i) Travel reservation service (ii) Airline reservation service (iii) Vehicle reservation service and (iv) Hotel reservation service. For example consider the airline reservation service which consists of following business rules, functions and parameters: The business rules of an online airline reservation service are: authentication, ticket booking and total amount. Therefore the authentication rule of airline reservation must validate the business functions namely, user name verification and password verification. Similarly the ticket booking rule has to validate the following business functions: check availability, get details and verify schedule. Now if a new function “concession” has to be added to the existing business rule “ticket booking” in order to offer concessions to senior citizens and children under age group 12, this method employs the generation of business logic which will be added dynamically to the existing source code. Hence the logic is automatically generated without the need of IT staff.

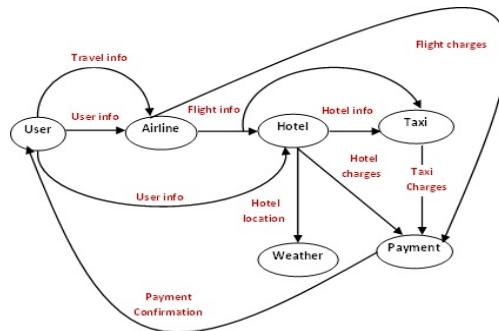


Fig. 2. Travel Agency Workflow

4.1 Airline_Reservation_Service.xsd [Before Change]

```
<?xml version="1.0" standalone="yes"?>
<airlinereservation>
<Book_ticket>type = "rule", id="R2"
<flight_details>type="function",id="F21"
<Flight_name>type="parameter", id="P211"
</Flight_name>
<Flight_id>type="parameter", id="P212"</Flight_id>
</flight_details>
<check_availability>type="function",id="F22"
<no_of_seats>type="parameter", id="P221"</no_of_seats>
<seat_number>type="parameter", id="P222"</seat_number>
</check_availability>
<verify_schedule>type="function",id="F23"
<Flight_name> type="parameter", id="P231"</Flight_name>
<Flight_id>type="parameter", id="P232"</Flight_id >
<arrival_time>type="parameter", id="P233"</arrival_time>
</verify_schedule>
</Book_ticket>
</airlinereservation>
```

4.2 Airline_Reservation_Service.xsd [After Change]

```
<?xml version="1.0" standalone="yes"?>
<airlinereservation>
<Book_ticket>type = "rule", id="R2"
<flight_details>type="function",id="F21"
<Flight_name>type="parameter", id="P211"
</Flight_name>
<Flight_id>type="parameter", id="P212"</Flight_id>
</flight_details>
```

```
<check_availability>type="function", id="F22"
<no_of_seats>type="parameter", id="P221"</no_of_seats>
<seat_number>type="parameter", id="P222"</seat_number>
</check_availability>
<verify_schedule>type="function", id="F23"
<Flight_name> type="parameter", id="P231"</Flight_name>
<Flight_id>type="parameter", id="P232"</Flight_id>
<arrival_time>type="parameter", id="P233"</arrival_time>
</verify_schedule>
</Book_ticket>
<concession>type="function" id="F24"
<age>type="parameter" id="P241"</age>
</concession>
</airlinereservation>
```

Now the changes done at the schema level are mapped to the corresponding source code without the intervention of source code developer. All the changes are done at business analyst side.

5 Implementation

We have implemented the above mentioned case study using java web services technology in Net Beans IDE and also incorporated changes to the airline reservation service schema which reflects the changes back to the source code level without the intervention of source code developer [Fig. 3]. A new business logic “concession” is added to the service and all the changes are implemented successfully at the business analyst level itself.

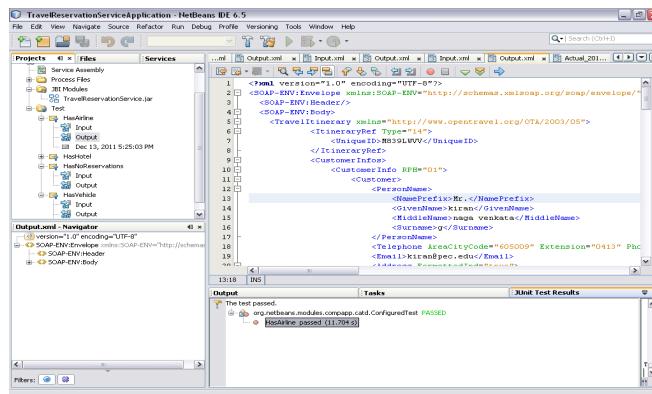
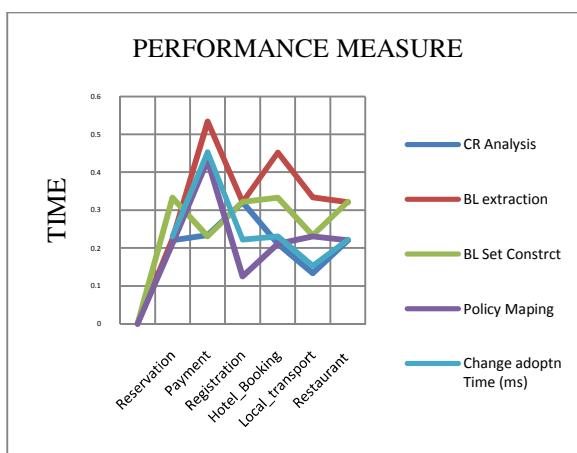


Fig. 3. Reservation Service Build and Redeployment after the change

Table 1. Evaluation of Change management

Business Logic	CR Analysis Time (ms)	BL extraction Time (ms)	BL Set Construc Time (ms)	Policy Mappi ng Time (ms)	Change adoptin g Time (ms)	Total Time
Reservation	0.22	0.22	0.33	0.21	0.23	1.22
Payment	0.23	0.53	0.23	0.43	0.45	1.88
Registration	0.32	0.32	0.32	0.12	0.22	1.31
Hotel_Booking	0.21	0.45	0.33	0.21	0.23	1.43
Local_transport	0.13	0.33	0.23	0.23	0.15	1.08
Restaurant	0.22	0.32	0.32	0.22	0.22	1.30

The above table gives the results of performance measures for various services implemented in our case study. They are identified with respect to change request analysis [CR Analysis], business logic extraction [BL Extraction], and business logic set construction time [BL Set Construct time], policy mapping time, and change adoption time. Also the total time is calculated for analyzing the time taken to redeploy the services after changes are done at business logic level. The performance measure for the above case study is given in the following graph depicted as Figure 4.

**Fig. 4.** Performance evaluation of various services implemented with respect to time

6 Related Works

Dimitris Apostolou et all presented an ontology-based approach for managing e-Government services that enables systematic response of e-Government systems to changes by applying formal methods for achieving consistency when a change is discovered; enables knowledgeable response of service designers and implementers to changes by utilizing design rationale knowledge [1]. Xumin Liu et all proposed a framework where managing changes in LCSs is modeled as a dual service query optimization process where in the first phase, the selection of Web services is based on using reputation as the key parameter and in the second phase, the non-functional QoS is used to narrow down the set to those Web services that are both reputable and best meet the QoS [2]. It is central to recognize that differences don't only exists among those reviewed methodologies; change management tasks are not the same even in one ERP project as different groups and individuals will be affected differently and therefore will need different strategies [3]. Combination of Ordinary Petri nets and Reconfigurable Petri nets were used to model the triggering changes and reactive changes, respectively by Salman Akram and proposed an automatic change management framework that is based on the Petri net models [4]. Christian Gerth and others described an approach that allows the semantic comparison of different business process models using a normal form [5]. The dependencies between business processes and services in service oriented environment for managing changes were proposed by Yi Wang, Jian Yang and Weiliang Zhao [6]. Praveen K. Muthuswamy developed change scheduling algorithms that seek to attain the "change capacity" of the system [7]. Bruno Wassermann proposes the Change 2.0 approach to cross-domain change management based on an inversion of responsibility for impact assessment and the facilitation of cross-domain service process integration [8]. Xiang Luo, Koushik Kar et al considered the Change Management Process for Enterprise IT services with the goal of improving the efficiency of this process, i.e., minimizing change completion time and maximizing the "change capacity" [9]. Uttam Kumar Tripathi presented a methodology and system for changing SOA-based business process implementation at two layers: the design layer processes are modeled in the ontology-based semantic markup language for web services OWL-S, and for execution, the processes are translated into BPEL [10].

7 Conclusion and Future Enhancements

Managing and implementing business logic changes during run-time are a critical problem. We have provided an efficient framework for monitoring and managing changes in web services business logic during run-time. A Finite State Machine based model was presented to show the improvements from the existing approaches and for simulating the business logic set which includes the business policy constraints with its associated mapping function in order ensure the changes made are policy enabled during run time. It facilitated Business Analyst directly to control and manage the business logic of the targeting web services dynamically and thereby eliminated the IT staff effort.

References

- [1] Apostolou, D., Mentzas, G., Stojanovic, L., Thoenissen, B., Lobo, T.P.: A collaborative decision framework for managing changes in e-Government services. ScienceDirect, Government Information Quarterly 28, 101–116 (2011)
- [2] Liu, X., Bouguettaya, A., Yu, Q., Malik, Z.: Efficient change management in long-term composed services. Original Research Paper. Springer-Verlag London Limited (2010)
- [3] Alballaa, H., Al-Mudimigh, A.S.: Change Management Strategies for Effective Enterprise Resource Planning Systems: A Case Study of a Saudi Company. International Journal of Computer Applications (0975 – 8887) 17(2) (March 2011)
- [4] Akram, S., Bouguettaya, A., Liu, X., Haller, A., Rosenberg, F., Wu, X.: A Change Management Framework for Service Oriented Enterprises. International Journal of Next-Generation Computing (IJNGC) 1(1), 1–77 (2010)
- [5] Gerth, C., Luckey, M., Kuster, J.M., Engels, G.: Detection of Semantically Equivalent Fragments for Business Process Model Change Management. In: IEEE International Conference on Services Computing 2010. IEEE Computer Society (2010)
- [6] Wang, Y., Yang, J., Zhao, W.: Managing Changes for Service Based Business Processes. In: IEEE Asia-Pacific Services Computing Conference 2010. IEEE Computer Society (2010)
- [7] Muthuswamy, P.K., Kar, K., Sahu, S., Pradhan, P., Sarkar, S.: Change Management in Enterprise IT Systems: Process Modeling and Capacity-optimal Scheduling. In: IEEE INFOCOM 2010 Proceedings. IEEE Communications Society (2010)
- [8] Wassermann, B., Ludwig, H., Laredo, J., Bhattacharya, K., Pasquale, L.: Distributed Cross-Domain Change Management. In: IEEE International Conference on Web Services 2009. IEEE Computer Society (2009)
- [9] Luo, X., Kar, K., Sahu, S., Pradhan, P., Shaikh, A.: On Improving Change Management Process for Enterprise IT Services. In: IEEE International Conference on Services Computing 2008. IEEE Computer Society (2008)
- [10] Tripathi, U.K., Hinkelmann, K., Feldkamp, D.: Life Cycle for Change Management in Business Processes using Semantic Technologies. Journal of Computers 3(1) (January 2008)

Author Index

- Afzali, Hammad 345
Agarwal, Rekha 589
Amalarethinam, D.I. George 969
Amritpal, 119
Analoui, Morteza 1041
Anitha, Avula 999
Anitha, R. 727, 773
Anitha, V. 813
Asha, S. 953
Aswathy, M.C. 1027
Atiea, Mohammed A. 913, 919
Avadhani, P.S. 703
Azeez, Arifa 47
- Balwan, Jai 119
Banerjee, P.K. 659
Basu, Atanu 713
Bate, Stephen 1
Bawa, Seema 323
Bhardwaj, Ved Prakash 979
Bhattacharyya, Debika 659
Bhattacherjee, Vandana 617
Bhosale, Varsharani 873
Bhushan, Bharat 519
Bose, S. 333
- Chaki, Nabendu 627
Chandra, Shalini 937
Chandrakala, C.B. 481
Charyulu, N.Ch. Bhatra 737
Chaudhury, Sankhayan 627
Chauhan, Piyush 1051
Chetan, S. 499
Chowdhury, Prasun 457
Chowdhury, Soumit 745
- Damodaram, A. 193, 737
Das, Indrajit 15
Dasgupta, Moitreyee 627
Dattagupta, Rana 181, 211
Deodhar, R.S. 843
Devaki, P. 897
Dey, Dhananjoy 803
Dhariwal, Sumit 793
Dhavachelvan, P. 1085
Dhavale, Sunita V. 843
Dorairaj, Prabhu 391
Doreswamy, 161
Dubey, Gaurav 149
- Ekpenyong, Moses E. 103
- Gad, R.S. 81
Gad, V.R. 81
Gajawada, Satish 267
Garg, Poonam 295
Gebril, Zahra Mohana 305
Ghosal, Nabin 863
Ghosal, S.K. 763
Ghoshal, Nabin 745
Ghoualmi, Nacira 853
Giri, Debasis 617
Girkar, Fazila 873
Gopalan, N.P. 257
Goswami, Anirban 863
Guha, Sutirtha Kumar 181
Gupta, Shailender 519
Gupta, Vishal 73
- H. Ahmed, Ali 447
Hareesha, K.S. 481

- Hedar, Abdel-Rahman 913, 919
 Hemanth, K.S. 161
 Hirai, Hiromi 469
 Holambe, Raghunath S. 599
 Ibrahim, HosnyM. 447
 Ilyas, Muhammed 647
 Jacob, K. Paulose 231
 Jain, Pranita 835
 Jisha, G. 63
 Joardar, Subhankar 617
 Jophin, Shany 507
 Joseph, Gnana Jayanthi 405
 Kadambi, Govind 1
 Kahya, Noudjoud 853
 Kamhoua, Charles A. 883
 Kanrar, Soumen 55
 Karangi, Javid K. 313
 Karthik, R. 773
 Karthika, S. 333
 Kaur, Gurpreet 803
 Kaushik, B.K. 125
 Kaushik, Sona 377
 Kayalvizhi, R. 953
 Khan, Raees Ahmad 91, 937
 Khola, R.K. 609
 Kiruthiga, A. 333
 Kodada, Basappa B. 1009
 Kohli, Sheena 275
 Kulkarni, P.J. 579
 Kumar, Ajay 323, 507
 Kumar, Brijesh 125
 Kumar, B. Vijay 257
 Kumar, Chirag 519
 Kumar, Jalesh 783
 Kumar, Neeraj 91
 Kumar, Sunit 221
 Kumar, T.V. Vijay 149
 Kundu, Anindita 457
 Kundu, Anirban 181
 Kushal, K.S. 499
 Kwiat, Kevin A. 883
 Lafourcade, Pascal 853
 Lashkari, Arash Habibi 305
 Lee, Young-Ran 695
 Li, Tiantian 989
 Liu, Xuebing 989
 Lobiyal, D.K. 551
 Maan, Jitendra 559
 Mahalingam, P.R. 63
 Mahdy, Yousef B. 913, 919
 Majumder, Abhishek 435
 Mala, C. 257
 Malviya, Vijay 367
 Mandal, Jyotsna Kumar 745, 753, 763
 Mankar, V.H. 169, 529
 Mannava, Vishnuvardhan 1017, 1075
 Manu, Vardhan 425
 Mao, Yingchi 637
 Marigowda, C.K. 313
 Meenakshi, A.V. 953
 Meghanathan, Natarajan 415
 Miki, Kaori 469
 Milidiú, Ruy Luiz 141
 Mishra, Mina 169, 529
 Mishra, Minati 221
 Mishra, Subhadra 221
 Misra, Iti Saha 457
 Mittal, Poornima 125
 Mohan, Akhil 241
 Mohanty, Ipsita 823
 Mokhtari, Hassan 345
 Mondal, Jyotsna Kumar 863
 Mondal, Uttam Kr. 753
 Motta, Eduardo 141
 Mukhopadhyay, Debajyoti 211
 Mustafa, K. 905
 Muthulakshmi, P. 969
 Muttanna Kadal, H.K. 499
 Nagpal, C.K. 519
 Naik, G.M. 81
 Naik, Neelima S. 1061
 Naik, Samedha S. 541
 Natarajan, V. 727
 Nayak, Pinki 589
 Nayak, Vidyavati S. 803
 Neal, David (DJ) 685
 Neelakantappa, M. 193
 Negi, Y.S. 125
 Nene, Manisha J. 541
 Nirmala, S. 783
 Nithya, B. 257
 Nitin, 979, 1051

- Oguchi, Masato 469
 Omar, NagwaM. 447
- Padmavathi, S. 285
 Pal, Debajyoti 489
 Pal, Dipankar 863
 Pal, S.K. 803
 Pande, Vijay 1
 Pandey, Soma K. 1, 905
 Paras, Gupta 425
 Parhizkar, Behrang 305
 Park, Joon S. 883
 Parwani, Kashish 569
 Patnaik, L.M. 843
 Philip, Priya 507
 Prabhu, Raghavendra 1009
 Pravin, V. 773
 Preetha, K.G. 47, 231
 Prema, K.V. 481
 Puri, Shalini 377
 Purohit, G.N. 275, 569
 Putta, Chandra Shekar Reddy 355
- Quadri, Mohammed A.R. 1017, 1075
- Rabara, S. Albert 405
 Raghavendra Rao, C. 999
 Raghavendra Rao, G. 897
 Raghuwanshi, Sandeep 793
 Rahman, Syed (Shawon) 685
 Raikwar, Amit Kumar 673
 Rajalaxmi, C. 285
 Rajput, Nupur 835
 Ramachandran, Anand 305
 Ramachandran, Baskaran 37
 Ramalingam, Ashok Kumar 391
 Ramamoorthy, Saranya 391
 Ramchand, V. 551
 Ramesh, T. 1017, 1075
 Rao, D. Sreenivasa 25
 Rao, M. Varaprasad 737
 Ravilla, Dilli 355
 Reddy, K. Ganesh 927
 Rohil, Mukesh Kumar 73
 Rupa, Ch. 703
- Saha, Himadri Nath 659
 Sanyal, Salil K. 457
 Saravanan, R. 15
- Satyanarayana, B. 193
 Saurabh, Praneet 367
 Sawant, Mrunal 873
 Schwabe, Daniel 141
 Sengupta, Indranil 713
 Shah, Medha A. 579
 Shaikh, Zulfa 295
 Shankar, Gori 119
 Sharma, Manoj 609
 Sheethal, M.S. 507
 Shet, Raghavendra M. 599
 Shivakumar, K.M. 1009
 Shivaputra, 499
 Shrivastava, Shailendra 793, 835
 Sing, Jamuna Kanta 713
 Singh, Anju 1061
 Singh, Archana 149
 Singh, Kushwaha Dharmender 425
 Sinha, Sukanta 211
 Soman, K.P. 285
 Song, Jie 989
 Soryani, Mohsen 1041
 Subramanian, Sridhar 37
 Sudarsan, Dhanya 63
 Sulthani, R. Mynuddin 25
- Tahir, Mohammed A. 305
 Taruna, S. 275
 Tayal, Shikha 1033
 Tele, Imam Musa Abiodunde 305
 Terrell, Michael 415
 Thilagam, P. Santhi 927
 Thirugnanam, K. 773
 Thirumaran, M. 1085
 Thriveni, J. 313
 Toshniwal, Durga 267
 Tripti, C. 1027
- Umana, Enobong 103
 Unnikrishnan, A. 231
 Upadhyay, Nitin 241
 Upadhyay, Vimal 119
- Velusamy, R. Leela 813, 823
 Venkata Kiran, G. Naga 1085
 Verma, Bhupendra 367
 Verma, Ravindra Kumar 1061
 Verma, Sangeeta 1061
 Verma, Seema 589

- Vijay, Sandip 1033
Vijayakumar, R. 647
Vyavhare, Ashvini 873
Wankar, Rajeev 999
- Yamaguchi, Saneyasu 469
Yin, Ting 637
Zarrinchian, Ghobad 1041
Zhu, Zhiliang 989