



International Journal of Advanced Computer Science and Applications

Volume 2 Issue 1

January 2011



ISSN 2156-5570(Online)
ISSN 2158-107X(Print)



www.ijacsa.thesai.org



INTERNATIONAL JOURNAL OF
ADVANCED COMPUTER SCIENCE AND APPLICATIONS



A Publication of
The Science and Information Organization



IJACSA Editorial

From the Desk of Managing Editor...

May 2011 bring something new in life. Let us bathe in a resolution to bring hopes and a spirit to overcome all the darkness of past year. We hope it provide a determination to start all things with a new beginning which we failed to complete in the past year, some new promises to make life more beautiful, peaceful and meaningful.

Happy New Year and welcome to the first issue of IJACSA in 2011.

With the onset of 2011, The Science and Information Organization enters into agreements for technical co-sponsorship with SETIT 2011 organizing committees as one means of promoting activities for the interest of its members. It's a great opportunity to explore and learn new things, and meet new people

The number of submissions we receive has increased dramatically over the last issues. Our ability to accommodate this growth is due in large part to the terrific work of our Editorial Board.

In order to publish high quality papers, Manuscripts are evaluated for scientific accuracy, logic, clarity, and general interest. Each Paper in the Journal not merely summarizes the target articles, but also evaluates them critically, place them in scientific context, and suggest further lines of inquiry. As a consequence only 30% of the received articles have been finally accepted for publication.

IJACSA emphasizes quality and relevance in its publications. In addition, IJACSA recognizes the importance of international influences on Computer Science education and seeks international input in all aspects of the journal, including content, authorship of papers, readership, paper reviewers, and Editorial Board membership

The success of authors and the journal is interdependent. While the Journal is advancing to a new phase, it is not only the Editor whose work is crucial to producing the journal. The editorial board members, the peer reviewers, scholars around the world who assess submissions, students, and institutions who generously give their expertise in factors small and large— their constant encouragement has helped a lot in the progress of the journal and earning credibility amongst all the reader members.

We hope to continue exploring the always diverse and often astonishing fields in Advanced Computer Science and Applications

Thank You for Sharing Wisdom!

Managing Editor
IJACSA
January 2011
editorijacsa@thesai.org

IJACSA Associate Editors

Dr. Zuqing Zhu

Service Provider Technology Group of Cisco Systems, San Jose

Domain of Research: Research and development of wideband access routers for hybrid fibre-coaxial (HFC) cable networks and passive optical networks (PON)

Dr. Jasvir Singh

Dean of Faculty of Engineering & Technology, Guru Nanak Dev University, India

Domain of Research: Digital Signal Processing, Digital/Wireless Mobile Communication, Adaptive Neuro-Fuzzy Wavelet Aided Intelligent Information Processing, Soft / Mobile Computing & Information Technology

Dr. Sasan Adibi

Technical Staff Member of Advanced Research, Research In Motion (RIM), Canada

Domain of Research: Security of wireless systems, Quality of Service (QoS), Ad-Hoc Networks, e-Health and m-Health (Mobile Health)

Dr. T. V. Prasad

Dean, Lingaya's University, India

Domain of Research: Bioinformatics, Natural Language Processing, Image Processing, Expert Systems, Robotics

Dr. Bremananth R

Research Fellow, Nanyang Technological University, Singapore

Domain of Research: Acoustic Holography, Pattern Recognition, Computer Vision, Image Processing, Biometrics, Multimedia and Soft Computing

IJACSA Reviewer Board

- **Abbas Karimi**
I.A.U_Arak Branch (Faculty Member) & Universiti Putra Malaysia
- **Dr. Abdul Wahid**
University level Teaching and Research, Gautam Buddha University, India
- **Abdur Rashid Khan**
- **Dr. Ahmed Nabih Zaki Rashed**
Menoufia University, Egypt
- **Md. Akbar Hossain**
Doctoral Candidate, Marie Curie Fellow, Aalborg University, Denmark and AIT, Greeceas
- **Albert Alexander**
Kongu Engineering College, India
- **Prof. Alc -nia Zita Sampaio**
Technical University of Lisbon
- **Amit Verma**
Rayat & Bahra Engineering College, Mohali, India
- **Arash Habibi Lashakri**
University Technology Malaysia (UTM), Malaysia
- **Aung Kyaw Oo**
- **B R SARATH KUMAR**
Principal of Lenora College of Engineering, India
- **Dr.C.Suresh Gnana Dhas**
Professor, Computer Science & Engg. Dept
- **Mr. Chakresh kumar**
Assistant professor, Manav Rachna International University, India
- **Chandrashekhar Meshram**
Shri Shankaracharya Engineering College, India
- **Prof. D. S. R. Murthy**
Professor in the Dept. of Information Technology (IT), SNIST, India.

- **Prof. Dhananjay R.Kalbande**

Assistant Professor at Department of Computer Engineering, Sardar Patel Institute of **Technology**,
Andheri (West), Mumbai, India

- **Dhirendra Mishra**

SVKM's NMIMS University, India

- **Hanumanthappa.J**

Research Scholar Department Of Computer Science University of Mangalore, Mangalore, India

- **Dr. Himanshu Aggarwal**

Associate Professor in Computer Engineering at Punjabi University, Patiala, India

- **Dr. Jamaiah Haji Yahaya**

Senior lecturer, College of Arts and Sciences, Northern University of Malaysia (UUM), Malaysia

- **Prof. Jue-Sam Chou**

Professor, Nanhua University, College of Science and Technology, Graduate Institute
and Department of Information Management, Taiwan

- **Dr. Juan José Martínez Castillo**

Yacambu University, Venezuela

- **Dr. Jui-Pin Yang**

Department of Information Technology and Communication at Shih Chien University, Taiwan

- **Dr. K.PRASADH**

METS SCHOOL OF ENGINEERING, India

- **Dr. Kamal Shah**

Associate Professor, Department of Information and Technology, St. Francis Institute of
Technology, India

- **Lai Khin Wee**

Technischen Universität Ilmenau, Germany

- **Mr. Lijian Sun**

Research associate, GIS research centre at the Chinese Academy of Surveying and Mapping, China

- **Long Chen**

Qualcomm Incorporated

- **M.V.Raghavendra**

Head, Dept of ECE at Swathi Institute of Technology & Sciences, India.

- **Mahesh Chandra**
B.I.T., Mesra, Ranchi, India
- **Md. Masud Rana**
Khunla University of Engineering & Technology, Bangladesh
- **Dr. Michael Watts**
Research fellow, Global Ecology Group at the School of Earth and Environmental Sciences,
University of Adelaide, Australia
- **Mohd Nazri Ismail**
University of Kuala Lumpur (UniKL)
- **Mueen Malik**
University Technology Malaysia (UTM)
- **Dr. N Murugesan**
Assistant Professor in the Post Graduate and Research Department of Mathematics, Government
Arts College (Autonomous), Coimbatore, India
- **Dr. Nitin Surajkishor**
Professor & Head, Computer Engineering Department, NMIMS, India
- **Dr. Poonam Garg**
Chairperson IT Infrastructure, Information Management and Technology Area, India
- **Pradip Jawandhiya**
Assistant Professor & Head of Department
- **Rajesh Kumar**
Malaviya National Institute of Technology (MNIT), INDIA
- **Dr. Rajiv Dharaskar**
Professor & Head, GH Rasoni College of Engineering, India
- **Prof. Rakesh L**
Professor, Department of Computer Science, Vijetha Institute of Technology, India
- **Prof. Rashid Sheikh**
Asst. Professor, Computer science and Engineering, Acropolis Institute of Technology and Research,
India
- **Rongrong Ji**
Harbin Institute of Technology

- **Dr. Ruchika Malhotra**
Delhi Technological University, India
- **Dr.Sagarmay Deb**
University Lecturer, Central Queensland University, Australia
- **Dr. Sana'a Wafa Al-Sayegh**
Assistant Professor of Computer Science at University College of Applied Sciences UCAS-Palestine
- **Dr. Smita Rajpal**
ITM University Gurgaon, India
- **Suhas J Manangi**
Program Manager, Microsoft India R&D Pvt Ltd
- **Sunil Taneja**
Smt. Aruna Asaf Ali Government Post Graduate College, India
- **Dr. Suresh Sankaranarayanan**
Department of Computing, Leader, Intelligent Networking Research Group, in the University of West Indies, Kingston, Jamaica
- **T V Narayana Rao**
Professor and Head, Department of C.S.E –Hyderabad Institute of Technology and Management, India
- **Totok R. Biyanto**
Engineering Physics Department - Industrial Technology Faculty, ITS Surabaya
- **Varun Kumar**
Institute of Technology and Management, India
- **Dr. V. U. K. Sastry**
Dean (R & D), SreeNidhi Institute of Science and Technology (SNIST), Hyderabad, India.
- **Vinayak Bairagi**
Sinhgad Academy of engineering, Pune
- **Vuda Sreenivasarao**
St.Mary's college of Engineering & Technology, Hyderabad, India.
- **Mr.Zhao Zhang**
Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong

CONTENTS

Paper 1: Computing knowledge and Skills Demand: A Content Analysis of Job Adverts in Botswana

Authors: Y. Ayalew, Z. A. Mbero, T. Z. Nkgau, P. Motlogelwa, A. Masizana-Katongo

PAGE 1 – 10

Paper 2: Open Source Software in Computer Science and IT Higher Education: A Case Study

Authors: Dan R. Lipşa, Robert S. Laramee

PAGE 10 – 17

Paper 3: Analyzing the Load Balance of Term-based Partitioning

Authors: Ahmad Abusukhon, Mohammad Talib

PAGE 18 – 25

Paper 4: A Genetic Algorithm for Solving Travelling Salesman Problem

Authors: Adewole Philip, Akinwale Adio Taofiki, Otunbanowo Kehinde

PAGE 26 – 29

Paper 5: Grid Approximation Based Inductive Charger Deployment Technique in Wireless Sensor Networks

Authors: Fariha Tasmin Jaigirdar, Mohammad Mahfuzul Islam, Sikder Rezwanul Huq

PAGE 30 – 37

Paper 6: PAV: Parallel Average Voting Algorithm for Fault-Tolerant Systems

Authors: Abbas Karimi, Faraneh Zarafshan, Adznan b. Jantan

PAGE 38 – 41

Paper 7: Solution of Electromagnetic and Velocity Fields for an Electrohydrodynamic Fluid Dynamical System

Authors: Rajveer S Yaduvanshi, Harish Parthasarathy

PAGE 42 – 49

Paper 8: Survey of Wireless MANET Application in Battlefield Operations

Authors: Dr. C. Rajabhushanam, Dr. A. Kathirvel

PAGE 50 – 58

Paper 9: An Efficient Resource Discovery Methodology for HPGRID Systems

Authors: D.Doreen Hephzibah Miriam, K.S.Easwarakumar

PAGE 59 – 68

Paper 10: Virtualization Implementation Model for Cost Effective & Efficient Data Centers

Authors: Mueen Uddin, Azizah Abdul Rahman

PAGE 69 - 74

Paper 11: 'L' Band Propagation Measurements for DAB Service Planning in INDIA

Authors: P.K.Chopra, S. Jain, K.M. Paul, S. Sharma

PAGE 75 – 82

Paper 12: Universal Simplest possible PLC using Personal Computer

Authors: B.K.Rana

PAGE 83 – 86

Paper 13: Simulation of Packet Telephony in Mobile Adhoc Networks Using Network Simulator

Authors: Dr. P.K.Suri, Sandeep Maan

PAGE 87 – 92

Paper 14: A Novel Approach to Implement Fixed to Mobile Convergence in Mobile Adhoc Networks

Authors: Dr. P.K.Suri, Sandeep Maan

PAGE 93 – 99

Paper 15: IPS: A new flexible framework for image processing

Authors: Otman ABDOUN, Jaafar ABOUCHABAKA

PAGE 100 – 105

Paper 16: Improved Off-Line Intrusion Detection Using A Genetic Algorithm And RMI

Authors: Ahmed AHMIM, Nacira GHOUALMI, Noujoud KAHYA

PAGE 106 – 112

Paper 17: Automatic Facial Feature Extraction and Expression Recognition based on Neural Network

Authors: S.P.Khandait, Dr. R.C.Thool, P.D.Khandait

PAGE 113 – 118

Paper 18: Coalesced Quality Management System

Authors: A. Pathanjali Sastri, K. Nageswara Rao

PAGE 119 – 126

Paper 19: Detection of Routing Misbehavior in MANETs with 2ACK scheme

Authors: Chinmaya Kumar Nayak, G K Abani Kumar Dash, Kharabela parida, Satyabrata Das

PAGE 127 – 130

Paper 20: Framework for Automatic Development of Type 2 Fuzzy, Neuro and Neuro-Fuzzy Systems

Authors: Mr. Jeegar A Trivedi, Dr. Priti Srinivas Sajja

PAGE 131 – 137

Paper 21: A study on Feature Selection Techniques in Bio-Informatics

Authors: S.Nirmala Devi, S.P. Rajagopalan

PAGE 138 – 144

Paper 22: Software Effort Prediction using Statistical and Machine Learning Methods

Authors: Ruchika Malhotra, Ankita Jain

PAGE 145 – 152

Computing knowledge and Skills Demand: A Content Analysis of Job Adverts in Botswana

Y. Ayalew, Z. A. Mbero, T. Z. Nkgau, P. Motlogelwa, A. Masizana-Katongo
Department of Computer Science, University of Botswana
{ayalew, mberoz, nkgautz, motlogel, masizana}@mopipi.ub.bw

Abstract - This paper presents the results of a content analysis of computing job adverts to assess the types of skills required by employers in Botswana. Through the study of job adverts for computing professionals for one year (i.e., January 2008 to December 2008), we identified the types of skills required by employers for early career positions. The job adverts were collected from 7 major newspapers (published both daily and weekly) that are circulated throughout the country. The findings of the survey have been used for the revision and development of curricula for undergraduate degree programmes at the Department of Computer Science, University of Botswana.

The content analysis focused on the identification of the most sought after types of qualifications (i.e., degree types), job titles, skills, and industry certifications. Our analysis reveals that the majority of the adverts did not set a preference to a particular type of computing degree. Furthermore, our findings indicate that those job titles and computing skills which are on high demand are not consistent with previous studies carried out in the developed countries. This requires further investigation to identify reasons for these differences from the perspective of the practices in the IT industry. It also requires further investigation regarding the degree of mismatch between the employers computing skills demands and the knowledge and skills provided by academic programmes in the country.

Keywords - computing job adverts; job adverts in Botswana; content analysis

I. INTRODUCTION

Computing professionals have enjoyed better employment packages and job opportunities due to high demand for their knowledge and skills worldwide [1, 2]. As most organizations and businesses are becoming highly dependent on Information Technology, the demand for such professionals is expected to increase. To satisfy this demand, the knowledge and skills possessed by graduates with computing degrees should match with the types and variety of skills needed by employers. One recent trend observed is that employers are continually demanding a mix of skills in addition to technical skills from computing graduates [1, 3]. Academic institutions need to scrutinize this trend so that they can align their programmes and curricula with industry needs. In Botswana, employers' needs and expectations from computing graduates have not been properly studied. As the skills demand of employers may change from time to time, studies should be carried out regularly so that the patterns of growth for the future can be identified.

This research was motivated by the need to better understand the current and future employers' needs for computing graduates in Botswana. As the only public University in the country, a lot is expected from this institution to produce workforce that can meet the demands of employers. Realizing this expectation, the Department of Computer Science planned to diversify its programmes both at the undergraduate and graduate levels. To do so, the department needs to understand the competencies that are needed and will be needed by the job market. Currently, the department offers undergraduate degree programmes in Computer Science and in Information Systems.

In the past curricula revision used to be carried out in the department based on the perceptions of academicians of what might be needed by the employers. However, this trend is no longer practical because there is a gap in the perceptions of employers and academicians about the knowledge, skills and competencies required of computing graduates [4, 5]. Different studies have also identified that employers are demanding a variety of new skills which were not found in the mainstream computing studies.

In the context of Botswana, there is also a perception that, recently, graduates with computing degrees are having difficulties in getting employment. This is in contrary to the global trend where the demand for computing professionals is still high. For example, according to the United States Bureau of Labour and Statistics[6], occupations in the area of network systems and communications, application development and systems software development will be on the rise in the years to 2012. The situation in Botswana calls for further investigation to get an understanding of where the problem lies. It could be that graduates are not well prepared in terms of the required knowledge and skills for employment or that the industry is not well developed to avail employment opportunities for entry-level graduates. In our search for past studies focusing on Botswana, we were unable to find a study describing the knowledge and skills in demand by employers for computing graduates. Therefore, this study aims to investigate the knowledge and skills required for computing graduates. Specifically, this study focuses on the identification of knowledge and skills demanded for early career/entry-level positions.

The objectives of the study are as follows:

1. Identify the types of degree qualifications that are in demand

2. Identify the types of knowledge and skills required for entry-level positions
3. Assess to what extent industry certifications are required for entry-level positions

In order to achieve our objectives, we collected job adverts from the 7 major newspapers (published both daily and weekly) in the country for the period January 2008 to December 2008. Job adverts are widely used as a means of expressing employers' needs regarding skills and competencies of prospective employees. A number of studies [3-5, 7-11] have used job adverts as a source of data for analysing skills in demand in different disciplines. Therefore, we assume that employers are able to express correctly the types of skills they want from their prospective employees when they advertise in newspapers.

Examples of job adverts included were those that require a bachelor degree in Computer Science, Information Systems, Information Technology or related discipline with less than or equal to two years of experience requirements. We consider that such job adverts describe early career positions. Kennan et al. [5] defined that "early career or entry level positions are those positions requiring new graduates and those with up to three years work experience". In the context of Botswana, most employers advertise jobs in newspapers and hence the adverts collected from the major newspapers should be a sufficient source to make any conclusions. Online job adverts are not common in the country as Internet is not yet the preferred medium for advertisement of jobs.

The rest of the paper is organized as follows: Section 2 describes the research method used in this research. The analysis and findings of the survey are presented in Sections 3 and 4 respectively. Section 4 includes also a discussion of the main issues resulted in from the data analysis. Finally, Section 5 presents the main points of the survey result.

II. RESEARCH METHOD

One major activity during curriculum design is the identification of knowledge and skills that are in demand and those job categories and skills that are anticipated to grow in the future. This identification can be achieved using surveys, interviews, content analysis of job adverts, etc. or any combinations of these techniques [3].

The content analysis method was chosen for the job advert analysis from major newspapers in the country. According to Krippendorff [12], content analysis is a "research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use". As a scientific tool, content analysis provides new insights and increases a researcher's understanding of particular phenomena, or informs practical actions. Content analysis has been applied in different fields [13, 14]. Different studies used content analysis for analysing job adverts [8, 11] and indicated that content analysis helped to identify skills in demand at present and identify shifts in the patterns of skill demands over time.

A content analysis of job adverts for software developers by Surakka [9] discovered that the requirement for technical skills more than doubled in the years 1990 to 2004. Gallivan et. al [4] conducted a content analysis study of job adverts in printed media from 1988 to 2003 and discovered that employers were increasingly requiring a variety of skills from new IT employees. Their study also revealed that there was a recruitment gap in that even though employers wanted all-rounded employees, their adverts demanded more IT technical skills. Other similar studies using content analysis were conducted to identify skills demand of employers in different countries [4, 5, 8, 10, 14].

For our study, job adverts were collected from 7 major newspapers (Mmegi, Daily News, Gazette, Guardian, Midweek Sun, Sundays Standard, and Voice) for the period January 2008 to December 2008. Two of the newspapers are published daily while the other five are published weekly. These newspapers are circulated country-wide. As online job adverts are not popular yet in the country, we believe that the job adverts in these Newspapers have a fair representation of job adverts in the country.

In order to carry out the data collection, 6 research assistants (i.e., demonstrators and teaching assistants in the department with at least a bachelor degree in Computer Science or Information Systems) were employed. The research assistants were given a template that contains information items that need to be recorded by scanning the newspapers they were assigned (sample template is attached as an appendix). The Newspapers were available at the University Library to the research assistants for the indicated period. Since we are planning to use the result of this survey for curricula revision of the existing undergraduate programmes (Computer Science and Information Systems) and introduce new undergraduate programmes in other Computing disciplines, the research assistants were asked to record any job advert requiring a degree in one or more of the computing disciplines (as defined in [15]).

III. ANALYSIS

In our analysis, we excluded some of the adverts for various reasons. The first group of adverts excluded were those requiring a qualification different from Bachelors degree. Since our main focus is to get an understanding of the skills requirement for early career employees in Computing (B.Sc. degree), we removed those requiring other than B.Sc. degree. We also excluded qualifications such as Degree/Diploma or MSc/BSc for an advert as these types of mixed-qualification adverts are not specifically targeting B.Sc. degree graduates. Finally, we excluded those adverts which were redundant; meaning those adverts appearing in multiple newspapers. This is common since the same advert can be listed in different newspapers. In addition, the same advert can appear in different issues of the same newspaper.

The criterion we used to remove redundant adverts is using vacancy number. Unfortunately, not all adverts have vacancy number. Therefore, to minimize the redundancy, we scanned

those adverts manually. Using all these elimination criteria, we reduced the number of job adverts from 494 (initially collected) to 131. The following table shows the number of adverts for each Newspaper.

TABLE I. ADVERTS PER NEWSPAPER

Newspaper	Adverts	Positions	Diploma	BSc(<= 2years)	BSc(> 2years)	MSc	PhD	Other*
Daily News	78	341	15	37	14	4	0	8
Sunday Standard	62	76	9	6	44	3	0	0
Gazette	58	81	6	15	30	7	0	0
Voice	18	249	3	6	8	1	0	0
Midweek Sun	36	44	0	12	22	0	0	2
Guardian	55	73	8	8	33	4	0	2
Mmegi	187	215	28	47	90	11	5	6
Total	494	1079	69	131	241	30	5	18

* refers to those that do not fall into one of the specified categories.

A. Job Categories

Since the department is, currently, running B.Sc. degree programmes in Computer Science and Information Systems and has a plan to launch new programmes in other computing disciplines, we opted to use generic computing job categories that can describe the knowledge and skills required of computing graduates. Gullivan et al. [3] developed computing skill categories for IT professionals in their study of changing demand patterns for IT professionals. In order to come up with those job categories, they used Computerworld job/skill classifications scheme, literature and evolved it through Delphi refinement approach using Faculty members in their institution. The job categories they introduced helped them to assess the level of demand of each job category. They identified the increase in demand for software engineers.

Litecky et al. [8] also developed job categories using web content mining techniques to identify job categories and the associated skills needs prevalent in the computing profession in the USA. They extracted job adverts of about 20 months from three online sources for jobs requiring a degree in Computer Science, Information Systems, and other computing programs. They analyzed the data collected using cluster analysis to come up with clusters of job categories and identified 19 job titles and 24 most frequently mentioned skills in computing job adverts. The job definitions and the associated skills required for these titles are provided in [8]. These job definitions were finally categorized into five larger job classifications to minimize the similarities among the job definitions. The classifications include Web developers, software developers, database developers, managers, and analysts. Nunamaker et al. [16] also provided job classification scheme used in the 1980s which provided the categories: programmers (software developers), technical specialists, business systems analysts, end-user support, and computer operators & data-entry clerks.

Another study by Liu et al. [10] provided five broad skill categories: Programming languages, Web development, Database, Networking, and Operating systems & Environments. Each category is further divided into different types of skills. For example, under programming languages category, skills such as C++, Java, VB, etc were listed. They reported the level of demand of each skill in each of the five categories. For example, in the programming languages category, C++/Visual C++ skill is in high demand with 38.64%.

From the above classification schemes, we can see that some are based on job titles [3, 8, 16] and others are based on skill categories [5, 10]. However, both approaches list the associated skills to the job titles or skill categories. In addition, we observed that some focus only on technical skills while others include other non-technical skills such as project management, business strategy, etc. As many recent studies [5, 8] indicated, the trend is that most employers demand a mix of technical and non-technical skills from computing graduates. Another difference among the different schemes is the level of detail in skills categorizations. For example, the job category developed by Litecky et al. [8] indicated how specialized the skill requirements are in the USA. Such categorization may not be appropriate in developing countries where the industry is not well developed. On the other hand, classification schemes like the one provided by Nunamaker et al. [16] indicate very generic classifications which were appropriate in the past but may not be applicable at present.

In Botswana, we were unable to find standardized job titles and their associated skills that can be used by most employers. In our preliminary analysis, we observed that different job titles are used for the same positions. For this study, using the various literature resources [8, 17, 18] and our own experience, we adopted a classification scheme as given below in Table 2.

TABLE II. **JOB TITLES AND ASSOCIATED SKILLS**

Job title	Job description	Major skills required
Systems analyst	researches problems, plans solutions, recommends software and systems, and coordinates development to meet business or other requirements	Programming, operating systems, hardware platforms
Education	covers a range of areas, including tertiary and secondary teaching; vendor training; corporate trainers and trainers / managers of specialist training organizations.	Not specific
User/technical support	Provide support regarding application packages, computer systems, networks to end users	Network, packages, operating systems, installation, upgrading
Systems administrator (systems manager)	Administration of end-user systems and workstations as well as networking and telecommunications	Operating systems (Windows, Linux, Unix), security, certification, networking
Network administrator	Responsible for the maintenance of computer hardware and software that comprises a computer network. Includes the deployment, configuration, maintenance and monitoring of networks equipment. It can also be called Network specialist, network designer, network support, or LAN administrator	Operating systems, security, protocols (TCP/IP), Cisco
Database administrator	Works with the administrative component of databases. The role includes developing and designing the database strategy, monitoring and improving database performance and capacity, and planning for future expansion requirements.	DBMS (Oracle/MS SQL Server/mysql, etc.), SQL, Security, certifications
IT manager	plan, administer and review the acquisition, development, maintenance and use of computer and telecommunication systems within an organization	Sound technical experience, leadership, strategy, finance, accounting, knowledge of administrative procedures such as budgeting, quality assurance and human resources
Web developer	Web application development using a variety of programming languages and tools	HTML, XML, JavaScript, AJAX, Java, ASP, SQL, PHP
Software developer	Involved in the specification, designing, and implementation of a software system and work with different languages and associated tools	C/C++, C#, .NET, Java, OOP, software development
Database developer	Working with SQL, and different DBMS, programming, and systems analysis	SQL, DBMSs, programming
Project management	involves the selection, approval and	Estimation, scheduling,

	initiation, planning, implementation and tracking, reporting and review of a project in the context of IT projects	controlling, resources management
Other (consultant, graphic designer, etc)	Different non-common job titles	

IV. FINDINGS & DISCUSSION

An analysis of our findings regarding the demand for the computing degrees, job titles, specific computing skills and industry certifications is provided below.

A. Education requirements

Under education requirement, we assessed the level of demand for the different B.Sc. degree (computing)

qualifications in the country. Since we are interested in to assess the types of B.Sc. degree qualifications that are in demand for entry level positions and their associated skills, we focus on those adverts requiring less than or equal to 2 (≤ 2) years of experience. The findings regarding the demand for the different degree qualifications are presented numerically in Table 3 and graphically in Figure 1.

TABLE III. DEMAND FOR COMPUTING DEGREES

Newspaper	Computer Science	Information Systems	Software Engineering	Computer Engineering	Computing (Any)	Total	Percent
Daily News	27	0	0	0	10	37	28.2%
Mmegi	28	3	0	1	15	47	35.9%
Sunday Standard	3	0	0	0	3	6	4.6%
Gazette	2	0	0	0	13	15	11.5%
Voice	1	1	0	0	4	6	4.6%
Midweek Sun	3	0	0	0	9	12	9.2%
Guardian	1	1	0	0	6	8	6.1%
Total	65	5	0	1	60	131	
Percent	49.6%	3.8%	0.0%	0.8%	45.8%		

In terms of the qualification type classification, we could have added Information Technology as this qualification is now a well established qualification according to the recent ACM/IEEE curricula guidelines. However, its use in the industry is still confused between its meaning as a Computing discipline and its generic meaning (i.e., referring to the IT industry). Therefore, we were not sure which meaning the advert is referring to when the education requirement is specified as "Degree in IT". Hence we classified such adverts under "Computing (any)". In addition, under education requirement, some adverts specify a list of alternative fields such as "Degree in Computer Science/ Information Systems" etc. We classified such requirements as "Computing (any)" since they did not indicate preference to a particular Computing field of study.

The result in Table 3 (or Figure 1) above shows that Computer Science degree is still in high demand (49.6%) for early career positions while Information Systems has a low demand (3.8%). We can also see that the demand for computing (any) degree is also high (45.8%). This could be an indicator that either employers do not care about the specific computing degree type as long as prospective employees are equipped with the desired skills or the tasks for which employers plan to recruit are so generic that anyone with a computing degree should be able to accomplish. Another interesting result is that there is no demand for software engineering degree in contrary to worldwide trend where the demand for software engineers is growing especially in developed countries. This could also be an indicator that there

is no much software development locally or a degree in software engineering is still not known to employers.

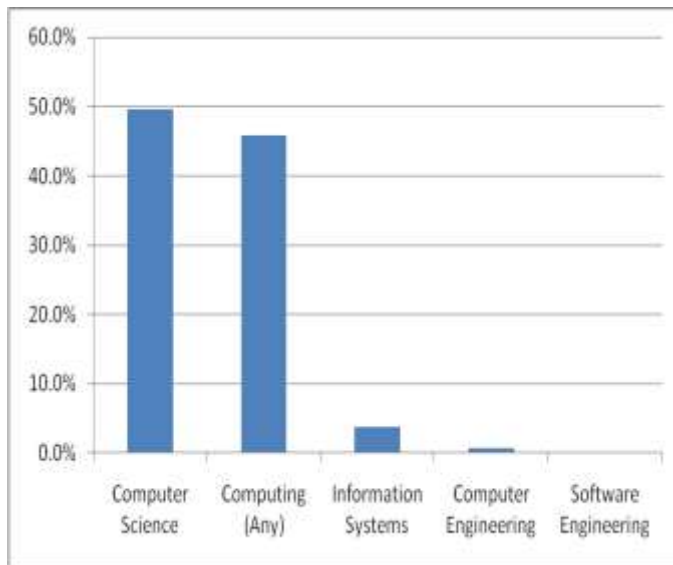


Figure 1. Demand for computing degrees

It is interesting to see that Computer Science degree has still a high demand compared to the other computing degrees. This information is particularly important for prospective students so that they can make their decisions regarding the choice of field of study based on job prospects. However, it is doubtful whether the demand for Computer Science degree

represents the actual demand for Computer Science graduates. In our observation, the job descriptions of those adverts that required Computer Science degree are not typical descriptions requiring a computer science degree. Most of those job descriptions looked appropriate for anyone who has a computing degree. It could be that employers are more familiar with the computer science degree than the other computing degrees because it has been offered by the University since 1992 while Information Systems degree started to be offered in 2002.

B. Job titles

Even though the classification of the job titles is based on the description given in Table 2 (Section 4.1), we realized that it was difficult to come up with the same classification by different people as some of the job titles appear to fit into different job titles in the classification provided. For example, by looking at the job description of a particular advert, some one might classify it as a software developer position while others might classify it as database developer position when the description equally emphasizes on the database and software development skills. To minimize such misclassifications, the authors independently classified the job adverts and then discussed together to reach a consensus as to the best classification.

The findings regarding the demand for the different job titles are presented numerically in Table 4 and graphically in Figure 4.

TABLE IV. DEMAND FOR JOB TITLES

Job Title	Daily News	Sunday Standard	Gazette & Voice	Midweek & Guardian	Mmegi	totals(by job title)	percent
Systems Analyst	8	1	6	2	5	22	16.8
Education	9	0	0	2	9	20	15.27
software developer	2	2	3	3	6	16	12.22
other(consultant, graphic designer)	7	0	0	1	5	13	9.93
network administrator	1	1	0	3	7	12	9.17
database administrator	2	1	3	4	2	12	9.17
User/technical support	2	1	5	1	2	11	8.4
systems administrator	3	0	1	1	3	8	6.11
IT manager	0	0	1	1	6	8	6.11
database developer	3	0	2	2	0	7	5.35
web developer	0	0	0	0	1	1	0.77
project management	0	0	0	0	1	1	0.77

totals (# of adverts by:)	37	6	21	20	47	131	100
----------------------------	----	---	----	----	----	-----	-----

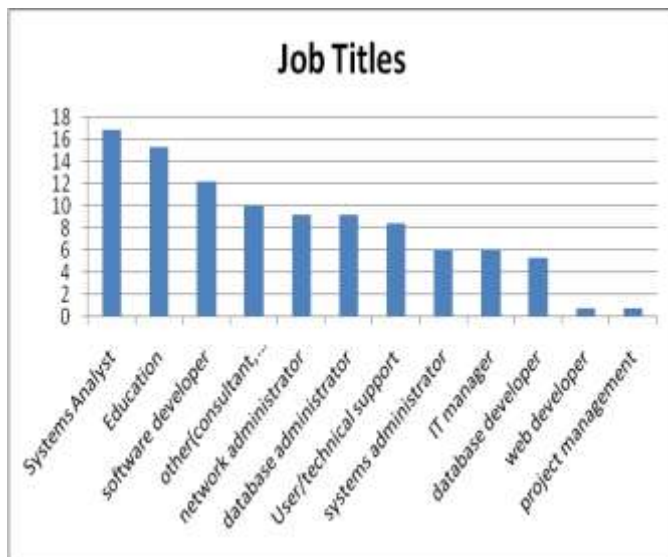


Figure 2. Demand for computing degrees

The highest job title in demand is Systems Analyst (16.8%) followed by Education (15.27%) and Software Developer (12.22%). The two findings that are in contrary to world wide trends are the demand for Education (15.27%) and Web developer (0.77%).

It is interesting to observe that the job title with the highest demand is systems analyst which is not a typical job title for computer science graduates. This title is usually more suitable for Information Systems graduates. Therefore, it is likely that the problem lies in that employers are not able to properly differentiate the different computing disciplines when they specify educational requirements in their adverts. It is important to note that the high ranking of systems analyst job title is similar to findings in other studies [5, 19].

The demand for education is the second highest (15.27%) compared to the other job titles. This is probably a finding which is in contrary to other studies carried elsewhere. It could be an indicator of the job situation in a developing country context. The following two reasons might explain the situation: 1) that many private higher learning institutions are being established in the country and these institutions usually employ new graduates to fill their positions as those graduates with M.Sc. and PhD are scarce; 2) a new subject called "Computer Studies" is offered in most government secondary schools and new graduates are usually employed to teach this subject.

With the growing importance of the web, the demand for experts in this area is continuously increasing. It is expected that Web technology will provide the foundation for most future software systems. For this reason some Universities actually started to design web-centric curriculum to satisfy the ever increasing demand [20]. However, our finding shows the

contrary where the demand for web developers is the least (0.77%). This could be an indicator that there is no much web development activity locally.

C. Skills

In order to analyze the various skills that are in demand, we adapted the skills categories provided in [10] as this scheme lists the most likely skill categories for entry-level Computing positions. This scheme basically aggregates the skills listed in Table 2 corresponding to the different job titles. The skill categories are Programming (with skill set of C/C++, C#, Java, .NET, VB, generic, others), Web Development (with skill set of Java/JavaScript, XML, PHP, ASP, HTML, generic, others), Database (with skill set of Oracle, MS SQL Server, MySQL, MS Access, generic, others), Operating Systems (Windows, Linux, UNIX, generic, others), and Networking (with skill set of network design, network management, network security, generic, others). In each of the skill categories, "generic" refers to generic skills without any preference to specific skills. For example, in the programming category, generic refers to generic programming skills without preference to a specific programming language. Similarly, in all the skill categories, "others" refers to specific skills specified different from the ones provided in each of the skill categories. For example, in the programming category, if one specifies skill in Perl, it is considered as "Others" as it is different from the specific skills listed in the programming category.

In the programming skill category, generic programming skill accounts for 83.3% of the programming skills specified followed by C/C++ (11.1%) and .NET (5.6%). In the web development category, generic web development skill constitutes 66.7% followed by HTML (13.3%) and Java/JavaScript (6.7%). In the database category, generic database skill constitutes 77.3% followed by Oracle (18.2%) and MS SQL Server (4.5%). In the operating systems category, generic operating systems skill constitutes 57.7% followed by Microsoft Windows (23.1%) and UNIX (19.2%). Finally, in the networking category, generic networking skill constitutes 77.4% followed by network management (19.4%) and security (3.2%).

In terms of the skill categories, more adverts specified skills in networking (27.7%) than the other categories. The other skill categories' demand is as follows: operating systems (23.2%), database (19.6%), programming (16.1%) and web development (13.4%).

A clear trend that can be observed from the adverts is that the majority of them specify generic skills as opposed to specific skills. Even though it is difficult to make conclusions why most of them prefer generic skills, this pattern might be an indicator that most jobs do not require specialized skills. This in turn could be an indicator that the industry is not yet mature enough to engage the services of specialised professionals. This finding might also indicate a correlation with types of computing degrees specified in adverts. We saw that many of the adverts did not set a preference to a particular type of

computing degree in their adverts. For example, many of the adverts specify, under education requirement, a list of alternatives such as degree in computer Science/Information Systems/... or related. This finding is also in contrary to other studies conducted especially in the developed countries where employers specify specific skills that are more relevant for their tasks. The actual reason needs further investigation. It could be that most job positions require generic skills or employers failed to include the specific skills desired in their adverts.

D. Industry certifications

Even though entry-level positions are expected to focus on the academic performance of graduates at their University education and their potential to succeed at work, we observed that some adverts actually require additional industry certifications in addition to academic degrees. Table 5 below presents the types of certifications that are required for entry level positions in the Botswana Computing job market.

TABLE V. DEMAND FOR INDUSTRY CERTIFICATIONS

Certification	Frequency	Percentage
MCSE	8	21.62%
A+	5	13.51%
K2 SAP	4	10.81%
OCP	3	8.11%
OCA	3	8.11%
ICDL	2	5.41%
MCP	2	5.41%
CCNA	2	5.41%
CCVP	1	2.70%
N+	1	2.70%
CISA	1	2.70%
ITIL	1	2.70%
ISEB	1	2.70%
IT SECURITY	1	2.70%
CCNP	1	2.70%
MCSA	1	2.70%
Total	37	

Table 5 shows that MCSE is required in 21.6% of the adverts followed by A+ (13.51%). Since Microsoft products are dominant in most organizations, the need for MCSE certification is understandable. However, the number of adverts requiring certifications are few compared to the total number of adverts for early career positions. An interesting finding is that the demand for skill in SAP is growing. This has also been observed in those adverts which required more than 2 years of experience.

The issue related to the demand for industry certifications is a new trend. Even though the demand for certifications for entry-level positions is low in this study, compared to the total number of adverts, it might grow in the future. From our experience, most companies in the country expect their new employees to be able to handle their tasks immediately after employment without much training investment on the incumbents. This is particularly true for small to medium-sized companies where training investment on new employees is not a priority. Such companies usually require certifications and product specific skills from their new employees. Most private companies fall in to this category of companies. Therefore, the demand for certifications for entry-level positions has a challenging implication for academic programmes as certifications and product specific skills are not the primary focus of University education. Nonetheless, it is important to find a balance to minimize the expectation gap between employers' demands and the knowledge and skills provided by academic programmes.

V. CONCLUSION

This study attempted to understand the types of computing knowledge and skills in demand in Botswana by analyzing job adverts from the 7 major newspapers in the country for the period January 2008 to December 2008. In particular, the study focused on the identification of the types of computing degrees, job titles, and skills in demand for entry-level positions. For this purpose, 131 (from the total 494 collected) job ads were extracted directly from the newspapers and analyzed.

The findings indicate that the demand for Computer Science degree is still high compared to the other computing degrees. It was also observed that a significant number of adverts specify a number of alternative degrees suggesting that the tasks for which the positions are advertised could be generic. In terms of job titles in demand, Systems Analyst is the highest ranking which is similar to the findings of other studies carried out elsewhere. The findings regarding those skills in demand show that generic skills have the highest ranking in all skills categories. From all the skills categories, web development skills are the least mentioned in the adverts which is in contrary to our expectation where web development is becoming common in other countries. It is also observed that some certifications are frequently required.

We believe that such studies should be conducted regularly by collecting data continuously so that skill demand patterns can be understood properly. This understanding can lead to informed curricula design that can prepare graduates equipped with the necessary skills for employment.

Even though the findings of this study are based only on one year advert data, the result can still be used as a baseline for further studies on the issue. We believe that, in the future, a comprehensive data covering a longer period of time (e.g., 5 – 10 years data) needs to be collected and analyzed in order to see the trend in the demand for the different computing skills. Once such studies are carried out, students can use the findings to select courses that focus on those skills which are in demand.

Academic institutions can use the findings so that those skills in demand can be taken into account during curriculum design.

REFERENCES

- [1] J. Liu, "Computing as an Evolving Discipline: 10 Observations," *IEEE Computer*, vol. 40, pp. 110 - 112, 2007.
- [2] K. L. Chin, "Issues and Challenges in Teaching SE, IT, IS and EC," in 19th Australian Conference on Software Engineering, 2008, pp. 604 - 610.
- [3] M. Gallivan, D. P. T. III, and L. Kvasny, "An Analysis of the Changing Demand Patterns for Information Technology Professionals," in SIGCPR, Kristiansand, Norway, 2002, pp. 1-13.
- [4] M. J. Gallivan, D. P. T. III, and L. Kvasny, "Changing Patterns in IT Skill Sets 1988-2003: A Content Analysis of Classified Advertising," *ACM SIGMIS Database*, vol. 35, pp. 64-87, 2004.
- [5] M. A. Kennan, D. Cecez-Kecmanovic, P. Willard, and C. S. Wilson, "IS Knowledge and Skills Sought by Employers: A Content Analysis of Australian Early Career Job Advertisements," *Australian Journal of Information Systems*, vol. 15, pp. 169-190, 2009.
- [6] U. B. o. L. a. Statistics, "Occupational Outlook Handbook," 2010 -11 ed. vol. 2010, U. S. D. o. L. Bureau of Labor Statistics, Ed., 2010.
- [7] Y. Choi and E. Rasmussen, "What Qualifications and Skills are Important for Digital Librarian Positions in Academic Libraries: A Job Advertisement Analysis," *The Journal of Academic Librarianship*, vol. 35, pp. 457-467, 2009.
- [8] C. Litecky, A. Aken, A. Ahmed, and H. J. Nelson, "Mining for Computing Jobs," *IEEE Software*, vol. 27, pp. 78-85, 2010.
- [9] S. Surakka, "Trend Analysis of Job Advertisements: What technical skills do software developers need?," Helsinki University of technology 2005.
- [10] L. C. Liu, K. S. Koong, and L. Rydl, "A Study of Information Technology Job Skills," in 37th Annual Southwest Region Decision Sciences Institute (SWDSI) Conference, Oklahoma City, 2006, pp. 419-426.
- [11] P. A. Todd, J. D. McKeen, and R. Brent, "The Evolution of IS Job Skills: a Content Analysis of IS Job Advertisement from 1970 to 1990," *MIS Quarterly*, vol. 19, pp. 1-27, 1995.
- [12] K. Krippendorff, *Content Analysis: An Introduction to its Methodology*, 2nd ed.: Sage Publications, Inc., 2004.
- [13] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*: SAGE Publications, 1980.
- [14] E. Ettinger, C. Wilderom, and H. Ruel, "Service-Quality Criteria of Web Recruiters: A Content Analysis," in 42nd Hawaii International Conference on Systems Sciences, 2009, pp. 1-10.
- [15] CC2005, "Computing Curricula 2005," A Volume of the Computing Curricula Series, pp. 1-62, 2005.
- [16] J. F. Nunamaker, J. D. Couger, and G. B. Davis, "Information Systems Curriculum Recommendations for the 1980s: Undergraduate and Graduate Programs - A Report of the ACM Curriculum Committee on Information Systems," *Communications of the ACM*, vol. 25, pp. 781-805, 1982.
- [17] ACS, "Australian Computer Society Job Descriptions." vol. 2010, 2010.
- [18] BLS, "Occupational Outlook Handbook," 2010 -11 ed. vol. 2010, U. S. D. o. L. Bureau of Labor Statistics, Ed., 2010.
- [19] C. K. Lee and H. J. Han, "Analysis of Skills Requirement for Entry-Level Programmer/Analysts in Fortune 500 Corporations," *Journal of Information Systems Education*, vol. 19, pp. 17 - 27, 2008.
- [20] J. Gorgone and V. Kanabar, "Masters in Information Systems: A Web-Centric Model Curriculum," in *Informing Science + IT Education*, 2002, pp. 553 - 563.

AUTHORS PROFILE



Yirsaw Ayalew is a Senior Lecturer at the Department of Computer Science, University of Botswana. He holds a PhD in computer science from the University of Klagenfurt, Austria and teaches both undergraduate and graduate courses in software engineering and database systems. His research interests are mainly in the area of software engineering; particularly software quality, requirements engineering, software engineering for healthcare systems and end-user software engineering.



Mr. Mbero is a Lecturer in the Department of Computer Science, University of Botswana, Botswana. He currently holds a Bachelor of Science degree, Kenyatta University, Kenya, Master of Science

degree (information systems) from University of Nairobi, Kenya. He is registered for a PhD in Computer Science at the University of Western Cape, Cape Town, South Africa. He is professionally qualified CCNA, CCAI, and CCNP in networks. He is also a fellow of Computer Society of Botswana. His research interests include: Mobile Ad Hoc networks, Wireless Sensor Networks, Routing in Networks, Social Networks, Neural Networks and Artificial intelligence, E-learning in networks.



Mr Tallman Z. Nkgau is a lecturer in the Department of Computer Science at the University of Botswana. He holds an H.BSc(Computer Science) from Lakehead University and an MSc(Computer Science) from McGill University. He is also a certified Cisco Academy Instructor and is CCNP certified. His research interests are in network security, algorithms and data structures, combinatorial optimization and effective use of ICT for development



Nkwebi P. Motlogelwa works as a Lecturer in the Department of Computer Science, University of Botswana. His research interests are in high performance computing for socio-economic development. This research aims to utilize prevalent low-cost clusters for socio-economic development. In addition, Motlogelwa has been involved for the past two years in a Microsoft funded project investigating how wireless and mobile technologies can enable under-served communities to benefit from opportunities afforded by the power of information and communication technologies to achieve development goals. His main focus is on how mobile phones can be used in healthcare information dissemination, particularly HIV/AIDS information dissemination to illiterate and semi-illiterate people in rural areas in Botswana.



Dr. Audrey Masizana-Katongo is a lecturer at University of Botswana, Computer Science Department. She holds a PhD in computer science from the UMIST in the UK, and teaches both undergraduate and graduate courses in areas with applications of Decision Support Systems and Expert Systems and Web Engineering. Her research interests are mainly in the area of Decision and Intelligent systems; particularly application of mathematical and computing techniques in decision problems.

Open Source Software in Computer Science and IT Higher Education: A Case Study

Dan R. Lipşa

Visual and Interactive Computing Group
Department of Computer Science, Swansea Univ.
Swansea, UK
d.lipsa@swansea.ac.uk

Robert S. Laramee

Visual and Interactive Computing Group
Department of Computer Science, Swansea Univ.
Swansea, UK
r.s.laramee@swansea.ac.uk

Abstract— The importance and popularity of open source software has increased rapidly over the last 20 years. This is due to a variety of advantages open source software has to offer and also the wide availability of the Internet in the early nineties. We identify and describe important open source software characteristics and then present a case study using open source software to teach three Computer Science and IT courses for one academic year. We compare fulfilling our educational requirements and goals with open source software and with proprietary software. We present some of the advantages of using Open Source Software (OSS). Finally we report on our experiences of using open source software in the classroom and describe the benefits and drawbacks of using this type of software over common proprietary software from both a financial and educational point of view.

Keywords—open source software (OSS), free software

I. INTRODUCTION

Open source software (OSS) has become widely used in IT departments, with large software vendors making a significant amount of revenue from activities that use OSS [1]. The emergence of the Internet in the early nineties has enabled collaboration between programmers at different locations in the world and easy distribution of software. That together with distinct advantages OSS offers has resulted in an increasing popularity of this type of software.

We briefly introduce of open source software, and describe its main proponents. We describe the main OSS licenses and explain how some licenses protect users' freedom and the ability to use OSS in the future. We describe the impact open source software has on the computer industry. We believe this knowledge is important for fully appreciating the value offered by open source software.

We present a case study in using open source software in teaching three Computer Science and IT classes for one academic year. We compare satisfying our educational requirements with open source software and with proprietary programs. We describe open source software used for infrastructure, user applications and development applications and compare it with proprietary software that achieves the same goals. We evaluate the two categories of software for

cost, student appeal and ease of use and we conclude with the main reasons we believe open source software should be more broadly integrated in Computer Science and IT education.

We believe our study presents a balanced comparison between open source and commercial products relevant to an educational environment. We contribute to a better awareness of the relative benefits and drawbacks of open source software versus commercial software and we help educators make informed decisions regarding the software used in their classrooms and in the infrastructure that supports classroom activities.

II. OPEN SOURCE SOFTWARE BACKGROUND

Open source software has a rich history with great achievements, spectacular falls, powerful players and colorful stories. We believe knowledge of the history of open source software is useful to understanding the software business and the computer industry as a whole. We present a brief introduction to open source software describe its achievements and introduce its main proponents. We describe common open source licenses and present the impact open source software has on the computer industry.

A. History of Open Source Software

When discussing open source software, two prominent figures stand out as what we call the creator and the enabler of today's events in this area.

Richard Stallman can be rightfully considered the father of Open Source Software (or Free Software as he calls it). He is the founder of the Free Software Foundation (FSF) a tax-exempt charity that raises funds for work on the GNU (Gnu's Not Unix) Project [4]. The GNU project started in 1983 with an email to a Unix newsgroup, in which Richard Stallman, states that he is going to write a complete Unix-compatible software system and share it with everybody. He asks for contributions of time, money, programs and equipment. With the help of thousands of programmers from around the world, and with the arrival of Linux, an operating system kernel, Richard Stallman succeeded in doing just that. He is the initial developer for many popular Open Source projects such as GNU C Compiler, GNU Emacs, GNU debugger, and GNU Make and FSF

developed Bourne Again Shell (bash) and GNU C library. The name GNU, comes from a recursive acronym for “Gnu’s Not Unix”, which was designed to show its relationship with the Unix operating system. The GNU operating system is a Unix compatible system but in the same time it was written from scratch so it is different than the proprietary Unix systems.

A more recognizable name than Stallman’s is Linus Torvalds and the Linux operating system kernel. By contributing the original version of Linux to the Open Source pool, Torvalds added the last piece missing from completing Stallman’s vision: a free Unix like system, and so, he enabled the widespread of the GNU/Linux systems as we see it today. Linux was initially created by Linus Torvalds when he was a student at University of Helsinki in Finland. The first version 0.01 was released in September 1991, and version 1.0 was release in March 1994 [9].

Open Source Software is a term promoted by the Open Source Initiative (OSI) [20], a non-profit organization launched in 1998. In their own words, the movement, is a marketing campaign to turn around the negative image Free Software had outside the hacker community. They argue for Free Software on pragmatic grounds of reliability, cost and business risks. They claim that development of Open Source Software happens at an astonishing pace compared with the conventional software [22]. This is a consequence of the fact that source code is available to everyone and can be changed, so anyone can find and fix bugs, and add their own improvements to the source code. This rapid evolutionary process produces better software than the traditional closed model. For this reason, and because of the lower development costs it makes business sense to choose open source software, and contribute to its development.

The official definition of Open Source Software is very close to how FSF defines Free Software. Still the two movements differ in the reason they argue why people should adopt Open Source/Free Software. For FSF, the reason is that people want and deserve freedom, as defined in Section II-B1. For OSI the motivation is that software produced in an open source environment is technically superior.

From now on, we will use the term Open Source Software because it appears to be much more popular than Free Software in the general press.

B. Open Source Licenses

This section describes the various license agreements under which open source software is available.

1) Free Software

Richard Stallman sees software as information, and he believes everyone should have the freedom to use it and to learn from it. In particular, for a program to be Free Software, everyone should be able to run it for any purpose, to study how the application works, adapt it to their needs, redistribute copies so that the user may assist others, and improve the program and release the improvements, so that the whole community benefits. A common misconception that FSF tries

to clarify, is that Free Software means no money for your work. Free refers to freedom, as in “free speech” not as in “free lunch”.

2) Copyleft

Copyleft is the use of a license to protect the rights of free software (as defined in Section II-B1) such that remains free software.

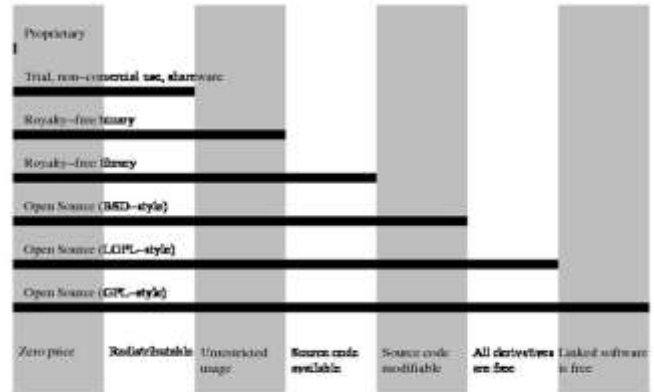


Figure 1: Software Licenses Classification. On the X axis we show possible license features. On the Y axis we show possible types of software.

X Windows is a good example of what happens when free software is not protected by copyleft. X Windows is a windowing system for Unix, developed at MIT, which was released as free software with a permissive license (without copyleft). It was adopted by many software companies, which shipped their improved versions of X Windows without the source code. In those releases, X Windows was no longer Free Software. The users lost the freedom they had for the initial release of X Windows. Copyleft was introduced to prevent this.

Copyleft uses copyright law, but flips it over to serve the opposite of its usual purpose. Instead of keeping the software proprietary, it becomes a mean to keep the software Free Software. Copyleft, gives everyone permission to run, copy, modify a program, and distribute modified versions, -- but not permission to add restrictions of their own. The term copyleft comes from a letter sent to Richard Stallman by Don Hopkins, in which the following phrase appears: “Copyleft – all rights reversed” [23].

3) Software Licenses Classification

An extended classification of software licenses, proprietary and Open Source, adapted from [6], is presented in Figure 1.

Proprietary software does not have any of the seven properties listed at the bottom of Figure 1.

Trial software, non-commercial software and shareware, do not cost anything and they are redistributable but they all have restricted usage. For trial software the time it may be used is

restricted or the features available limited. Non-commercial software cannot be used for any purpose, and shareware has an unenforced limited time usage (for instance WinZip [27]).

A royalty-free binary allows unrestricted usage and a royalty-free library is usually distributed with the source code.

Open Source (BSD-style, where BSD stands for Berkeley Software Distribution) license allows you to modify the source of the program and redistribute the improved version. This is the non-copyleft open source software distributed before the apparition of the Free Software Foundation. Some examples of projects distributed under this kind of license are the X Windows windowing system [29], the FreeBSD operating system [5] and the Apache web server [1].

The software protected by the last two licenses is copylefted, so it is guaranteed to remain Free Software. GPL stands for General Public License and LGPL stands for Library (Lesser) General Public License. Both were created by the Free Software Foundation. The difference between the two licenses is that only a library protected by GPL requires that all programs that link with it, should be GPL programs as well. LGPL protected libraries allow proprietary programs to link with it as well. Most of libraries on GNU/Linux system are protected by LGPL, or less strict licenses, which means that the user may release proprietary programs on GNU/Linux, and link with the libraries available. Many companies have done so (see [25]). Linux, GNU Compiler Collection and Emacs are example of programs protected by GPL.

C. Impact of Open Source Software

International Data Corporation (IDC) forecasts that revenues from open source software will grow at a 22.4% rate to reach \$8.1 billion by 2013 [11].

In June 2000, Netcraft Web Server Survey~\cite{netcraft} found that GNU/Linux runs on about 29.9% of the active websites, Microsoft OS runs on about 28.32%, and Solaris is third with 16.33%. Companies like IBM, Oracle and Intel fully support GNU/Linux systems.

Apache is a powerful, full-featured and efficient open source web server. Apache is also the most popular web server on the Internet The July 2009 Netcraft Web Server Survey [17] found that over 66% of the million busiest sites on the Internet are using Apache, thus making it more widely used than all other web servers combined. Apache Web Server is based on National Center for Supercomputing Applications (NCSA), University of Illinois, Urbana-Champaign public domain HTTP daemon, and the first release was in 1995. It is released under a simple, non-copyleft open source software license. Examples of sites which run on Apache are: Apple (<http://www.apple.com>), Financial Times (<http://www.ft.com>), Sony (<http://www.sony.com>), Palm (<http://www.palm.com>), Cnet (<http://www.cnet.com>) and Amazon (<http://www.amazon.com>).

With success comes competition. The company that has the most to lose from a wide acceptance of GNU/Linux systems is Microsoft. Their monopoly on the Operating System market is threatened. So, they have increased the propaganda against GPL and GNU/Linux.

Windows operating-system chief Jim Allchin has declared in 2001 that Open Source (GPL-style) will result in "the demise of both intellectual property rights and the incentive to spend on research and development" [15]. The Initiative For Software Choice organization [24] was created to fight against governments that mandate use of Open Source in government agencies and against licensing publicly funded projects with GPL.

On the other hand many companies have found that GNU/Linux fits well in their business plans. Linux is certified on all IBM Systems [10]. Oracle is the first commercial database on Linux in 1998 and it invests significant resources in developing, optimizing and testing many open source technologies [10]. Intel works on a wide variety of open source projects to enable a broad range of programs and environments to run best on their hardware [12].

A recent attack on GNU/Linux and GPL is the SCO Group (SCO stands for Santa Cruz Operation) lawsuit accusing IBM of adding copyrighted Unix code into Linux. SCO is asking for 1 billion dollars in damages, and credible speculations surfaced recently that Microsoft is financing SCO through a third party venture capital firm (see [7]).

An interesting use of GPL in promoting a proprietary product is that of QT library by Trolltech [21] which was later acquired by Nokia. QT provides a platform-independent interface to all central computer functionality: GUI, database access, networking, file handling, etc. The library became popular with its use in the KDE desktop, and is included in Suse, a German distribution of Linux which is currently owned by Novell. The Open Source community started a campaign against KDE (because of their use of proprietary QT library) and Red Hat didn't include KDE desktop in their distribution of Linux. In 2000, QT on Unix was release under dual-license GPL and proprietary, ending the quarrel with the Open Source community. By releasing their library under GPL, Trolltech continues to receive the free marketing from the use of the library in KDE. In the same time, they don't lose any business because GPL won't allow a proprietary program to link with QT. This is just one example of successfully combining open source and generating a profit.

III. OPEN SOURCE IN COMPUTER SCIENCE AND IT HIGHER EDUCATION: A CASE STUDY

We present the infrastructure, user and development applications used for one academic year in teaching three classes: Data Structures and Algorithms using Java, Rapid Java Application Development (an advanced Java class) and Design and Analysis of Algorithms. For these classes our goals were to:

- Present information about classes, present assessment methods and post student grades. Teach web applications development (web server recommended)
- Use a database to store students enrolled in classes and grades assigned to students and teach database access from Java (database server needed)
- Use both a desktop and a laptop (a method to synchronize between the two needed)
- Maintain the security of the two computers (a method to encrypt the communication and a firewall is

- Browse the Internet (browser needed)
- Read/write email (email client needed)
- Create documents containing mathematical formulas for the Algorithms class (word processor with support for mathematical formulas needed)
- Create presentations for classes (presentation program needed)
- Create diagrams for use in classes (drawing tool needed)
- Use an IDE for writing and debugging Java programs

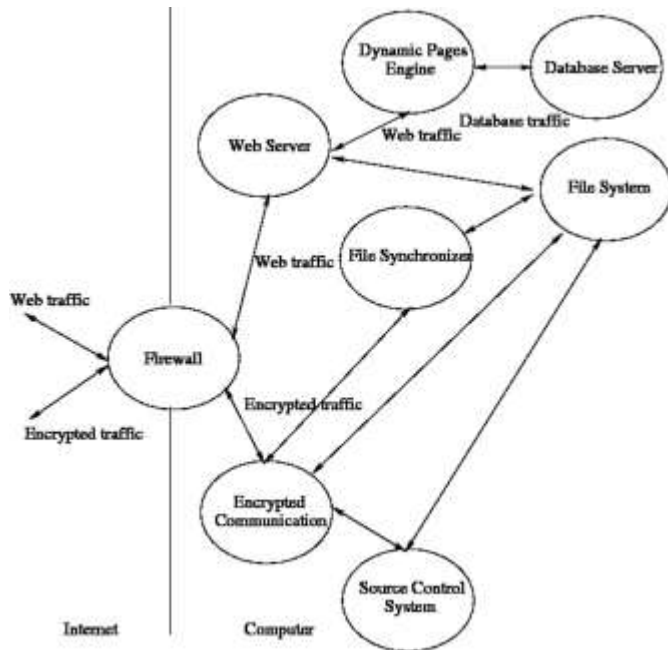


Figure 2: IT Infrastructure. We display course materials on a Web Server, and we use a Dynamic Page Engine and a Database Server to keep track of students grades. We use a File Synchronizer to maintain the same content on our desktop and laptop and we use a Source Control System to keep track of changes made to our classes. Our server is protected by a Firewall and all communication through the Internet is encrypted.

needed)

- Maintain history of changes to course files and web site (source control system needed)

A. Infrastructure

We used the server infrastructure described in Figure 2 for providing information about the educational institution and the classes taught, for using a database server to store grades for students and display them once the proper credentials were provided, for allowing students to submit homework through the website, and for allowing us to synchronize our work between the laptop and the server. A Firewall is protecting the server allowing only two types of communication: Web traffic for serving the website and Encrypted traffic for remote sessions on the server and file copying and synchronizing. A Web Server provides information about the professor and classes taught (static pages) and information about the grades assigned (dynamic pages). The static pages are read from the File System and the dynamic pages are built by programs run by a Dynamic Pages Engine which uses information stored in the Database Server. The Encrypted Communication server is used to encrypt any communication with the server. We can either synchronize files between the laptop and the server or access the Source Control System.

Figure 3 shows the Open Source implementation of the abstract infrastructure presented in Figure 2. We used the same setup on our laptop and our desktop. An identical setup on both computers enables the mobility of the teacher as they work either on their desktop or on their laptop. Almost all the applications used to perform the desired functions come standard in most GNU/Linux distributions (We used RedHat 9.0). The exception is Unison File Synchronizer a tool built at University of Pennsylvania~\cite{unison}. The GNU Head (mascot of the GNU Project) and the Penguin (the mascot of Linux) show the source of the components used in our setup.

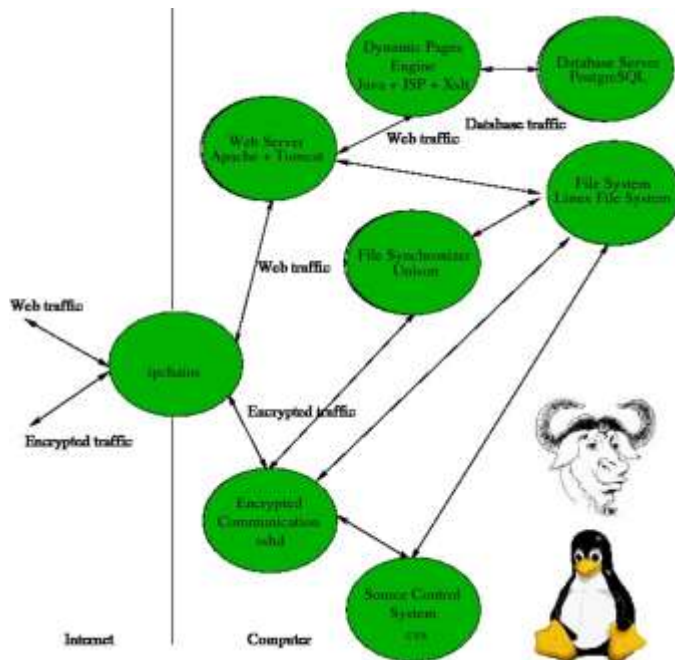


Figure 3: IT Infrastructure using Open Source Components. We use Apache as a web server, PostgreSQL as a database engine, Java, Java Server Pages (JSP) and XSL Transformation (XSLT) to generate dynamic web pages. We use Unison as a file synchronizer and cvs as a source control system. Communication is encrypted using ssh and we use ipchains as a firewall. All these components come standard on a Linux operating system.

We categorize the applications used in two groups: user applications and developer applications.

B. User Applications

- **Web Browser:** We used Mozilla which allows us to browse the Internet and to read newsgroup content.
- **Email:** We used Evolution [3] which allows email, address book and calendar management, and synchronizes with a Palm compatible device.
- **Word Processor:** We used LaTeX [13], a powerful formatting and typesetting processor with strong capabilities for writing mathematical formulas. For writing LaTeX files we used Emacs. Another option is Open Office Writer, which is especially useful for reading or writing Microsoft Word files.
- **Presentation:** We used Prosper [14], a LaTeX based presentation package with features such as: incremental display, overlays, transition effects, predefined slide styles. Another choice is Open Office Impress which is useful for reading Microsoft PowerPoint presentations.
- **Drawing Tool:** We used Xfig [28], a drawing tool for creating vector graphics.

C. Development Applications

The development tools used reflect the programming language taught (Java). However powerful tools exist for C/C++ development and many other languages.

- **Text Editor:** We used Emacs [2], a text editor offering intelligent text editing for Java, HTML, LaTeX, XSLT and many other languages.
- **Integrated Development Environment (IDE):** We used the NetBeans IDE [16] for editing and debugging Java, JSP, HTML, XSLT and other.

IV. OPEN SOURCE VERSUS PROPRIETARY SOFTWARE

This section compares open source software we used in our experiment with an equivalent setup using proprietary software. We did not compare individual features of open source software versus proprietary software. Our benchmark for listing an individual piece of software was to satisfy our educational objective. Both open source and proprietary software satisfied that criterion. We compare the open source versus proprietary software for cost, appeal to students and ease of use.

A. Cost

The Open Source programs used do not cost anything, so we calculate the cost of using common proprietary programs for an equivalent setup. This is presented in Table I. In parenthesis we present a package that contains the listed program. Microsoft uses a licensing model where Client Access Licenses (CALs) - which can be users or devices - are used to regulate access to their server programs. We used server programs with 5 CALs as they serve to make our point and they were the cheapest alternative. Wherever possible we applied an Education discount to the prices. We didn't used a volume discount as we were interested in the price that a student would get if he wants to install the given software on their own machine. We did not use the Express editions for certain pieces of software that Microsoft make available at no cost. While those versions can be used for education, they have reduced functionality that prevents their use in developing a business. In recent years, many companies including Microsoft began offering limited functionality of their products at no cost, we believe as a direct consequence of the strong competition open source products provide.

Using Open Source products may result in increased administration costs, but that cost is difficult to calculate and depends on individual circumstances. The extra administration cost may range from zero if the system administrator is familiar with the particular open source product to being prohibitive if significant training is required. Administration cost is influenced by the maturity of the open source product and also by the number of products that the administrator has to manage. If open source products are used alongside with commercial products the system administrator has more work to do just as a result of the number of software products she administers.

TABLE I. COST OF PROPRIETARY PROGRAMS USED TO ACHIEVE OUR EDUCATIONAL GOALS

Function	Application	Cost (USD)
Operating System	MS Windows XP Professional	300
Web Server	IIS (Windows Server 2003, Std.)	1000
Firewall	(Windows Server 2003, Std.)	0
Encrypted Communic.	(Windows Server 2003, Std.)	0
Database Server	MS SQL Server	1500
Source Control	(Visual Studio .NET 2003 Enterprise)	0
Web Browser	MS Internet Explorer	0
Email Client	Outlook (MS Office 2003)	0
Word Processor	Word (MS Office 2003)	150
Presentation Program	Powerpoint (MS Office 2003)	0
Drawing Tool	MS Visio Standard 2003	200
IDE	Visual Studio .NET 2003 Enterprise	1800
	Total	4950

B. Student Appeal

There are several reasons for which Open Source might appeal more to students than proprietary programs.

First many Open Source projects have their roots in academia. Great examples are X Windows which was an MIT project, the Unison File Synchronizer which was created at University of Pennsylvania, the BSD Unix which was created at University of California at Berkeley and Linux which started at the University of Helsinki.

Second, the availability of source code and documentation for the programs students work with, and the possibility for them to improve those programs could be very beneficial in attracting them to the IT field. In this respect the quote from [8] is revealing: "I'm a poorly skilled UNIX programmer but it was immediately obvious to me how to incrementally extend the DHCP client code (the feeling was exhilarating and addictive)."

Third, the costs detailed in the previous section would affect not only the professor and the school but the student as well. Students like to install the software used in school and work with it on their home computer. They may even want to start a business using the same software. When using open source software no additional costs are required. This is a big advantage for students and for promoting entrepreneurship.

Proprietary software appeals to students because they get direct experience with something they might use at their work place. While this might be beneficial, the computer industry is notorious for fast changes and for many competing products on the same market segment. It is impossible for a school to train students in all competing products, so we believe market share should not be the main criteria for selecting the software product to be used in class.

C. Ease of Use

Individual open source applications are comparable with proprietary applications when trying to achieve common tasks.

However, open source software operating systems are not as user friendly as their commercial software counterparts. This is the case mainly because of lack of hardware drivers from the hardware manufacturers. It is still difficult to use GNU/Linux on a laptop because of lack of wireless drivers and missing support for suspend and hibernate functionality. We see this as the major reason why we have not seen a widespread of open source software in the consumer market.

V. CONCLUSIONS: WHY OPEN SOURCE SOFTWARE?

We present our experience in using entirely open source tools for teaching and research, and we examine at why open source projects might appeal to students and professors.

Some advantages in using open source software in Computer Science and IT education that we have seen from our experience are:

- The cost for the university and the cost for students may be lower.
- Open source projects are advantageous for research as the user can get the source and is free to implement new ideas. They are great for teaching as students have the opportunity to make a difference in a project used by many people.
- Open source software allows the developer to port an application to another operating system. Proprietary software, is usually shipped on specific platforms. In our experience we used open source software on Linux as students used it on Windows. No problems were observed.
- In many cases, an open source project is the de facto standard for that particular type of application. So, the user is working with the best application possible. Some examples are Apache Web Server, Linux Operating System, sendmail Mail Server.
- Open source encourages entrepreneurship, as students can directly use open source tools in order to develop a business idea without the up-front costs of proprietary programs.

The main disadvantage in using open source software is the fact that Linux usability on laptops is seriously affected by the lack hardware drivers especially for wireless, graphic cards and suspend/sleep functionality in laptops.

Student feedback from this experiment was mixed, many students were excited to use Linux and open source tools some students thought that learning a proprietary tool will give them a better chance to get a job. A student who worked in an IT department commented that he is glad that we use and cover some Linux and open source software because he is using it at his job and he had to learn it all by himself. He thought we should cover open source software in other classes as well.

Adopting open source software for your courses is a challenge. Here are a few misconceptions and challenges that have to be overcome:

- Open source software is a niche market used only by a small group of hobbyists. In fact the opposite is true. There is wide adoption in the computer industry for open source software.
- There is no space on the lab computers to install this piece of open-source software. A decision at the department level to use a certain open source software instead of a proprietary product helps in this case.
- Proprietary software is better and students learn more by using better software. Some open source projects are leaders in their market segment (see Section II-C) and many got great reviews from publications in the field (see [18]). So it can be argued that there is no significant difference in what can be taught using open source software or proprietary software.

We believe that both open source software and proprietary software have an important role to play in the computer industry of the future. While we do not advocate only using open source software for education we believe exposure to open source software is essential for student's success. As future work we plan to develop questionnaires that evaluate specific commercial software products and their open source counter-parts. The evaluation criteria will be how well each product helps in reaching the educational objective of the course.

ACKNOWLEDGMENT

This research was partially funded by the Welsh Institute of Visual Computing (WIVC).

REFERENCES

- [1] The Apache Software Foundation. Online document. Accessed Aug. 31, 2009, <http://www.apache.org>.
- [2] GNU Emacs. Online document. Accessed Aug. 31, 2009, <http://www.gnu.org/software/emacs/>.
- [3] Evolution. Online document. Accessed Aug. 31, 2009, <http://projects.gnome.org/evolution/>.
- [4] Free Software Foundation. The GNU Project and the Free Software Foundation. Online document. Accessed Aug. 31, 2009, <http://www.gnu.org>.
- [5] Free BSD. Online document. Accessed Aug. 31, 2009, <http://www.freebsd.org/>.
- [6] Halloween Document I - Open Source Software: A (New?) Development Methodology. Online document. Accessed Aug. 31, 2009, Published Nov. 1, 1998, <http://www.catb.org/~esr/halloween/halloween1.html>
- [7] Halloween X: Follow The Money. Online document. Accessed Aug. 31, 2009, Published Mar. 3, 2004, <http://www.catb.org/~esr/halloween/halloween10.html>.
- [8] Halloween Document II - Linux OS Competitive Analysis: The next Java VM. Online document. Accessed Aug. 31, 2009, Published Nov 3, 1998 <http://catb.org/~esr/halloween/halloween2.html>.
- [9] Ragib Hasan. History of Linux. Online document. Accessed Aug. 31, 2009, Published Jul. 2002, <https://netfiles.uiuc.edu/rhasan/linux/>.
- [10] IBM and Linux. Online document. Accessed Aug. 31, 2009, <http://www.ibm.com/linux/>.
- [11] Open Source Software Market Accelerated by Economy and Increased Acceptance From Enterprise Buyers, IDC Finds. Online document. Accessed Sep. 03, 2009, Published Jul. 29, 2009, http://www.businesswire.com/portal/site/home/permalink/?ndmViewId=news_view&newsId=20090729005107&newsLang=en.
- [12] Open Source at Intel. Online document. Accessed Sep. 03, 2009, <http://oss.intel.com/>.
- [13] LaTeX - A Document Preparation System. Online document. Accessed Aug. 31, 2009, <http://www.latex-project.org/>.
- [14] LaTeX Prosper Class. Online document. Accessed Aug. 31, 2009, <http://amath.colorado.edu/documentation/LaTeX/prosper/>.
- [15] Ralph Nader and James Love. RE: US v. Microsoft proposed final order. Online document. Accessed Sep. 22, 2009, Published Nov. 5, 2001, <http://www.nader.org/releases/msfinalorder.html>.
- [16] NetBeans IDE. Online document. Accessed Aug. 31, 2009, <http://www.netbeans.org/>.
- [17] Netcraft. Netcraft Web Server Survey. Online document. Accessed August 31 2009, <http://www.netcraft.com/survey>.
- [18] OpenOffice.org Product Reviews. Online document. Accessed Sep. 2nd, 2009, <http://www.openoffice.org/product/reviews.html>.
- [19] Oracle's Technical Contributions to Linux. Online document. Accessed Aug. 31, 2009, <http://www.oracle.com/us/technologies/linux/026042.htm>.
- [20] Open Source Initiative. Online document. Accessed August 31, 2009, <http://www.opensource.org>.
- [21] Qt Development Frameworks a Nokia unit. Online Document. Accessed Aug. 31, 2009, <http://qt.nokia.com>.
- [22] Eric S. Raymond. The Cathedral and the Bazaar. Online Document. Accessed Aug. 31, 2009, Published August 02, 2002 <http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/>.
- [23] Richard Stallman. The GNU Project. Online document. Accessed Aug. 31, 2009, <http://www.gnu.org/gnu/thegnuproject.html>
- [24] Initiative for Software Choice. Online document. Accessed Aug. 31, 2009, <http://softwarechoice.org/>.
- [25] Richard Stallman. Why you shouldn't use the Lesser GPL for your next library. Online document. Accessed Aug. 31, 2009, <http://www.gnu.org/licenses/why-not-lgpl.html>.
- [26] Unison File Synchronizer. Online document. Accessed Aug. 2009, <http://www.cis.upenn.edu/~bcpierce/unison/>.
- [27] Winzip. A Corel Company. Online document. Accessed Aug. 31, 2009, <http://www.winzip.com>.
- [28] Xfig Drawing Program for the X Windows System. Online document. Accessed Aug. 31, 2009, <http://www.xfig.org/>.
- [29] X.org Foundation. Online document. Accessed August 31, 2009, <http://www.x.org>.

AUTHORS PROFILE

Dan Lipsa received a bachelor's degree in computer science, from Polytechnic University, Bucharest in 1994. In 1998, he received a master's degree in computer science from University of New Hampshire, Durham. He held positions of technical lead and senior software engineer at various companies in US and Europe and he was an Assistant Professor at Armstrong University, Savannah between 2003-2009. He has been a Research Assistant at Swansea University in the Department of Computer Science since 2010 where he is also working on his PhD. His research interests are in the areas of scientific visualization, computer graphics and software engineering. Lipsa has published six peer-reviewed, scientific papers and a book chapter.

Robert S. Laramée received a bachelor's degree in physics, cum laude, from the University of Massachusetts, Amherst in 1997. In 2000, he received a master's degree in computer science from the University of New Hampshire, Durham. He was awarded a PhD from the Vienna University of Technology, Austria at the Institute of Computer Graphics in 2005. He was a Senior Researcher at the VRVis Research Center and at the same time a software engineer at AVL (www.avl.com) in the department of Advanced Simulation Technologies from 2001-2006. He

has been a Lecturer in Visualization at Swansea University in the Department of Computer Science since 2006. His research interests are in the areas of scientific visualization, computer graphics, and human-computer interaction. Laramee has published over 50 peer-reviewed, scientific papers, including 20 journal papers, in visualization, human-

computer interaction, software engineering, and simulation. Dr Laramee has served on over 20 International Programme Committees (IPCs) organized 4 international workshops, and serves as a reviewer for more than 25 journals, conferences, workshops, and institutions.

Analyzing the Load Balance of Term-based Partitioning

Ahmad Abusukhon
Faculty of Science & IT
Al-Zaytoonah Private University of Jordan
Amman Jordan
ce4aab@student.sunderland.ac.uk

Mohammad Talib
Department of Computer Science
University of Botswana
Private Bag UB 00704, Gaborone, BOTSWANA
talib@mopipi.ub.bw

Abstract— In parallel (IR) systems, where a large-scale collection is indexed and searched, the query response time is limited by the time of the slowest node in the system. Thus distributing the load equally across the nodes is very important issue. Mainly there are two methods for collection indexing, namely document-based and term-based indexing. In term-based partitioning, the terms of the global index of a large-scale data collection are distributed or partitioned equally among nodes, and then a given query is divided into sub-queries and each sub-query is then directed to the relevant node. This provides high query throughput and concurrency but poor parallelism and load balance. In this paper, we introduce new methods for terms partitioning and then we compare the results from our methods with the results from the previous work with respect to load balance and query response time.

Keywords- *Term-partitioning schemes, Term-frequency partitioning, Term-lengthpartitioning, Node utilization, Load balance*

I. INTRODUCTION

The number of pages (documents) available online is increasing rapidly. Gulli and Signorini [17] estimated the current size of the web. They mentioned that Google claims to index more than 8 billion pages. They estimated the indexable web to be at least 11.5 billion pages. Beside the huge document collection, we have a large number of information requests (queries) that are submitted by clients. Sullivan [18] reported that the number of searches per day performed by Google is 250 million. In order to the users to effectively retrieve documents that are relevant to their needs, the IR systems must provide effective, efficient, and concurrent access to large document collections. Thus, the first step in developing information retrieval system is to decide on what access method should be used in order to access large-scale collection efficiently. In IR systems the indices of documents must be built to perform timely information retrieval. The most known structures for building the index of large-scale collection are inverted files and signature files. The most common and most efficient structure for building the index of large-scale collection is the inverted file [1,2].

Zobel[3] compared inverted files and signature files with respect to query responsetime and space requirements. They

found that inverted files evaluate queries in less time than signature files and need less space, thus for efficiency reasons, we use the inverted files in our research.

In general, inverted files consist of vocabulary and a set of inverted lists. The vocabulary contains all unique terms in the whole data collection; while the inverted lists composed of a list of pointers and each pointer consists of document identifier and term frequency. The term frequency in each pair represents how many times term i appears in document j ($F_{i,j}$). Let's suppose that the inverted list for term "world" is:

World 2:5, 6:3, 12:1, 15:1

This means that term world appears five times in document 2, three times in document 6, one time in document 12, and one time in document 15. The numbers 2,6, 12, and 15 are called the document identifiers while the numbers 5, 3, 1, and 1 are called the term frequencies.

In parallel IR system when term partitioning scheme is used all unique terms in the data collection and their inverted lists reside on a single node called the broker. The broker distributes all terms and their inverted lists across nodes using different approaches. The terms, for instance, may be distributed in round robin fashion. In this case the broker iterates over all terms in the inverted file, and distributes them sequentially across nodes. The aim of round robin partitioning scheme is to balance the load over the nodes by storing nearly equal number of terms on all nodes.

Moffat[4] showed that distributing the terms of the term-based index in round robin fashion results in load imbalance especially when there are heavy loaded terms. Because of this the round robin partitioning scheme does not take care of those terms to be distributed equally across nodes.

Xi[5] proposed the hybrid partitioning scheme in order to achieve load balance. In hybrid partitioning scheme the inverted lists of the term-based index are split into chunks then chunks are distributed across nodes. They investigated partitioning the inverted list into different sizes and they

concluded that hybrid partitioning scheme achieves better load balance than the other schemes (document-based and term-based partitioning) when the chunk size is small (1024 posting). But when the chunk size is large the hybrid partitioning is worse than the document partitioning. In this paper, we propose two methods for term partitioning scheme - term length partitioning and term frequency partitioning.

Abusukhon et al. [13, 14, 16] proposed improving the load balance of hybrid partitioning using hybrid queries. In their work, they divided the nodes into clusters then the inverted lists of all terms were divided into a number of chunks, the chunks of a given term that start with a certain letter were distributed equally among the nodes of a certain cluster. A hybrid query was generated from a set of queries and then this query was divided into streams with respect to the first letter of each term. Each stream was directed to the relevant cluster.

II. RELATED WORK

Inverted files can be partitioned by different approaches. Different approaches of data partitioning leads to different load balance and different query response time as described by Abusukhon et al. [15]. In this section we shed light on various strategies for term-partitioning schemes as described in the previous work.

Cambazoglu[6] demonstrated two main types for inverted file partitioning - term-based partitioning and document-based partitioning. In term-based partitioning all unique terms in the data collection and their inverted lists reside on a single node. In document-based partitioning the data collection is divided into sub-collections, sub-collections are distributed across nodes, and then each node builds its own index.

Jeong [7] proposed two methods for load balancing when using term-based partitioning scheme. In the first method they proposed to split the inverted list into equal parts and then distribute those parts across nodes instead of distributing equal number of terms across nodes in order to achieve better load balance.

In the second method they proposed to partition the inverted lists based on the access frequency of terms in the user query and the inverted list size for each term appears in the query. They studied the performance of the above schemes by simulation under different workloads.

Marin and Costa [19] stated that load balance is sensitive to queries that include high frequency terms that refer to inverted lists of different sizes.

Moffat[4] examined different methods to balance the load for term-distributed parallel architecture and proposed different techniques in order to reduce the net querying costs. They defined the workload as follows:

$$W_t = Q_t * S_t$$

Where W_t is the workload caused by the term t that appears in a query batch Q_t and has an inverted list of length equals S_t bytes. The workload for a given node is the sum of W_t of the terms distributed over that node. In one of their experiments the terms of the queries were distributed over the nodes randomly. The simulation result showed that some nodes were heavily-loaded because they retrieved very large size inverted lists; therefore, some of the nodes in the system were half-idle and affect the system throughput. In order to improve the load balance they proposed distributing the inverted lists equally among the nodes based on the number of pointers P in each inverted list.

Jeong and Omiecinski [20] concluded that partitioning by term resulted in load imbalance because some terms were more frequently requested in a query. Thus, nodes where these terms associated with their inverted lists were stored would be heavily utilized.

Xi[5] proposed a hybrid partitioning scheme in order to distribute terms across the nodes. Hybrid partitioning scheme avoids storing terms with long posting lists on one node instead of the inverted list of a given term is split into a number of equal size chunks and then distributed randomly across the nodes. They measured the load balance and concluded that the hybrid-partitioning scheme outperforms other schemes when the chunk size is small. In this paper, we propose Term Length partitioning and Term Frequency partitioning for improving the load balance of term-based partitioning.

III. SYSTEM ARCHITECTURE

Fig.1 shows our system architecture. It consists of six nodes and one broker. All nodes are connected to the broker via Ethernet switch.

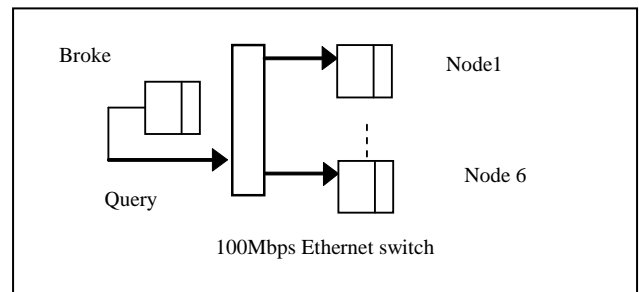


Figure 1. Distributed IR Architecture

The machine specifications for five nodes are: CPU 2.80Ghz RAM 256MB whereas the specification for the last and the broker are: CPU 3.00Ghz, RAM 512MB. All machines are running in Windows XP environment.

IV. RESEARCH METHODOLOGY

We carried-out a set of real experiments using six nodes and one broker as shown in Fig. 1. In all of our experiments we use the data collection WT10G from TREC-9 and 10,000

queries extracted from the start of Excite-97 log file. The chronology of our research methodology is traced below:

1. We build the global index (called the term-based partitioning) in the following way:

a. Broker sends the documents across nodes in round robin fashion.

b. Each node when receiving its document performs these activities-

- Filters the document it receives from stop words (the stop word list consists of 30 words), HTML tags, and all noncharacters and non-digit terms.

- Accumulates the posing lists in main memory until a given memory threshold is reached. At this point the data stored in memory is flushed to on-disk file [8,9, 2,11]. This process is repeated until all documents in the data collection are indexed.

- Merge all on-disk files together into one on-disk file called the local index or the document-based partitioning.

c. Finally, the broker collects all local indices from all nodes and merges them together in order to produce the global index.

2. We partition the terms of the global index across nodes using four different approaches, viz., round robin partitioning, partitioning based on the length of the inverted list, term length partitioning and term frequency partitioning.

Next, we demonstrate the above approaches and then run a set of real experiments in order to compare them with respect to the node utilization.

A. Round Robin Partitioning

In round robin partitioning, we distribute the terms of the global index across nodes. If we have three nodes and four terms A, B, C, and D associated with their posting lists then term A may reside on node 1, term B on node 2, term C on node 3, and term D on node 1, and so on [10].

B. Term Partitioning Based on the Length of the Inverted List

In this method of partitioning, we pass over the terms of the global index twice. In the first pass, we calculate the length of the inverted list L for each term T, store T and L in a look up file PL after sorting them on L in ascending order. In the second pass, we distribute the terms and the inverted lists of the global index across the nodes using the PL in round robin fashion in the follow order:

1. Read one record (L, T) from PL
2. Search T in the global index and retrieve its inverted list
3. Send T and its inverted list to a certain node in round robin fashion.
4. If no more records then EXIT else go to step1.

We use the above algorithm in order to guarantee that all inverted lists of the same length reside on all nodes equally. Fig. 2 shows an example of the PL file.

<u>Inverted List Length</u>	<u>Term</u>
100	a
100	b
.	.
.	.
1200	z

Figure 2. Sorted look up file (PL)

C. Term Length Partitioning

Case[12] described Zipf's principle of least effort. He stated that:

"According to Zipf's law (1949) each individual will adopt a course of action that will involve the expenditure of the probable least average of his work, in other words, the least efforts"

He wrote that the statistical distribution of words in the text of James Joyce's Ulysses follows the type of pattern on which Zipf based his theory. The 10th most common word appears 2,653 times; the 100th most common word, 265 times; and the 1,000th, 26 times. This relation is called "harmonic distribution". He stated that humans try to use short, common words whenever they can rather than longer words that take more effort. This is the first motivation for the term-length partitioning. In this section, we propose to partition the terms of the global index associated with their inverted lists with respect to the term length (in letters).

Our research hypothesis for term length partitioning is based on statistical information collected from the query log file Excite-97. This information is stored into a look up file as it is shown in Fig.2. In Fig. 3, we see that the term lengths are not distributed equally in Excite-97 (i.e. have very skewed distribution). For example, the number of terms of length 5 equals 360093 while the number of terms of length 11 equals 59927. The total number of terms in Excite-97 is 2235620. Thus the percentage of the terms of length 5 to the total number of terms = $360093 / 2235620 = 0.16\%$ while the percentage of terms of length 11 = $59927 / 2235620 = 0.03\%$. This is the second motivation for the term-length partitioning. When users submitted their queries, the queries contain terms of different length. Suppose that the majority of those terms are of length 4 and 5 as it is shown in Fig. 3. In addition, we partitioned the terms of the global index in round robin fashion and all terms of length 4 and 5 resided on one or two nodes. This way of partitioning will result in load imbalance because most of the work will be carried out by one or two nodes only while other nodes doing less work or may be idle. Thus our

hypothesis is that all terms of the same length must be distributed equally across nodes in order to achieve more load balance.

Our partitioning method requires passing over all terms of the global index twice. In the first pass, we calculate the length WL of each term T, store WL and T in a look up file PL after sorting them on WL in ascending order. In the second pass, we distribute the terms and the inverted lists of the global index across nodes using the PL in round robin fashion in the following order:

1. Read one record (WL, T) from PL
2. Search T in the global index and retrieve its inverted list
3. Send T and its inverted list to a certain node in round robin fashion.
4. If no more records then EXIT else go to step1.

We use the above algorithm in order to guarantee that all terms of the same length reside on all nodes equally.

Our proposed partitioning algorithm differs from round robin partitioning in that it distributes the terms across the nodes equally, and it also guarantees that all nodes get the same number of terms of the same length. The round robin algorithm distributes the terms across the nodes equally regardless of the term length. Therefore, we expect that the term length-partitioning scheme achieves better load balance than the round robin partitioning. To the best of our knowledge, no previous work investigated partitioning the global index based on the term length, or measured the nodes utilization when using the term length partitioning.

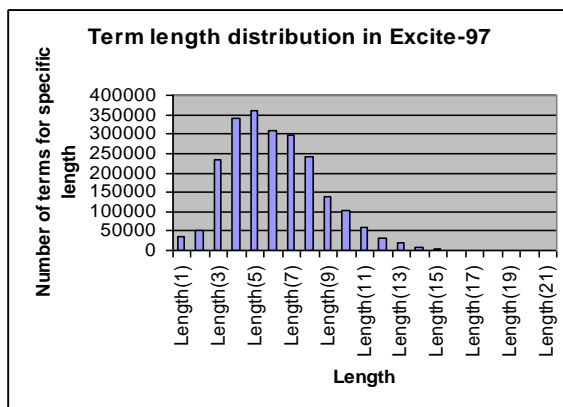


Figure 3. Term length distribution for Excite-97

D. Term Frequency Partitioning

Baeza[1] demonstrated Zipf's law (Fig.4), which is used to capture the distribution of i -th most frequent word is $1/i^r$ times that the most frequent word" (r between 1.5 and 2.0), thus the frequency of any word is inversely proportional to its rank (i.e. i -th position) in the frequency table. They

showed that the distribution of sorted frequencies (decreasing order) is very skewed (i.e. there were a few hundred words which take up 50% of the text) thus words that are too frequent like stop words can be ignored.

In Fig. 4, graph (A) shows the skewed distribution of the sorted frequencies while graph (B) is the same as graph (A) but we divided the curve into six clusters (A, B, C, D, E, F) after ignoring the stop words. Cluster (A) has the most frequent terms, then cluster B, then C, and so on. In addition, in graph (B) we assume that most or all of the query terms appear in cluster (A), that all or most of the terms in cluster (A) may not reside on all nodes but on one or two nodes in the system. In this case, we may have one or two nodes busy answering the query terms while other nodes are idle, and thus cause the load imbalance.

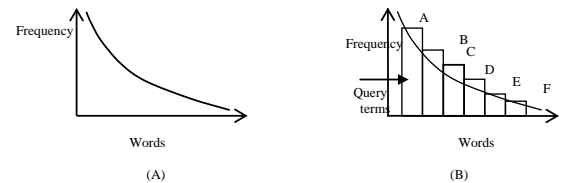


Figure 4. Zipf's Law

Our hypothesis is that if we filter the data collection from stop words (words like the, in, on, ..., etc), then the terms with high total frequency (for example the terms in cluster A) are more likely to appear in the user query (i.e. have higher probability to appear in the user query) than the terms with low total frequency. Thus, the terms with high frequency must be distributed equally across the nodes in order to achieve more load balance. Here we propose to partition the terms of the global index with respect to the total term frequency calculated from their inverted lists. To make it clear what we mean by term frequency, we demonstrate the following example: Let's suppose we have two terms (A, B) associated with their inverted lists as it is shown in table 1.

TABLE I. INVERTED LISTS

term	Inverted list
A	1:3, 4:1, 6:2, 9:5, 12:5, 13:2, 15:3
B	2:1, 4:1, 7:2, 9:2, 11:1, 12:1

Then, the total frequency F for each term is calculated as follows:

$$F_A = 3+1+2+5+5+2+3 = 21$$

$$F_B = 1+1+2+2+1+1 = 8$$

Based on the above calculations, the total frequency of term A is higher than the total frequency of term B and thus we expect that term A has higher probability to appear in the user query than term B. We expect that the load imbalance may occur, if the majority of the terms with higher total frequency reside on one or two nodes, as a result of performing some techniques like round robin partitioning, in this case, most of the user query terms are answered by one or two nodes and thus cause the load imbalance. In the next section, we show how to calculate the probability of a given term to appear in the user query terms.

1) Calculate the Term Probability

Suppose that we have the document collection Ω where:

$$\Omega = \{D_0, D_1, D_2, \dots, D_n\}$$

Let $T = \{t_0, t_1, t_2, \dots, t_n\}$ be the set of terms appears in any document D_i in any combination. Let the term t_j occurs m times in D_i , then we assume that the probability (P_{t_j}) that the term t_j appears in the query terms is equivalent to how many times it occurs in the whole data collection.

$$p_{tj} = \sum_{i=1}^n m_{j,i} \quad (1)$$

Where, n is the total number of documents in the data collection and $m_{j,i}$ is how many times the term j appears in document i .

For example, suppose we have a data collection contains 4 documents (d_1, d_2, d_3 , and d_4) and three terms (t_1, t_2 , and t_3) and that t_1 appears in these documents (5, 10, 2, 3) times, t_2 appears (1, 3, 0, 1) and t_3 appears (1, 1, 1, 2). We assume that the probability of term t_1 to appear in the query terms is 20, t_2 is 5 and t_3 is 5.

To normalize the value of P_{tj} , we divided it by the summation of the total frequencies of all distinct terms in the data collection, i.e.

$$p_{tj} = \frac{\sum_{i=1}^n m_{j,i}}{\sum_{k=1}^n \sum_{l=1}^s m_{l,k}} \quad (2)$$

Where, n is the total number of documents in the data collection and s is total number of distinct terms in the data collection. In the above example, after normalization, the probability of term $t_1 = 20 / 30$ (i.e. 0.7) while the probability of term $t_2 = 5/30$ (i.e. 0.17).

2) Term Distribution Based on the Total Term Frequency

We pass over all terms of the global index twice. In the first pass, we calculate the total frequency F for each term T using equation 1 then store F and T in a look up file PL after sorting them on F in ascending order.

In the second pass, we distribute the terms and the inverted lists of the global index using the PL in round robin fashion in the following order:

1. Read one record (F, T) from PL
2. Search T in the global index and retrieve its inverted list
3. Send T and its inverted list to a certain node in round robin fashion
4. If no more records then EXIT else go to step1.

The above algorithm is used in order to guarantee that all terms of the same total frequency F are distributed across all nodes equally.

To the best of our knowledge, no previous work investigated partitioning the global index based on the total term frequency or measured the nodes utilization when using the term frequency partitioning.

V. EXPERIMENTS

In this research, we carried out a set of real experiments using the system architecture shown in Fig. 1. We used the data collection WT10G from TREC-9 in order to build the global index and 10,000 queries extracted from the start of the Excite-97 log file to measure the node utilization for each node.

Xi[5] defined the node utilization as – “the total amount of time the node is serving requests from the IR server divided by the total amount of time of the entire experiment”.

We distributed the terms of the global index using the four approaches mentioned in sections A, B, C, and D. We carried out 10,000 queries, each query is sent across all nodes. Each node retrieves and sends the inverted lists of the query terms to the broker for evaluation. We considered the time the node serving the query S_i to be the time required to retrieve all inverted lists of query terms and send them to the broker. We considered the node to be idle if the query term does not exist on its hard disk in that case, the searching time is excluded from S_i . The total time for each experiment is shown in table VII. For each partitioning scheme mentioned in sections A, B, C, and D, we calculated ΔU :

$$\Delta U = \text{Maximum node utilization} - \text{Minimum node utilization} \\ = \text{MaxU} - \text{MinU} \quad (3)$$

Tables II, III, IV, and V show the time taken by each node to serve 10,000 queries sent by the broker as well as the node utilization. We calculated the node utilization by dividing the time the node serving queries by the total time of the experiment. For example, the node utilization for node 1 (Table II) is calculated as follows:

$$\text{Node utilization} = 598195 / 3349515 = 0.1785.$$

This calculation step is carried out for all tables (II, III, IV, and V). Next, we produce table VI from the above tables. For each table we got the minimum and the maximum node

utilization (MinU, MaxU). For table II, MinU = 0.1104 and MaxU= 0.1947 then we calculate ΔU :

$$\begin{aligned}\Delta U &= \text{MaxU} - \text{MinU} \\ &= 0.1947 - 0.1104 \\ &= 0.0843\end{aligned}$$

We use table VI to produce Fig. 5 and we calculate the average query response time for each of the above partitioning algorithms by dividing the total time of the experiment by the total number of the executed queries. For round robin partitioning scheme:

$$\begin{aligned}\text{The average query response time} &= 3349515 / 10000 \\ &= 334.95 \text{ milliseconds}\end{aligned}$$

Table VIII and Fig. 6, show the partitioning methods and the average query response time.

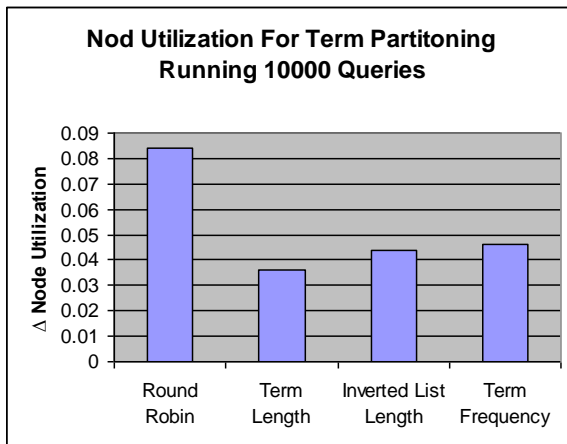


Figure 5. Comparison between four approaches for term partitioning scheme (Round robin, Term Length, inverted List Length, and Term frequency) with respect to ΔU

TABLE II. NODE UTILIZATION FOR ROUND ROBIN PARTITIONING

Node #	Time serving queries (milliseconds)	Node utilization
1	598195	0.1785
2	541421	0.1616
3	369798	0.1104
4	652215	0.1947
5	628682	0.1876
6	604870	0.1805

TABLE III. NODE UTILIZATION FOR PARTITIONING BASED ON THE LENGTH OF INVERTED LIST.

Node #	Time serving queries (milliseconds)	Node utilization
1	598195	0.1785
2	541421	0.1616
3	369798	0.1104
4	652215	0.1947
5	628682	0.1876
6	604870	0.1805

1	667148	0.2046
2	596106	0.1828
3	640117	0.1963
4	699275	0.2145
5	647914	0.1987
6	581225	0.1782

TABLE IV. NODE UTILIZATION FOR PARTITIONING BASED ON THE TOTAL TERM FREQUENCY

Node #	Time serving queries (milliseconds)	Node utilization
1	559151	0.1692
2	640726	0.1939
3	590804	0.1788
4	594265	0.1799
5	667375	0.202
6	711796	0.2154

TABLE V. NODE UTILIZATION FOR PARTITIONING BASED ON TERM LENGTH

Node #	Time serving queries (milliseconds)	Node utilization
1	596510	0.1690
2	535703	0.1518
3	689623	0.1954
4	625451	0.1772
5	629086	0.1783
6	618969	0.1754

TABLE VI. Δ NODE UTILIZATION

Term Partitioning Scheme	Δ Node Utilization (Max - Min)
Round Robin	0.0843
Term Length	0.0363
Inverted List Length	0.0436
Term Frequency	0.0462

TABLE VII. TOTAL TIME OF EXPERIMENTS

Term partitioning method	Total time of experiment (milliseconds)
Round Robin	3349515
Length of inverted list	3259879
Term frequency	3303175
Term length	3528047

TABLE VIII. AVERAGE QUERY RESPONSE TIME (MILLISECONDS)

Partitioning Method	Average Query Response Time (milliseconds)
Length of inverted list	325.9879

Round Robin	334.9515
Term frequency	330.3175
Term length	352.8047

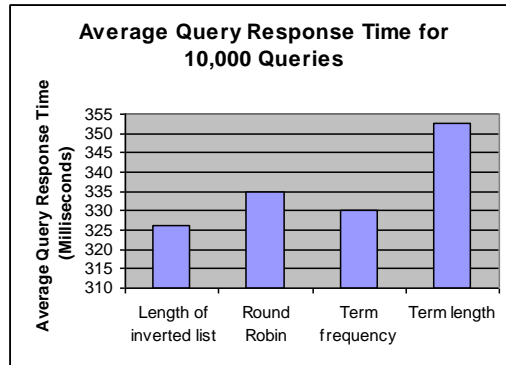


Figure 6. Average query response time

VI. CONCLUSION AND FUTURE WORK

In this paper, we carried out a set of real experiments using our parallel IR system in order to improve the load balance for term partitioning scheme. We proposed to partition the terms of the global index based on term length and the total term frequency extracted from the inverted lists.

We compared our proposed methods with round robin partitioning scheme and the partitioning scheme based on the length of the inverted list. Our results showed that the term length-partitioning scheme performed slightly better than other schemes with respect to node utilization (Table VI). On the other hand, partitioning terms based on the length of the inverted list achieved slightly less average query response time than other schemes (Table VIII).

ACKNOWLEDGMENT

The authors would like to thank Al-Zaytoonah University for their support.

REFERENCES

- [1] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press, Addison-Wesley, New York (1999)
- [2] Zobel, J., Moffat, A.: Inverted Files for Text Search Engines, ACM Computing Surveys (CSUR) (2006)
- [3] Zobel, J., Moffat, A., Ramamohanarao, k.: Inverted Files Versus Signature Files for Text Indexing, ACM Transactions on Database systems, 453-490 (1998)
- [4] Moffat, A., Webber, W., Zobel, J.: Load Balancing for Term-Distributed Parallel Retrieval, The 29th annual international ACM SIGIR conference on Research and development in information, 348-355. ACM, New York (2006)
- [5] Xi, W., Somil, O., Luo, M., and Fox, E.: Hybrid partition inverted files for large-scale digital libraries. In Proc. Digital Library: IT Opportunities and Challenges in the New Millennium, Beijing, China, Beijing Library Press (2002)
- [6] Cambazoglu, B., Catal, A., Aykanat, C.: Effect of Inverted Index Partitioning Schemes on Performance of Query Processing in Parallel

- Text Retrieval Systems. A. Levi et al. (Eds.): ISCIS 2006, LNCS, 4263(6), 717-725. Springer, Heidelberg (2006)
- [7] Jeong, B.S., Omiecinski, E.: Inverted File Partitioning Schemes in Multiple Disk Systems, IEEE, Transactions on Parallel and Distributed Systems, 6(2), 142-153. IEEE press, Piscataway, NJ, USA (1995)
 - [8] Heinz, S., Zobel, J.: Efficient Single-Pass Index Construction for Text Databases. Journal of the American Society for Information Science and Technology, 54(8), 713-729 (2003)
 - [9] Jaruskulchai, C., Kruegkrai, C.: Building Inverted Files Through Efficient Dynamic Hashing (2002)
 - [10] Badue, C., Baeza-Yates, R., Ribeiro-Neto, B., Ziviani, N.: Distributed Query Processing Using Partitioned Inverted Files, 10-20 (2001)
 - [11] Lester, N., Moffat, A., Zobel, J.: Fast On-Line Index Construction by Geometric Partitioning, CIKM'05, October 31 and November 5, Proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, 776-783 ACM (2005)
 - [12] Case, D. *Looking for Information: A survey of research on Information Seeking, Needs, and Behavior*. USA: Elsevier Science. pp:140-141 (2002)
 - [13] Abusukhon, A., Talib, M. and Oakes, M.P. Improving the Load Balance for Hybrid Partitioning Scheme by Directing Hybrid Queries. In: Burkhart, H. (Eds.). *Proceedings of the IASTED International Conference on Parallel and Distributed Computing and Networks as part of the 26th IASTED International Multi-Conference on APPLIED INFORMATICS*. Innsbruck, Austria 12-14 February 2008, pp. 238-244. ACTA press: USA. (2008a).
 - [14] Abusukhon, A. and Oakes, M.P. An Investigation into Query Throughput and Load Balance Using Grid IR. In *Proceedings of the 2nd BCS- IRSG Symposium on Future Directions in Information Access FDIA 2008*. BCS London Office, UK, 22nd September 2008, pp. 38-44. eWic: UK. (2008b)
 - [15] Abusukhon, A., Oakes, M. Talib, M. and Abdalla, A. Comparison Between Document-based, Term-based and Hybrid Partitioning. In Snael, V. et al. (Eds.), *Proceedings of the First IEEE International Conference on the Application of Digital Information and Web Technologies*. Ostrava, Czech Republic, 4-6 August, pp. 90-95. IEEE, (2008c)
 - [16] Abusukhon, A. and Talib, M. Improving Load Balance and Query Throughput of Distributed IR Systems. International Journal of Computing and ICT Research (IJCIR), 4(1), pp 20-29, (2010)
 - [17] Gulli, A. and Signorini, A., The indexable web is more than 11.5 billion pages. The 14th international conference on World Wide Web ACM, New York, USA, pp 902-903, (2005).
 - [18] Sullivan, D., Searches per day. Search Engine. Watch, <http://searchenginewatch.com/reports/article.php/2156461>, (2003)
 - [19] Marin, M., and Costa, G.V. High-Performance Distributed Inverted Files. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management CIKM'07*. Lisbon, Portugal, 6-9 November 2007, pp. 935-938, ACM: New York, USA, (2007).
 - [20] Jeong, B.S., and Omiecinski, E. Inverted File Partitioning Schemes in Multiple Disk Systems. *IEEE Transactions on Parallel and Distributed Systems*, 6(2), pp. 142-153. IEEE Press, USA. (1995).

AUTHORS PROFILE

Dr. Ahmad Abusukhon got his Bachelor degree in Computer Science from Mu'tah University in 1990, his M.Sc degree in Computer Science from the University of Jordan in 2001 and he got his PhD degree in Computer Science from the University of Sunderland in 2009. He is now working as assistant Prof. at Al-Zaytoonah University. Dr. Abusukhon is intersted in Computer networks, Distributed systems, and Distributed computing.

Professor M. Talib has, presently, been associated with the computer science department of the university of Botswana and has also been an adjunct professor at the Touro University International (TUI), USA. He has worked at a number of universities all across the globe in different capacities besides India where he remained the Head of the Department of Computer Scienc. He

has an excellent industrial relevance and has worked as Software Engineer at the silicon valley in California for a significant period of time. He has been a Consultant for several software development companies and handled various small and big projects all across the world. He was conferred upon a degree of the Doctor of Philosophy (Ph.D.) in computer science with specialization in computer vision from the prestigious University of Lucknow in India with Certificate of Honor. Besides PhD, he is also flanked by an M.S. in computer science, MSc in statistics and PG Diploma in Computing. He has supervised over a dozen Master and four PhD students in different areas of Computer Science, Business and IT. His research areas include Bio informatics, Computer Vision, and Robotics. Presently, he is working on a two way interactive video communication through the virtual screen with the essence

of smell. He has over sixty five research papers published in different world class journals and conferences besides a book. He is also credited with over 300 publications including (under)graduate project reports, thesis, extension articles, study guides, edited research papers, books, etc. besides a minimum of 50 Industrial training supervision reports all across the world. He has chaired and remained member of various Academic Councils, Board of Studies, Academic and Advisory Boards, Examination Committees, Moderation and Evaluation Committees worldwide. He is the Member of the Editorial Board of about a dozen International Journals. He has also been associated with a number of international computer societies, associations, forums etc. in various capacities.

A Genetic Algorithm for Solving Travelling Salesman Problem

Adewole Philip

Department of Computer Science
University of Agriculture,
Abeokuta, Nigeria
philipwale@yahoo.com

Akinwale Adio Taofiki

Department of Computer Science
University of Agriculture,
Abeokuta, Nigeria
aatakinwale@yahoo.com

Otunbanowo Kehinde

Department of Computer Science
University of Agriculture,
Abeokuta, Nigeria
kenny_csc@yahoo.com

Abstract— In this paper we present a Genetic Algorithm for solving the Travelling Salesman problem (TSP). Genetic Algorithm which is a very good local search algorithm is employed to solve the TSP by generating a preset number of random tours and then improving the population until a stop condition is satisfied and the best chromosome which is a tour is returned as the solution. Analysis of the algorithmic parameters (Population, Mutation Rate and Cut Length) was done so as to know how to tune the algorithm for various problem instances.

Keywords- Genetic Algorithm, Generation, Mutation rate, Population, Travelling Salesman Problem

I. INTRODUCTION

The traveling salesman problem (TSP) is a well-known and important combinatorial optimization problem. The goal is to find the shortest tour that visits each city in a given list exactly once and then returns to the starting city. In contrast to its simple definition, solving the TSP is difficult since it is an NP-complete problem [4]. Apart from its theoretical approach, the TSP has many applications. Some typical applications of TSP include vehicle routing, computer wiring, cutting wallpaper and job sequencing. The main application in statistics is combinatorial data analysis, e.g., reordering rows and columns of data matrices or identifying clusters.

The NP-completeness of the TSP already makes it more time efficient for small-to-medium size TSP instances to rely on heuristics in case a good but not necessarily optimal solution is sufficient.

In this paper genetic algorithm is used to solve Travelling Salesman Problem. Genetic algorithm is a technique used for estimating computer models based on methods adapted from the field of genetics in biology. To use this technique, one encodes possible model behaviors into "genes". After each generation, the current models are rated and allowed to mate and breed based on their fitness. In the process of mating, the genes are exchanged, crossovers and mutations can occur. The current population is discarded and its offspring forms the next generation. Also, Genetic Algorithm describes a variety of modeling or optimization techniques that claim to mimic some

aspects of biological modeling in choosing an optimum. Typically, the object being modeled is represented in a fashion that is easy to modify automatically. Then a large number of candidate models are generated and tested against the current data. Each model is scored and the "best" models are retained for the next generation. The retention can be deterministic (choose the best k models) or random (choose the k models with probability proportional to the score). These models are then randomly perturbed (as in asexual reproduction) and the process is repeated until it converges. If the model is constructed so that they have "genes," the winners can "mate" to produce the next generation.

II. TRAVELLING SALESMAN PROBLEM

The TSP is probably the most widely studied combinatorial optimization problem because it is a conceptually simple problem but hard to solve. It is an NP complete problem. A Classical Traveling Salesman Problem (TSP) can be defined as a problem where starting from a node is required to visit every other node only once in a way that the total distance covered is minimized. This can be mathematically stated as follows:

$$\text{Min} \quad \sum_{i,j} c_{ij}x_{ij} \quad (1)$$

$$\text{s.t} \quad \sum_j x_{ij} = 1 \quad \forall i \neq j \quad (2)$$

$$\sum_i x_{ij} = 1 \quad \forall j \neq i \quad (3)$$

$$u_i = 1 \quad (4)$$

$$2 \leq u_i \leq n \quad \forall i \neq 1 \quad (5)$$

$$u_i - u_j + 1 \leq (n-1)(1 - x_{ij}) \quad \forall i \neq j \quad \forall j \neq 1 \quad (6)$$

$$u_i \geq 0 \quad \forall i \quad (7)$$

$$x_{ij} \in \{0,1\} \quad \forall i,j \quad (8)$$

Constraints set (4), (5), (6) and (7), are used to eliminate any sub tour in the solution. Without the additional constraints for sub tour elimination, the problem reduces to a simple assignment problem which can be solved as an Linear Programming without binary constraints on x_{ij} and will still result in binary solution for x_{ij} . Introduction of additional constraints for sub tour elimination, however, makes the problem a Mixed Integer Problem with n^2 integer variables for a problem of size n , which may become very difficult to solve for a moderate size of problem [7].

III. METHODOLOGY

Genetic algorithm is a part of evolutionary computing, which is a rapidly growing area of artificial intelligence. Genetic algorithm is inspired by Darwin's theory about evolution. It is not too hard to program or understand, since they are biological based. The general algorithm for a GA:

1) Create a Random Initial State:

An initial population is created from a random selection of solutions (which are analogous to chromosomes). This is unlike the situation for symbolic artificial intelligence systems where the initial state in a problem is already given instead.

2) Evaluate Fitness:

A value for fitness is assigned to each solution (chromosome) depending on how close it actually is to solving the problem (thus arriving to the answer of the desired problem). These "solutions" are not to be confused with "answers" to the problem, think of them as possible characteristics that the system would employ in order to reach the answer.

3) Reproduce (ChildrenMutate):

Those chromosomes with a higher fitness value are more likely to reproduce offspring which can mutate after reproduction. The offspring is a product of the father and mother, whose composition consists of a combination of genes from them (this process is known as "crossing over").

4) Next Generation:

If the new generation contains a solution that produces an output that is close enough or equal to the desired answer then the problem has been solved. If this is not the case, then the new generation will go through the same process as their parents did. This will continue until a solution is reached.

B. Algorithm

1. Initialization: Generate N random candidate routes and calculate fitness value for each route.

2. Repeat following steps Number of iteration times:

- Selection: Select two best candidate routes.*
- Reproduction: Reproduce two routes from the best routes.*
- Generate new population: Replace the two worst routes with the new routes.*

3. Return the best result

IV. IMPLEMENTATION

1) Program Details

The program was written with Java. In genetic algorithm, a class "Chromosome" is needed. The Chromosome class generates random tours and makes them population members when its object is instantiated in the TSP class. The TSP class uses the Chromosomes "mate" method to reproduce new offspring from favoured Population of the previous generations. The TSP class in this case has two methods that use methods in Chromosome, the two methods are described below.

start(): This method initializes the cities and creates new chromosomes by creating an array of Chromosome objects. It also sorts the chromosomes by calling the method sortChromosomes() in Chromosomes then it sets the generation to 0

run(): Gets the favoured population from all the chromosomes created and mates them using mate() after this it sorts the chromosomes and then calculates the cost of the tour of the best chromosome. It repeats this procedure until the cost of the best tour can't be further improved.

In this program we have three algorithmic parameters that can be altered at each run of the program so as to vary the evolutionary strategies. The two parameters are Population and Mutation rate. The parameters go long way in determining the result of the algorithm. The program generates n random tours where n is the population size. These n tours are then sorted based on their fitness where the fitness function is basically the cost of tour. The best two tours gotten after sorting are mated to produce two new tours. And some randomly selected tours are also mutated. The worst tours are removed from the population and replaced with the new ones gotten from mating and mutation. This continues until best candidate can no longer be improved.

To test the algorithm a geometric city with 7 nodes is used. The optimal tour of the geometric city is 6 -> 5 -> 4 -> 3 -> 2 -> 1 -> 0 or 0 -> 1 -> 2 -> 3 -> 4 -> 5 -> 6, both have the same cost. Genetic Algorithm was tested and the result is shown on the screen capture figure 1 below.



Figure 1: Using the Algorithm to solve a Geometric City.

The program is used to solve a geometric city with clear optimal solution so as to be sure that the algorithm can arrive at optimal solution. As we can see in the figure 1 above, the path is optimal and the run time is fair at 218 milliseconds.

V. RESULTS AND DISCUSSIONS

The data used in this paper is the distances between Nigerian major cities. The data was used because it is an average sized problem with 31 cities and its distances are moderate. The data was stored in a text file which can be imported into the program by clicking the load button on the GUI as shown in figure 2.



Figure 2: Loading cities from text file

It is possible to alter the data in the file or add another different data. Table 1 shows the results of experiments carried out on the algorithm using different parameters. The table shows all the parameters used and the results. Performance of the result is based in the runtime and distance (cost).

TABLE 1 PARAMETERS AND RESULTS

Pop.	Mut. Rate	Cut Length	Run Time	Distance	Individuals
1000	0.1	0.2	546	5918.0	221000
1000	0.2	0.2	640	5896.0	226000
1000	0.3	0.2	748	5829.0	232000
1000	0.4	0.2	577	5886.0	192000
1000	0.5	0.2	577	5700.0	97000
1000	0.6	0.2	453	5981.0	190000
1000	0.7	0.2	577	5973.0	191000
1000	0.8	0.2	500	6100.0	195000
1000	0.9	0.2	608	5925.0	211000
1000	1	0.2	562	6010.0	209000
1000	0.01	0.2	532	9048.0	139000
100	0.1	0.2	31	10584.0	14400
100	0.2	0.2	47	10581.0	14600

100	0.3	0.2	31	11141.0	13600
100	0.4	0.2	31	12221.0	13500
100	0.5	0.2	32	10564.0	13900
100	0.6	0.2	32	9668.0	15200
100	0.7	0.2	31	9888.0	13900
100	0.8	0.2	31	10634.0	13700
100	0.9	0.2	31	11778.0	14000
100	1	0.2	31	10335.0	13000
100	0.01	0.2	46	9642.0	13600

It is important to note that change in the mutaton rate affects the runtime of the program. Figure 3 and 4 show the effect of mutation rate on runtime. Figure 3 is the plot using population size of 1000 while figure 4 illustrates the plot using a population size of 100.

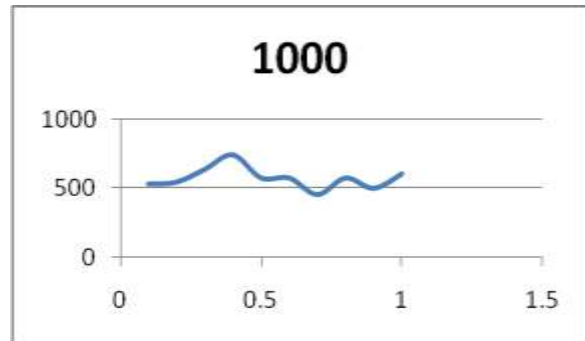


Figure 3: Plot of runtime against mutation rate for Population size of 1000

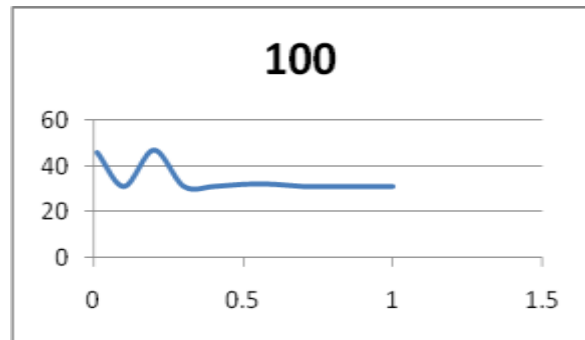


Figure 4: Plot of runtime against mutation rate Population size of 100

Another major thing to note about the algorithm is the number of Individuals which is the result of the population size and the ability to improve candidate solution. Also the more the number Individuals used the higher the likelihood of getting a better solution. Figure 5 shows the plot of Individuals against the distance of tour gotten.

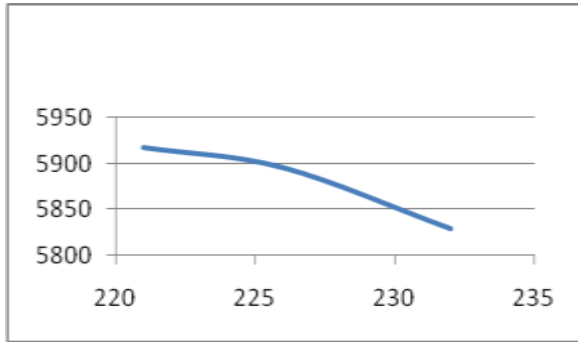


Figure 5: Plot of Individuals against distance

VI. CONCLUSION

We presented an efficient Genetic Algorithm program for solving the Travelling Salesman Problem. The program produced good results for various problem sizes but run time increases with increasing number of cities. We found that the population should be tuned to match the problem size (not arithmetically). To get very good solutions a tradeoff must be made between runtime and the solution quality.

REFERENCES

- [1] R. Anthony and E. D. Reilly (1993), Encyclopedia of Computer Science, *Chapman & Hall*.
- [2] U. Aybars, K. Serdar, C. Ali, Muhammed C. and Ali A (2009), Genetic Algorithm based solution of TSP on a sphere, *Mathematical and Computational Applications*, Vol. 14, No. 3, pp. 219-228.
- [3] B. Korte (1988), Applications of combinatorial optimization, *talk at the 13th International Mathematical Programming Symposium, Tokyo*.
- [4] E. L. Lawler, J. K. Lenstra, A. H. G. RinnooyKan, and D. B. Shmoys (1985), The Traveling Salesman Problem, *John Wiley & Sons, Chichester*.
- [5] H. Holland (1992), Adaptation in natural and artificial systems, *Cambridge, MA, USA: MIT Press*.
- [6] H. Nazif and L.S. Lee (2010), Optimized Crossover Genetic Algorithm for Vehicle Routing Problem with Time Windows, *American Journal of Applied Sciences* 7 (1): pg. 95-101.
- [7] R. Braune, S. Wagner and M. Affenzeller (2005), Applying Genetic Algorithms to the Optimization of Production Planning in a real world Manufacturing Environment, *Institute of Systems Theory and Simulation Johannes Kepler University*.
- [8] Z. Ismail, W. Rohaizad and W. Ibrahim (2008), Travelling Salesman problem for solving Petrol Distribution using Simulated Annealing, *American Journal of Applied Sciences* 5(11): 1543-1546.

Grid Approximation Based Inductive Charger Deployment Technique in Wireless Sensor Networks

Fariha Tasmin Jaigirdar
Dept. of Computer Science
Stamford University, Bangladesh
Dhaka, Bangladesh.
farihajaiigirdar@yahoo.com

Mohammad Mahfuzul Islam
Dept. of Computer Science &
Engineering
Bangladesh University of
Engineering & Technology (BUET)
Dhaka, Bangladesh.
mahfuz@cse.buet.ac.bd

Sikder Rezwanul Huq
Dept. of Computer Science &
Engineering
Islamic University of Technology
(IUT)
Dhaka, Bangladesh.
rezwan_cit_2005@yahoo.com

Abstract— Ensuring sufficient power in a sensor node is a challenging problem now-a-days to provide required level of security and data processing capability demanded by various applications scampered in a wireless sensor network. The size of sensor nodes and the limitations of battery technologies do not allow inclusion of high energy in a sensor. Recent technologies suggest that the deployment of inductive charger can solve the power problem of sensor nodes by recharging the batteries of sensors in a complex and sensitive environment. This paper provides a novel grid approximation algorithm for efficient and low cost deployment of inductive charger so that the minimum number of chargers along with their placement locations can charge all the sensors of the network. The algorithm proposed in this paper is a generalized one and can also be used in various applications including the measurement of network security strength by estimating the minimum number of malicious nodes that can destroy the communication of all the sensors. Experimental results show the effectiveness of the proposed algorithm and impacts of the different parameters used in it on the performance measures.

Keywords- *wireless sensor network; energy efficiency; network security; grid approximation; inductive charger.*

I. INTRODUCTION

With the rapid development of computing and communication technologies, Wireless Sensor Networks (WSNs) are now being used in various applications including environment monitoring, healthcare, complex and sensitive surveillance systems and military purposes. These applications, however, demands high level of system security and longer lifespan of the sensors. WSNs are self-organized ad-hoc networks consisting of a large number of tiny sensor nodes having small sized low capacity batteries and able to communicate over wireless media. The low power battery limits a sensor node in providing a desired level of processing and buffering capabilities and hence refrains from providing necessary securities [1] [2] [3].

Large sized batteries and expensive sensors can mitigate the low energy problem of WSNs. However, these solutions are not feasible because of the nature of WSN deployment

especially when it is deployed in an unfriendly environment like battlefield or underwater investigation. On the other hand, the sensors of a WSN run unassisted for a long time and they can neither be disposable nor its battery can be replaced or recharged. Another possibility of dealing with the low energy problem is to deploy new external sensors in the network region periodically to keep the network alive. Some locations of the network (e.g., deployment of sensor nodes inside a machine or in a dangerous/unreachable forest, etc) do not even permit to deployment of new sensors even if cost is ignored. Thus, recharging from a distance is the only effective and generalized approach for dealing with the power problem of WSNs. Many research works [4][5][6] have been proposed to keep WSNs alive or prolong the network lifetime by recharging the batteries using the distance based recharging technique, which is also called the inductive charging technology [7]. This technology is convenient, safe, efficient and green [8].

There are different ways of recharging the sensors inductively. Some research works have been proposed to recharge sensor nodes by extracting energy directly from the deployment environment. These are called “scavenging” techniques [4]. The scavenging techniques recharge the sensors by collecting energies from solar power [9], kinetic energy [10], floor vibration [11], acoustic noise and so on. Due to the requirement of large exposure area of sensing devices, these techniques are not feasible to accommodate in tiny sensors. Moreover, the power generated by using these techniques is not sufficient enough for sustaining the regular operation of sensor nodes. Recent research works investigate that mobile nodes can deliver power to the sensors inductively (i.e., cordlessly) by creating an electromagnetic channel [8]. This amount of power supplied by this technology is sufficient to meet the power requirement of TV, laptop, mobile phone, digital camera, PDA and so on[8]. Sensors nodes can collect sufficient powers from the infrequent visit of active mobile nodes and hence prolong the WSN lifetime.

Efficient and cost-effective deployment of active mobile nodes is one of the key problems for introducing inductive sensor charging technique. To the best of our knowledge, there

is no existing technique to solve this problem. In this paper, we propose a solution by finding the minimum number of active mobile nodes (also known as charging nodes (i.e., charger)) along with their locations, which can charge all the sensors of a network. Here, our goal is to maintain continuous energy supply to the sensing nodes by placing the chargers in appropriate locations that will gain the efficiency of the work and also for ensuring the cost effectiveness, our approach will find the minimum number of chargers as well. The solution seems close to that of the coverage problem, but in real sense it is totally different. The coverage problem schemes are mainly of two types: area coverage and target or point coverage. The area coverage schemes explore the solution to cover the entire area of a WSN, while point coverage schemes, a special case of area coverage problem, focus on determining the exact position of sensor nodes to provide efficient coverage application for a limited number of targets [12]. Finding the optimum location of active nodes is an NP-hard problem. Therefore, our solution focuses on finding out the best locations of placing active nodes by exploiting and merging the techniques of grid approximation [13], minimum set cover problem [14] and the greedy approach [15]. The experimental results clearly reveal that, by setting appropriate parameters, the proposed solution can efficiently find out the best locations of the chargers.

The rest of the paper is organized as follows: Section 2 discusses the background and related works of the problem, while the details of the proposed inductive charger deployment problem have been given in Section 3. The experimental setup, parameters and results along with comparative analysis has been given in Section 4. Some concluding remarks and future works are given in Section 5.

II. BACKGROUND AND RELATED WORKS

To the very best of our knowledge, the charger deployment strategy that we approach here is unique and does not directly relate to any other work. Moreover, to better understand the concepts and clarify the novelty of our work we discuss different coverage problems in this section. As we are concerning with the energy issue, different power factor related works are also discussed in this section. The coverage problem deals with the quality of service (QoS) that can be provided by a particular sensor network. The term *coverage* means how well a sensor network is monitored or tracked by the sensors. As the purpose of WSN is to collect relevant data for processing, it can be done properly by covering the overall network to achieve the required goal. Many research works [16] [17] have been proposed in this area and researchers define the problems from different angles and also gave different design view.

In this part, we have studied different coverage approaches and classified them considering the coverage concept. As stated before, the two main parts of coverage problems along with some design choices [18] have added here for better understanding the problem and its applications.

- **Sensor Deployment Method:** Deterministic versus random. A deterministic sensor placement is the method in which the location and the number of sensors needed are predetermined and is feasible in friendly, previously organized and accessible environments. But in the scenario of deployment in remote and inhospitable areas, deterministic node placement can be impractical for some extent and the solution is random sensor distribution where having knowledge about the number and location of sensor nodes will be ignored.
- **Communication Range:** An important factor that relates to connectivity is communication range, which can be equal or not equal to the sensing range.
- **Additional Critical Requirements:** Energy-efficiency, connectivity and fault tolerance.
- **Algorithm Characteristics:** centralized versus distributed/localized.

The most studied coverage problem is the *area coverage problem*. As an important research issue many researchers have been studied extensively on this topic and different sensor network applications have revealed a new era to solve their area coverage problem in different scenarios by varying design choices and other factors. To better understand the area coverage scheme and also how it differs from our work, here we have shown a brief discussion on area coverage.

As related to our energy efficiency issue, we have discussed here the power efficient coverage with random deployment. In the previous section, we have already discussed about power factors and consideration in WSNs and we have come to know that replacing the sensor node's battery is not feasible in many applications, so, approaches that has power conserve facility, are highly desirable. The works in [19] and [20] consider a large amount of sensors, deployed randomly for area monitoring. The contribution in [19] explains why energy is a factor of consideration in sensor network. In these papers, the goal is to achieve an energy-efficient design that maintains area coverage. As the number of sensors deployed is greater than the optimum required to perform the monitoring task, the solution proposed is to divide the sensor nodes into disjoint sets, such that every set can individually perform the area monitoring tasks. These set areas are then activated successively to perform the area monitoring task, and while the current sensor set is active, all other nodes are in a low-energy sleep mode to sustain the battery lifetime for further activations. The goal of this approach is to determine a maximum number of disjoint sets, as this has a direct impact on the network lifetime.

Another special case issue of area coverage is *point coverage*. An example of it showed in [21] has military applicability. In this paper the authors addressed the problem of energy efficiency in wireless sensor applications for surveillance of a set of limited number of targets with known locations. Here, a large number of sensors are dispersed

randomly in close proximity to the monitor the targets; and finally the sensors send the monitored information to a central processing node. The objective is that every target must be monitored at all times by at least one sensor and on the road to reach the required goal the authors have given a solution of preparing disjoint sets of sensor nodes and maintaining maximum number of such sets to be activated successively. Here, the authors prove that the disjoint set coverage problem is NP-complete and propose an efficient heuristic for set covers computation using a mixed integer programming formulation.

One more work of point coverage proposed in [22], where energy efficiency still maintained by covering targets with sensors. In this paper the author model the solution as the maximum set cover problem and design two heuristics that efficiently compute the sets, using linear programming and greedy approach. Here, in maximum set cover (MSC) definition, C denotes the set of sensors and R the set of targets, such that each sensor covers a subset of targets. In the greedy approach, at each step, a critical target is selected to be covered. This can be, for example, the target most sparsely covered, both in terms of sensors as well as the residual energy of those sensors. Upon selecting the critical target, the heuristic selects the sensor with the greatest contribution that covers the critical target. Once a sensor has been selected it is added to the current set cover and all additionally covered targets are removed from the TARGETS set, which contains the targets that still have to be covered by the current set cover. When all targets are covered, the new set cover was formed. Here simulation results are presented to verify the approaches.

III. OUR PROPOSED GRID APPROXIMATION STRATEGY

Our objective here is to strategically place the mobile chargers in spite of deploy them blindly to charge all the nodes in the network using inductive charging technique. Here, the word “strategically” means the way of deployment of the chargers that we have showed here to be best. Actually, a blind deployment strategy may show better result in some scenarios, but it will not always happen, and not guaranteed to be any time, moreover, it would take away the solution to an infinite searching that should not be our goal. By our proposed deployment strategy, it would be possible for us to find the best locations and as well as the minimum amount of charging nodes or chargers to charge all the sensor nodes in the network and also we would gain the efficiency and cost effectiveness of the deployment methodology and reach the required goal as well.

A. The Strategy

Here to cover (i.e., to charge) all the sensor nodes in the network, at first the search space for the solution will be identified. Obviously an infinite searching can achieve the optimum solution. However, to devise any practical implement able solution the search space needs to be reduced and made finite at the cost of optimality.

Our approach for making search space finite is the well known *grid approximation technique*. In this methodology, we will divide the entire network deployment area into grid points. The sensing nodes at first will be placed randomly over this deployment area. Then from every grid points, all the deployed sensing node's distances will be calculated by using the famous distance calculation equation. Finally these distances will be compared with the transmission range of the nodes. Note that a sensing node may be within the range of several grid points. As the target is to place the mobile chargers in grid points only so that all the sensor nodes can be charged and finding the minimum number of such grid points, hence the problem can be easily mapped to a minimum set cover problem, because, in minimum set cover problem the target is to find the minimum number of sets to cover every element and that is actually our required goal. Greedily, we can at first place a mobile charger to the grid point which can cover maximum number of sensing nodes, then remove all the covered nodes and reapply the greedy approach until all the sensing nodes are being charged. Here, we use the greedy approach, as algorithms of this type often leads to very efficient and simple solution [15] and that tends us to fulfill our desired solution. Upon finding the first grid point that will cover the maximum number of sensing nodes in its transmission range, we will keep the location of that point also to find out the position of the first charger and processing in this way we will find all the chargers position and the number of chargers needed to charge all the sensing nodes in the network using inductive charging technique.

B. Mathematical Representation

Given a WSN of N sensing nodes deployed in a two dimensional field of size $D \times D$, where the transmission range of every node is considered as R . Here, we assume that we have full knowledge about the network topology in the way that each node has a low-power Global Position System (GPS) receiver, which provides the position information of the node itself. If GPS is not available, the distance between neighboring nodes can be estimated on the basis of incoming signal strengths. To cover a wireless node v (i, j), a mobile node or charger j (x, y) should be placed inside an R radius circle centered at v . Here, at first the distance between j and v will be calculated using Eq. (1)

$$\text{dis}(j, v) = \sqrt{(x-i)^2 + (y-j)^2}. \quad (1)$$

Now, this value of distance will be compared with the transmission range of nodes, i.e., if $\text{dis}(j, v) < R$, we can say that by placing a charging node or charger at grid point $j(x, y)$ we can charge the node $v(i, j)$. We will continue on finding the distance of all the sensor nodes from all the grid points in the total networking topology and hence we will apply our greedy approach, that is, upon finding the grid point that will have maximum number of sensing nodes in its transmission range, we will place our first charger to that point and remove the sensors (that have been charged by that point) from the entire set of sensing nodes by marking them charged and

reapply the greedy method to find the minimum number of chargers needed to charge all the sensing nodes in the network. Here, our goal is to find minimum number of grid point $j(x,y)$ where to place the chargers to charge all the sensors N in the network. So, it can be easily expressed by the following equation, i.e., Eq. (2).

$$\min(j(x,y)) = V(\text{all the nodes, } N). \quad (2)$$

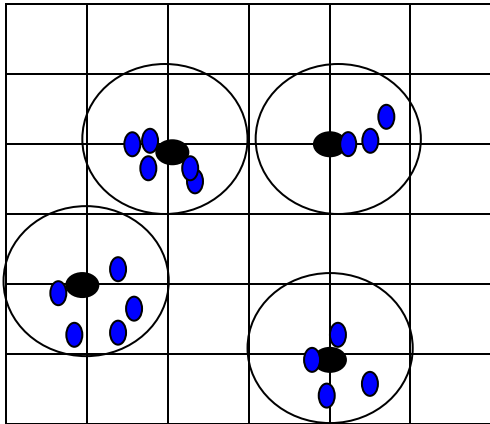


Figure 1. Minimum number of chargers needed to charge all the nodes in the network using Grid Approximation strategy.

The scenario can be understood easily from Fig. 1. In the figure, the blue circles represent the sensor nodes and the black circles represent the mobile charging nodes or chargers that we will deploy strategically by using our Grid Approximation strategy. The outline of the black nodes represents their transmission range. Here, our target is to deliver continuous energy supply to the sensing nodes and we have decided to use inductive charging technique for this purpose. If we place the external chargers which we call active mobile nodes blindly to charge all the nodes in the network it is possible that we would need seventeen chargers as in this case to charge a sensor we would need one charger. On the contrary, we can see from Fig. 1, if we use the Grid Approximation technique, to charge all the seventeen sensing nodes, we need only four chargers and it is also possible that the number of chargers will be lessened for different network topology. Here, the positions of the chargers are (1, 2), (2, 4), (4, 1) and (4, 4).

Here, in Grid Approximation technique, we have taken unit distance between the grid points and that we call step size 1. Actually a unit distanced approximation technique will give a best result of our problem but the processing steps and time will increase, which is not desirable. So we tried to increase the distance between the grid points by a pruning strategy, which will guarantee not to leave any points inside the working area and also reduce the processing steps and time as it will help in solving the brute force technique that we stated before. To cover a square shaped area, which we use here just to represent a small structural view of a grid, and here it is

recommended that no small point can be missed, we can find our desired step size value or maximum distance between the grid points by placing four circles in four corners. In the Fig. 2, we can see that if the sum of the radius of two circles (that are placed in opposite corner of a square) is exactly equals the diagonal measurement of the square area, and then it is guaranteed not to leave any point inside the area. Moreover, it is the maximum distance between the grid points as in this scenario the maximum value of the diagonal can be $2R$, where R is the radius of the circles. In Fig. 2, the maximum value of the diagonal PQ can be $2R$, so using Pythagoras theorem [23], we can find our desired value and the value is derived in Eq. (3). Here we consider the value of P to X -axis and P to Y -axis is same and that is X , so by Pythagoras theorem we can write, $X^2 + X^2 = (2R)^2$. So,

$$X = \sqrt{2}R. \quad (3)$$

Hence, the highest step size or maximum distance between two grid points can be $\sqrt{2}R$, without leaving any point of interest within the network area untouched.

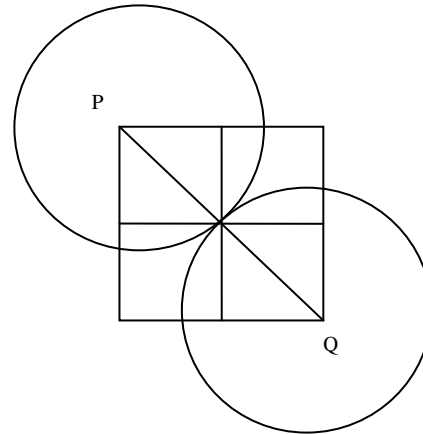


Figure 2. The maximum distance between two grid points to cover all sensors.

C. Algorithm

We have placed our proposed Grid Approximation algorithm in Fig. 3. The different parameters used in the algorithms are number of nodes, step size value, transmission range, maximum value in X -axis and Y -axis. Here, we start by considering that all the sensing nodes are uncharged at first iteration and the algorithm will stop its iteration upon charge all the sensing nodes. In this algorithm, we have made a comparison between the distance and transmission range or radius of sensing nodes. We have counted the number of neighbors a grid point has by considering the transmission range, and hence our algorithm's goal is to find the grid point that has highest number of sensing nodes as its neighbors. This is the point that we call `high_freq(x, y)` in our algorithm. Upon finding the point we will place our first charger at that point

and remove all its neighbors of that point marking as being charged and will continue the algorithm with the sensing nodes that have not been charged yet. We reapply the strategy until all the sensing nodes are being charged. So, finally we get the minimum number of chargers and their locations that we need to charge all the sensing nodes in the network.

```
Algorithm Grid (NN, R, XMAX, YMAX, Δ)
//NN=Number of sensor nodes,
//num_not_charged_yet=number of not_charged
//nodes, high_freq=highest number of nodes covered,
//high_freq_x, high_freq_y=represents the coordinates
//of highest frequency point, i.e., the points that have
//covered maximum number of nodes in its
//transmission range, XMAX=maximum dimension in
//x-axis, YMAX=maximum dimension in y-axis,
//R=transmission range or radius, Δ= distance between
//the grid points.
num_not_charged_yet=NN
while num_not_charged_yet>0 do
    high_freq:=0
    for i := 0 to YMAX step size Δ
        for j := 0 to XMAX step size Δ
            count:= 0
            for k := 0 to num_not_charged_yet
                if distance (i, j ,
                    nodes[uncharged_nodes [k]][0],
                    nodes[uncharged_nodes
                        [k]][1])<=R
                    count++
            if count>high_freq
                high_freq:=count
                high_freq_x:=j
                high_freq_y:=i
//Here, we will find the highest frequency point in the
//grid that will be the point, where we need to place
//our first charging node to charge all the nodes in the
//network.
i:=0
for j := 0 to num_not_charged_yet
    if distance(high_freq_x, high_freq_y,
        nodes[uncharged_nodes[j]][0],
        nodes[uncharged_nodes[j]][1])>R
        temp[i++]:=uncharged_nodes[j]
    for j:=0 to i
        uncharged_nodes[j]:=temp[j]
```

Figure 3. Grid Approximation algorithm

D. Complexity Analysis

Here, we assume that the Dimension or Grid size, $G=M*M$, Number of sensors= N , C_D = Cost for calculating distance, C_G = Cost of each comparison. As mentioned above, we can say that there are mainly two steps in this algorithm. In the first step, we have to find out the complexity intended for finding the distance of all sensor nodes from all grid points. Complexity in this step is $GN C_D$. In second step, the algorithm will go by finding the point that will cover the maximum number of sensing nodes, for each comparison, and the complexity is GC_G . These steps must be regulated for each sensor nodes or grid points. So, In **Worst case**, the complexity is: $G (N C_D + C_G) * \min (G, N)$. In **Best case**, the complexity is: $G (N C_D + C_G)$. In **Average case**, the complexity is: $\frac{1}{2} * G (N C_D + C_G) * \min (G, N) = \frac{1}{2} * G (NP C_G + C_G) * \min (G, N)$ [using $C_D = PC_G$, (where P is a given value)], so the average complexity finally is $\frac{1}{2} * GC_G (NP+1) * \min (G, N)$.

IV. EXPERIMENTAL RESULTS

In this section, we apply the approximation algorithm on various network topologies to demonstrate the algorithm's efficiency. Different experimental results have shown here by verifying different parameters of the network. In our work, the changing parameters for better understanding the scenario are transmission range R , i.e., a sensor's covering range of receiving and transmitting signals, number of sensing nodes N and total networking area or dimension, D . As we have deployed the sensors randomly in the network, for having the best result and gain the accuracy, we have taken the same snapshot of experimental result for five times and finally have taken the average.

We have showed our experimental results in three charts by verifying the concerning parameters which has shown in Fig. 4, 5 and 6. In these figures we consider the distance between the grid points to be unit and have showed the other parameters behavior to better understand the scenario. In Fig. 7, 8 and 9 we have shown the impacts of changing the different distances between the grid points. In Fig. 4, we have shown the result for the dimension $50*50$, whereas in Fig. 5 and Fig. 6 that value will be changed to $150*150$ and $200*200$. The number of nodes used in this paper differs from 50 to 300 and if necessary can be expanded as well. While taking different values of transmission range we would prefer to maintain a margin in low, medium and high range of measurement, and so, we preferred to take the values as 25, 40 and 70 in Fig. 4, 5 and 6.

Here, we can see that for different dimensions and number of nodes, when the radius or transmission range of the nodes increases, the amount of charger requires decreases. This is because, with increasing radius more nodes can be in the range of the grid points (i.e., by placing a charger at that point we can charge all the sensors that are within the transmission range of the point) and number of charger is lessening accordingly. Here, it can be noted that the charger required is minimum for highest valued transmission range, i.e., 70 in every charts and lowest for the low valued transmission range, 25.

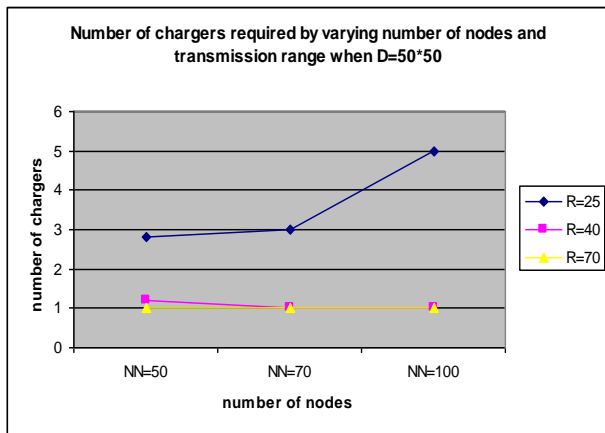


Figure 4. Results by changing transmission range and number of nodes for dimension 50

Another important parameter of the network is number of sensing nodes, N . In most of the cases, we need a sensor network with maximum number of nodes deployed in the network for establishing WSNs different kinds of applications. If we have a look in the Fig. 4, 5 and 6, we can see that, in most of the cases, as number of sensing nodes increase the number of chargers require increase accordingly. This is because, with higher number of nodes, more charging nodes are necessary to embrace them meeting the corresponding criteria.

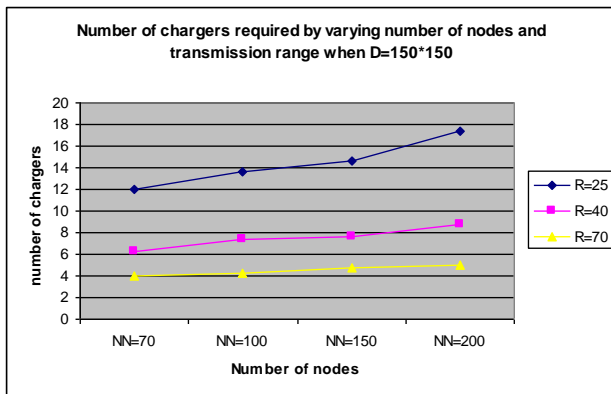


Figure 5. Results by changing transmission range and number of nodes for dimension 150

A last concerning criterion of the network is its dimension, D . In a large area, where nodes are placed randomly it is normally happens that they are placed in a scattered manner and that's why more chargers are required to charge all the sensing nodes in the network for such a network. The Fig. 4 is designed for low sized area network, 50*50, whereas in Fig. 5 and 6 we take different results for the area 150*150 and 200*200 accordingly. In Fig. 4, with the transmission range 25 and number of sensing nodes 100, we can see that the numbers of chargers required are approximately 5. Whereas, for the same parameters in Fig. 5, which is designed for dimension

150, the chargers required are approximately 13 and in Fig. 6 it is 18 for dimension 200.

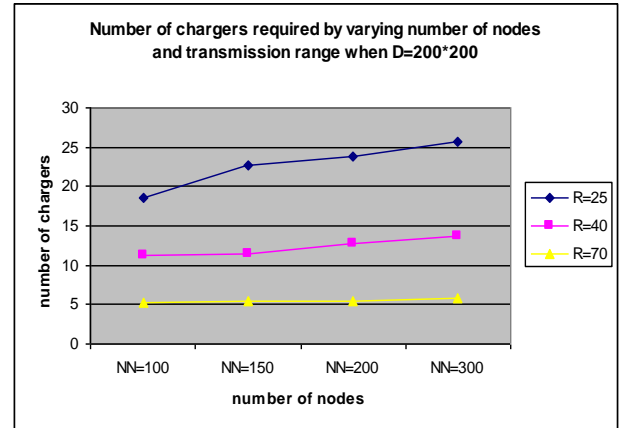


Figure 6. Results by changing transmission range and number of nodes for dimension 200

Now, we will discuss about changing the step size, i.e., the distance between the grid points that have shown in Fig. 7 and 8. In Fig. 7, we have shown the result for different number of sensing nodes, whereas in Fig. 8, we have changed the values of transmission range to show different results. We can see in Fig. 7, the value of R has taken 20 here, so the highest value of the step size as stated before is $\sqrt{2} \cdot R$, means $\sqrt{2} \cdot 20$, i.e., 28. Here, we can see that for different step sized value, as the distance between the grid points increase, the amount of chargers needed also increase. The reason is, as the distances between the grid points increase, the grid points are placing in far distances than previous unit distanced grid point placement and their searching area also increase accordingly, whereas, in lower step size valued grid, the more grid points are placed in close distance and checked accordingly and thus the number of chargers required are minimum for such a scenario. Moreover, to gain the efficiency, processing steps and time need to be low and for this reason it is necessary to increase the step size of the grid approximation.

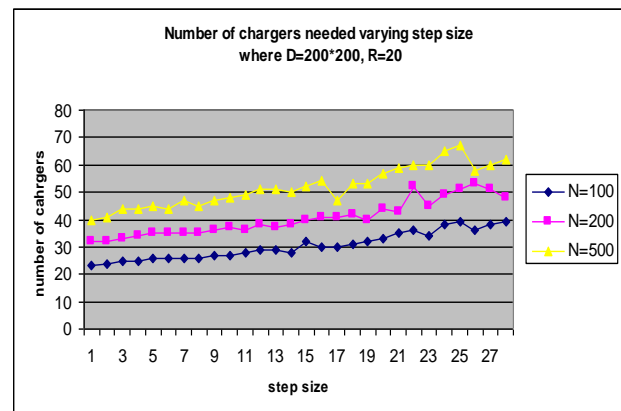


Figure 7. Results for different step size value by changing N and keeping D and R fixed.

In Fig. 8, we have taken same consideration for different transmission range and here, we have kept the number of nodes to fix. Here, we can see that for higher valued transmission range, the numbers of chargers needed are low compare to the lower valued transmission range. The reason behind this is same as stated before for Fig. 4, 5 and 6.

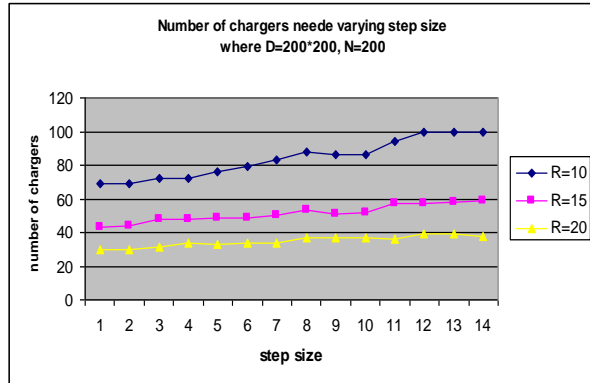


Figure 8. Results for different step size value by changing R and keeping D and N fixed.

V. CONCLUSION AND FUTURE WORK

Limitation of continuous energy or power supply is a prime concern in Wireless Sensor Networks (WSNs). To meet different applications of WSNs, it is very important for the sensors to have the longer lifespan. In this paper, we propose a solution to power limitation problem of WSN by an inductive charger deployment scheme to charge all the sensing nodes in the network using inductive charging technique. Here, we present a grid approximation algorithm to find out a deployment strategy for chargers with the goal to achieve the cost effectiveness and efficiency. The proposed algorithm will find out the location as well as the minimum number of chargers needed to charge all the sensing nodes in the network in order to maintain a continuous energy supply in the network which will help a sensor network to meet its various applications in an unfriendly environment. Moreover, the algorithm is designed for different step sized valued grid approximation and hence ensure the reduction of processing steps and time to make the algorithm much stronger and flexible. Different experimental results by changing different parameters have shown the strength of the algorithm for the proposed scheme.

As an advance to our work, in future, we have desire to work on different deployment approaches by developing more strong and innovative algorithms to solve the energy limitation problem of WSNs. Moreover, as our proposed algorithm is a generalized one, we have plan to expand our idea in the field of security for calculating minimum number of malicious nodes necessary to corrupt or jam the overall network and with this regard to measure a network strength of security. Moreover, we have aim to explore some more methodologies to implement the concept of this paper in real world and also explore for

intelligent agent based deployment policies to achieve the goals.

REFERENCES

- [1] A. Azim and M. M. Islam, "Hybrid LEACH: A relay node base low energy adaptive clustering hierarchy for wireless sensor networks", *IEEE 9th Malaysia International Conference on Communications (MICC)*, pp.911-916, 2009.
- [2] A. Azim and M. M. Islam, "Dynamic service policy-based clustered wireless sensor networks", *IEEE 6th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pp. 196-202, 2010.
- [3] A. Azim and M. M. Islam, "A dynamic round-time based fixed low energy adaptive clustering hierarchy for wireless sensor networks", *IEEE 9th Malaysia International Conference on Communications (MICC)*, pp.922-926, 2009.
- [4] W. Yao, M. Li and M-Y Wu, "Inductive charging with multiple charger nodes in wireless sensor networks", *APWeb Workshops, LNCS 3842*, pp 262-270, 2006.
- [5] Chevalerias, O. O. Donnell, T. Power, D. O Donovan, N. Duffy, G. Grant and G. O Mathuna, "Inductive telemetry of multiple sensor modules", *Pervasive Computing*, Vol. 4, pp. 46- 52, 2005.
- [6] A. LaMarca, D. Koizumi, M. Lease, S. Sigurdsson, G. Borriello, W. Brunette, K. Sikorski and D. Fox, "Making sensor networks practical with robots," *Lecture Notes on Computer Science (LNCS)* 2414, pp. 152-166, 2002.
- [7] What is Inductive Charging, <http://www.wisegEEK.com/what-is-inductive-charging.htm>, last visited: 05 January 2011.
- [8] Splashpower Wireless Power in now eCoupled, <http://www.splashpower.com>, last visited: 05 January 2011.
- [9] J. M. Kahn, R. H. Katz and K. S. J. Pister. "Next century challenges: mobile networking for smart dust", *Proc. MobiCom*, pp. 271-278, 1999.
- [10] J. Paradiso, M. Feldmeier, "A Compact, Wireless, Self-Powered Pushbutton Controller", *Proc. International Conference on Ubiquitous Computing*, pp. 299-304, 2001.
- [11] S. Meninger, J. O. Mur-Mirands, R. Amirtharajah, A. P. Chandrakasan and J. H. Lang, "Vibration-to-electric energy conversion", *Proc. ACM/IEEE International Symposium on Low Power Electronics and Design*, pp. 48-53, 1999.
- [12] F. GaoJun and J. ShiYao, "Coverage Problem in Wireless Sensor Network: A Survey", *Journal of Networks*, Vol.5, No.9, September 2010.
- [13] Ramachandran, "Real-atom grid approximation (RAGA) a new technique of crystal structure analysis using only amplitudes without determination of phases of reflections", *Acta Crystallographica, Section A, Foundations of Crystallography*, 46 (5). Pp 359-365.
- [14] Minimum Set Cover, <http://www.sprklab.com/notes/16-minimum-set-cover/>, last visited: 04 January 2011.
- [15] Greedy Approach, http://ist.marshall.edu/ist238/greedy_appr.html, last visited: 04 January 2011.
- [16] M. Cardei and D. -Z. Du, "Improving wireless sensor network lifetime through power aware organization", *Wireless Networks 11*, pp. 333-340, 2005.
- [17] F. Ye, G. Zhong, S. Lu and L. Zhang, "Energy efficient robust sensing coverage in large sensor networks", *Technical Report ULCA*, 2002.
- [18] M. Cardei and J. Wu, "Coverage in wireless sensor networks", Department of Computer Science and Engineering, Florida Atlantic University, Boca Raton, FL 33431, www.cse.fau.edu/~jie/coverage.ps, last visited: 05 January 2011.
- [19] S. Slijepcevic and M. Potkonjak, "Power efficient organization of wireless sensor networks", *Proc. of IEEE International Conference on Communications*, Vol-2, pp 472-476, Helsinki, Finland, June 2001.

- [20] M. Cardei, D. MacCallum, X. Cheng, M. Min, X. Jia, D. Li and D.-Z. Du, "Wireless sensor network with energy Efficient Organization", *Journal of Interconnection Networks*, Vol 3, No 3-4, pp 213-229, Dec 2002.
- [21] M. Cardei and D. -Z. Du, "Improving wireless sensor network lifetime through power aware organization", in press.
- [22] M. Cardei, M. T. Thai, Y. Li and W. Wu, "Energy-Efficient target coverage in wireless sensor networks", *In Proc. of INFOCOM*, pp.1976-1984, 2005.
- [23] Pythagoras Theorem, <http://www.mathsisfun.com/pythagoras.html>, last visited: 05 January, 2011.

AUTHORS PROFILE

Fariha Tasmin Jaigirdar is working in Stamford University, Bangladesh as a Lecturer in the dept. of Computer Science. She has completed her Bsc Engineering from Chittagong University of Engineering and Technology

(CUET) securing 5th position in merit list. Now She is doing her Msc in Bangladesh University of Engineering & Technology (BUET). She is young, energetic and hard working. Her research interest includes wireless networks, digital signal processing and bangla language display in digital form. Recently she is working on various research works to prove her best with her academic excellence as well as her analytical ability.

Mohammad Mahfuzul Islam is working in Bangladesh University of Engineering & Technology (BUET) as an Associate Professor in the dept. of Computer Science & Engineering. He has completed his Phd from Australia. He has a huge amount of publications in international journals and conferences. His research interest is wireless resource management, network security, image processing and artificial intelligence.

Sikder Rezwanaul Huq is working in Stamford University Bangladesh as a Lecturer in the dept. of Computer Science. He has completed his Bsc Engineering from Islamic University of Technology (IUT). His research interest is grid computing and wireless networks.

PAV: Parallel Average Voting Algorithm for Fault-Tolerant Systems

Abbas Karimi^{1,2,*}, Faraneh Zarafshan^{1,2}, Adznan b. Jantan²

¹ Department of Computer Engineering, Faculty of Engineering, Islamic Azad University, Arak Branch, Iran

² Departments of Computer and Communication Systems Engineering, Faculty of Engineering, UPM, Malaysia

*Akarimi@ieee.org

Abstract—Fault-tolerant systems are such systems that can continue their operation, even in presence of faults. Redundancy as one of the main techniques in implementation of fault-tolerant control systems uses voting algorithms to choose the most appropriate value among multiple redundant and probably faulty results. Average (mean) voter is one of the commonest voting methods which is suitable for decision making in highly-available and long-missions applications in which the availability and speed of the system is critical. In this paper we introduce a new generation of average voter based on parallel algorithms which is called as parallel average voter. The analysis shows that this algorithm has a better time complexity ($\log n$) in comparison with its sequential algorithm and is especially appropriate for applications where the size of input space is large.

Keywords- Fault-tolerant; Voting Algorithm; Parallel- Algorithm; Divide and Conquer.

I. INTRODUCTION

Fault-tolerance is the knowledge of manufacturing the computing systems which are able to function properly even in the presence of faults. These systems compromise wide range of applications such as embedded real-time systems, commercial interaction systems and e-commerce systems, Ad-hoc networks, transportation (including rail-way, aircrafts and automobiles), nuclear power plants, aerospace and military systems, and industrial environments in all of which a precise inspection or correctness validation of the operations must occur (e.g. where poisonous or flammable materials are kept)[1]. In these systems, the aim is to decrease the probability of system hazardous behavior and keep the systems functioning even in occurrence of one or more faults.

One of the mechanisms to achieve fault tolerance is fault masking which is used in many fault-tolerant systems [2]. In fault masking, hardware modules or software versions are replicated and then voting is used to arbitrate among their results to mask the effect of one or more run time errors.

Replication of hardware modules is the most applicable form of hardware redundancy in control systems which can be in forms of passive (static), active (dynamic) and hybrid.

The aim in static redundancy is masking the effect of fault in the output of system. N-Modular Redundancy (NMR) and N-Version Programming (NVP) are two principal methods of

static redundancy in hardware and software respectively. Three modular redundancies (TMR) is the simplest form of NMR which is formed from $N=3$ redundant modules and a voter unit which arbitrates among modules' outputs (figure 1).

Voter performs a voting algorithm in order to arbitrate among different outputs of redundant modules or versions and mask the effect of fault(s) from the system output. Based on the application, we can use different types of voting algorithms.

Average voter is one of several voting algorithms which are applied in fault-tolerant control systems. Main advantages of this voter are its high availability and its potentiality to extend to large scale systems. Furthermore, in contradict with many voters like majority, smoothing and predictive; it does not need any threshold. The main problem of this voter is that whatever the number of inputs increases, the complexity of its formula increases. Hence, more calculations overhead imposes and the processing speed will decrease. In this paper, we use parallel algorithms on EREW shared-memory systems to present a new generation of average voter – we call as parallel average voter- which provides the average voter extension without enlarging the calculations, suitable for large scale systems and with optimal processing time. Basically there are two architectures for multi-processor systems. One is shared-memory multi processor system and the other is message passing[3]. In a shared-memory parallel system it is assumed n processor has either shared their public working space or has a common public memory.

The current paper is organized as follows: in section 2, background and related works are described. In Section 3, the sequential average voting algorithm and the parallel average voting algorithms are presented. Section 4, deals with performance analysis of new parallel algorithms and its comparison with sequential algorithm. Finally, the conclusions and future works are explained in section 5.

II. RELATED WORKS

Voting algorithms have been extensively applied in situations where choosing an accurate result out of the outputs of several redundant modules is required. Generalized voters including majority, plurality, median and weighted average have been first introduced in [4].

Majority voter is perhaps the most applicable voter that chooses a module output as the output of voting if majority of voter inputs has been produced that value but if less than majority of modules are in agreement, plurality voter can make an agreement. Plurality and majority are actually extended forms of m-out-of-n voting in which at least m modules out of n modules should be in agreement; otherwise, voter cannot produce the output. This voting method is a suitable choice for the systems where the number of voter inputs is large. The other generalized voter is median voter that always chooses the mid-value of voter inputs as the system output. The most significant limitation of this algorithm is that the number of the voter inputs is assumed to be odd [4]. In weighted average algorithm, the weighted mean of the input values is calculated as the voting result. The weight value is assigned to each voter input in various methods [2, 4-6], then, calculated weights, w_i , are used to provide voter output, $y = \sum w_i \cdot x_i / \sum w_i$, where x_i s are the voter inputs and y is the voter output. Average voter is a special case of weighted average voter in which all weights are assumed to be equal to $\frac{1}{n}$. In two latest methods, the voting results may be clearly different from input values, while some voters like majority, plurality and median always choose a value among their input values as the voter output.

One difficulty with majority voter and alike is their need to threshold, while so far not any general approaches have been achieved to calculate fair value of them; however, average voter is free of this issue. Furthermore, average voter can always produce output. So the availability of this voter and voter's alike including median and weighted average is 100 percent which makes them the choicest voters for highly available missions.

One critical issue about the voters is their performance in large scale systems. In [7, 8], the above mentioned algorithms along with their operation and time complexity for small and large number of inputs are analyzed and it has demonstrated that the complexity of them depends on the structure of the input space. The main problem with all the weighted methods and consequently average voter is the increasing in the complexity of voter output calculations while the number of voter inputs increases. It also has harmful effects on speed of processing in control system.

To address this problem for average voter, by using parallel algorithms, we have proposed an effective parallel average algorithm based on shared memory EREW. So far, parallel voters have not been taken into account and only two references [2, 9] have covered this issue. In [3], an efficient parallel algorithm has been proposed to find the majority element in shared-memory and message passing parallel systems and its time complexity was determined, while an

approach for parallelized m-out-of-n voting through divide-and-conquer strategy has been presented and analyzed in [9].

III. SHARED MEMORY SYSTEM PARALLEL ALGORITHM

In this section, we propose an optimal parallel average voting algorithm on EREW shared-memory systems for large object space applications such as public health systems, geographical information systems, data fusion, mobile robots, sensor networks, etc.

First, we introduce sequential average voting. Then we proceed with introducing and describing the parallel average algorithm with inspirations from the functions of this algorithm and using Divide-and-conquer method and Brent's theorem [10-12].

A. Sequential Average Voting

As mentioned in the previous section, in sequential average voting, the mean of the modules output will be chosen as the output. This will be simply gained through the Lorzak relation mentioned in [4] considering $w_1 = w_2 = \dots = w_n = \frac{1}{n}$, provided in (1) in which x_i is output of the i^{th} redundant module; w_i , weight of i^{th} module; and X is the output of voter.

$$X = \frac{\sum_{i=1}^n x_i w_i}{\sum_{i=1}^n w_i} \quad (1)$$

B. Parallel Average Voting

In this section, an effective parallel algorithm is presented for calculating average voting in PRAM machines with EREW shared-memory technology. To do so, the following assumptions are taken into account:

- Array A [1...n] with n elements, comprises a_1, a_2, \dots, a_n , where each a_i is the output of i^{th} module.
- Number of redundant modules, n, is considered as the power of 2.
- Array A is divided to $p = \frac{n}{2}$ sub-arrays each of which contains at most $\log n$ element.
- We assume $\sum w_i = 1$, $\sum w_i * x_i = \sum x_i$.
- For enhancing the algorithm, the number of required processors is assumed equal to the number of sub-arrays i.e. p.

The Pseudo-code of our optimal parallel average voter is presented in fig. 2.

Procedure PAV (PRAM-EREW)

Input: A is an array of n elements a_1, a_2, \dots, a_n where n is a power of 2.

Output: Return X as the output of parallel average voting.

1. A is subdivided into $p=n/2$ subsequences A_i of length $\log n$ where $1 \leq i \leq n$;
2. $\forall i \in [1 .. n/2]$
Create array A with corresponding elements a_i in Parallel.
3. **End Par.**
4. $j \leftarrow p$;
5. **While** $j > 1$ **do**
6. **For** $i=1$ **to** j **Do in Parallel**
7. $A[i] \leftarrow A[2i-1] + A[2i]$;
8. **End Par.**
9. $j \leftarrow j/2$;
10. **End While.**
11. $X \leftarrow A[1]/n$;
12. **End.**

Figure 1: Pseudo- code of parallel average voter

IV. ANALYSIS AND COMPARISON

In this section, step by step, we try to analyze both parallel and sequential average voting algorithms introduced in sections 3.A and 3.B through using the rule of complexity of the computation of the algorithms in order to highlight the efficiency of the new parallel algorithm.

To describe the time complexity of the two algorithms we define $T_s(n)$, the function of executing time of the average sequential voting algorithm and $T_p(n)$, the function of the executing time of the parallel voting algorithm in which p is the number of the processors.

Definitely as a result of using \sum operator, sequential average voter needs time complexity equal to $T_s(n) = O(n)$, while parallel algorithm needs constant time of $O(1)$ to divide array A into sub-arrays having maximal length of $(\log n)$. Line 2 of PAV uses $O(\log n)$ time in order to copy and transfer the information. Since in lines 4-10, we do calculation (adding odd and even nodes) in each sub-array by using tree structure, the overall time complexity of these lines will be equal to $O(\log n)$. Finally in line 11, we need an $O(1)$ time to calculate the average voting output.

Hence, the total time complexity of our parallel average voting algorithm is:

$$T_p(n) = O(\log n). \quad (2)$$

By comparing the time complexities of sequential and parallel algorithms we can conclude that since the execution time of parallel average voter is logarithmic, it is able to run faster than sequential average voter. Also, it can be seen obviously that the total number of required processors in parallel algorithm does not exceed $\frac{n}{2}$. So taking into account

the execution time and number of processors needed, the cost and time complexity of the proposed algorithm is better than sequential algorithm. We also have good Speedup (S_p) and Efficiency (E_p) which are indicated in equations (3) and (4).

$$S_p = \frac{T_s}{T_p} = \frac{n}{\log n} \quad (2)$$

$$E_p = \frac{S_p}{P} = \frac{n / \log n}{n / 2} = \frac{2}{\log n} \quad (3)$$

For large scale system i.e. for big n we have good speed up and efficiency.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed an effective parallel algorithm for finding average voting among the results of n redundant modules in parallel shared-memory systems in EREW model. As seen in section 3 the execution time of the sequential algorithm is linear whereas it is logarithmic in our proposed parallel algorithm. Since the parallel average voter can always make result, it has more availability than other parallel voters including parallel majority and parallel m-out-of-n.

Furthermore, in contradict with many voters like majority, smoothing and predictive; it doesn't need any threshold. It also resolves the problem associated with sequential average voter in dealing with large number of inputs.

This algorithm can be implemented in future on parallel on Bus, Hyper Cube and Mesh typologies in message passing systems. Additionally, it can be developed for generating parallel Weighted Average Voting algorithm in which the weights are unequal.

REFERENCES

- [1] G. Latif-Shabgahi, *et al.*, "A Taxonomy for Software Voting Algorithms Used in Safety-Critical Systems," *IEEE Transactions on Reliability*, vol. 53, pp. 319- 328, 2004.
- [2] G. Latif-Shabgahi, "A Novel Algorithm for Weighted Average Voting Used in Fault-Tolerant Computing Systems," *Microprocessors and Microsystems*, vol. 28, pp. 357-361, 2004.
- [3] C.-L. Lei and H.-T. Liaw, "Efficient Parallel Algorithms for Finding the Majority Element," *Journal of Information Science and Engineering*, vol. 9, pp. 319-334, 1993.
- [4] P. R. Lorzak, *et al.*, "A Theoretical Investigation of Generalized Voters for Redundant Systems," in *FTCS-19. Digest of Papers., Nineteenth International Symposium on Fault-Tolerant Computing.*, Chicago, USA, 1989, pp. 444-451.
- [5] Z. Tong and R. Y. Kain, "Vote Assignments in Weighted Voting Mechanisms," *IEEE Transactions on Computers*, vol. 40, pp. 664-667, 1991.

- [6] G. Latif-Shabgahi, *et al.*, "A Novel Family of Weighted Average Voters for Fault Tolerant Computer Systems," in *Proceedings of ECC03: European Control Conference*, Cambridge, UK, 2003.
- [7] B. Parhami, "Voting Algorithms," *IEEE Transactions on Reliability*, vol. 43, pp. 617-629, 1994.
- [8] B. Parhami, "Optimal Algorithms for Exact, Inexact and Approval Voting " presented at the 22nd International Symposium on Fault-Tolerant Computing (FTCS-22), Boston, N.A, USA, 1992.
- [9] B. Parhami, "Parallel Threshold Voting," *The Computer Journal*, vol. 39, pp. 692-700, 1996.
- [10] R. P. Brent, *Algorithms for Minimization without Derivatives*. Englewood Cliffs, NJ.: Prentice-Hall, 1973.
- [11] P. B. Richard, "The Parallel Evaluation of Arithmetic Expressions in Logarithmic Time," ed: Academic Press, 1973.
- [12] R. P. Brent, "The Parallel Evaluation of General Arithmetic Expressions," *J. ACM*, vol. 21, pp. 201-206, 1974.

AUTHORS PROFILE



Abbas Karimi was born in Ahwaz, Iran, in 1976. He received the B.S. degree and M.S. degree in computer hardware and software engineering from Iran. He is currently Ph.D. candidate of computer system engineering, UPM, Malaysia. He has been working as a lecturer and a faculty member in the department of computer engineering in I.A.U-Arak Branch. He was leader of multiple research projects, and author of three textbooks, multiple journals and conference papers. He is senior members of IACSIT, member of IEEE, IAENG, SDIWC, WASET and reviewer in multiple journals. His research interests include load balancing algorithms, and real time, distributed, parallel and fault-tolerant systems.



Faraneh Zarafshan was born in Ahwaz, Iran. She received the B.S. degree and M.S. degree in computer hardware engineering from Iran. She is currently Ph.D. candidate of computer system engineering, UPM, Malaysia. She was leader of multiple research projects, author of three textbooks, multiple journals and conference papers. She is senior members of IACSIT, and member of SDIWC. Her research interests include sensor network, real time systems, and fault-tolerant systems.



Adznan b. Jantan Obtained his PhD from University College of Swansea, Wales, UK. He is currently associate professor in Universiti Putra Malaysia (UPM) under the Faculty of Engineering. Before that, he had been collaborating with Universiti Sains Malaysia (USM), Multimedia University of Malaysia (MMU), Universiti Islam Malaysia (IIUM) and King Fahd University Petroleum Minerals (KFUPM), Saudi Arabia as a lecturer. He has published many papers in international conferences and journals and is the author of several books in the field on engineering. His research interests include speech recognition systems, data compression systems, human computer interaction systems, medical imaging, and smart school design systems.

Solution of Electromagnetic and Velocity Fields for an Electrohydrodynamic Fluid Dynamical System

Rajveer S Yaduvanshi
AIT, Govt. of Delhi
Delhi, India
yaduvanshirs@yahoo.co.in

Harish Parthasarathy
NSIT, Govt of Delhi
Delhi, India
harishp@nsit.ac.in

Abstract— We studied the temporal evolution of the electromagnetic and velocity fields in an incompressible conducting fluid by means of computer simulations from the Navier Stokes and Maxwell's equations. We then derived the set of coupled partial differential equations for the stream function vector field and the electromagnetic field. These equations are first order difference equations in time and fetch simplicity in discretization. The spatial partial derivatives get converted into partial difference equations. The fluid system of equations is thus approximated by a nonlinear state variable system. This system makes use of the Kronecker Tensor product. The final system has taken account of anisotropic permittivity. The conductivity and magnetic permeability of the fluid are assumed to be homogeneous and isotropic. Present work in this paper describes characterization of magneto hydrodynamic anisotropic medium due to permittivity. Also an efficient and modified novel numerical solution using Tensor product has been proposed. This numerical technique seems to be potentially much faster and provide compatibility in matrices operation. Application of our characterization technique shall be very useful in tuning of permittivity in Liquid crystal polymer, Plasma and Dielectric lens antennas for obtaining wide bandwidth, resonance frequency reconfigure ability and better beam control.

Keywords- *Permittivity tuning, Incompressible fluid, Navier-Maxwell's coupled equations, resonance frequency reconfigure ability.*

I. INTRODUCTION

Electro magneto hydrodynamic equation solutions in the field of fluid dynamic have been very recent in many of the applications in the current era [1]. In recent years, there has been a growing interest of research in the interaction between ionic currents in electrolyte solutions and magnetic fields. Electro magneto hydrodynamics (EMHD) is the academic discipline which studies the dynamics of electrically conducting fluids [1-2]. Examples of such fluids include plasmas, liquid metals, and salty water.

Large magnetic field B produces enhanced spectral transfer in the direction perpendicular to field B_0 , which changes the polarization. Wave turbulence (WT) theory was developed for deviations from a strong uniform external magnetic field within the incompressible EMHD model. EMHD turbulence can be characterized as counter propagating

wave. According to Maxwell's equation direction of energy flow of a plane wave is given by $E \times B$ [3-5].

Material medium, in which an EM field exists is characterized by its constitutive parameters σ , μ , ϵ . The medium is said to linear if σ, μ, ϵ are independent of E, H or nonlinear otherwise. It is homogeneous if σ, μ, ϵ are not function of space variable or inhomogeneous otherwise. Hence σ, μ, ϵ are independent of direction or anisotropic otherwise. In isotropic case ϵ, μ are constants and can be extracted from curl term. In an anisotropic medium, properties are different from isotropic. Anisotropy is inherent in fluid. Anisotropic property provide great flexibility for tuning the spatial variations of electromagnetic waves in a desired manner by manipulating their structure features [13-14]. Based on this principle we investigate all components of turbulence applying theoretical approach and validate the same with FDM numerical method. Investigations on electrical properties and physical characteristics of this medium have been a problem for electromagnetic wave transport. As there is a greater need exists for studying interaction of electromagnetic field and anisotropic medium [9-10], when permittivity takes tensor form. The anisotropy can be realized by solution of 3D MHD equations, fluctuations due to uniform applied magnetic field. When strong field is applied, fluid motion and motion of magnetic field lines are closely coupled [15-17]. Thus anisotropic wave becomes Partial Differential Equations, which are difficult to solve analytically. Studies of these non linear effects in our defined problem have been possible, with customized numerical approach using Kronecker Tensor. If we take anisotropy due to permittivity, hence ϵ becomes an matrix, but μ is scalar number. The permittivity tensor $\epsilon = \begin{bmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_2 & 0 \\ 0 & 0 & \epsilon_3 \end{bmatrix}$ can be solved for x,y,z directions to have solution for dielectric constant.

We study the event when there exists an external electric and magnetic field acting on conducting fluid, exhibits anisotropic behavior and its permittivity takes on tensor form. Here we take on inherent characteristics of conducting fluid which has anisotropic permittivity [6-7]. We take this opportunity of anisotropic nature and use for electromagnetic wave propagation, which can be suitable for antenna bandwidth

increment and size reduction applications. This requires tuning of medium permittivity which could be possible by tensor analysis method [8]. For this we shall explore solution of dynamic fluid equations for which there is a need to develop fluid dynamic equations and solve them, for this we shall be using perturbation theory [11-12] to evaluate stream function ψ_α , magnetic field B_α , Electric field E_α and Current density J_α as function of time by taking x, y, z as Cartesian coordinates. We have used Kronecker Tensor for numerical solution and simulations have been carried out by MATLAB to validate parameters viz ψ_α , B_α , E_α and J_α . We have worked for non linear solution of fluid flow taking medium as anisotropic. We have solved for permittivity tensor with conductivity and permeability are kept as constant. Here we used Kronecker Product, discrete values from 0 to N-1 of x, y, z components to generate sparse matrix. This technique provides approximate solution of different dimensional matrices. Hence all nonlinear parameters can be validated with this method. We have worked for solution of B, E, J and ψ with time in x,y,z coordinates.

We have extended the concept cited in reference [1] where solution of 2D MHD have been proposed, it uses conventional method of numerical solution instead of the proposed work. This paper is organized in five main sections. Section I deals with introduction and survey part. Section II consists of all formulation developed. Sections III have MATLAB simulation results. Section IV conveys the possible applications of our research work. Section V concludes this paper with future implementations.

Abbreviations :

E - Electric intensity (Volts per meter)
H -Magnetic intensity (ampere per meter)
D- Electric flux density (coulomb per square meter)
B - Magnetic flux density (Weber per square meter)
J - Electric current density (ampere per square meter)
 q_v - Electric charge density (coulomb per cubic meter)
q - Charge electric
 σ - Conductivity
 μ - Permeability
 ϵ - Permittivity
 ψ - Stream function
v- Velocity of fluid (meter/second)
 ρ - Mass density
 p - Pressure,
 $\nu = \eta / \rho$ Kinetic viscosity
 $F = J \times B$ Lorentz force

II. FORMULATIONS

MHD is an anisotropic medium. The basic equations are

$$v, t + v \cdot \nabla v = -\frac{\nabla p}{\rho} + v \nabla^2 v + J \times \frac{B}{\rho} \quad (1)$$

And the above equation(1) is based on Navier stokes equation

$$\nabla \cdot D = 0 \quad (2)$$

$$\nabla \cdot B = 0 \quad (3)$$

$$\nabla \times H = J + D, t \quad (4)$$

$$\nabla \cdot E = -B, t \quad (5)$$

$$B = \mu H \quad (6)$$

$$D = \epsilon E \quad (7)$$

where μ and ϵ are now 3x3 matrices. In term of components, they are

$$B_a = \sum_b \mu_{ab} H_b \quad (8)$$

$$D_a = \sum_b \epsilon_{ab} E_b, a = 1,2,3 \quad (9)$$

$$J = \sigma (E + v \times B) \quad (10)$$

If σ is also a 3x3 matrix, then the last equation can be expressed as

$$J_a = \sum_b \sigma_{ab} E_b + \sum_{b,c,p} \sigma_{ab} \epsilon_{bcp} v_c B_p \quad (11)$$

Here e_{abc} is a cyclic unit vector. We can solve V and B cross product as

$$(v \times B)_1 = v_2 B_3 - v_3 B_2 \quad (12)$$

$$(v \times B)_2 = v_1 B_3 - v_3 B_1 \quad (13)$$

$$(v \times B)_3 = v_2 B_1 - v_1 B_2 \quad (14)$$

hence in general ,

$$(v \times B)_a = \sum_{b,c} e_{abc} v_b B_c \quad (15)$$

Where $e_{123} = e_{231} = e_{312} = -1$

$$e_{213} = e_{132} = e_{321} = 1$$

Others zero from the cyclic unit vector. Initially let $J = 0$ and $\mu_{ab} = -\mu_0 \delta_{ab}$

Then we derive

$$\nabla \cdot (\epsilon E) = 0 \quad (16)$$

$$\nabla^2 E - \nabla (\nabla \cdot E) - \mu_0 (\epsilon E)_{,tt} = 0 \quad (17)$$

This equation describes the propagation of EM waves in a medium having isotropic permeability but anisotropic permittivity. Assume that ρ the density of the medium is a constant i.e. the fluid is incompressible.

Then the mass conservation equation is $\nabla \cdot v = 0$ and hence there is a vector field ψ such that, as we know , $v = \nabla \times \psi$

The curl of the Navier Stokes equation gives

$$\Omega, t + \nabla \times (\Omega \times v) = v \nabla^2 \Omega + \rho^{-1} (\nabla \times (J \times B)) \quad (18)$$

Where $\Omega = \nabla \times v = -\nabla^2 \psi$

Since we can assume without loss of generality

$$\nabla \cdot \psi = 0$$

$$\text{Thus } \nabla^2 \psi, t + \nabla \times ((\nabla^2 \psi) \times (\nabla \times \psi)) = \\ v \nabla^2 \nabla^2 \psi - \rho^{-1} \nabla \times (J \times B) \quad (19)$$

Note that

$$(\nabla^2 \psi)_a = \nabla^2 \psi_a \\ (\nabla \times \psi)_a = \sum_{b,c} e_{abc} \psi_{c,b} \\ J_a = \sum_{b,c,d,p} e_{abc} (\nabla^2 \psi_b) e_{cdp} \psi_{p,d} (\nabla^2 \psi_b) \cdot \psi_{p,d} \quad (20)$$

$$\text{Since } (J \times B)_a = \sum_{b,c} e_{abc} J_b B_c \quad (21)$$

Hence

$$J_a = \sum_{b,c,p} e_{abc} \sigma_{bp} \left(E_p + \sum_{q,f} e_{pqf} v_q B_f \right) \cdot B_c \quad (22)$$

Here we have worked for presenting a nonlinear solution of fluid flow taking medium as anisotropic due to permittivity and conductivity and permeability are assumed to be fixed.

Now we have to Solve for ψ_a, E_a, B_a w.r.t. time from the above equations (1-28). The above equations are self explanatory hence does not need more elaborations. Also we can describe

$$J \times B$$

$$= \mu^{-1} (\nabla \times B) \times B = -\frac{\mu^{-1} \nabla |B|^2}{2} + \mu^{-1} (B \cdot \nabla) \cdot B \quad (23)$$

Hence equation (1) can be written as

$$\nabla^2 \psi, t + \nabla \times ((\nabla^2 \psi) \times (\nabla \times \psi)) = \\ v \nabla^2 \nabla^2 \psi - \rho^{-1} \nabla \times \left(-\frac{\mu^{-1} \nabla |B|^2}{2} \right) \\ + \mu^{-1} (B \cdot \nabla) \cdot B \quad (24)$$

$$\text{Let } \nabla^2 = A(\text{Matrix})$$

Hence

$$\psi, t = A^{-1} [v \nabla^2 \nabla^2 \psi - \rho^{-1} \nabla \times \left(-\frac{\mu^{-1} \nabla |B|^2}{2} \right) \\ + \mu^{-1} (B \cdot \nabla) \cdot B - \nabla \times ((\nabla^2 \psi) \times (\nabla \times \psi))] \quad (25)$$

From Tensor product analysis $\psi_{(t,x,y,z)}$ when; $0 \leq x, y, z \leq N-1$; we can formulate table for generating block matrix from the coefficients as follows

$$\psi_{(t, 0, 0, 0)}$$

$$\psi_{(t, 0, 0, 1)}$$

$$\psi_{(t, 0, 0, 2)}$$

⋮

$$\psi_{(t, 0, 0, N-1)}$$

$$\psi_{(t, 0, 1, 0)}$$

$$\psi_{(t, 0, 1, 1)}$$

$$\psi_{(t, 0, 1, N-1)}$$

$$\psi_{(t, 0, 2, 0)}$$

$$\psi_{(t, 0, 2, N-1)}$$

⋮

$$\psi_{(t, 0, N-1, N-1)}$$

$$\psi_{(t, 1, 0, 0)}$$

⋮

$$\psi_{(t, 1, 0, N-1)}$$

⋮

$$\psi_{(t, 0, N-1, N-1)}$$

⋮

$$\psi_{(t, N-1, N-1, N-1)}$$

From finite difference method it can be discretized as given below

$$\nabla^2 = \frac{\psi_{(t,x+1,y,z)} + \psi_{(t,x-1,y,z)} + \psi_{(t,x,y+1,z)} + \psi_{(t,x,y-1,z)} + \psi_{(t,x,y,z+1)} + \psi_{(t,x,y,z-1)} - 6\psi_{(t,x,y,z)}}{\Delta^2}$$

Similarly we can complete column of ψ_y, ψ_z of table 1 to form this block matrix form coefficients. This can be generated by computer. The computer generated table 2 in triangular form matrix has been named as matrix A, which shall be used to operate with all the elements of the matrices placed at right side in equation (25). This will enable us to solve the problem. Thus we can now estimate approximate value of stream function. Hence we thus gets the value of ψ, t , results are presented in figure 1-10.

Here electric field E can be expressed as

$$E, t = \frac{1}{\mu \epsilon} \nabla \times B - \frac{\sigma}{\epsilon} (E - (\nabla \times \psi) \times B) \quad (26)$$

Similarly we can evaluate E, t, simulated results are presented in fig 18. We have to work for getting solution of Matrix [A] to get its inverse we take help of MATLAB.

$$\psi_{(0)}, E_{(0)}, B_{(0)}$$

⋮

$$\psi_{(n)}, E_{(n)}, B_{(n)}, \text{ Also}$$

$$\vec{\nabla} \times \vec{\psi} =$$

$$(\psi_{z,y} - \psi_{y,z}, \psi_{x,z} - \psi_{z,x}, \psi_{y,x} - \psi_{x,y}) \quad (27)$$

$$(\nabla \times \psi)_x = \psi_{z,y} - \psi_{y,z}$$

$$(\nabla \times \psi)_y = \psi_{x,z} - \psi_{z,x}$$

$$(\nabla \times \psi)_z = \psi_{y,x} - \psi_{x,y}$$

From Finite difference method, we have

$$\Delta \psi_x(n, k, l, m) = \frac{1}{\Delta^2} [\psi_x(n, k+1, l, m) + \psi_x(n, k-1, l, m) - 2\psi_x(n, k, l+1, m) - 2\psi_x(n, k, l-1, m) + \psi_x(n, k, l, m+1) + \psi_x(n, k, l, m-1) - 2\psi_x(n, k, l, m)]$$

$$\Delta \psi_x(n, k, l, m) = \frac{1}{\Delta^2} [\psi_x(n, k+1, l, m) + \psi_x(n, k-1, l, m) - 6\psi_x(n, k, l, m) + \psi_x(n, k, l+1, m) + \psi_x(n, k, l-1, m) + \psi_x(n, k, l, m+1) + \psi_x(n, k, l, m-1)]$$

(28)

We get this Matrix A from Kronecker Tensor Product which is generated by computer.

$$\psi = \begin{bmatrix} \psi_x \\ \psi_y \\ \psi_z \end{bmatrix}, \text{ writing this in lexicographical order for solution.}$$

Thus, Matrix A can be expressed as follows:-

$$\psi(t, x, y, z) =$$

$$\begin{matrix} \psi_x(t, 0, 0, 0) & \psi_y(t, 0, 0, 0) & \psi_z(t, 0, 0, 0) \\ \psi_x(t, 0, 0, 1) & \psi_y(t, 0, 0, 1) & \psi_z(t, 0, 0, 1) \\ \psi_x(t, 0, 0, 2) & \psi_y(t, 0, 0, 2) & \psi_z(t, 0, 0, 2) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \end{matrix}$$

$$\begin{matrix} \psi_x(t, 0, 0, N-1) & \psi_y(t, 0, 0, N-1) & \psi_z(t, 0, 0, N-1) \\ \psi_x(t, 0, 1, 0) & \psi_y(t, 0, 1, 0) & \psi_z(t, 0, 1, 0) \\ \psi_x(t, 0, 1, 1) & \psi_y(t, 0, 1, 1) & \psi_z(t, 0, 1, 1) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \end{matrix}$$

$$\begin{matrix} \psi_x(t, 0, 1, N-1) & \psi_y(t, 0, 1, N-1) & \psi_z(t, 0, 1, N-1) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \end{matrix}$$

$$\psi_x(t, N-1, 0, 0) \quad \psi_y(t, N-1, 0, 0) \quad \psi_z(t, N-1, 0, 0)$$

$$\vdots \quad \vdots \quad \vdots$$

$$\psi_x(t, N-1, N-1, N-1) \quad \psi_y(t, N-1, N-1, N-1) \quad \psi_z(t, N-1, N-1, N-1)$$

Table 1 Generated from Kronecker Tensor

$$A \psi_t = A \begin{bmatrix} \psi_x, t \\ \psi_y, t \\ \psi_z, t \end{bmatrix} \text{ where A Matrix =}$$

TABLE 1. A – BLOCK MATRIX GENERATED BY COMPUTER

$$\text{or } \begin{bmatrix} A & \psi_{x,t} \\ A & \psi_{y,t} \\ A & \psi_{z,t} \end{bmatrix} = \begin{bmatrix} k & 0 & 0 \\ 0 & k & 0 \\ 0 & 0 & k \end{bmatrix} \begin{bmatrix} \psi_x, t \\ \psi_y, t \\ \psi_z, t \end{bmatrix}$$

Hence this will generate $3N^3 \times 3N^3$ order Matrix. For example tensor product will look like the given below matrix.

$$\begin{bmatrix} B & \psi_x \\ B & \psi_y \\ B & \psi_z \end{bmatrix} \times \begin{bmatrix} (C\psi)_x \\ (C\psi)_y \\ (C\psi)_z \end{bmatrix} = \begin{bmatrix} (B\psi_y) \cdot (C\psi)_z - (B\psi_z) \cdot (C\psi)_y \\ (B\psi_z) \cdot (C\psi)_x - (B\psi_x) \cdot (C\psi)_y \\ (B\psi_y) \cdot (C\psi)_z - (B\psi_z) \cdot (C\psi)_y \end{bmatrix}$$

Here we have taken value of $N=10$, Which is significantly small and convenient to run on the computer as large value of N require much amount of memory, more time to compute. Matrix generated has array of elements and with most of the elements are zero. In this way we notice that these non zero elements values get shifted toward one column right side, when moved from first row to second row and second to third row and so on. It forms a triangular matrix; same is shown in table 2. With the help of Kronecker tensor we are able solve PDE equations in matrix form. Results of the solution are discussed in section 3 of this paper.

III. SIMULATION RESULTS

We have solved for stream function from given equation (25) using Matlab at some fixed length of t and z values. ψ, X, Y have been taken variables as shown in figure [1-10]. We keenly observe turbulence as we proceed from figure 1 to 10. Plot [11-12] presents clear view of turbulence of y specific components and fig [13-14] depicts total stream function plots. Plots [15-17] presents x, y, z components of J . Plots [18] presents electric field which seems to be sinusoidal. Plot 19 presents $J \times B$ features. We have taken plots of stream function in x, y, z coordinates with respect to time and intensity of the figure in colors represents stream function as shown in fig [13-14]. We have first generated one matrix A using Kronecker product and inverse of this matrix has been computed. This generated inverse matrix gets multiplied with all other values in right hand side terms in equation (25) with each element. Results have revealed stream function turbulence due to presence of high magnetic field.

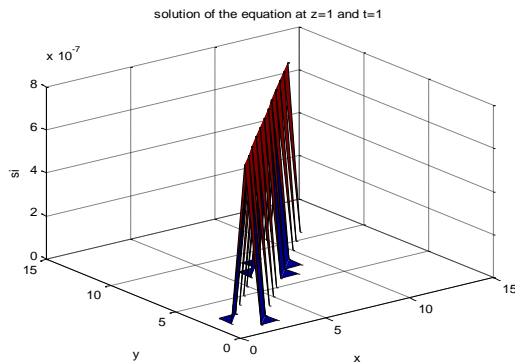


Figure 1. Plot of ψ stream function in x and y components at given fixed value of t, z

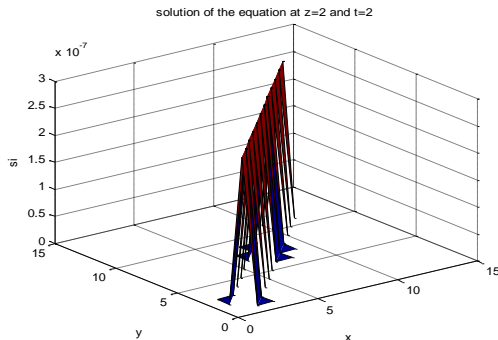


Figure 2. Plot of ψ stream function in x and y components at given fixed value of t, z

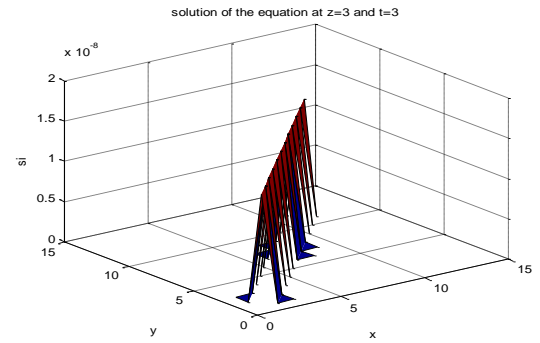


Figure 3. Plot ψ of stream function in x and y components at given fixed value of t, z

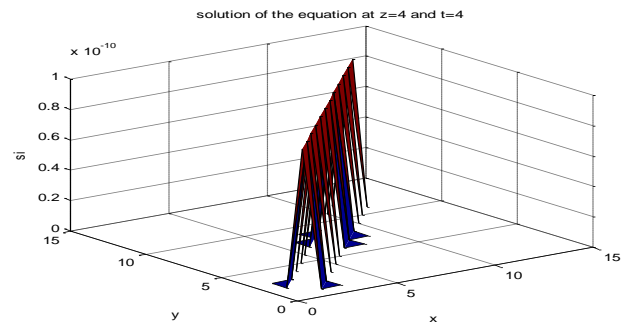


Figure 4. Plot of ψ stream function in x and y components at given fixed value of t, z

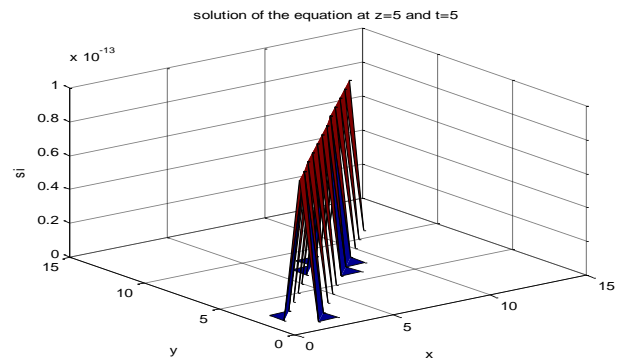


Figure 5. Plot of stream function in x and y components at given fixed value of t, z

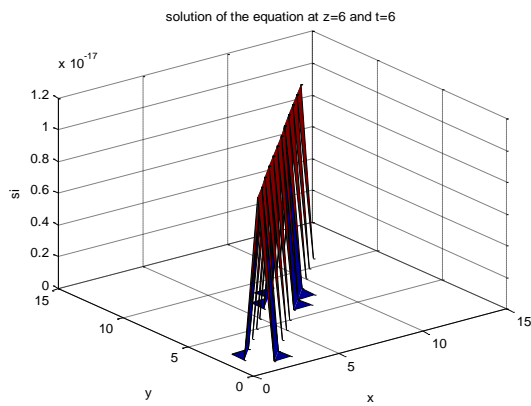


Figure 6. Plot of stream function in x and y components at given fixed value of t, z

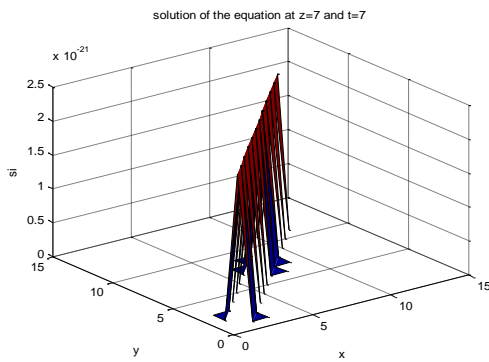


Figure 7. Plot of ψ stream function in x and y components at given fixed value of t, z

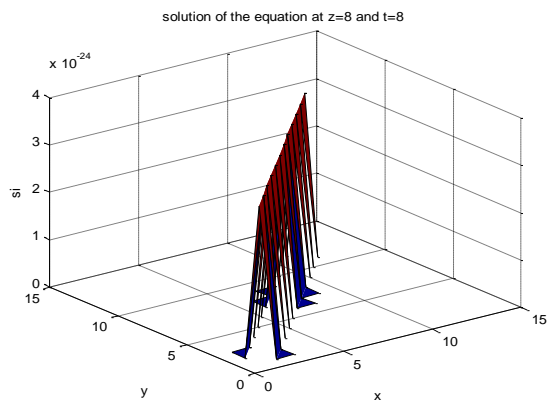


Figure 8. Plot of stream function in x and y components at given fixed value of t, z

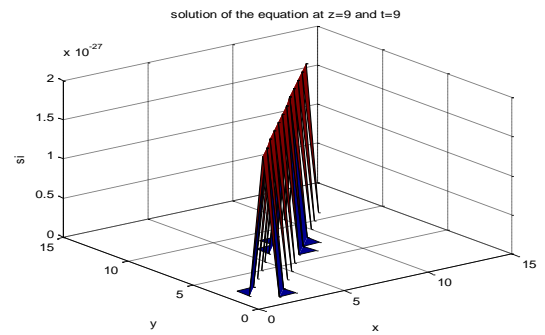


Figure 9. Plot of stream function in x and y components at given fixed value of t, z

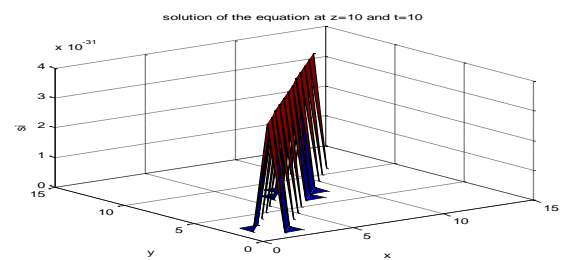


Figure 10. Plot of ψ stream function in x and y components at given fixed value of t, z

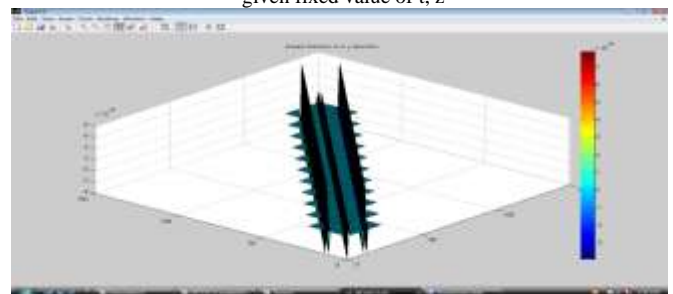


Figure 11 Plot ψ of stream function showing turbulence

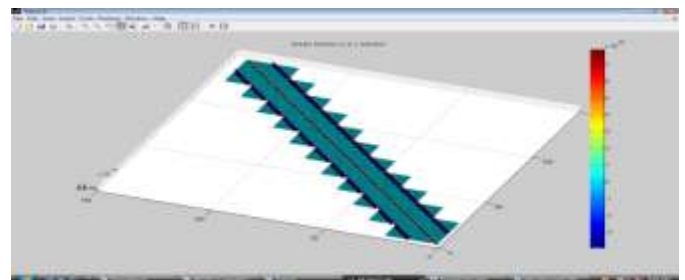


Figure 12 Plot ψ of stream function in y direction

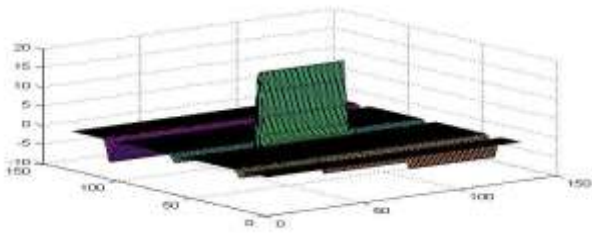


Figure 13. 3D Plot of ψ stream function intensity showing turbulence

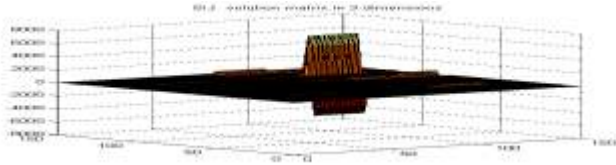


Figure 14. 3D plot ψ stream function showing turbulence

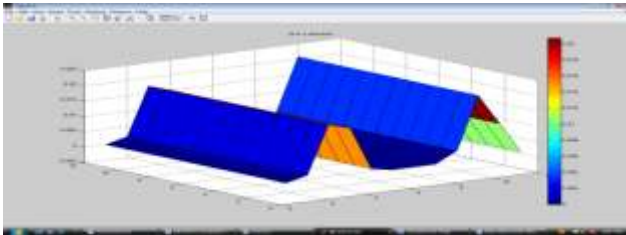


Figure15 Plot of J in x axis

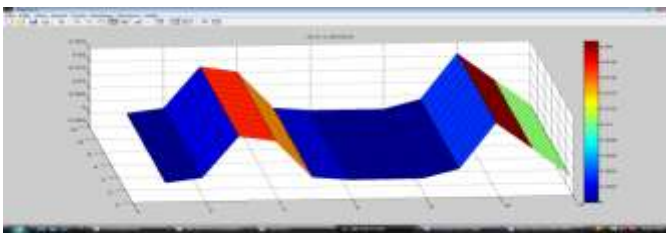


Figure16 Plot of J in y axis

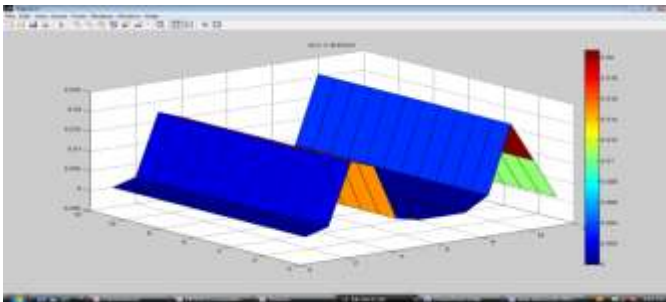


Figure 17. Plot of J in z axis

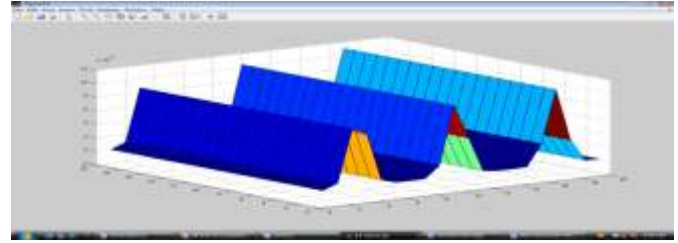


Figure 18. Plot of Electric E field with time

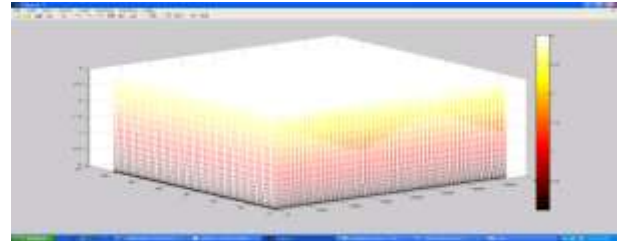


Figure 19 plot of JXB

IV. APPLICATIONS

Permittivity tensor imparts directional dependent properties. Positive permittivity may play significant role for bandwidth and gain in energy propagation. Significant application of this feature can be in designing dielectric lens and LCP antennas. Application of our approach can be an electromagnetically flow control [18] of fluid metal for steel casting with varying magnetic field strength create transition of turbulence due to perturbation. Evaluating field gradient drag effect can predict and control conducting fluid flow. Also in meta material antenna with tuning permittivity may effect in reduction of size of the antenna and these facts are yet to be established as per best of author knowledge. Tele robot, actuator, sensor, EM brakes, Steel casting, Seismic wave, Dielectric lens antenna, Liquid Crystal Polymer antenna, Meta material antenna.

V. CONCLUSION

We have used Kronecker tensor product for numerical solution to avoid matrices mismatch problem. This numerical approach has not only given us the solution to operate on each other having different order of matrices but also this approach has fetched us the fast computations. An efficient numerical solution technique has been proposed. In this work we have developed 3D MHD equations for turbulence of fluid flow and stream function has been validated in x, y, z direction. Analysis of anisotropic medium for permittivity has been proposed to obtain better transmission parameters. As compared to [1] detailed analyses have been worked with 3D features. Also Numerical techniques developed and used for the present work have been much efficient in computations and time as compared to the previous work cited in reference [1]. Turbulence results have been validated with precision control as compared to reference[1].

In future works we are busy in working for devising permittivity control and tuning mechanism for anisotropic medium by permittivity tensor analysis, in x, y, z directions to obtain higher dielectric constant which can be the improve efficiency of dielectric lens and liquid crystal antenna. This high dielectric constant or permittivity can also enhance bandwidth and can reduce length of these antennas.

ACKNOWLEDGEMENT AND BIOGRAPHY

My Director Prof Asok De and Director NSIT Prof Raj Senani, who inspired me for this research and enriched me with all necessary resources required for the research. Dr RK Sharma, Associate Professor ECE for fruitful discussions. I am indebted to my wife Sujata, daughter Monica and son Rahul for giving me their time in my research work making home an excellent place for the research.

REFERENCES

- [1] Rajveer S Yaduvanshi Harish Parthasarathy, "Design, Development and Simulations of MHD Equations with its proto type implementations"(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 1, No. 4, October 2010.
- [2] EM Lifshitz and LD Landau, "Classical theory of fields", 4th edition Elsevier.
- [3] Matthew N.O.Sadiku, "Numerical techniques in Electromagnetic with MATLAB" CRC press, third edition.
- [4] EM Lifshitz and LD Landau, "Electrodynamics of continuous media" Butterworth-Heinemann.
- [5] EM Lifshitz and LD Landau, "Theory of Fields" Vol 2 Butterworth-Heinemann.
- [6] EM Lifshitz and LD Landau, "Fluid Mechanics" Vol6, Butterworth-Heinemann [7] Chorin, A.J., 1986, "Numerical solution of the Navier-Stokes equations", *Math. Comp.*, Vol. 22, pp. 745-762.
- [8] Cramer, K.R. & Pai, S., 1973, "Magnetofluid Dynamics For engineers and Applied Physicists", McGraw-Hill Book.
- [9] Feynman, R.P., Leighton, R.B. and Sands, M., 1964, "The Feynman Lectures on Physics", Addison-Wesley, Reading, USA, Chapter 13-1.
- [10] Gerbeau, J.F. & Le Bris, C., 1997, "Existence of solution for a density dependant magnetohydrodynamic equation", *Advances in Differential Equations*, Vol. 2, No. 3, pp. 427-452.
- [11] Guermond, J.L. & Shen, J., 2003, "Velocity-correction projection methods for incompressible flows", *SIAM J. Numer. Anal.* Vol. 41, No. 1, pp. 112-134.
- [12] Holman, J.P., 1990, "Heat Transfer" (7th edition), Chapter 12, Special Topics in Heat Transfer, MacGraw-Hill Publishing Company, New York.
- [13] PY Picard, "Some spherical solutions of ideal magnetohydrodynamic equation", *Journal of non linear physics* vol 14 no4 (2007), 578-588.
- [14] Temam, R., 1977, "Navier Stokes Equations", *Stud. Math. Appl.* 2, North-Holland, Amsterdam.
- [15] Tillack, M.S. & Morley, N.B., 1998, "MagnetohydroDynamics", McGraw Hill, Standard Handbook for Electrical Engineers, 14th Edition.
- [16] Wait, R. (1979). The numerical solution of algebraic equations. A Wiley-interscience publication.
- [17] Fermigier, M. (1999). *Hydrodynamique physique*. Problèmes résolus avec rappels de cours, collections sciences sup. physique, edition Dunod.
- [18] Bahadir, A.R. and T. Abbasov (2005). A numerical investigation of the liquid flow velocity over an infinity plate which is taking place in a magnetic field. *International journal of applied electromagnetic and mechanics* 21, 1-10.

AUTHORS PROFILE

Author: Rajveer S Yaduvanshi, Asst Professor has 21 years of teaching and research work. He is fellow member of IETE. He is author of two international journal paper and five international /national conference papers. He has successfully implemented fighter aircraft arresting barrier projects at select flying stations of Indian Air Force. He has worked on Indigenization projects of 3D radars at DGAQA, Min of defense and visited France for 3D radar modernization program as Senior Scientific Officer. Currently he is working on MHD Antenna prototype implementations. At present, he is teaching in ECE Deptt. of AIT, Govt of Delhi-110031.

Co-Author : Prof Harish Parthasarathy is an eminent academician and great researcher. He is professor and dean at NSIT, Govt. College of Engineering at Dwarka, Delhi. He has extraordinary research instinct and a great book writer in the field of DSP. He has published more than ten books and have produces more than seven PhDs in ECE Deptt of NSIT, Delhi.

Survey of Wireless MANET Application in Battlefield Operations

Dr. C. Rajabhushanam and Dr. A. Kathirvel
Department of Computer Science & Engineering
Tagore Engineering College
Chennai, India
rajcheruk@gmail.com, kathir.tagore@gmail.com

Abstract—In this paper, we present a framework for performance analysis of wireless MANET in combat/battle field environment. The framework uses a cross-layer design approach where four different kinds of routing protocols are compared and evaluated in the area of security operations. The resulting scenarios are then carried out in a simulation environment using NS-2 simulator. Research efforts also focus on issues such as Quality of Service (QoS), energy efficiency, and security, which already exist in the wired networks and are worsened in MANET. This paper examines the routing protocols and their newest improvements. Classification of routing protocols by source routing and hop-by-hop routing are described in detail and four major categories of state routing are elaborated and compared. We will discuss the metrics used to evaluate these protocols and highlight the essential problems in the evaluation process itself. The results would show better performance with respect to the performance parameters such as network throughput, end-to-end delay and routing overhead when compared to the network architecture which uses a standard routing protocol. Due to the nature of node distribution the performance measure of path reliability which distinguishes ad hoc networks from other types of networks in battlefield conditions, is given more significance in our research work.

Keywords- MANET; routing; protocols; wireless; simulation

I. INTRODUCTION

As the importance of computers in our daily life increases it also sets new demands for connectivity. Wired solutions have been around for a long time but there is increasing demand on working wireless solutions for connecting to the Internet, reading and sending E-mail messages, changing information in a meeting and so on. There are solutions to these needs, one being wireless local area network that is based on IEEE 802.11 standard. However, there is increasing need for connectivity in situations where there is no base station (i.e. backbone connection) available (for example two or more PDAs need to be connected). This is where ad hoc networks step in.

A. AD-HOC NETWORK

The “Ad Hoc Networks” are wireless networks characterized by the absence of fixed infrastructures. This

allows the use of this kind of network in special circumstances, such as disastrous events, the reduction or elimination of the wiring costs and the exchange of information among users independently from the environment. The devices belonging to the network must be able not only to transmit and receive data, but also to manage all the functions of the network in a distributed way, as routing of the packets, security, Quality Of Service (QoS), etc; so these are not only terminals, but they become sheer nodes. These devices have a wireless interface and are usually in mobile systems of several types, from those simple ones like PDA to notebooks.

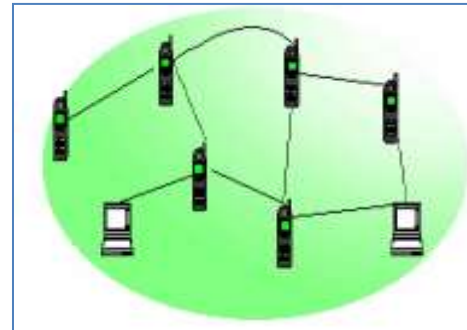


Figure 1. Example of a wireless MANET

A mobile ad-hoc network (MANET) is a self-configuring network of mobile routers (and associated hosts) connected by wireless links—the union of which form an arbitrary topology. The routers are free to move randomly and organize themselves arbitrarily; thus, the network's wireless topology may change rapidly and unpredictably. There are several applications of mobile ad hoc networks such as disaster recovery operations, battle field communications, data sharing in conference halls, etc [1]. One of the main issues in such networks is performance- in a dynamically changing topology; the nodes are expected to be power-aware due to the bandwidth constrained network. Another issue in such networks is security - The goals of confidentiality, integrity, authenticity, availability and non-repudiability are very difficult to achieve in MANETs since every node participates in the operation of the network equally and malicious nodes are difficult to detect. The addition of security layers also adds to the performance overhead drastically. We investigate these

related issues and study the tradeoffs involved so that an optimal solution may be achieved.

In Latin, ad hoc means "for this," further meaning "for this purpose only"- It is a good and emblematic description of the idea why ad hoc networks are needed. They can be set up anywhere without any need for external infrastructure (like wires or base stations). They are often mobile and that's why a term MANET is often used when talking about Mobile Ad hoc NETworks [2]. MANETs are often defined as follows: A "mobile ad hoc network" (MANET) is an autonomous system of mobile routers (and associated hosts) connected by wireless links - the union of which forms an arbitrary graph. The routers are free to move randomly and organize themselves arbitrarily; thus, the network's wireless topology may change rapidly and unpredictably. Such a network may operate in a standalone fashion, or may be connected to the larger Internet. The strength of the connection can change rapidly in time or even disappear completely. Nodes can appear, disappear and re-appear as the time goes on and all the time the network connections should work between the nodes that are part of it. As one can easily imagine, the situation in ad hoc networks with respect to ensuring connectivity and robustness is much more demanding than in the wired case [1].

Ad hoc networks are networks are not (necessarily) connected to any static (i.e. wired) infrastructure. An ad-hoc network is a LAN or other small network, especially one with wireless connections, in which some of the network devices are part of the network only for the duration of a communications session or, in the case of mobile or portable devices, while in some close proximity to the rest of the network.

The ad hoc network is a communication network without a pre-existing network infrastructure. In cellular networks, there is a network infrastructure represented by the base-stations, Radio network controllers, etc. In ad hoc networks every communication terminal (or radio terminal RT) communicates with its partner to perform peer to peer communication. If the required RT is not a neighbor to the initiated call RT, (outside the coverage area of the RT) then the other intermediate RTs are used to perform the communication link. This is called multi-hop peer to peer communication. This collaboration between the RTs is very important in the ad hoc networks. In ad hoc networks all the communication network protocols should be distributed throughout the communication terminals (i.e. the communication terminals should be independent and highly cooperative).

In wireless communication, Ad hoc mobile network is a collection of mobile nodes that are dynamically and arbitrarily located in such a manner that the interconnections between nodes are capable of changing on a continual basis. In order to facilitate communication within the network, a routing protocol is used to discover have been proposed for MANET and some of them have been widely used. This project paper utilizes network scoping to model MANET routing for four different routing protocols: Dynamic Source Routing (DSR), Ad hoc on-demand Distance Vector (AODV) and Optimized

Link State Routing (OLSR) and Location Aided Routing (LAR) and compare them with the traditional mathematical equation models [1].

B. Application Areas

Some of the applications of MANETs are

- Military or police exercises.
- Disaster relief operations.
- Mine site operations.
- Urgent Business meetings
- Personal area network

Such networks can be used to enable next generation of battlefield applications envisioned by the military including situation awareness systems for maneuvering war fighters, and remotely deployed unmanned micro-sensor networks. Ad Hoc networks can provide communication for civilian applications, such as disaster recovery and message exchanges among medical and security personnel involved in rescue missions.

Many examples of MANETs can be found in real life where an access point and existing infrastructure is not available. For example, battlefields and emergencies where no one has the time to establish access points are the prime examples of MANETs. Common examples are:

Battlefield situations where each jeep and even each soldiers gun has a wireless card. These "nodes" form a MANET and communicate with each other in the battlefield. In addition, MANETs can be used to detect hostile movements in remote areas instead of land mines.

Emergency situations where, for example, a building has been destroyed due to fire, earthquake, or bombs. In such a case, it is important to set up a quick network. MANETs are ideal for such situations. For example, in emergency operation, police and fire fighters can communicate through a MANET and perform their operations without adequate wireless coverage.

II. PREVIOUS WORK

Most of the previous work have reviewed and implemented the vast literature using various routing protocols employing the constant bit rate for their analysis [3], [4]. Most of the previous is limited on performing simulations for ad hoc networks. Our work differs in that we use variable bit ratio in a combat situation with battlefield like conditions. We will observe and comment on the behavior of each protocol.

A. Routing protocols and Algorithm

An ad hoc mobile network is a collection of mobile nodes that are dynamically and arbitrarily located in such a manner that the interconnections between nodes are capable of changing on a continual basis. In order to facilitate communication within the network, a routing protocol is used to discover routes between nodes. The primary goal of the routing protocol is to obtain correct and efficient route establishment between a pair of source and destination nodes

so that messages may be delivered in a timely manner. The general performance measures are the delay and throughput of information [4]. Due to the limited bandwidth associated with most wireless channels, it is obvious that route construction should be done with a minimum of overhead and bandwidth consumption. And due to the nature of node distribution, another performance measure is path reliability, which distinguishes ad hoc networks from other types of networks. Much work has appeared in these areas, but advances in wireless research are focusing more and more on the adaptation capability of routing protocols due to the interrelationship among these performance measures.

Routing algorithms, it is well known, determine the optimum path between n senders and receivers based on specific metrics, such as shortest time delay or minimum cost. Determination of optimal paths in large networks has been an area of active research for many years, with applications for travelling salesman, school bus routing, flight routings and others. An important factor in routing algorithm design is the time T it takes to develop a solution. If T is more than the average time between topology changes, then the algorithm cannot update the routing table fast enough. For example, if the topology changes every 20 seconds, but it takes a minute to find a route, then the routing tables will not have correct routing information and the whole routing system will collapse. This is the main challenge in MANET routing. In MANET, the internal routing algorithms do not work well because they assume that the topology will change very infrequently, thus an optimal path can be found almost at leisure [4].

For mobile ad hoc networks, the core routing functionalities include:

- Path generation to generate possible paths between the senders and receivers.
- Path selection to determine the appropriate paths based on a selection criteria (minimal time).
- Data forwarding to transmit user traffic along the selected paths.
- Path maintenance to make sure that the selected route is maintained and to find alternates in case of problems.
- Due to the nature of MANETs, the routing protocols should be highly adaptive, fast, and energy/bandwidth efficient.

B. MANET Routing Algorithms

Many routing protocols for MANET have been published. Although there are different ways of classifying them, a convenient approach is to view them in terms of small or large networks [6].

For smaller networks, the following are well known:

Dynamic Source Routing (DSR) uses a source (versus hop-by-hop) algorithm. Thus there is no need for intermediate nodes to maintain routing information.

Ad Hoc On-Demand Distance Vector (AODV) combines DSR with sequence numbering and other features to make it more efficient

In recent years, research efforts have been focusing on improving the performance of routing protocols in MANET. The MANET working group coordinates the development of several candidates among the protocols including OLSR and AODV. These protocols are classified into four classes based on the time when routing information is updated, the Proactive Routing Protocol (PRP), Reactive Routing Protocols (RRP), Hybrid Routing Protocol (HRP) and the Geographical Routing Protocol (GRP) [4], [6].

There are other classifications of routing protocols such as the distance vector (DV) class and link state (LS) class based on the content of the routing table. The DV protocols broadcast a list (vector) of distances to the destinations and each node maintains the routing table of the shortest paths to each known destination. On the other hand, the LS protocols maintain the topology of the network (links state). Each entry in LS routing table represents a known link. In LS routing, each node needs to calculate the routing table based on the local (links state) information in order to obtain a route to destination. Normally, the link state protocols are more stable and robust but much more complex than distance vector protocols. There are also instances of the above two family In MANET. The OLSR is the most widely used link state protocol, while AODV is the most popular distance vector protocol. We provide a general analysis of link state routing and distance vector routing in MANET respectively.

Another classification of routing protocols is source routing and hop-by-hop routing. In source routing, the source computes the complete path towards the destination, which consequently leads to loop-free routing. In hop-by-hop routing, each intermediate node computes the next hop itself. The nature of hop-by-hop routing reduces the chance of failed route in MANET, which suffers much faster topology changes than wired networks. Consequently, the source routing protocol in MANET, DSR, allows the intermediate nodes and even overhearing nodes to modify the route in order to adapt to the nature of MANET. Most MANET routing protocols such as OLSR and AODV have the hop-by-hop nature.

C. Proactive Routing Protocols (PRP)

In proactive (table-driven) protocols, nodes periodically search for routing information within a network. The control overhead of these protocols is foreseeable, because it is independent to the traffic profiles and has a fixed upper bound. This is a general advantage of proactive routing protocols.

DSDV: The Destination-Sequenced Distance-Vector (DSDV) Routing protocol is based on the idea of the classical Bellman-Ford Routing Algorithm with certain improvements such as making it loop-free. The basic improvements made include freedom from loops in routing tables, more dynamic and less convergence time. Every node in the MANET maintains a routing table which contains list of all known destination nodes within the network along with number of

hops required to reach to particular node [4], [7]. Each entry is marked with a sequence number assigned by the destination node. The sequence numbers are used to identify stale routes thus avoiding formation of loops. To maintain consistency in routing table data in a continuously varying topology, routing table updates are broadcasted to neighbor's periodically or when significant new information is available. In addition to its time difference between arrival of first and arrival of the best route to a destination is also stored so that advertising of routes, which are likely to change soon, can be delayed. Thus avoiding the advertisement of routes, which are not stabilized yet, so as to avoid rebroadcast of route entries that arrive with node is supposed to keep the track of settling time for each route so that fluctuations can be damped by delaying advertisement of new route to already known and reachable destination thus reducing traffic. Fluctuating routes occurs as a node may always receive two routes to a destination with same sequence number but one with better metric later. But new routes received which take to a previously unreachable node must be advertised soon. Mobiles also keep track of the settling time of routes, or the weighted average time that routes to a destination will fluctuate before the route with the best metric is received. By delaying the broadcast of a routing update by the length of the settling time, mobiles can reduce network traffic and optimize routes by eliminating those broadcasts that would occur if a better route was discovered in the very near future. Consequently, the proactive routing protocols prefer link state routing because additional route calculation of link state routing doesn't contribute to delay.

OLSR: Optimized Link State Routing (OLSR) is a proactive, link state routing protocol specially designed for ad hoc networks. OLSR maintains Multipoint Relays (MPRs), which minimizes the control flooding by only declaring the links of neighbors within its MPRs instead of all links [4], [7]. The multicast nature of OLSR route discovery procedure can be integrated with the mobile IP management by embedding the mobile-IP agent advertisement into the OLSR MPR-flooding. This is important for the 4G global ubiquitous networks, which requires the wireless access network to be fully adhoc. Several extensions of OLSR are available that correspond to different network scenario. For fast changing MANET, provides a fast-OLSR version which reacts faster to topology changes than standard OLSR by enabling the fast moving nodes to quickly discover its neighbors and select a subset of their MPRs to establish connection to other nodes.

D. Reactive Routing Protocol (RRP)

The reactive (on-demand) routing protocols represent the true nature of ad hoc network, which is much more dynamic than infrastructure networks. Instead of periodically updating the routing information, the reactive routing protocols update routing information when a routing requirement is presented, consequently reducing the control overhead, especially in high mobility networks where the periodical update will lead to significant useless overhead.

AODV: Ad hoc On-demand Distance Vector Routing (AODV) is an improvement of the DSDV algorithm. AODV

minimizes the number of broadcasts by creating routes on-demand as opposed to DSDV that maintains the list of all the routes. The on-demand routing protocols suffer more from frequent broken source-to-destination links than table driven routing due to the delay caused by on-demand route recalculation. AODV avoids such additional delay by using distance vector routing. There are some improved versions of AODV. A "source route accumulation" version of AODV which modifies the Routing REquest (RREQ) and Routing REPLY (RREP) messages in order to speed up the convergence of route discovery [4]. In order to reduce control overhead, a controlled flooding (CF) mechanism to reduce overlapped flooding messages for AODV is used.

The AODV algorithm is an improvement of DSDV protocol described above. It reduces number of broadcast by creating routes on demand basis, as against DSDV that maintains routes to each known destination. When source requires sending data to a destination and if route to that destination is not known then it initiates route discovery. AODV allows nodes to respond to link breakages and changes in network topology in a timely manner. Routes, which are not in use for long time, are deleted from the table. Also AODV uses Destination Sequence Numbers to avoid loop formation and Count to Infinity Problem. An important feature of AODV is the maintenance of timer based states in each node, regarding utilization of individual routing table entries. A routing table entry is expired if not used recently [4], [7]. A set of predecessor nodes is maintained for each routing table entry, indicating the set of neighboring nodes which use that entry to route data packets. Each predecessor node, in turn, forwards the RERR to its own set of predecessors, thus effectively erasing all routes using the broken link. Route error propagation in AODV can be visualized conceptually as a tree whose root is the node at the point of failure and all sources using the failed link as the leaves.

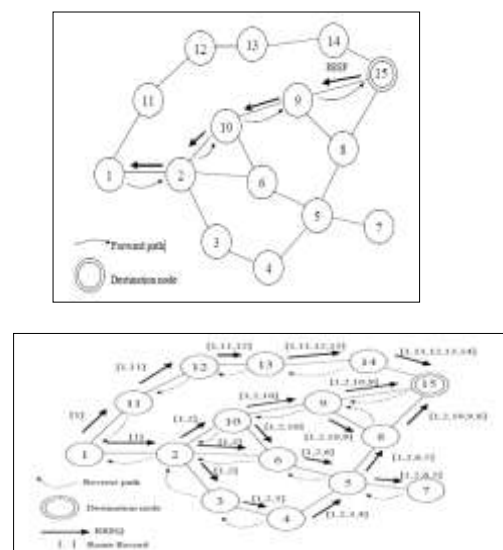


Figure 2. AODV routing protocol

DSR: The key feature of DSR is the use of source routing, which means the sender knows the complete hop-by-hop route to the destination. The node maintains route caches containing the source routes that it is aware of. Each node updates entries in the route cache as and when it learns about new routes [4], [6]. The data packets carry the source route in the packet headers. The delay and throughput penalties of DSR are mainly attributed to aggressive use of caching and lack of any mechanism to detect expired stale routes or to determine the freshness of routes when multiple choices are available. Aggressive caching, however, helps DSR at low loads and also keeps its routing load down.

The DSR is a simple and efficient routing protocol designed specifically for use in multi-hop wireless ad hoc networks of mobile nodes. DSR allows the network to be completely self-organizing and self-configuring, without the need for any existing network infrastructure or administration. The protocol is composed of the two main mechanisms of "Route Discovery" and "Route Maintenance", which work together to allow nodes to discover and maintain routes to arbitrary destinations in the ad hoc network. All aspects of the protocol operate entirely on DSR protocol include easily guaranteed loop-free routing, operation in networks containing unidirectional links, use of only "soft state" in routing, and very rapid recovery when routes in the network change. In DSR, Route Discovery and Route Maintenance each operate entirely "on demand" [4], [7]. In particular, unlike other protocols, DSR requires no periodic packets of any kind at any layer within the network.

The sender of a packet selects and controls the route used for its own packets, which together with support for multiple routes also allows features such as load balancing to be defined. In addition, all routes used are easily guaranteed to be loop-free, since the sender can avoid duplicate hops in the routes selected. The operation of both Route Discovery and Route Maintenance in DSR are designed to allow unidirectional links and asymmetric routes.

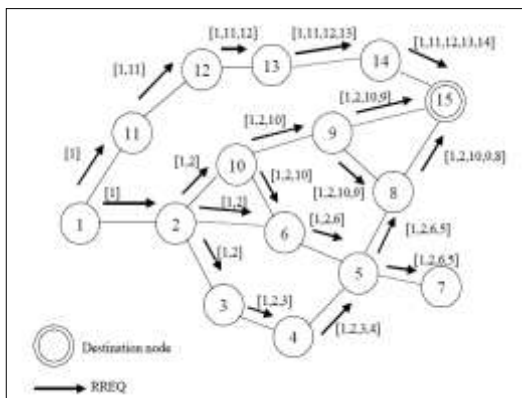


Figure 3. DSR Routing protocol

E. Hybrid Routing Protocol

The Ad Hoc network can use the hybrid routing protocols that have the advantage of both proactive and reactive routing

protocols to balance the delay and control overhead (in terms of control packages). Hybrid routing protocols try to maximize the benefit of proactive routing and reactive routing by utilizing proactive routing in small networks (in order to reduce delay), and reactive routing in large-scale networks (in order to reduce control overhead) [4]. In the literature survey, hybrid routing protocols are compared with proactive routing protocol OLSR. The results show the hybrid routing protocols can achieve the same performance as the OLSR and are simpler to maintain due to its scalable feature. The difficulty of all hybrid routing protocols is how to organize the network according to network parameters. The common disadvantage of hybrid routing protocols is that the nodes that have high level topological information maintains more routing information, which leads to more memory and power consumption.

ZRP: The Zone Routing Protocol (ZRP) localizes the nodes into sub-networks (zones). Within each zone, proactive routing is adapted to speed up communication among neighbors. The inter-zone communication uses on-demand routing to reduce unnecessary communication [4], [7]. An improved mathematic model of topology management to

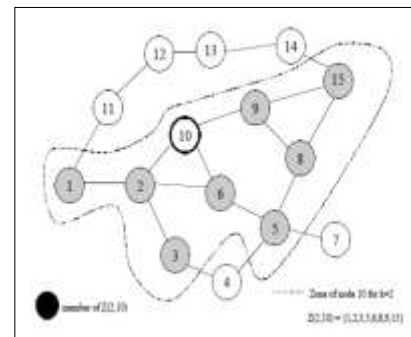


Figure 4. Zone Routing protocol

organize the network as a forest, in which each tree is a zone, was introduced. This algorithm guarantees overlap-free zones. Furthermore, the concept introduced in this algorithm also works with QoS control because the topology model is also an approach to estimate the link quality. An important issue of zone routing is to determine the size of the zone. An enhanced zone routing protocol, is the Independent Zone Routing (IZR), which allows adaptive and distributed reconfiguration of the optimized size of zone. Furthermore, the adaptive nature of the IZR enhances the scalability of the ad hoc network.

F. Geographical Routing

LAR: Location-Aided Routing (LAR) protocol is an approach that decreases overhead of route discovery by utilizing location information of mobile hosts. Such location information may be obtained using the global positioning system (GPS). LAR uses two flooding regions, the forwarded region and the expected region. LAR protocol uses location information to reduce the search space for a desired route, limiting the search space results in fewer route discovery messages. When a source node wants to send data packets to a

destination, the source node first should get the position of the destination mobile node by contacting a location service which is responsible of mobile node positions. This causes a connection and tracking problems. Two different LAR algorithms have been presented. LAR scheme 1 and LAR scheme 2 [5]. LAR scheme 1 uses expected location of the destination (so-called expected zone) at the time of route discovery in order to determine the request zone. The request zone used in LAR scheme 1 is the smallest rectangle including current location of the source and the expected zone for the destination. The sides of the rectangular request zone are parallel to the X and Y axes. When a source needs a route discovery phase for a destination, it includes the four corners of the request zone with the route request message transmitted. Any intermediate nodes receiving the route request then make a decision whether to forward it or not, by using this explicitly specified request zone. Note that the request zone in the basic LAR scheme 1 is not modified by any intermediate nodes. On the other hand, LAR scheme 2 uses distance from the previous location of the destination as a parameter for defining the request zone. Thus, any intermediate node j receiving the route request forwards it if j is closer to or not much farther from the destination's previous location than node i transmitting the request packet to j . Therefore the implicit request zone of LAR scheme 2 becomes adapted as the route request packet is propagated to various nodes [8].

III. CHARACTERISTICS OF MANET

Mobile Adhoc Network (MANET) is a collection of independent mobile nodes that can communicate to each other via radio waves. The mobile nodes that are in radio range of each other can directly communicate, whereas others need the aid of intermediate nodes to route their packets. These networks are fully distributed, and can work at any place without the help of any infrastructure. This property makes these networks highly flexible and robust.

It does not require fixed infrastructure components such as access points or base stations. In a MANET, two or more devices are equipped with wireless communications and networking capability. Such devices can communicate with another node that is immediately within their radio range (peer-to-peer communication) or one that is outside their radio range by using intermediate nodes to relay the packets from the source to destination [4].

It causes route changes, sources may need to traverse multiple and different links to reach the destination every time because all nodes may be moving. Due to this, the traditional routing protocols fail because they assume fixed network topology.

It is self-organizing and adaptive. This means that a formed network can be formed on-the-fly without the need for any system administration. This allows rapid deployment of networks when needed and a quick teardown when not needed

It can consist of heterogeneous devices: i.e., the nodes can be of different types (PDAs, laptops, mobile phones, routers, printers etc) with different computation, storage and

communication capabilities. The only requirement is that the basic MANET software has to be able to run in the devices.

Power consumption can be high because nodes have to be kept alive to forward data packets sent by other nodes that just happen to be in the neighborhood. This especially presents a challenge to the tiny sensors that participate in MANETs.

The characteristics of these networks are summarized as follows:

- Communication via wireless means.
- Nodes can perform the roles of both hosts and routers.
- No centralized controller and infrastructure. Intrinsic mutual trust.
- Dynamic network topology. Frequent routing updates.
- Autonomous, no infrastructure needed.
- Can be set up anywhere.
- Energy constraints
- Limited security

Generally, the communication terminals have a mobility nature which makes the topology of the distributed networks time varying. The dynamical nature of the network topology increases the challenges of the design of ad hoc networks. Each radio terminal is usually powered by energy limited power source (as rechargeable batteries). The power consumption of each radio terminal could be divided generally into three parts, power consumption for data processing inside the RT, power consumption to transmit its own information to the destination, and finally the power consumption when the RT is used as a router, i.e. forwarding the information to another RT in the network [9]. The energy consumption is a critical issue in the design of the ad hoc networks. The mobile devices usually have limited storage and low computational capabilities. They heavily depend on other hosts and resources for data access and information processing. A reliable network topology must be assured through efficient and secure routing protocols for Ad Hoc networks.

A. Network Security

Network security extends computer security, thus all the things in computer security are still valid, but there are other things to consider as well. Network security is then computer security plus secure communication between the computers or other devices. Not all nodes are computers in an Ad Hoc network, thus nodes cannot be assumed to implement the security services normally existent in computers' operating systems. That is why network security should be defined as - Making sure that the nodes enforce a proper computer security and then securing the communication between them-. Different variables have different impact on security issues and design [4]. Especially environment, origin, range, quality of service and security criticality are variables that affect the security in the network. If the environment is concerned, networks can operate in hostile or friendly environments. A battlefield has totally different requirements for security if

compared with home networks. On a battlefield also physical security and durability might be needed to ensure the functionality of the network.

The ways to implement security vary if the range of the network varies. If the nodes are very far from each other, the risk of security attacks increases. On the other hand, if the nodes are so close to each other that they actually can have a physical contact, some secret information (e.g. secret keys) can be transmitted between the nodes without sending them on air. That would increase the level of security, because the physical communication lines are more secure than wireless communication lines. Quality of service issues deal with questions like -Is it crucial that all messages reach their destination?- or -Is it crucial that some information reaches the destination in certain time?-. QoS is generally demanded in real time applications where reliable and deterministic communication is required. These issues refer to security e.g. in process control applications. We could have an Ad Hoc network in some process and all the measurements and control signals could be transmitted through the network. In order to have secure and reliable control of the process, quality of service requirements need to be met [4].

The last variable of Ad Hoc networks described with respect to security is security criticality. This means that before we think of the ways to implement security, we must consider carefully whether security is required at all or whether it matters or not if someone outside can see what packets are sent and what they contain. Is the network threatened if false packets are inserted and old packets are retransmitted? Security issues are not always critical, but it might cost a lot to ensure it. Sometimes there is trade-off between security and costs.

B. Security Problems in MANETs

MANETs are much more vulnerable to attack than wired network. This is because of the following reasons:

- Open Medium - Eavesdropping is easier than in wired network.
- Dynamically Changing Network Topology - Mobile Nodes comes and goes from the network, thereby allowing any malicious node to join the network without being detected.
- Cooperative Algorithms - The routing algorithm of MANETs requires mutual trust between nodes which violates the principles of Network Security.
- Lack of Centralized Monitoring - Absence of any centralized infrastructure prohibits any monitoring agent in the system.
- Lack of Clear Line of Defense - The only use of first line of defense - attack prevention may not succeed. Experience of security research in wired world has taught us that we need to deploy layered security mechanisms because security is a process that is as secure as its weakest link. In addition to prevention, we need second line of defense - detection and response.

Advantages

- They provide access to information and services regardless of geographic position.
- These networks can be set up at any place and time.
- These networks work without any pre-existing infrastructure.

Disadvantages

Some of the disadvantages of MANETs are:

- Limited resources. Limited physical security.
- Intrinsic mutual trust vulnerable to attacks. Lack of authorization facilities.
- Volatile network topology makes it hard to detect malicious nodes.
- Security protocols for wired networks cannot work for ad hoc networks.

IV. PROPOSED APPROACH

We will evaluate the performance of our algorithm using ns-2 simulation [8]. Our institution has extended ns-2 with some wireless supports, including new elements at the physical, link, and routing layers of the simulation environment. Using these elements, it is possible to construct detailed and accurate simulations of ad hoc networks. For scenario creation, two kinds of scenario files are used to specify the wireless environment. The first is a movement pattern file that describes the movement that all nodes should undergo during the simulation. The second is a communication pattern file that describes the packet workload that is offered to the network layer during the simulation. To obtain the performance of MSR at different moving speeds, we will use two simulation sets with speeds of 1 and 25 m/sec, respectively. Our simulations model a network of 50 mobile hosts placed randomly within a 1500 m \times 300 m area, both with zero pause time. To evaluate the performance of MSR, we experimented with different application traffic, including CBR and file transfer protocol (FTP). CBR uses UDP as its transport protocol, and FTP uses TCP. The channel is assumed error free except for the presence of collision.

We chose the following metrics for our evaluation –

- Network size: presented as number of nodes;
- Connectivity: the average degree of a node, normally presented as neighbors;
- Mobility: the topology of the network, relative position and speed of the node;
- Link capacity: bandwidth, bit error rate (BER), etc.
- Queue size: The size of the IFQ (Interface Priority Queue) object at a node
- Packet delivery ratio: The ratio between the number of packets originated by the “application layer” CBR sources

and the number of packets received by the CBR sink at the final destination

- Data throughput: The total number of packets received during a measurement interval divided by the measurement interval
- End-to-end delay
- Packet drop probability
- Average Delay or end-to-end delay
- Hop count
- Message Delivery Ratio
- Normalized routing overload

It is necessary to fine tune the performance measures outlined above such that our proposed criteria has significant advantages over already developed techniques. We have described in great detail the performance measure of path reliability. Path generation to generate possible paths between the senders and receivers and path selection to determine the appropriate paths based on minimal time, will be highlighted in our research. Furthermore, our proposed technique will be highly adaptive, fast and energy/bandwidth efficient.

V. CONCLUSION

As this paper is a research in progress, there has not been a section for discussions laid out here. Nevertheless the issues and highlights about MANET in battlefield zones have been expressed in great detail. Ad hoc networks can be implemented using various techniques like Bluetooth or WLAN for example. The definition itself does not imply any restrictions to the implementing devices. Ad Hoc networks need very specialized security methods. There is no approach fitting all networks, because the nodes can be any devices. The computer security in the nodes depends on the type of node and no assumptions on security can be made. In this paper, the computer security issues have not been discussed, because the emphasis has been on network security.

But with the current MAC layer and routing solutions the true and working ad hoc network is just a dream for now. However, it can be used with relatively small networks and potentially some very nice applications can be realized. Although some peer-to-peer type of solutions work nicely already today, it would be nice to see that some new and innovative solutions would be seen in the arena of ad hoc networks since it is not hard for one to imagine a countless number of nice and ad hoc based applications.

Advances in ad hoc network research have opened the door to an assortment of promising military and commercial applications for ad hoc networks. However, because each application has unique characteristics, (such as traffic behavior, device capabilities, mobility patterns, operating environments, etc.) routing in such a versatile environment is a challenging task, and numerous protocols have been developed to address it. While many protocols excel for certain types of ad hoc networks, it is

clear that a single basic protocol cannot perform well over the entire space of ad hoc networks. To conform to any arbitrary ad hoc network, the basic protocols designed for the edges of the ad hoc network design space need to be integrated into a tunable framework.

As such there has been a lot of research on routing protocols and their impact on network transmission rates and delay. These protocols have significant advantages in their own right and are best suitable for each circumstance in their mode or operating environment. There cannot be a single protocol that is judged as the best of its class and so we have made sure that each category of routing protocol is elaborately described and characterized.

In addition, more research has to be done regarding to network size, mobility, queue size and normalized routing overload parameters. We will continue our work into this regard in future publications and will attempt our best to discuss more about the physical, link and routing layers of the simulation environment.

VI. FUTURE RESEARCH

The emphasis in this paper has been on garnering knowledge in the areas of wireless MANET and their applications in battlefield operations. We have made the best effort to keep abreast of technical developments in this area and so our efforts in documenting our research results and discussions have not been made. In phase II of our future research, we will deploy the settings and performance parameters as outlined in the text, to our research goals. Future scope will include simulating the network parameters such as network size, connectivity, bit error rate, bandwidth and queue size. Spatial location issues such as geolocation will be emphasized using geographical routing algorithms. Using the Global Positioning System (GPS), location aided routing will be used to reduce the search space in fewer route discovery messages. Also, in order to have secure and reliable control of the process, Quality of Service (QoS) requirements will be met.

REFERENCES

- [1]. N.H. Saeed, M.F. Abbod, H.S. Al-Raweshidy, "Modeling MANET Utilizing Artificial Intelligent," *Second UKSIM European Symposium on Computer Modeling and Simulation*, 2008, pp. 117-122.
- [2]. Mobile Ad-hoc Networks (manet) Working Group (<http://www.ietf.org/html.charters/manet-charter.html>).
- [3]. Mobile Computing and Wireless Communications by Amjad Umar, NGE Solutions, Inc. 2004.
- [4]. *The Handbook of Ad Hoc Wireless Networks* by Mohammad Ilyas, CRC Press, 2003.
- [5]. Routing protocols for mobile ad-hoc networks: Current development and evaluation by Zhijiang Chang, Georgi Gadadijev, Stamatis Vassiliadis, 2007.
- [6]. Kniess, J; Loques, O; Albuquerque, C.V.N, "Location aware discovery services and selection protocol in cooperative mobile wireless ad hoc networks", *IEEE Infocom workshop* 2009, pp. 1-2.
- [7]. Ashtrani, H; Nikpour, M; Moradipour, H, "A comprehensive evaluation of routing protocols for ordinary and large-scale wireless MANETs", *IEEE International conference on networking, architecture, and storage*, 2009, 167-170.
- [8]. "Network Simulator," <http://www.isi.edu/nsnam/ns>.
- [9]. Mohammad A. Mikki, "Energy Efficient Location Aided Routing Protocol for wireless MANETs", in *International Journal of Computer Science and Information Security*, Vol 4, No. 1&2, 2009.

AUTHORS PROFILE

Dr. C. Rajabhushanam is with the computer science engineering department at Tagore Engineering College, Rathinamangalam, Chennai, Tamilnadu 600048, India (e-mail: rajcheruk@gmail.com).

Dr. A. Kathirvel is with the computer science and engineering department at Tagore Engineering College, Rathinamangalam, Chennai, Tamilnadu 600048, India (e-mail:kathir.tagore@gmail.com).

An Efficient Resource Discovery Methodology for HPGRID Systems

D.Doreen Hephzibah Miriam

Department of Computer Science and Engineering,
Anna University
Chennai, India
Email: doreenhm@gmail.com

K.S.Easwarakumar

Department of Computer Science and Engineering,
Anna University
Chennai, India
easwara@cs.annauniv.edu

Abstract — An efficient resource discovery mechanism is one of the fundamental requirements for grid computing systems, as it aids in resource management and scheduling of applications. Resource discovery activity involves searching for the appropriate resource types that match the user's application requirements. Classical approaches to Grid resource discovery are either centralized or hierarchical, and it becomes inefficient when the scale of Grid systems increases rapidly. On the other hand, the Peer-to-Peer (P2P) paradigm emerged as a successful model as it achieves scalability in distributed systems. Grid system using P2P technology can improve the central control of the traditional grid and restricts single point of failure. In this paper, we propose a new approach based on P2P techniques for resource discovery in grids using Hypercubic P2P Grid (HPGRID) topology connecting the grid nodes. A scalable, fault-tolerant, self-configuring search algorithm is proposed as Parameterized HPGRID algorithm, using isomorphic partitioning scheme. By design, the algorithm improves the probability of reaching all the working nodes in the system, even in the presence of non-alive nodes (inaccessible, crashed or nodes loaded by heavy traffic). The scheme can adapt to a complex, heterogeneous and dynamic resources of the grid environment, and has a better scalability

Keywords- *Peer-to-Peer; Grid; Hypercube; Isomorphic partitioning; Resource Discovery*

I. INTRODUCTION

Computational Grids and Peer-to-Peer (P2P) computing are the two popular distributed computing paradigms that have been converging in recent years. Computational Grid is an infrastructure that can integrate the computational resources of almost all kinds of computing devices to form a global problem-solving environment. On the other hand, P2P systems aim at resource sharing and collaboration through direct communication between computers without a centralized server as a medium. Computational Grids and P2P are both resource sharing systems having as their ultimate goal the harnessing of resources across multiple administrative domains. These two distributed systems have some commonalities as well as some conflicting goals as discussed in [4]. They have many common characteristics such as dynamic behavior and heterogeneity of the involved components. Apart

from their similarities, Grid and P2P systems exhibit essential differences reflected mostly by the behavior of the involved users, the dynamic nature of Grid resources (i.e., CPU load, available memory, network bandwidth, software versions) as opposed to pure file sharing which is by far the most common service in P2P systems. Although Grid and P2P systems emerged from different communities in order to serve different needs and to provide different functionalities, they both constitute successful resource sharing paradigms. It has been argued in the literature that Grid and P2P systems will eventually converge [12, 17]. The techniques used in each of these two different types of systems will result to a mutual benefit.

Resource discovery is the key requirements in large heterogeneous grid environments, and an effective and efficient resource discovery mechanism is crucial. Traditionally, resource discovery in grids was mainly based on centralized or hierarchical models. Resource discovery could be the potential performance and security bottleneck and single point of failure. The Peer-to-peer systems for discovering resources in a dynamic grid discussed in [19]. Using P2P technology, the resource can be discovered quickly and effectively in grid environment, scalability and robustness can also be improved in P2P Grid.

In this paper, we propose a P2P based Grid resource discovery model, HPGRID system which uses Parameterized HPGRID algorithm, to optimize grid resource discovery and reaches all the grid nodes during searching process even in the presence of non-alive (crashed, inaccessible, experiencing heavy traffic, etc.). The model overcomes the defects of central resource discovery mechanism. The HPGRID model can adapt to the distributed and dynamic grid environments, and has a better scalability. The HPGRID nodes are partitioned isomorphically listing the available resources according to their zones which aids the user to select the needed resource to execute its job rather than traversing the whole grid nodes.

The rest of this paper is structured as follows. Section 2 surveys the related work. Section 3 describes the Hypercubic P2P grid topology. Section 4 describes the HPGRID resource discovery algorithm integrated with Isomorphic partitioning.

Section 5 presents the performance evaluation. Finally, section 6 concludes the paper and presents the future work

II. RELATED WORK

The taxonomy of resource discovery discussed in [21] has identified four main classes of Resource Discovery systems namely centralized, distributed third party, multicast discovery and P2P resource discovery. P2P-based resource discovery systems allow nodes participating in the system to share both the storage load and the query load [18]. In addition, they provide a robust communication overlay. P2P-based Grid resource discovery mechanisms that appear in the literature can be divided into two categories: structured and unstructured [11]. Most proposed systems depend on a structured P2P underlying layer. A structured system however assumes that all pieces of information are stored in an orderly fashion according to their values in a DHT. This is the reason structured systems support efficient resource discovery. However, apart from static resources, Grids include dynamic resources whose values change over time. Whenever the value of a resource attribute stored in a structured system changes, it needs to be republished. If this occurs too often, the cost of republishing becomes prohibitively high.

Iamnitchi et al. proposes resource discovery approach in [7] based on an unstructured network similar to Gnutella combined with more sophisticated query of forwarding strategies which is taken from the Freenet overlay network. Requests are forwarded to one neighbor which are only based on experiences obtained from previous requests, thus trying to reduce network traffic and the number of requests per peer compared to simple query flooding as used by Gnutella. Iamnitchi improves the central control of the traditional grid and adapts fully the decentralized resource discovery in grid environments. However, the limitations are still there in this approach.

Felix Heine et al. propose grid resource discovery approach based ontology and structured P2P technologies in [6]. The approach tackles the semantic problem, but the maintenance of Peer is too high cost because Peer joins and leaves dynamically in structured P2P grid environments. Moreover, the approach focuses on the inherited relationship among grid resource classes and have not discussed the unstructured P2P technologies.

Several P2P schemes, e.g. MAAN [2], NodeWiz [1] and SWORD [8], [16] have been proposed to index and discover Grid resources in a structured P2P network. By using appropriate routing schemes, search queries are routed to the nodes that are responsible for indexing the corresponding resources. Therefore, these schemes scale well to large number of participating nodes. On the other hand, their flat indexing structures pose a major challenge to the global resource monitoring in Grids due to its large-scale and decentralized nature

The HyperCuP system used ontology to organize peers into groups of similar interests using a hypercube topology network [9]. Search queries were forwarded to interest groups to

produce a better hit rate and reduce redundant query messages. This approach required complex construction of the structured hypercube topology network. When joining the network, a peer declared its interest so that the network could put the peer into the cluster of its interest. As P2P is a dynamic environment, a peer might change its interest over time. Constantly updating the network would result in high cost. Furthermore, it would be more complicated if peers had more than one interest. A super-peer model for resource discovery services in large-scale grids discussed in [20]. Zheng [22] describes a model for resource discovery among Grids based on the community categorized by application domain. Rozlina [23] discussed the issues related to matrix for measuring the cost and benefit for choosing the right resource discovery mechanism for a P2P systems. The main purpose of the resource discovery Strategy [24] is to improve the efficiency of the implementation of grid system. Abdelkader Hameurlain [25] provides a survey and a qualitative comparison of the most promising approaches (P2P techniques and agent systems) for RD. Viability of Grid systems relies mainly on efficient integration of P2P techniques and mobile agent (MA) systems to bring scaling and decentralized control properties to Grids.

III. HYPERCUBIC P2P GRID

A. Hypercubic P2P Grid Topology

The Hypercubic P2P Grid Topology is the hypercube structure with additional neighborhood links. In short, we refer Hypercubic P2P Grid as HPGRID.

The Hypercubic P2P Grid nodes have $(k-1) \times (n+1)$ neighbors. Let l , $1 \leq l \leq k^{n-2}$, be the layer of the HPGRID. Let d be the set of nodes at each layer of the HPGRID, then $d = 0, 1, 2, 3$. Also, the number of nodes in HPGRID is k^n , and the number of edges are $n2^{n-1} + k^{n-1}$. The HPGRID Topology for $n = 3$ is depicted in Figure 1. There in, the dashed lines are the additional neighborhood links.

The HPGRID system can be represented by an undirected graph $G = (V, E)$ where $V = \{v_{l,d}, \dots, v_{0,0}\}$.

$$E = \bigcup_{l=1}^{n-2} \{(v_{l,0}, v_{(l+1),2}), (v_{l,1}, v_{(l+1),0}), (v_{l,2}, v_{(l+1),3}), (v_{l,3}, v_{(l+1),1})\}$$

$$\bigcup \{(p, q) : |p \oplus q|_1 = 1\}$$

where p and q are the binary values of the nodes of HPGRID for a 3D HPGRID (000) denotes node 0 and (001) denotes node 1 and so on., and $|p|_1$ denotes the number of ones in p . Here

$$V = \left\{ \begin{array}{l} v_{1,0}(=000), v_{1,1}(=001), v_{1,2}(=010), v_{1,3}(=011) \\ v_{2,0}(=100), v_{2,1}(=101), v_{2,2}(=110), v_{2,3}(=111) \end{array} \right\}$$

and

$$E = \left\{ \begin{array}{l} (v_{1.0}, v_{2.2}), (v_{1.1}, v_{2.0}), (v_{1.2}, v_{2.3}), (v_{1.3}, v_{2.1}) \\ (v_{1.0}, v_{1.1}), (v_{1.1}, v_{1.3}), (v_{1.3}, v_{1.2}), (v_{1.2}, v_{1.0}) \\ (v_{2.0}, v_{2.1}), (v_{2.1}, v_{2.3}), (v_{2.3}, v_{2.2}), (v_{2.2}, v_{2.0}) \\ (v_{1.0}, v_{2.0}), (v_{1.1}, v_{2.1}), (v_{1.2}, v_{2.2}), (v_{1.3}, v_{2.3}) \end{array} \right\}$$

In E , the first four edges are the additional neighborhood links, and the remaining edges are the hypercubic edges.

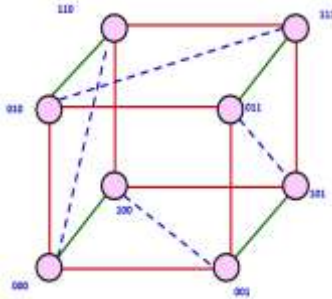


Figure 1. A 3D Hypercubic P2P Grid Topology

B. Representation of HPGRID

Generally, the grid model integrated with P2P mode is composed of many Cubic GridPeers [13]. Each Cubic GridPeer represents a super management domain. Each Cubic GridPeer controls the access of a group of local computing resources. It plays two roles: one as the resource provider and the other as resource consumer. The resource provider allows its free resources to other Cubic GridPeer (consumer), while the consumer arbitrarily uses its local resources or the free resources of other Cubic GridPeers to carry out its task. The resource discovery model for HPGRID is shown in figure 2. The bottom communities of the model using the traditional grid technologies, and the P2P mode are adapted to interact the information between Cubic GridPeers. Here, Cubic GridPeer (CGP) is equivalent to a super node. When they search resources, the users first query the resources in the domain of Cubic GridPeer. If no query result, the search will be carried out through Cubic GridPeer to query the other Cubic GridPeers with P2P way.

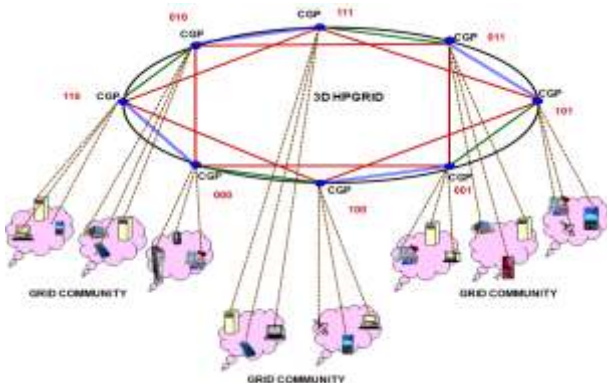


Figure 2. Overview of HPGRID

In HPGRID, each node represents a CGP where each CGP is a collection of Grid Nodes GNs. The GN that belongs to a particular CGP is called Grid Community GC . Each Grid node is represented using its own identifier and the identifier of the corresponding CGP. That is, grid node g is represented as $(ID_g, ID_{CGP(g)})$

At each CGP in the HPGRID system, it contains a CGP Header in the format represented in figure 3.

CGP_ID	Node_ID ₁	Node_ID ₂	...	Node_ID _n
DF	Resrc_Avail	Resrc_Avail		Resrc_Avail
RD	Ptr to RT	Ptr to RT	...	Ptr to RT
LF	Flag	No of CPUs		No of CPUs

Figure 3. CGP Header Format

The CGP Header field description is as follows,

- CGP_{ID} : Cubic Grid Peer Identifier,
- DF : Distance factor,
- RD : Resource Density,
- LF : Load Factor,
- $Flag = 0, CGP$ is non-alive, $Flag = 1, CGP$ is alive,
- $Node_{ID_i}$: Nodes Identifier where $i = 1, 2 \dots n$.
- $Resrc_Avail$: The total number resources available at the node,
- $Ptr\ to\ RT$: Pointer to Resource Table,
- $No.\ of\ CPUs$: Total number of processing element at the node.

Each CGP contains a common resource table which has the details of all the available resources in its own GC . The Resource table format is described in Figure 4. It contains the resource name and number of resources corresponding to its resource number.

Resrc. No	Resource	Total Number
1	Processor	(1-5)
2	Printer	(1-5)
.	.	.
R	.	(1-5)

Figure 4. Resource Table

C. Parameter at Cubic Grid Peer

The parameters represent the state of a Cubic Grid Peer (CGP) that one must meet the following criteria: they must be small, they must facilitate identifying the fertility of a Cubic GridPeer and they must not divulge resource structure information. Based on parallel application characterization experience, we identified the following parameters.

1) Distance Factor (DF)

This gives an idea of how far the target CGP is from the home CGP. A home CGP is defined to be the CGP in which the program and input data are present and to which the output data will go. If it is separated by a large network distance, i.e., high latency and low bandwidth, the staging files and the arriving program and the input files to that CGP will be costly. Another reason why such a factor is important is that tasks in parallel programs might be scheduled on different CGP. Thus there will be some communication between CGP, even though such a situation will be reduced as far as possible by the search algorithm. For tightly coupled applications this may not always be possible and the scheduler might be forced to schedule them on different CGP. This parameter will make CGP between which there is large latency or low bandwidth less desirable to the CGP selector. A high value of this factor makes a CGP less desirable for scheduling.

$$DF = \text{Min_dis}\{h(\text{CGP}), n(\text{CGP})\}$$

where $h(\text{CGP})$ denotes home CGP and $n(\text{CGP})$ denotes neighbor CGP.

2) Resource Density (RD):

This parameter represents the intensity of computing power per unit communication bandwidth. The lower the value of RD, the more will be the bandwidth between every pair of nodes. This signifies that the resources in those CGPs are tightly coupled. For parallel programs that have a communicator in which a small group of processes communicate a lot, a CGP with a low value of RD is important. For example, a SMP will have low RD whereas a network of workstations will have high RD. A similar parameter has been used to represent the computation to communication ratio in schedulers of parallel programs.

$$\text{Resource Density} = \frac{\sum \text{CommunicationBandwidth}}{\sum \text{ProcessorSpeed}}$$

3) Load Factor (LF)

This gives the overall load at some instant in that CGP. This is important to take care of the computation component of the parallel program. Thus parallel processes have a high computation aspect compared to communication which would prefer a better value for LF than for RD.

$$\text{Load Factor} = \sum_i \% \text{ProcessorUtilization}$$

These parameters would be number calculated from information about the state of resources in a particular .CGP

IV. RESOURCE DISCOVERY

In this section, a HPGRID resource discovery model is proposed, and is described as isomorphic partitioning and resource search algorithm.

A. Isomorphic Partitioning

The basic idea of isomorphic partitioning is to partition the HPGRID into $k^n/2^i$ number of hyper cubes, where i is the partition step, where $i = 1$ for $n = 3$, $i = 2$ for $n = 4$, and so on. After Partitioning, the HPGRID is divided into 4 zones namely Z_1, Z_2, Z_3 and Z_4 . Each zones differ in their higher order bit as shown in figure 5. The processor space is partitioned into higher dimensional isomorphic sub-cubes and keeping the same order of dimension. Isomorphic partitioning strategy for HPGRID systems significantly improves the Subcube recognition capability, fragmentation, and complexity compared to existing methods as discusses in [3].

The following zones show the results of isomorphically partition of the 3D HPGRID into 4 zones containing the following nodes at each zone.

$$Z_1(v_{0.0}, v_{1.0}), Z_2(v_{0.1}, v_{1.1}) Z_3(v_{0.2}, v_{1.2}) Z_4(v_{0.3}, v_{1.3})$$

Thus, there is one bit difference between the neighboring zones. The partitioned HPGRID for 3D has been depicted in figure 6. The resulting partitioned sub-cubes are said to be isomorphic in the sense that they are also n-cubes, and for this reason, they retain many attractive properties of Hypercube networks, includes symmetry, low node degree ($2n$) and low diameter (kn).

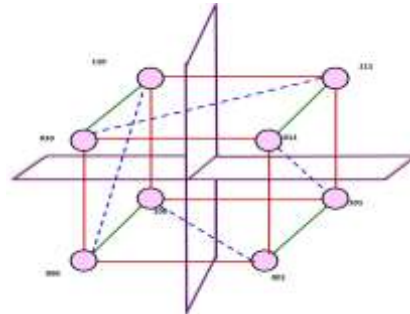


Figure 5. Isomorphic Partitioning of 3D HPGRID System

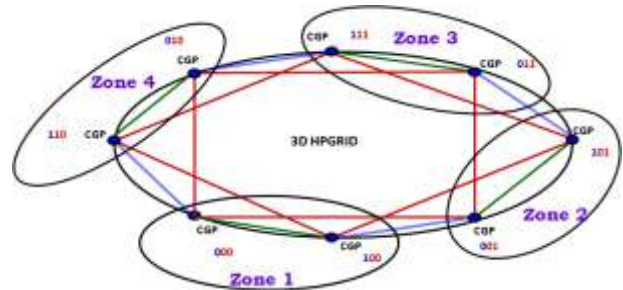


Figure 6. Partitioning 3D HPGRID System.

The 3D HPGRID is isomorphically partitioned into 4 zones containing the following nodes at each zone.

B. Resource discovery algorithms

In this section we present a scalable self configuring resource search algorithm (named Parameterized HPGRID

Algorithm) that is able to adapt to complex environments. It is possible to initiate a search request from any of the live nodes. For the reasons of clarity, however the examples used from now on assume that node 0 is the start node, without a loss of generality.

1) The Search Procedure in an HPGRID using Parameterized HPGRID Algorithm

The search procedure starts when a consumer wants to discover a CGP service. The consumer connects to one of the CGP nodes of the system and requests a service (a resource or some resources). The service discovery is tried first inside the requester's own CGP depending on the parameters explained in Section 3.3. If there is no provider, then the request is redirected to other CGPs. Parameterized HPGRID Algorithm gives the pseudo-code in a node when a new resource request message arrives

Algorithm 1: Parameterized HPGRID Algorithm

```

begin
    satisfyRequest=procRequest(message.request);
    if(satisfyrequest)then
        Calculate DF,RD,LF at its CGP;
        if (DF < Threshold) & (RD is lowerbound)
        &(LF < Acceptable Value) then
            Resource found at the present CGP.
        end
    end
end
if(NOT satisfyrequest)then
    if (startNode) THEN
         $v_d = \{0, 1, 2, \dots, n-1\}$ 
         $v_a = \{ \};$ 
         $vst = \{0000\};$ 
    else
         $v_d = \text{message. } v_d;$ 
         $v_a = \text{message. } v_a;$ 
         $vst = \text{message. } vst;$ 
    end
     $v_d, d_{\text{alive}}, n_{\text{non-alive}} = \text{statusNeighbour}(v_d);$ 
    if( $n_{\text{non-alive}} > 1$ ) then
         $v_{a2} = \text{addtoList}(v_a, d_{\text{alive}});$ 
    else
         $d_{\text{alive}} = \{ \};$ 
    for mk=size(vst), k=0 to (( $v_d.\text{size}()$ ) -  $n_{\text{nonalive}}$  - 1) do
        if ( $v_d$  is not in vst) then
             $\text{message. } v_d = \text{creatList}(k, v_d);$ 
        if ( $v_d[k] = d_{\text{alive}}$ ) then
             $\text{message. } v_a = v_{a2};$ 
        else
             $\text{message. } v_a = v_a;$ 
             $\text{addToList}(vst, v_d[k]);$ 
        end
    end
     $\text{msg.vst} = vst;$ 
    if (mk < size(vst)) then prop=1;
    for (k=mk to(vst.size() - 1)) do
         $\text{sendToNeighbor}(v_d[k], \text{message});$ 
    end
     $v_a, n_{\text{non-alive}} = \text{statusNeighbour}(v_a);$ 

```

```

for (j=0 to (va.size()-nnonalive -1) ) do
    if (neighbor va [j] is not parent
        node) then
        if (vd is not in vst) then
            EmptyList(va [j]);
            prop=1;
        sendToNeighbor(va[j], message);
        end
    end
end

if (prop = 0) then
    for (d = 0) to (d < di) do
        if (d is alive and d is not in
            vst) then
            EmptyList(d);
            sendToNeighbor(d,message);
        end
    end
end
end

```

Algorithm 2: EmptyList(k)

```
begin
  message.vd = { };
  message.va = { };
  message.vst = vst;
  addToList(vst, k);
```

end

2) *Description of the Parameterized HPGRID Algorithm*

1. When a new service request message is received by a node in the HPGRID system, the function `procRequest(message.request)` is called. If the request included in the message cannot be satisfied, the node sets the value of `satisfyRequest` to false and the request will be propagated. Otherwise, `satisfyRequest` is set to true. Here, it calculates the Distance factor (DF), Resource Density (RD) and Load Factor (LF) to check whether resource is available at CGPs Grid Community. If so, no propagation is performed. The message forwarded is composed of the request (`message.request`) and two vectors of dimensions (`message.vd` and `message.va`). In case the request cannot be satisfied and the node that receives the message is the start node (`startNode` is true), the list `vd` is initialized to `vd = {0, 1, ..., n - 1}` (the complete set of dimensions) and `va` is initialized to `va = {}` (an empty list). Otherwise, `vd` and `va` are initialized to the lists received along with the request message. In both the cases the lists represent the set of dimensions (neighbors) along which the message must be propagated.
2. The node calls the `statusneighbors` function and reorders the list `vd` in such a way that the dimensions corresponding to the nonalive neighbors are located at

the last positions of the list. For example, if $v_d = \{0,1,2\}$ and the neighbor along dimension 0 and 1 is not responding, then v_d is reordered to $\{2,1,0\}$. The `statusneighbors` also returns with two integer values $n_{\text{non-alive}}$ and d_{alive} . The integer value $n_{\text{non-alive}}$ represents the number of non-alive nodes in the reordered list v_d . The integer value d_{alive} represents the dimension of the last alive neighbor stored in v_d . For example, if $v_d = \{2,1,0\}$ and its neighbors in dimensions 0 and 1 are non-alive nodes, $n_{\text{non-alive}} = \{2\}$ and $d_{\text{alive}} = \{2\}$.

3. If the number of non-alive neighbors ($n_{\text{non-alive}}$) is more than one, the node calls the `addToList(v_a , d_{alive})` function. This function appends d_{alive} to the end of the list v_a and returns to the new list (v_{a2}) else it make the d_{alive} empty.
4. When a new service request message is received by a node in the HPGRID system, the function `procRequest(message.request)` is called. If the request included in the message cannot be satisfied, the node sets the value of `satisfyRequest` to false and the request will be propagated. Otherwise, `satisfyRequest` is set to true. Here, it calculates the Distance factor (DF), Resource Density (RD) and Load Factor (LF) to check whether resource is available at CGPs Grid Community. If so, no propagation is performed. The message forwarded is composed of the request (`message.request`) and two vectors of dimensions (`message.v_d` and `message.v_a`). In case the request cannot be satisfied and the node that receives the message is the start node (`startNode` is true), the list v_d is initialized to $v_d = \{0, 1, \dots, n - 1\}$ (the complete set of dimensions) and v_a is initialized to $v_a = \{\}$ (an empty list). Otherwise, v_d and v_a are initialized to the lists received along with the request message. In both the cases the lists represent the set of dimensions (neighbors) along which the message must be propagated.
5. The node calls the `statusneighbors` function and reorders the list v_d in such a way that the dimensions corresponding to the nonalive neighbors are located at the last positions of the list. For example, if $v_d = \{0,1,2\}$ and the neighbor along dimension 0 and 1 is not responding, then v_d is reordered to $\{2,1,0\}$. The `statusneighbors` also returns with two integer values $n_{\text{non-alive}}$ and d_{alive} . The integer value $n_{\text{non-alive}}$ represents the number of non-alive nodes in the reordered list v_d . The integer value d_{alive} represents the dimension of the last alive neighbor stored in v_d . For example, if $v_d = \{2,1,0\}$ and its neighbors in dimensions 0 and 1 are non-alive nodes, $n_{\text{non-alive}} = \{2\}$ and $d_{\text{alive}} = \{2\}$.
6. If the number of non-alive neighbors ($n_{\text{non-alive}}$) is more than one, the node calls the `addToList(v_a , d_{alive})` function. This function appends d_{alive} to the end of the list v_a and returns to the new list (v_{a2}) else it make the d_{alive} empty.

7. For each position k in the list v_d represents an live neighbor node, the node calls the `createList(k , v_d)` function which creates a new list composed of all the dimensions located after position k in the ordered list v_d . In other words, if the number of elements in v_d (`v_d.size()`) is q , the function returns $[\{ v_d[k+1], \dots, v_d[q-1] \}]$. For example, if $v_d = \{2,1,0\}$ and $k = 1$, the call to `createList(k , v_d)` will return $\{1,0\}$. Also for each alive neighbor, the v_a list is initialized. The request, v_d , and v_a are sent to the corresponding neighbor in the $v_d[k]$ dimension inside a new message by calling the `sendToNeighbor($v_d[k]$, message)` function. See Figure 7 (a complete example using Parameterized HPGRID Algorithm) where the start node (000) sends $v_d = \{1,0\}$ and $v_a = \{2\}$ to its last alive neighbor (the only one in this case).
8. Finally, the node propagates the request to each of the neighbors along with v_a dimensions only if the corresponding neighbor is not its parent node. Now this propagation takes place under two cases.

Case 1: If the number of elements in v_d is not equal to 0 i.e., $v_d.size() \neq 0$, then the request travels inside a message together with v_a and v_d as empty lists.

Case 2: If the number of elements in v_d is equal to 0 i.e., $v_d.size() = 0$, then the node calls the `statusneighbors(v_a)` function and reorders the list v_a in such a way that the dimensions corresponding to the $n_{\text{non-alive}}$ neighbors are located at the last positions of the list. The `status neighbors(v_a)` also returns two integer values $n_{\text{non-alive}}$ and d_{alive} . The integer value $n_{\text{non-alive}}$ represents the number of non-alive nodes in the reordered list v_a . The integer value d_{alive} represents the dimension of the last alive neighbor stored in v_a .

9. For each position k in the list v_a that represents a live neighbor node, the node calls the `createList(k , v_a)` function which creates a new list composed of all the dimensions located after position k in the ordered list v_a . Also, for each alive neighbor, the v_d list is initialized as empty. The request, v_d and v_a are sent to the corresponding neighbor in the $v_d[k]$ dimension inside a new message by calling the `sendToNeighbor($v_a[k]$, message)` function. For the remaining elements in the list v_a represents non-alive neighbor node, the request travels inside the message together with v_a and v_d as empty lists.
10. Propagating the requests in this way, the effect of non-alive nodes is reduced. Consequently, the algorithm tries to isolate the nodes that are in a non-alive state so that they become leaf nodes (if it is possible) under such circumstances, each node has only one non-alive neighbor, and then all live nodes can be reached. On the other hand, the nodes that are unreachable because of inaccessible or crashed nodes along a path to them, can be reached eventually via other nodes - using the v_a list.

V. PERFORMANCE EVALUATION

In this section, we present the simulation results using the GridSim simulator [14]. The Parameterized HPGRID algorithm has been implemented for three and four dimensional HPGRID system. This algorithm has been evaluated in comparison with the existing algorithm described in [5]. In order to evaluate we describe the following test cases.

- **BEST CASE:** All the nodes in the network are in alive state and the request has been satisfied at the first hop itself.
- **AVERAGE CASE:** Some of the nodes are in alive and some are in non-alive state and the request is satisfied at the next hop.
- **WORST CASE:** Most of the nodes are in non-alive state and the request is satisfied at the last zone.

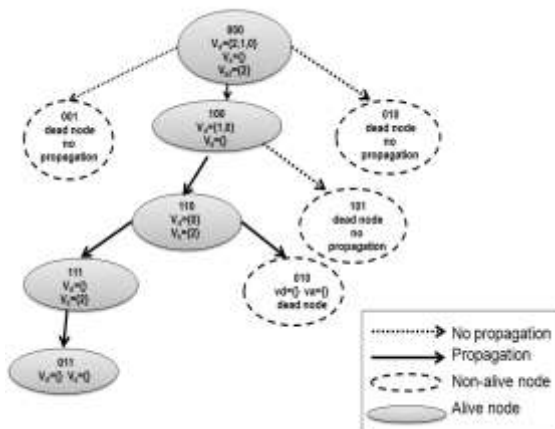


Figure 7. A complete example using Parameterized HPGRID Algorithm. A request of resource started at node 000 in a three-dimensional hypercube

Simulation has been done on a 3D HPGRID for the worst case keeping the nodes source nodes as (000) making the zeroth, first and second dimension nodes as non alive depicted in Figure 8, the following figure 9 gives the complete traversal example for 3D HPGRID system starting from node (000), having all its neighbor non alive namely (001,010,100) except the node present in the additional link node 110 is alive. Sample part of the gridsim output for resource discovery on a best cast 4D HPGRID is shown in Figure 10. Resource search path traversal for the best case in the 4D HPGRID system is depicted in figure 11.



Figure 8. Gridsim output of Parameterized HPGRID Resource discovery Algorithm for a worst case 3D HPGRID

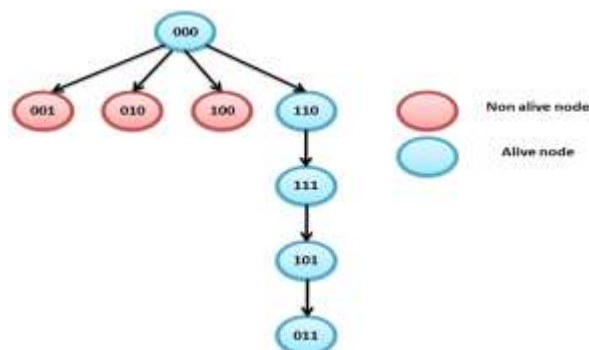


Figure 9. Parameterized HPGRID Resource discovery Algorithm for a worst case 3D HPGRID

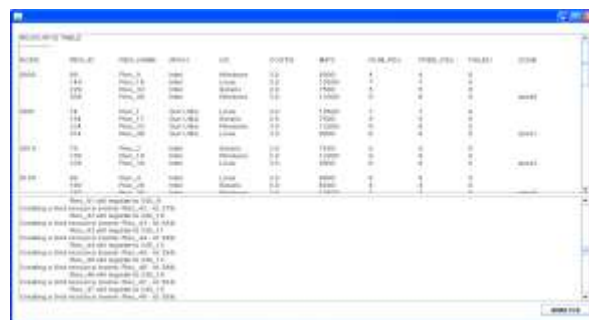


Figure 10. Part of Gridsim output of Parameterized HPGRID Resource discovery Algorithm for a best case 4D HPGRID

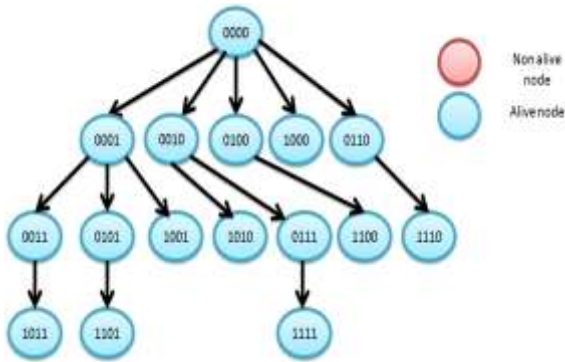


Figure 11. Parameterized HPGRID Resource discovery Algorithm for a best case 4D HPGRID

Figure 12 and 13 depicts that HPGRID algorithm outperforms the existing algorithm in comparison with the number of hops needed to complete the resource discovery process simulated using Gridsim [15] for both the 3D and 4D HPGRID systems in comparison with the HGRID system which does traversals in a normal hypercube.

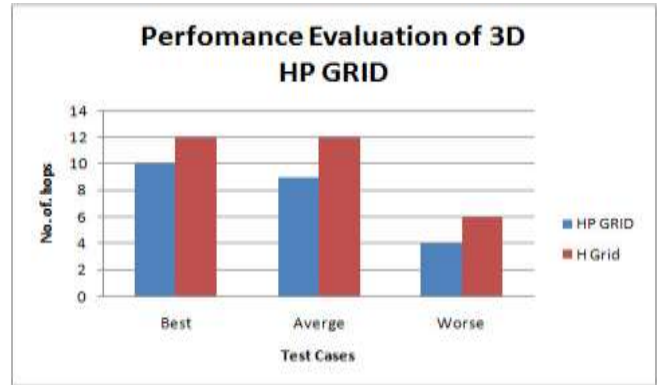


Figure 12. Performance Evaluation of 3D HPGRID

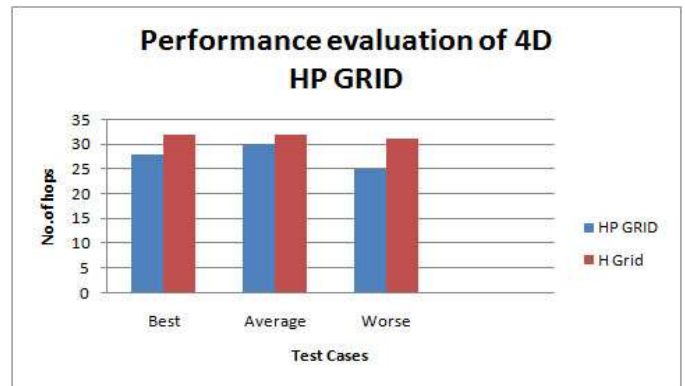


Figure 13. Performance Evaluation of 4D HPGRID

TABLE I. COMPARISON OF RESOURCE DISCOVERY ALGORITHM

Approach/ Property	Peer-to-Peer	Ontology Description	Routing Transferring	Parameter	QoS	Request Forwarding	Hypercubic P2P Grid
Base Approach	agent/query	agent	query	agent	query	agent/query	query
Scalability	More scalable as it uses the four axes framework	Limited scalability centralized broker	Uses routing protocols for scalability hence scalable	Scalable due to grid potential concept used by it	Uses different time map strategies in centralized system to increase the scalability	Nodes are chosen randomly which makes it scalable	More scalable as it uses the structured systems
Reliability	Based on graph theory so reliability increases	Failures are detected as soon as they occurs so more reliable	Quite reliable as it uses the routing concept	Reliable as we can add or delete a node from anywhere	Considers parameters like network bandwidth, required CPU, storage capacity etc. that make it less reliable	Random walk Approach make it reliable in case the resources are equally distributed	Considers Parameters Resource Density, Load Factor, Distance Factor so reliability increases
Adaptability	Multiple platforms environment make it more adaptive	Can be made adaptable by providing manager information about different platforms	Routing table is used to make records of different platforms.	Adaptable due to universal, network and distractive awareness parameters	Depends upon the Service Level Agreement (SLA) sign with user for providing adaptability	Using Best neighbour Approach adaptability is easy	Using Hypercube network so make it more adaptable
Manageability	Complex architecture hence difficult to manage.	Quite easy to manage as a lot of its working is dependent on single node.	Management is easy due to SDRT algorithms as it deals with different topologies	Manage the consistency by using the data dissemination algorithms	Uses algorithm like DIAR for the resource discovery	Better Management can be achieved by combining its two sub Approaches.	Managing the network is easy as it uses structured Hypercube topology
Complexity	$O(\log n)$	$O(\log \log n)$	$O(n)^{1/2}$	$O(n)$	$O(\log 2n)$	$O(n)$	$O(\log_2 n)$

The hypercubic P2P grid approach for resource discovery has been compared with the existing approaches discussed in [10]. The following table gives the comparison study of resource discovery algorithm described in Table I.

VI. CONCLUSIONS AND FUTURE WORK

Our resource discovery scheme in HPGRID system uses Parameterized HPGRID algorithm which reaches all the alive nodes with minimum number of hops. The proposed algorithm is scalable in terms of time, because it keeps the maximum number of time steps required to resolve a resource request, to a logarithmic scale with respect to the total number of nodes. Moreover, each node has knowledge of the overlay CGP using the parameters defined. Therefore, our approach is also scalable and reaches all the alive nodes even in the lesser dimension of its search. Furthermore, scalability is also maintained by querying each node only once at the most (if possible). This important property (scalability) also extends to the number of nodes in the CGP. By using the deep multidimensional interconnection of a hypercube with additional neighborhood links, we provide enough connectivity so that resource requests can always be propagated in spite of non alive nodes. This makes our proposed algorithm much more fault-tolerant when it is compared with other topologies such as centralized, hierarchical or trees. In the absence of non alive nodes, it is able to offer lookup guarantees. Using isomorphic partitioning

scheme if the resource needed not in the start node zones, then the number of resources and the number of tasks under examination are reduced by a single hop, thereby reducing resource discovery time. The future work could be integrating of other resource management issues in this topology which could be extended to generalized topology like k ary n-cube systems. It could be extended considering the scheduling, security, QoS issues and also design and maintenance of new protocols in HPGRID.

REFERENCES

- [1] S. Basu, S. Banerjee, P. Sharma, and S.J. Lee. NodeWiz: Peer-to-peer resource discovery for grids. In *Proceeding of Cluster Computing and the Grid (CCGrid)*, 2005.
- [2] M. Cai, M. Frank, J. Chen, and P. Szekely. MAAN: A multi-attribute addressable network for grid information services. *Journal of Grid Computing*, 2(1):3–14, 2004.
- [3] D. Doreen Hephzibah Miriam, T. Srinivasan, and R. Deepa. An efficient SRA based isomorphic task allocation scheme for k-ary ncube massively parallel processors. In *Proceedings of the International Symposium on Parallel Computing in Electrical Engineering (PARELEC' 06)*, 2006.
- [4] I Foster and Adriana Iamnitchi. On death, taxes, and the convergence of peer-to-peer and grid computing. In *2nd International Workshop on Peer-to-Peer Systems (IPTPS03)*, pages 118–128, 2003.
- [5] A. Gallardo, L. Daz de Cerio, and K. Sanjeevan. HGRID: A hypercube based grid resource discovery. In *Proceeding of the 2nd International Conference on Complex, Intelligent and software Intensive Systems (CISIS)*, pages 411–416, 2008.

- [6] F. Heine, M. Hovestadt, and O. Kao. Ontology-driven p2p grid resource discovery. In Proceedings of the Fifth IEEE/ACM International Workshop on Grid Computing (GRID'04), pages 76–83, 2004.
- [7] A. Iamnitchi and Ian Foster. On fully decentralized resource discovery in grid environments. In Proceedings of the Second International Workshop on Grid Computing, pages 51–62. Springer-verlage, 2001.
- [8] D. Oppenheimer, J. Albrecht, D. Patterson, and A. A. Vahdat. Design and implementation tradeoffs for wide-area resource discovery. In Proceeding of the International Symposium of High Performance Distributed Computing(HPDC), 2005.
- [9] M. T. Schlosser, M. Sintek, S. Decker, and W. Nejdl. HyperCuP - hypercubes, ontologies, and efficient search on peer-to-peer networks. Lecture Notes in Computer Science, 2530:112–124, 2002.
- [10] A. Sharma and S. Bawa. Comparative analysis of resource discovery approaches in grid computing. Journal of Computers, 3(5):60–64, May 2008.
- [11] P. Trunfio, D. Talia, C. Papadakis, P. Fragopoulou, M. Mordacchini, M. Pennanen, K. Popov, V. Vlassov, and S. Haridi. Peer-to-peer resource discovery in grids: Models and systems. Future Generation Computer Systems, 23(7), 2007.
- [12] P. Trunfio and Domenico Talia. Toward a synergy between p2p and grids. IEEE Internet Computing, 7(4):94–96, 2003.
- [13] Z. Xiong, Xuemin Zhang, and Jianxin Chen. Research on grid resource discovery scheme integrated p2p mode. In International Symposium on Electronic Commerce and Security, pages 105–109, 2008.
- [14] Rajkumar Buyya and Manzur Murshed. Gridsim: A toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing. CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE (CCPE), 14(13):1175–1220, 2002.
- [15] Agustin Caminero, Anthony Sulistio, Blanca Caminero, Carmen Carrion, and Rajkumar Buyya. Extending gridsim with an architecture for failure detection. Parallel and Distributed Systems, International Conference on, 1:1–8, 2007.
- [16] Manfred Hauswirth and Roman Schmidt. R.: An overlay network for resource discovery in grids. In In: 2nd International Workshop on Grid and Peer-to-Peer Computing Impacts on Large Scale Heterogeneous Distributed Database Systems, 2005.
- [17] A. Iamnitchi and Ian Foster. On death, taxes, and the convergence of peer-to-peer and grid computing. Lecture Notes in Computer Science, 2735:118–128, 2003.
- [18] Adriana Iamnitchi, Ian Foster, and Daniel C. Nurmi. A peer-to-peer approach to resource discovery in grid environments. In In High Performance Distributed Computing. IEEE, 2002.
- [19] Moreno Marzolla, Matteo Mordacchini, and Salvatore Orlando. Peer-to-peer systems for discovering resources in a dynamic grid. Parallel Computing, 33(4-5):339–358, 2007.
- [20] Carlo Mastroianni, Domenico Talia, and Oreste Verta. A super-peer model for resource discovery services in large-scale grids. Future Generation Computer Systems, 21(8):1235–1248, 2005.
- [21] Koen Vanthournout, Geert Deconinck, and Ronnie Belmans. A taxonomy for resource discovery. Personal Ubiquitous Computing, 9(2):81–89, 2005.
- [22] ZHENG Xiu-Ying, CHANG Gui-Ran, TIAN Cui-Hua and LI Zhen. A Model for Resource Discovery Among Grids. The 10th IEEE International Conference on High Performance Computing and Communications, 678-682, 2008.
- [23] Rozlina Mohamed and Siti Zanariah Satari. Resource Discovery Mechanisms for Peer-to-peer Systems. Second International Conference on Computer and Electrical Engineering, 100-104, 2009
- [24] Honggang Xia and Hongwei Zhao A Novel Resource Discovery Mechanism in Grid. International Conference on Future BioMedical Information Engineering, 493-495, 2009.
- [25] Abdelkader Hameurlain, Deniz Cokuslu, Kayhan Erciyes. Resource discovery in grid systems: a survey. International Journal of Metadata, Semantics and Ontologies , 5(3): 251-263,2010

AUTHORS PROFILE

D. Doreen Hephzibah Miriam is currently a Research Scholar at the Department of Computer Science and Engineering at Anna University, Chennai. She received her B.Tech in Information Technology from Madras University, Chennai, and M.E degree in Computer Science and Engineering from Anna University, Chennai. Her research interests include parallel and distributed computing, peer to peer computing and grid computing.

K. S. Easwarakumar is a Professor & Head at the Department of Computer Science and Engineering at Anna University, Chennai. He received his M.Tech in Computer and Information Sciences from Cochin University of Science and Technology, Cochin and Ph.D in Computer Science and Engineering from Indian Institute of Technology, Madras. His research interests include parallel and distributed computing, Data Structures and Algorithms, Graph Algorithms, Parallel Algorithms, Computational Geometry, Theoretical Computer Science and Molecular computing.

Virtualization Implementation Model for Cost Effective & Efficient Data Centers

Mueen Uddin¹

Department of Information Systems,
Universiti Teknologi Malaysia
Mueenmalik9516@gmail.com

Azizah Abdul Rahman²

Department of Information Systems,
Universiti Teknologi Malaysia
azizahar@utm.my

ABSTRACT: - Data centers form a key part of the infrastructure upon which a variety of information technology services are built. They provide the capabilities of centralized repository for storage, management, networking and dissemination of data. With the rapid increase in the capacity and size of data centers, there is a continuous increase in the demand for energy consumption. These data centers not only consume a tremendous amount of energy but are riddled with IT inefficiencies. Data center are plagued with thousands of servers as major components. These servers consume huge energy without performing useful work. In an average server environment, 30% of the servers are “dead” only consuming energy, without being properly utilized. This paper proposes a five step model using an emerging technology called virtualization to achieve energy efficient data centers. The proposed model helps Data Center managers to properly implement virtualization technology in their data centers to make them green and energy efficient so as to ensure that IT infrastructure contributes as little as possible to the emission of greenhouse gases, and helps to regain power and cooling capacity, recapture resilience and dramatically reducing energy costs and total cost of ownership.

Keywords: *Virtualization; Energy Efficient Data Centre; Green IT; Carbon Footprints; Physical to Live Migration, Server Consolidation.*

I. INTRODUCTION

Data centers form the backbone of a wide variety of services offered via the Internet including Web-hosting, e-commerce, social networking, and a variety of more general services such as software as a service (SAAS), platform as a service (PAAS), and grid/cloud computing [1]. They consist of concentrated equipment to perform different functions like Store, manage, process, and exchange digital data and information. They support the informational needs of large institutions, such as corporations and educational institutions, and provide application services or management for various types of data processing, such as web hosting, Internet, intranet, telecommunication, and information technology [2]. Data centers are found in nearly every sector of the economy, ranging from financial services, media, high-tech, universities, government institutions, and many others. They use and operate data centers to aid business processes, information management and communication functions [3]. Due to rapid growth in the size of the data centers there is a continuous increase in the demand for both the physical infrastructure and

IT equipment, resulting in continuous increase in energy consumption.

Data center IT equipment consists of many individual devices like Storage devices, Servers, chillers, generators, cooling towers and many more. Servers are the main consumers of energy because they are in huge number and their size continuously increases with the increase in the size of data centers. This increased consumption of energy causes an increase in the production of greenhouse gases which are hazardous for environmental health.

Virtualization technology is now becoming an important advancement in IT especially for business organizations and has become a top to bottom overhaul of the computing industry. Virtualization combines or divides the computing resources of a server based environment to provide different operating environments using different methodologies and techniques like hardware and software partitioning or aggregation, partial or complete machine simulation, emulation and time sharing [4].

It enables running two or more operating systems simultaneously on a single machine. Virtual machine monitor (VMM) or hypervisor is a software that provides platform to host multiple operating Systems running concurrently and sharing different resources among each other to provide services to the end users depending on the service levels defined before the processes. Virtualization and server consolidation techniques are proposed to increase the utilization of underutilized servers so as to decrease the energy consumption by data centers and hence reducing the carbon footprints.

This paper identifies some of the necessary requirements to be fulfilled before implementing virtualization in any firm. Section 2 describes a complete description about data centers. Section 3 emphasizes the need for implementing virtualization technology in a data center and provides a five step model of implementing it. Section 4 discusses some of the advantages of virtualization after being implemented. In the end conclusions and recommendations are given.

II. PROBLEM STATEMENT

Data Centers are the main culprits of consuming huge energy and emitting huge amount of CO₂, which is very hazardous for global warming. Virtualization technology provides the solution but it has many overheads, like single point of failure, total cost of ownership, energy and efficiency calculations and return of investment. The other problem faced by IT managers related with the proper implementation of Virtualization technology in data centers to cop up with the above defined problems. This paper comes up with a model to be followed by IT managers to properly implement virtualization in their data centers to achieve efficiency and reduce carbon footprints.

III. LITERATURE REVIEW

In recent years the commercial, organizational and political landscape has changed fundamentally for data center operators due to a confluence of apparently incompatible demands and constraints. The energy use and environmental impact of data centers has recently become a significant issue for both operators and policy makers. Global warming forecasts that rising temperatures, melting ice and population dislocations due to the accumulation of greenhouse gases in our atmosphere from use of carbon-based energy. Unfortunately, data centers represent a relatively easy target due to the very high density of energy consumption and ease of measurement in comparison to other, possibly more significant areas of IT energy use. Policy makers have identified IT and specifically data centre energy use as one of the fastest rising sectors. At the same time the commodity price of energy has risen faster than many expectations. This rapid rise in energy cost has substantially impacted the business models for many data centers. Energy security and availability is also becoming an issue for data centre operators as the combined pressures of fossil fuel availability, generation and distribution infrastructure capacity and environmental energy policy make prediction of energy availability and cost difficult [5].

As corporations look to become more energy efficient, they are examining their operations more closely. Data centers are found a major culprit in consuming a lot of energy in their overall operations. In order to handle the sheer magnitude of today's data, data centers have grown themselves significantly by continuous addition of thousands of servers. These servers are consuming much more power, and have become larger, denser, hotter, and significantly more costly to operate [6]. An EPA Report to Congress on Server and Data Center Energy Efficiency completed in 2007 estimates that data centers in USA consume 1.5 percent of the total USA electricity consumption for a cost of \$4.5 billion [7]. From the year 2000 to 2006, data center electricity consumption has doubled in the USA and is currently on a pace to double again by 2011 to more than 100 billion kWh, equal to \$7.4 billion in annual electricity costs [8].

Gartner group emphasizes on the rising cost of energy by pointing out that, there is a continuous increase in IT budget from 10% to over 50% in the next few years. Energy increase

will be doubled in next two years in data centers [9]. The statistics Cleary shows that the yearly cost of power and cooling bill for servers in data centers are around \$14billion and if this trend persists, it will rise to \$50billion by the end of decade [10].

With the increase in infrastructure and IT equipment, there is a considerable increase in the energy consumption by the data centers, and this energy consumption is doubling after every five years. [11]. Today's data centers are big consumer of energy and are filled with high density, power hungry equipment. If data center managers remain unaware of these energy problems then the energy costs will be doubled between 2005 and 2011. If these costs continue to double every five years, then data center energy costs will increase to 1600 % between 2005 and 2025 [12]. Currently USA and Europe have largest data center power usage but Asia pacific region is rapidly catching up. [13].

IV. PROPOSED WORK

This paper comes up with a virtualization implementation model to be followed by IT managers before implementation of Virtualization technology in their data centers. This technology has many overheads like single point of failure, total cost of ownership, energy and efficiency calculations and return of investment. In this paper I proposed a five step model can be called as pre requisites for the proper implementation of virtualization. The proposed model signifies the importance of categorizing the resources of data center into different resource pools and then selecting and applying the proper virtualization where needed, to maximize the performance of different components as whole in the data center.

Before implementing server virtualization in any firm it is important to seriously plan and consider virtualization risks associated with it. It is also important for the data center to check whether it has the necessary infrastructure to handle the increased power and cooling densities arise due to the implementation of virtualization. It is also important to consider the failure of single consolidated server, because it is handling the workload of multiple applications. In order to properly implement virtualization there is a need to answer some of the questions:

- What is virtualization?
- Why we need it?
- How it can improve our businesses?
- Types of virtualization technologies exist?
- What is cost/benefit ratio of virtualization?
- What new challenges it will bring to business firms?
- Structure of virtualization solution being implemented?
- Which applications or services are good virtualization candidates?
- Server platforms best suited to support virtualization?

A. Proposed Model

We define a layered model consisting of five layers and further each layer comprising of more detailed processes. These components provide a detailed treatment of state of the

art and emerging challenges faced by data centers managers to implement and manage virtualization properly in their data centers to achieve desired objectives. The proposed model defines that, the process of virtualization should be structured and designed in such a way that it must fulfill the necessary requirements and should be within the scope & infrastructure domain already installed in the data center. It is therefore much more than simply loading a virtualization technology on different servers and transforming one or two workloads into virtual machines. Rather it is a complex and rigorous process that need to be implemented and monitored properly. The proposed model defines five key steps need to be followed at different stages in a structural way to achieve the efficiency required in the data center. The components of proposed model are listed below:

- Inventory Process
- Type & Nature of Virtualization
- Hardware Maximization
- Architecture
- Manage Virtualization

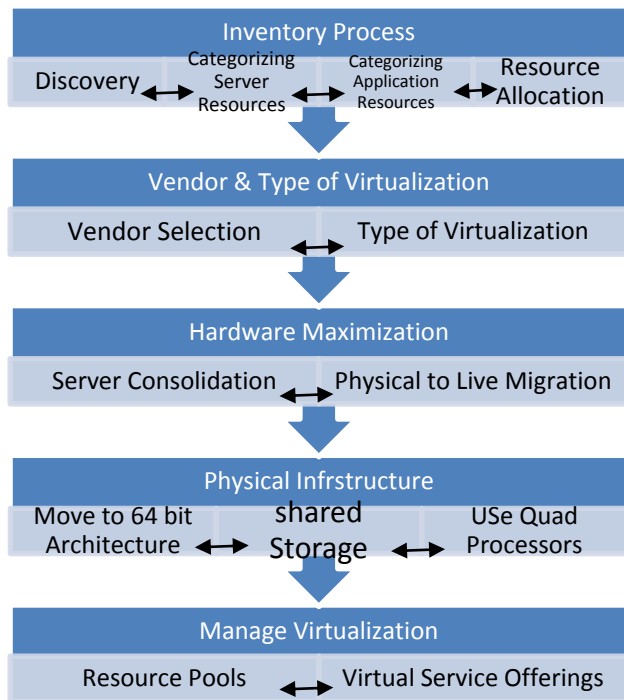


Figure1. Process of Virtualization

B. Inventory Process

The process of virtualization starts by creating an inventory of all hardware and software resources including servers and their associated resources and workloads they require for processing, storage components, networking components etc. The inventory process includes both utilized and idle servers. This process also includes information related to:

- Make and Model of the Processor
- Types of processors (socket, Core, Threads, Cache)
- Memory size and speed
- Network type (Number of ports, speed of each port)
- Local storage (number of disk drives, capacity, RAID)
- Operating system and their patch levels (service levels)
- Applications installed
- Running services
- Running Application
- Storage Devices etc.

The inventory process also discovers, identifies and analyzes an organizations network before it is being virtualized. It consists of following phases:

a) Discovery: It is very important for an organization to know in advance the total content of its infrastructure before implementing virtualization. This is the most important step in any virtualization project. There are many tools available from different vendors for performing initial analysis of an organization.

Microsoft Baseline Security Analyzer (MBSA) tool provides different information like IP addressing, Operating System, installed applications and most importantly vulnerabilities of every scanned system. After analyzing, all generated values are linked to MS Visio, which generates a complete inventory diagram of all components and also provides details about each component being analyzed. Microsoft Assessment and Planning toolkit (MAP) is another tool for the assessment of network resources. It works with windows management instrumentation (WMI), the remote registry service or with simple network management protocol to identify systems on network. VMware, the founder of X-86 virtualization, also offers different tools for the assessment of servers that could be transformed into virtual machines. VMware Guided Consolidation (VGC) a powerful tool assesses network with fewer than 100 physical servers. Since VGC is an agent less tool it doesn't add any overhead over production server's workload.

b) Categorize Server Resources: After creating server inventory information, the next step is to categorize the servers and their associated resources and workloads into resource pools. This process is performed to avoid any technical political, security, privacy and regulatory concern between servers, which prevent them from sharing resources. Once analysis is performed, we can categorize each server roles into groups. Server roles are categorized into following service types:

- Network infrastructure servers
- Identity Management servers
- Terminal servers
- File and print servers
- Application servers
- Dedicated web servers
- Collaboration servers
- Web servers
- Database servers

c) Categorizing Application Resources: After categorizing servers into different resource pools, applications will also be categorized as:

- Commercial versus in-house
- Custom applications
- Legacy versus updated applications
- Infrastructure applications
- Support to business applications
- Line of business applications
- Mission critical applications

d) Allocation of Resources: After creating the workloads, the next process is to allocate computing resources required by these different workloads and then arranging them in normalized form, but for normalization the processor utilization should be at least 50%. It is very important to normalize workloads so as to achieve maximum efficiency in terms of energy, cost and utilization. The formula proposed in this paper for normalization is to multiply utilization ratio of each server by total processor capacity that is (maximum processor efficiency * number of processors * number of cores).

C. Type & Nature of Virtualization

After analyzing and categorizing servers and other resources, the second step defines virtualization in more detail like its advantages, its types, layers and most importantly vendor identification and selection whose product most suits and fulfills all criteria for data gathered in first step.

VMware Capacity Planner (VCP) tool can be used when network size extends over 100 physical servers. It generates reports on server processor utilization including CPU, Memory, and network and disk utilization on server by server basis and finally identifies potential virtualization candidates. Other tools like CIRBA's Power Recon and Plate Spin's are also very useful tools which analyze technical and non-technical factors in data centers and generate reports for the consolidation of servers. It should be noted that all analysis should be done on time for a period of at least one month; this will generate high and low utilization ratios for each server.

D. Hardware Maximization

This is the most important step of virtualization process. Since servers are now going to run multiple virtual workloads, it is important to consider hardware issues because already available hardware is not enough and suitable for providing high availability of virtual workloads. A change is required to install new hardware that supports and delivers the best price and performance. This process ensures high availability of virtual workloads and also provides leaner and meaner resource pool of resources for these virtual workloads. Hardware maximization can also be achieved by purchasing new quad core processors which have better hardware utilization capability along with less consumption of energy hence emissions of CO₂ is greatly reduced.

a) Server Consolidation: Server consolidation is an approach to efficiently use Server resources in order to reduce the total number of servers or server locations to maximize the hardware utilization. This technique is used to overcome the problems of server sprawl, a situation in which multiple, under-utilized servers take up more space and consume more resources than can be justified by their workload. The process of server consolidation always begins from servers that are mostly underutilized and remain idle for long durations of time. The other most important reason for applying server consolidation is that these servers are in huge quantity while the available resources are very much limited. Servers in many companies typically run at 15-20% of their capacity, which may not be a sustainable ratio in the current economic environment. Businesses are increasingly turning to server consolidation as one means of cutting unnecessary costs and maximizing return on investment (ROI) in the data centers [5].

b) Physical to Virtual Live Migration: Physical to Live Migration is the most critical, time-consuming and painful operation when performed manually. Mostly data center managers find this process much more complex and rigorous. It includes cloning existing operating system and restoring it on an identical machine, but at the same time changing the whole underlying hardware, which can lead to driver reinstallation or possibly the dreadful blue screen of death. To avoid these ambiguities, virtualization vendors started to offer different physical to virtual (P2V) migration utilities. These software's speeds up the movement of operation and solve on the fly driver incompatibilities, by removing physical hardware dependencies from server operating systems and allowing them to be moved and recovered. Instead of having to perform scheduled hardware maintenance at some obscure hour over the weekend, server administrators can now live migrate a VM to another physical resource and perform physical server hardware maintenance in the middle of the business day. Virtuozzo for Windows 3.5.1 SWsoft itself introduced a physical to virtual (P2V) migration tool called VZP2V. This tool can remotely install P2V knowing machine administrative username and password.

E. Physical Infrastructure

Data Center physical infrastructure is the foundation upon which Information Technology and telecommunication network resides. It is the backbone of businesses, and its elements provide the power, cooling, physical housing, security, fire protection and cablings which allow information technology to function. This physical infrastructure as whole helps to design, deploy and integrate a complete system that helps to achieve desirable objectives [15].

a) Move To 64-Bit Architecture: The architecture of a machine consists of set of different instructions that allow inspecting or modifying machine state trapped when executed in any or most probably the privileged mode. To support proper hardware utilization, it is important to update and revise whole datacenter architecture. To protect virtual workloads, x-64 systems should be linked to shared storage and arranged into some form of high availability clusters so as

to minimize the single point of failure. One of the major issues in hardware maximization is the proper utilization and availability of RAM for each virtual machine. For this reason it is important to consider 64 bit architecture, which provides more utilization and availability of RAM for all virtual and physical systems.

b) **Rely On Shared Storage:** It is also important to consider single point of failure because one server is now running the workloads of multiple servers. If this server goes down the whole process of virtualization becomes fail. To remove the chances of single point of failure at any stage can be achieved by using redundancy and clustering services to protect virtual workloads. These services are mostly provided by Microsoft and Citrix. While VMware on the other hand uses custom configuration approach called High availability (HA).

F. Manage Virtualization

This is the most important step that involves end users and the top management to make the decision whether to implement the virtualization or not. It involves many factors like cost, return on investment, security, and service level agreements. Virtualized data centers are managed by dividing the functionalities of data center into two layers.

- Resource Pool (RP)
- Virtual Service Offering (VSO)

It is very much important to consider the available resources and what other resources are required to fulfill the requirements of proper implementation of virtualization in data center. It is also important to note that conversion should always be preferred when servers are offline to protect existing services and maintain service level agreements (SLA) with end users.

V. BENEFITS OF VIRTUALIZATION

Virtualization promises to radically transform computing for the better utilization of resources available in the data center reducing overall costs and increasing agility. It reduces operational complexity, maintains flexibility in selecting software and hardware platforms and product vendors. It also increases agility in managing heterogeneous virtual environments. Some of the benefits of virtualization are

A. Server & Application Consolidation

Virtual machines can be used to consolidate the workloads of under-utilized servers on to fewer machines, perhaps a single machine. The benefits include savings on hardware and software, environmental costs, management, and administration of the server infrastructure. The execution of legacy applications is well served by virtual machines. A legacy application may not run on newer hardware or operating systems. Even if it does, it may under-utilize the server, hence virtualization consolidates several such applications, which are usually not written to co-exist within a single execution environment. Virtual machines provide secure, isolated sandboxes for running entrusted applications.

Examples include address obfuscation. Hence Virtualization is an important concept in building secure computing platforms.

B. Multiple Execution Environments

Virtual machines can be used to create operating systems or execution environments that guarantee resource management by using resource management schedulers with resource limitations. Virtual machines provide the illusion of hardware configuration such as SCSI devices. It can also be used to simulate networks of independent computers. It enables to run multiple operating systems simultaneously having different versions, or even different vendors.

C. Debugging and Performance

Virtual machines allow powerful debugging and performance monitoring tools that can be installed in the virtual machine monitor to debug operating systems without losing productivity. Virtual machines provide fault and error containment by isolating applications and services they run. They also provide behavior of these different faults. Virtual machines aid application and system mobility by making software's easier to migrate, thus large application suites can be treated as appliances by "packaging" and running each in a virtual machine. Virtual machines are great tools for research and academic experiments. They provide isolation, and encapsulate the entire state of a running system. Since we can save the state, examine, modify and reload it. Hence it provides an abstraction of the workload being run.

D. Resource Sharing

Virtualization enables the existing operating systems to run on shared memory multiprocessors. Virtual machines can be used to create arbitrary test scenarios, and thus lead to very imaginative and effective quality assurance. Virtualization can also be used to retrofit new features in existing operating systems without "too much" work. Virtualization makes tasks such as system migration, backup, and recovery easier and more manageable. Virtualization provides an effective means of binary compatibility across all hardware and software platforms to enhance manageability among different components of virtualization process.

VI. CONCLUSION

This paper highlights the importance of virtualization technology being implemented in data centers to save the cost and maximize the efficiency of different resources available. We proposed a five step model to properly implement virtualization. It starts by categorizing servers and their associated applications and resources into different resource pools. It is important to consider that virtualization not only needs to characterize the workloads that are planned to be virtualized, but also target the environments into which the workloads are to be applied. It is important to determine the type of servers, their current status whether idle or busy, how much it will cost to implement server virtualization, the type of technology needed to achieve the service levels required and finally meet the security/privacy objectives. It is also important for the data center to check whether it has the

necessary infrastructure to handle the increased power and cooling densities arise due to the implementation of virtualization.

It is also important to consider the failure of single consolidated server, because it is handling the workload of multiple applications. It poses many challenges to the data center physical infrastructure like dynamic high density, under-loading of power/cooling systems, and the need for real-time rack-level management. These challenges can be met by row-based cooling, scalable power and predictive management tools. These solutions are based on design principles that simultaneously resolve functional challenges and increase efficiency.

REFERENCES

- [1] Kant.K, "Data Center Evolution A tutorial on state of the art issues and challenges", Computer Networks, Vol.53, 2009.
- [2] Data Center Journal (n.d.), "What is a data center", Data Center Journal, available at: http://datacenterjournal.com/index.php?option=com_content&task=view&id=63&Itemid=147 (accessed March 16, 2009).
- [3] T. Daim, J. Justice, M. Krampits, M. Letts, G. Subramanian, M. Thirumalai, "Data center metrics An energy efficiency model for information technologists managers", Management of Environmental Quality, Vol.20 No.6, 2009.
- [4] Amit singh,, "An introduction to virtualization", <http://www.kernelthread.com/publications/virtualization>, 2004.
- [5] M. Uddin, A.A. Rahman, "Server consolidation: An Approach to make Data Centers Energy Efficient & Green", International journal of Scientific and Engineering Research, Vol. 1, No.1, 2010.
- [6] Green Grid, "Using virtualization to improve data center efficiency", available at: www.thegreengrid.org/ (accessed February 2009).
- [7] L. Newcombe, "Data centre energy efficiency metrics", 2009.
- [8] W. McNamara, G. Seimetz, K. A. Vales, "Best Practices for Creating The Green Data Center", 2008.
- [9] EPA Report to Congress on Server and Data Center Energy Efficiency – Public Law109-431, Environmental Protection Agency, Washington, DC, 2007.
- [10] Tung, T. Data Center Energy Forecast, Silicon Valley Leadership Group, San Jose, CA, 2008.
- [11] Gartner, "Cost Optimization and Beyond: Enabling Business Change and the Path to Growth", A Gartner Briefing, Gartner, London, 2009.
- [12] EPA Report to Congress on Server and Data Center Energy Efficiency – Public Law109-431, Environmental Protection Agency, Washington, DC, 2009.
- [13] Gartner, "Sustainable IT", A Gartner Briefing, Gartner, Dublin, 2008.
- [14] Caldow, J. "The greening of Government: a study of how governments define the green agenda", p. 8, available at: www01.ibm.com/industries/government/ieg/pdf/green_gov_agenda.pdf, 2008.
- [15] Kumar, R. Media Relations, Gartner, available at: www.gartner.com/it/page.jsp?id/4781012, 2008.
- [16] Wendy Torell, "Data Center Physical Infrastructure: Optimizing business value", APC by Schneider Electric, 2008.

AUTHOR BIOGRAPHIES



Mueen Uddin is a PhD student at University Technology Malaysia UTM. His research interests include digital content protection and deep packet inspection, intrusion detection and prevention systems, analysis of MANET routing protocols, green IT, energy efficient data centers and Virtualization. Mueen has a BS & MS in Computer Science from Isra University Pakistan with specialty in Information networks. Mueen has over ten international publications in various journals and conferences.



Azizah Abdul Rahman is an Associate Professor at University Technology Malaysia. His research interests include designing and implementing techniques for information systems in an organizational perspective, knowledge management, designing networking systems in reconfigurable hardware and software, and implementing security protocols needed for E-businesses. Azizah has BS and MS from USA, and PhD in information systems from University Technology Malaysia. She is a member of the IEEE, AIS, ACM, and JIS. Azizah is a renowned researcher with over 40 publications in journals and conferences.

‘L’ Band Propagation Measurements for DAB Service Planning in INDIA

P.K.Chopra¹, S. Jain², K.M. Paul³, S. Sharma⁴

HoD ECE^{1,2}, Former Engineer in chief³

Ajay Kumar Garg Engineering College¹, IGIT Kashmiri Gate², All India radio³, BECIL⁴
Ghaziabad¹, New Delhi²

India

*prajyot_chopra@indiatimes.com, prajyotchopra@gmail.com*¹

Abstract— The nature of variations of L band satellite signal strength for direct reception -both in fixed as well as in mobile reception are important technical parameters for the planning of satellite broadcast and communication services network. These parameters have been assessed through a field experiment using simulated satellite conditions. Variation of signal strength due to vegetation; urban structures; etc. as well as the building penetration loss along with the Standard Deviation of each of these variations has been assessed based on the data collected during the fixed and mobile reception. This paper gives an insight into the propagation in ‘L’ band under the simulated satellite conditions.

Keywords- Satellite, L-band, Signal, Mobile, Antenna, Attenuation.

I. INTRODUCTION

The ITU World Administrative Radio Conference-92 (WARC-92) held in Torremolinos, Spain, during February/March 1992, allocated frequency bands between 1000 and 3000 MHz (L Band) to Broadcast Satellite Service(BSS)-Sound and complementary Sound broadcasting. These bands are limited to the use of digital sound broadcasting. A frequency band between 1452 and 1492 MHz was allocated on a world-wide basis, with the exception of the USA, to the Broadcast Satellite Service (BSS) and the Broadcast Service (BS) for Digital Sound Broadcasting (DSB).

Before WARC-92 there was no specific spectrum assignment available for BSS (Sound). Therefore WARC-92 represents a major milestone in the development and hence future deployment of DSB services world-wide. With the WARC-92 assignment, the efforts made for some 20 years by broadcasters, administrations, manufacturers, sound programme providers and researchers ultimately met with great success. The 1.5 GHz frequency (L band) allocation is technically close to the optimum for BSS (Sound) and complementary terrestrial services.

The world Administrative Radio Conference (WARC) organized by the ITU in Feb. 1992 inter alia allocated a 40 MHz spectrum in ‘L’ band from 1452 to 1492 MHz for satellite sound Broadcasting service. Report ITU-R BS.955-2 provides details of a new high quality advanced digital sound

broadcasting service (Popularly termed as DAB) which could be supported through satellite in WARC’ 92 allocated ‘L’ band spectrum for large area coverage. The DAB service is designed to provide high quality, multi-channel radio broadcasting for reception by vehicular, portable and fixed receivers. The service is rugged and yet highly spectrum efficient and consumes much lower power.

The planning of a terrestrial and satellite sound broadcasting system is strongly dependent on the factors affecting the characteristics of the propagation channel. The propagation path is subject to attenuation by shadowing due to buildings, different structures, trees, and other foliage and to the multipath fading due to diffuse scattering from the ground and nearby obstacles such as trees and buildings. The degree of impairment to the received signal level depends on the operating frequency, the receiving antenna height, the elevation angle of the satellite and the type of environment in which the receiver is operating: whether it is an open, rural, wooded or mountainous, suburban or dense urban environment.

For moderate satellite elevation angles, it is known that over large areas (of the order of several hundred wavelengths) the mean value of field strength follows a log-normal distribution. However within small areas (of the order of a few wavelengths) two distribution models are applicable:

- Rayleigh distribution where there is no direct line-of sight (LOS) to the satellite.
The only signal is the scattered multipath signal.
- Rice distribution where there is direct line-of sight (LOS) to the satellite, giving
one component of constant amplitude as well as scattered multipath signal.

In urban areas the environment demands application of both these models. So for planning the broadcast service it is necessary to assess the nature of signal level variation in small areas under various receiving environments as stated above.

II. BROADCAST SERVICE PLANNING

A. Location and Time Percentage Correction Factor:

For the planning of broadcast service with analog carrier in the VHF/UHF bands, the field strength prediction is done using ITU-R Recommendation 370 (50% locations, 50% time for the wanted signal and 1% time for the unwanted signal). When planning the Digital Audio Broadcasting (DAB), the same ITU-R Rec. 370 is used but with some correction factors.

In digital broadcasting, due to sharp degradation of quality when the required C/N ratio is not obtained, calculations involving high percentage of time and locations are required for the wanted field (and low percentages for the interfering signals). Therefore an extra correction to the value derived from the ITU-R Rec. 370 curves is required. In digital broadcasting the signal availability is planned for 99% of locations, instead of 50% in case of analog broadcasting. The correction factor for enhancement of location probability from 50% to 99% locations depends on the standard deviation (SD) of the field strength and is taken as $2.33 \times \text{SD}$ dB. This SD has been given in the ITU Rec. 370 as 8.3 dB for VHF and is appropriate for the narrow band signal as in normal analog broadcasting. However a value of $\text{SD}=5.5$ dB has been suggested in ITU-R texts as being more suitable for wideband signals like DAB due to a number of factors including the broadband nature of the DAB signal (COFDM modulation with 1.5 MHz bandwidth), low receiving antenna height and use of omni-directional receiving antenna etc. Taking this value of SD (5.5dB), the correction factor for the 99% locations probability comes out to be $2.33 \times 5.5 = 13\text{dB}$. Measurements in Germany and the UK indicate that still lower values of SD i.e. 3.5 dB and 4 dB on average were available. Field trials have shown that at 1.5 GHz the same $\text{SD}=5.5$ dB can be used.

However more field measurement are necessary in order to confirm the value of $\text{SD}=5.5$ dB.

It was suggested that the DAB service must be provided for 99% time as well. ITU-R Rec.370 does not provide propagation curves for 99% time, however, for short distances the difference between 50% time and 99% time signal strength will be negligible and thus for convenience the existing 50% time curves can be used.

B. Receiving Antenna Height Gain Correction Factor:

ITU-R Rec.370 curves relate to a receiving antenna height of 10 Meters above ground, whereas the DAB service will primarily be planned for mobile and portable reception, i.e. with an effective receiving antenna height of about 1.5 Meters. Measurements performed in Germany on vertically polarized Band III signals showed that an allowance of 10 dB is necessary to convert the field strength at 10 Meters height for use at mobile vehicle antenna height of 1.5 Meters. **There are scopes to reassess this 'Antenna Height Gain Correction Factor' by further field measurements.**

C. Building Penetration Loss Factor:

DAB services are primarily planned for vehicular reception as well as reception on portable receivers at home, without relying on fixed antennas. It follows therefore that an allowance or 'correction factor' to overcome the building penetration losses will be required in the DAB planning process.

Measurements within the UK in the VHF bands indicate that the median value of the building penetration loss appears to be about 8dB with a standard deviation of approximately 4dB. These values are contained in ITU-R Report 1203.

In the Netherlands, the building penetration loss measurements were made at 1.5 Metres antenna height, as part of research on digital television in the UHF band at 762 MHz. The loss was measured to be approximately 7 dB with a low standard deviation. However, according to CCIR Report 567-4, the building penetration losses at 900 MHz, range from 10 to 25 dB. Measurements by DBP-Telekom in Germany showed median values for penetration loss in the 500-600 MHz range to be 13dB for brick structures and 20 dB for concrete.

Work in the 570-670 MHz range by the BBC in the UK showed the median building penetration loss value in the ground floor rooms to be 19 dB referenced to external measurements at 9 Meters above ground.

Building penetration loss values for L-Band are more difficult to quantify. Research on DAB in Canada showed that the median value of the loss in L-Band for an average ground floor location is 17 dB and that for the best ground floor location is 8.4dB (ITU-R Report 1203)

Measurements made by the University of Liverpool in 1988 gave building penetration loss of between 7.5 and 15 dB for L-Band (higher value-for Rayleigh i.e. non-LOS channel; and the lower value- for Rice or partial LOS channel). The mean value is 11.25 dB.

In the late 1993 a number of measurements were made by the BBC in order to quantify the building penetration losses in L-Band (1.5MHz). From the limited no. of measurements made, a median value of 12 dB for the building penetration loss, at a height of about 1.5 Meters, was derived. The buildings measured were of conventional brick-built construction.

A mean of the above figures, gives a mean building penetration loss at ground floor level of 12 dB for L-Band.

However it is to be remembered that the building penetration loss will vary as a function of - not only its structure but also the no. of windows in the building and their size. **Obviously there is a scope of further measurements of this 'building penetration loss' parameter.**

III. L-BAND PROPAGATION MEASUREMENTS WITH 'NARROW-BAND' (CW) AND 'WIDE BAND' (COFDM) SIGNAL:

One of the issues which comes up in the propagation measurements is to examine how different are the fade statistics for the wide band signal (COFDM), compared to that of the narrow band (CW) signal. It has been observed in many field experiments that – the low level of multipath reflections also mean that frequency diversity and/or receiver processing techniques designed to enhance reception in a frequency selective(i.e. multipath) environment are of little consequence. This leads to the conclusion that, for the environments with low multipath reflections, the narrow band (CW) propagation measurements may also provide useful measure of the performance of wide band COFDM systems.

In the case of measurements of 'building attenuation'; 'shadow loss' or 'antenna height gain'-these are the parameters which give the difference between two figures. For example, the difference of signal levels outside and inside the building gives the 'building attenuation'; the difference of signal levels in a given environment taken without shadow and with shadow of an obstacle, gives the 'shadow loss'; similarly the difference between the field strengths measured at 10 meters antenna height and that measured at the required antenna height gives the antenna height correction factor. So these parameters being the difference signal, the effect of multipath or in other words the effect of 'wide band' signal do not make the measured parameters very much different from that of the narrow band (CW) signal.

As such as above mentioned parameters are concerned, the L band propagation measurements with CW signal give quite a fair picture of the propagation channel characteristics using broad band (COFDM) signal.

A field experiment under simulated satellite emission of CW L band signal was planned in Delhi. Through the experiment an attempt was made to assess – the building attenuation; shadow loss and antenna height gain in a CW signal environment.

A. Transmitting System

In order to simulate the satellite, a small L band transmitter of 3 watts power operating at 1440 MHz, along with a circularly polarized transmitting antenna was installed on the top of a high rise building (nearly 70 metres high) in New Delhi Fig. 1 . The antenna was tilted downwards in order to hit the adjoining low height buildings. The transmitting antenna used was of helical type, with a forward gain of 16.2 dBi and Beam width of 30°

Side View of Helical Transmitting
Antenna Used During Measurements

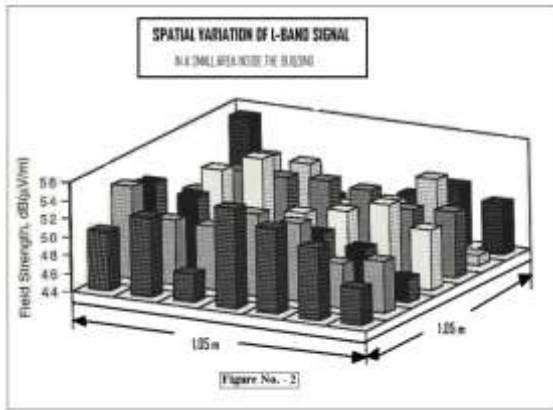


Figure - 1

B. Receiving Setup

The receiving set up used an omni directional vertically polarized ground plane antenna having 0 dBd gain. The signal from the receiving antenna was fed to the spectrum analyzer. The field strength values in dB (uv/m) were obtained after applying the necessary antenna correction factor etc. The height of the receiving antenna was fixed at 1.5 meters above the ground for all indoor as well as outdoor measurements including mobile measurements. Provision was made to change the height of receiving antenna while investigating the variation of received signal with the height of the receiving antenna. A mobile set up in a motor vehicle equipped with spectrum analyzer and computer controlled data acquisition system was used for the experiment.

Owing to multipath, the signal varies significantly with very small distances. As such while measuring signal strength at a particular location inside the building, instead of taking a single observation, a large set of readings (49 Nos.) separated by $\frac{3}{4}$ of the wavelength (about 15 cm) spread over an area of 1.05mX1.05m were taken and the average value worked out corresponding to each site. Fig.2 provides a sample of fluctuations of field strength at 1440 MHz within the small area.



While measuring the building attenuation, the locations inside the building have been divided into three types – ‘Best Locations’; ‘Average Locations’ and ‘Worst Locations’. The Best Locations are sites very close to the doors/windows on the outer building wall facing the transmitter. The ‘Average Locations’ are the sites inside the building which are slightly inside the building and away from the ‘Best Locations’. The third type is the ‘Worst Locations’. These are the sites deep inside the building and obstructed from transmitter by multiple attenuating walls.

C. Measurement of Building Attenuation

To ascertain the amount of attenuation of L-Band signal inside multistoried concrete buildings, a set of measurement were planned. Three typical buildings in near vicinity of the transmitting site were selected for the purpose. The transmitting antenna was made to direct towards the middle floor of the building selected for measurements. A set of measurements were carried out at the First building which is a 6 storied cement concrete building. The transmitter was making an elevation angle of 32 degrees from the measuring site at the building. In case of the Second and Third buildings the similar elevation of the transmitter from the respective measuring sites were 38.9 degrees and 47.7 degrees respectively. The sample of ‘L’ band field strength measurements taken at ground floor of a multistoried building is given in Annexure 1.

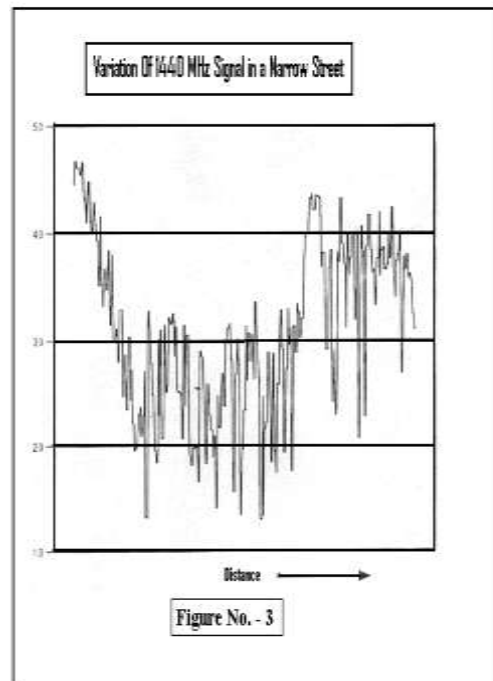
D. Measurement of Shadow Loss

A set of field strength measurement were taken in direct line of sight and in various shadow conditions of the buildings to evaluate shadow losses encountered in typical urban environment. These measurements were taken with the help of the Mobile set up, by moving the vehicle at a very slow speed (3-5 km/hour). A sample of field strength variation measured during the experiment is given in Fig.3. The receiving antenna was mounted on the rear side of the vehicle at a height of 1.5m from the ground. Continuous field strength recordings were made for line of sight as well as for the shadow areas as encountered on the way. The measurements can be classified into the following three types.

- Measuring points in direct line of sight of the transmitting antenna.
- Measuring points located in the shadow of a multistoried building with large open area behind (no reflecting structures).
- Measuring points located in Narrow Street between multistoried concrete buildings.

E. Field Strength Measurement with Variation of Receiving Antenna Height

In order to determine the variation of field strength with height of the receiving antenna above the ground, the field strengths were measured at a fixed location by lowering the receiving antenna from 8m to 1.5m. Measurements were made in direct line-of-sight as well as in obstructed visibility.



IV. ANALYSIS OF DATA

In order to assess building attenuation, the field strength values measured inside the building at each floor for various sites were compared with the respective line-of –sight values measured outside the windows. Thus, for each building and floor, building attenuation was worked out in dB as given in Table-1.

The building attenuations for ‘Best’, ‘Average’ and ‘Worst’ sites were found to be 7.4 to 15.5 dB; 9.5 to 19.0 dB and 14.0 to 24.9 dB respectively with a standard deviation in the range of 3 to 5 dB.

For assessment of shadow losses, the predicted field strength for free space propagation for various measuring conditions were calculated. Different percentile values of the measured field strengths relative to their respective predicted free space values were plotted. Fig. 4 (a) & 4 (b) show the statistical distribution of the relative field strengths in various environmental conditions. The shadow loss in dB was obtained by comparing the relative field strength values of direct line-of-sight with those obtained in shadow areas. The shadow loss of multistoried building with large open area behind was found to be 21 dB for an elevation angle of 26° . The shadow loss of multistoried buildings in Narrow street in typical urban environment was found to be 12 dB for average elevation angle of 42° .

Fig.5 provides the fluctuation of the field strength as a function of the receiving antenna height at a fixed point in direct line of sight as well as in obstructed visibility. The field strengths have been expressed relative to the predicted field strength for free space propagation.

V. CONCLUSION

The present study provides some insight into the radio wave propagation in 'L' band under the simulated satellite conditions. The attenuation inside multistoried buildings was found to vary from 7.4 to 15.5 dB for 'Best' locations and from 14.0 to 24.9 dB for 'Worst' locations. The shadow loss of multistoried buildings with open fields behind was found to be 21 dB whereas in Narrow Street between multistoried buildings it was found to be 12 dB. The received field strength was also found to be fluctuating around its mean value with the variation of receiving antenna height. The fluctuation was more prominent in shadow area as compared to direct line of sight conditions.

ACKNOWLEDGEMENT

I sincerely thank UNIVERSITY SCHOOL OF ENGINEERING & TECHNOLOGY, Guru Govind Singh Indraprastha University, New Delhi for providing the opportunity and guidance for research work.

REFERENCES

- [1] ITU-R special publication on "Terrestrial and satellite digital sound broadcasting to Vehicular, portable and fixed receivers in the VHF/UHF bands", Radio communication Bureau, Geneva, 1995.
- [2] Butt, G., Evans, B.G. Richharia, M.: "Narrowband channel statistics from multiband propagation measurement applicable to High Elevation Angle land mobile satellite Systems", IEEE journal on selected areas in communications, vol 10, no, 8, October, 1992, pp.1219-1226
- [3] Goldhrish, J. Vogel, W.J.: "Mobile Satellite system Fade Statistics for Shadowing and Multipath from Roadside Trees at UHF and L-band", IEE Transactions on Antennas & propagation, vol. 37, no. 4 April, 1989, pp 489-498

AUTHORS PROFILE



Prof. P. K. Chopra, Dean and HoD-ECE

prajyotchopra@gmail.com, prajyot_chopra@indiatimes.com

AJAY KUMAR GARG ENGINEERING COLLEGE, GHAZIABAD, U.P. (India)

Prof. Pradeep Kumar Chopra entered the field of education in the year 2004 after 24 years of exemplary service in the technical branch of the **Indian Air Force**. He earned his Bachelors degree in Engineering (Electronics) from **Delhi college of Engineering** in the year 1979 and Masters in Technology from **IIT Delhi** in the year 1985. He also has a Masters degree in Defence Studies from **Madras University**. While he was in the Indian Air Force he was part of, and headed a number of important technical projects. For his exemplary services he was awarded "**Vishist Seva Medal**" by the **President of India** in the year 1993. He took premature retirement from the IAF in the year 2004 and entered the field of education. He is the **Dean and Head of Dept.** (Electronics and Communication) in Ajay Kumar Garg Engineering College in Ghaziabad. AKGEC is rated as the No 1 Engineering Colleges in U.P. (India) affiliated to Uttar Pradesh Technical University, Lucknow (U.P.) India.

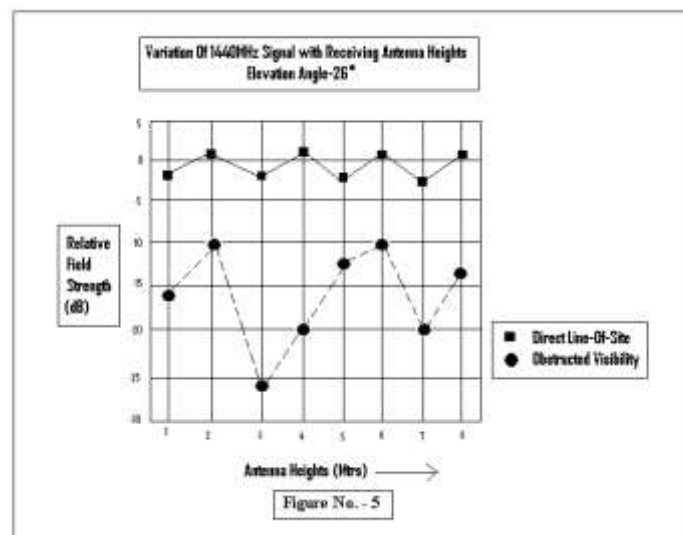
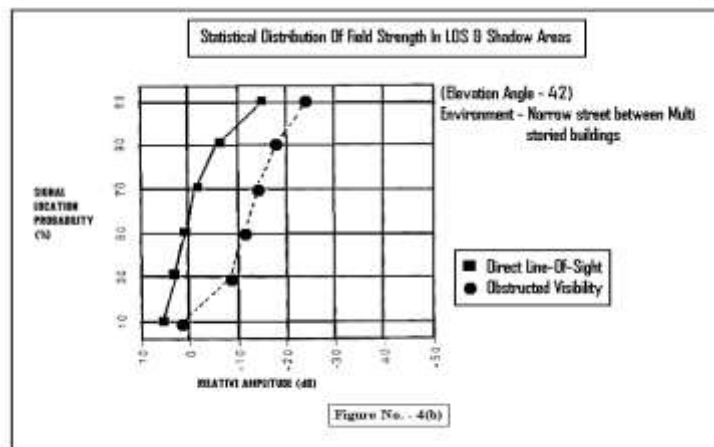
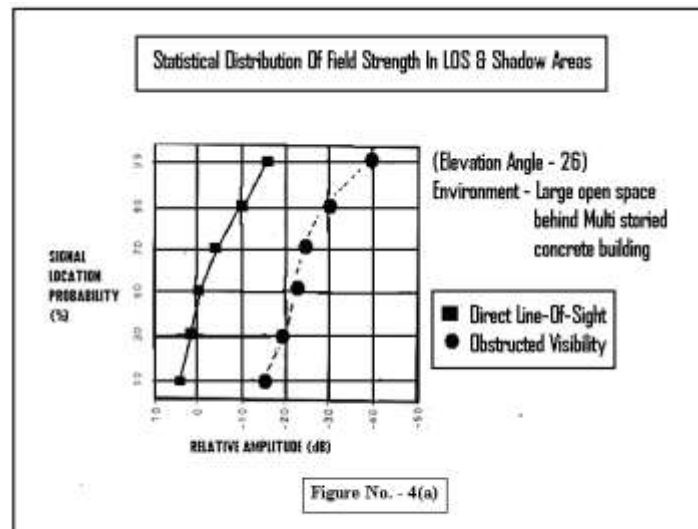


TABLE – 1

Building Attenuation in dB

Sl. No.	Name of Building/ Elevation Angle	Floor	Building Attenuation (dB)		
			Best Location	Average location	Worst location
1.	Y-Shape/32°	Ground	15.5	16.0	24.9
		1 st	7.4	14.2	23.6
		2 nd	10.2	14.5	20.2
		3 rd	7.8	16.1	19.4
		4 th	12.8	18.9	23.0
		5 th	13.0	19.0	24.9
2.	SPA/38.9°	Ground	12.2	15.5	17.9
		1 st	7.6	-	18.0
		2 nd	8.4	9.5	14.0
		3 rd	10.1	12.8	16.7
3.	AVM/47.7°	Ground	10.6	16.7	20.4
		1 st	9.9	14.4	14.3

ANNEXURE – 1

SAMPLE OF ‘L’ BAND FIELD STRENGTH MEASUREMENTS TAKEN AT GROUND FLOOR OF A MULTISTORIED BUILDING

“-----
“05-04-2010”,”,”, “AT AVM – GW1”, “# “, 1
“START TIME=”, “16:28:02”, “STOP TIME=”, “16:28:04”, “ON STATIONARY VEHI
“AVERAGE SIGNAL=”, 25.94333, “ dBUV / M”, “STD.DEV.=”, .5938197, “dBUV / M”
25.85,25.5,25.5,25.5,25.5,26.05,26.05,26.05,27.05,27.05,27.05,25.5,25
“-----
“05-04-2010”,”,”, “AT AVM – GW1”, “# “, 2
“START TIME=”, “16:28:43”, “STOP TIME=”, “16:28:46”, “ON STATIONARY VEHI
“AVERAGE SIGNAL=”, 23.73667, “ dBUV / M”, “STD.DEV. =”, .7605847, “dBUV / M”
24.35, 24.35, 24.3, 24.3, 24.3, 24.3, 24.65, 24.65, 22.55, 22.55, 22.55, 23.3,23
“-----
“05-04-2010”,”,”, “AT AVM – GW1”, “# “, 3
“START TIME=”, “16:29:09”, “STOP TIME=”, “16:29:11”, “ON STATIONARY VEHI
“AVERAGE SIGNAL=”, 28.98, “ dBUV / M”, “STD.DEV. =”, .4925444, “dBUV / M”
29.5,29.5,29.5,29.5,28.4,28.4,28.4,28.4,29.65,29.65,28.7,28.7,28.7,28
“-----
“05-04-2010”,”,”, “AT AVM – GW1”, “# “, 4
“START TIME=”, “16:29:36”, “STOP TIME=”, “16:29:38”, “ON STATIONARY VEHI
“AVERAGE SIGNAL=”, 25.53333, “ dBUV / M”, “STD.DEV. =”, .2958977, “dBUV / M”
25.65, 25.65, 25.8, 25.8, 25.8, 25.8, 25.05, 25.05, 25.25, 25.25, 25.25, 25.25, 25
“-----
“05-04-2010”,”,”, “AT AVM – GW1”, “# “, 5
“START TIME=”, “16:30:21”, “STOP TIME=”, “16:30:23”, “ON STATIONARY VEHI
“AVERAGE SIGNAL=”, 16.25, “ dBUV / M”, “STD.DEV. =”, .7722261, “dBUV / M”
15.05, 15.05, 15.05, 15.05, 16.6, 16.6, 16.6, 16.6, 16.9, 16.9, 16.45, 16.45, 16
“-----
“05-04-2010”,”,”, “AT AVM – GW1”, “# “, 6

“START TIME=”, “16:30:47”, “STOP TIME=”, “16:30:49”, “ON STATIONARY VEHI
“AVERAGE SIGNAL=”, 23.99333, “dBUV / M”, “STD.DEV. =”, .6096078, “dBUV / M”
24.4, 24.4, 24.4, 23.2, 23.2, 23.2, 23.2, 23.4, 23.4, 24.65, 24.65, 24.65, 24

“-----

“05-04-2010”, “”, “AT AVM – GW1”, “# “, 7
“START TIME=”, “16:31:17”, “STOP TIME=”, “16:31:19”, “ON STATIONARY VEHI
“AVERAGE SIGNAL=”, 30.62, “dBUV / M”, “STD.DEV. =”, .2372061, “dBUV / M”
30.3, 30.3, 30.3, 30.9, 30.9, 30.9, 30.9, 30.75, 30.75, 30.75, 30.35, 30.3, 30

“-----

“05-04-2010”, “”, “AT AVM – GW1”, “# “, 8
“START TIME=”, “16:31:36”, “STOP TIME=”, “16:31:38”, “ON STATIONARY VEHI
“AVERAGE SIGNAL=”, 30.71, “dBUV / M”, “STD.DEV. =”, .3791219, “dBUV / M”
30.3, 30.25, 30.25, 30.25, 30.25, 30.65, 30.65, 30.65, 30.65, 31, 31, 31, 31.25, 31

“-----

“05-04-2010”, “”, “AT AVM – GW1”, “# “, 9
“START TIME=”, “16:32:02”, “STOP TIME=”, “16:32:04”, “ON STATIONARY VEHI
“AVERAGE SIGNAL=”, 23.16667, “dBUV / M”, “STD.DEV. =”, .7196444, “dBUV / M”
23.05, 23.05, 23.05, 23.05, 22.45, 22.45, 22.45, 22.45, 24.3, 24.3, 24.3, 24.3, 24

“-----

“05-04-2010”, “”, “AT AVM – GW1”, “# “, 10
“START TIME=”, “16:32:17”, “STOP TIME=”, “16:32:20”, “ON STATIONARY VEHI
“AVERAGE SIGNAL=”, 30.59667, “dBUV / M”, “STD.DEV. =”, .1309794, “dBUV / M”
30.6, 30.6, 30.6, 30.6, 30.7, 30.7, 30.7, 30.7, 30.35, 30.35, 30.35, 30.7, 30

30.6, 30.6, 30.6, 30.6, 30.6, 30.7, 30.7, 30.7, 30.7, 30.35, 30.35, 30.35, 30.7, 30

Universal Simplest possible PLC using Personal Computer

B.K.Rana

Electrical Engg.,

VSSUT, Sambalpur Orissa , India

bkrana@rediffmail.com

Abstract— Need of industrial automation and control is not closed yet. PLC, the programmable logic controller as available in 2009 with all standardized possible features, discussed here concisely. This work on PLC gives a simplest form built in any computer hardware and software for immediate flexible need of the product in a small environment any number inputs and output logic permitted by a computer. At first, the product logic is implemented using simple available computer using a PCIMCIA card having 8255 parallel I/O device, RS232 etc. software like Java, C, C++, and Visual Basic. Further work continuing with all coming generations and variation in the mentioned technology. Expert engineer having these skills may fabricate a small operating PLC within one month. This is a view of preliminary work. Further up gradation of this work comes in term of Visual programming tool (computer aided design) and also universal make in terms of all operating system and computer available in the world. Sensors and module for data to logic conversion is not emphasized in this work. Connectivity has to be as per standard. Immediate use of this work as an education technology kit. The work in this huge technology area should not be accused because this is need in some way.

Keywords- PLC, prototype, Java, Visual Basic, education technology

I. INTRODUCTION

Everybody will agree that there is no need to describe the 25 years of technology and concepts of PLC. It is surveyed and many small key points are discussed towards another offshoot of this technology. The objective to develop a low cost trainer or education technology kit is realized. This work further can be extended to a small PLC prototype. As a trainer it should be safe, low cost, programming flexibility, multi input and multi output system. Logic system at front end computer (display, editor, logic evaluation), universal communication for commercial PLCs are the work done here. Universal make for these components are considered so that it works with any computer system. At present the work is carried in Personal computer. Java being a universally portable front end environment is considered here.

II. PRESENT PLC FEATURES

A. A. Manufacturing Companies

It is manufactured by companies like Brands - Lab-Volt, Allen-Bradley, Rockwell Automation, Allen-Bradley, Panel View Operator Terminals, Aromatic PLC , Automation Direct, Cutler Hammer, GE Series, Exor, IDEC, Koyo, Maple Systems, Mitsubishi, Madison, Omron PLC, Reliance Automate, Siemens, Square D, Texas Instruments, Toshiba.

B. Use of PLC at this stage

Precisely PLC is used in industrial environment for area like, Equipment status Process Control Chemical Processing, Equipment Interlocks, Machine protection, Smoke detection, Gas monitoring, Envelope monitoring, Personal monitoring, System setup and device monitoring , Bench marks, maximum I/O ports etc.

C. Lingering concerns about PLC

- Security and robustness. Failure in any part of the computer system such as crash, rebooting, disk problem gives inaccuracy to plant output functions and hence uncertainty in real plant dynamics.
- PLC in networked computers. Operating PLC from now days generic Ethernet based computer network is one of studies going now.

D. Standards for PLC

Contrary to proprietary or closed design open system design is carried out 1980 onward. This gives control over technology and flexibility to customer.

- International engineering consumerism IEC 61131-3 standardize ladder logic (graphical) , function block diagram (graphical) , structured text (textual) and instruction set (textual) . IEC 61850 process bus interface.
- In detail IEC 61131-3 is given as overview, requirement and test procedures, data type and programming, user guide lines, communication and fuzzy control.

- Organize programs for sequential and parallel control processing.

E. Fundamentals of PLC software

The list of software elements in PLC are method of representing logic, instruction, code, graphical presentation (functional logic diagram and ladder logic) Fundamental ladder logic instruction set comparison of different manufacturers, processor, memory and instruction code, Real Time ability particularly real time clock.

F. Fundamentals of PLC hardware

The list of hardware elements in PLC are block diagram of typical PLC, PLC – processor module memory organization, Input /Output section: module types, Power supply .

III. CONCEPT OF THIS PLC

This work of PLC, after a long evolution, aims at producing an education technology kit at first with very minimum cost. Also this work in later version will produce cream of application of Software Engineering in industrial automation. The simple possible PLC work is done with present software available for present personal computers.

The cost of producing this simple work comprises of a personal computer, Visual Basic, Java (free), C++, Printer & com port cable etc. may be \$ 5000.00. Excessive optimization is done to make this PLC as simplest possible.

IV. PRELIMINARY DESIGN

A. Software Engineering Requirements

Since this product logic is based on matured software languages and tools for visual concepts no more to bother for this. For example Java being a Object Oriented Programming Language nothing mean in terms of software engineering for this project. Communication and Device driver part requires some special Software Engineering for which plenty of international works are available in IEEE [2, 3].

B. Connection in this PLC

Parallel or Serial port driver for PC to PLC connection may be built with Java front end and Java Native–Win32 driver. Further work is under progress. Figure 1 gives the software architecture stack for communication module. User interface of the communication stack can be high level language like java with native interface. Device driver layer may be developed with C or C++. This communication stack is built with the concept of OSI (open system interconnection) [4].

Table 1 lists the possible use of communication devices like Intel 8255 (parallel device), Intel 8251(serial device) and Personal Computer's serial communication port like COM1 and COM2. Communication hardware module may be designed with standard PC bus interface specification like PCMIMCIA.

C. Report generation

Report of activity and data may be generated in this experimental PLC model using Microsoft Excel for Visual Basic front end design tool , simple text file is first for all universal form of design . In windows environment simple database like MS Access is sufficient and XML may be a universal form.

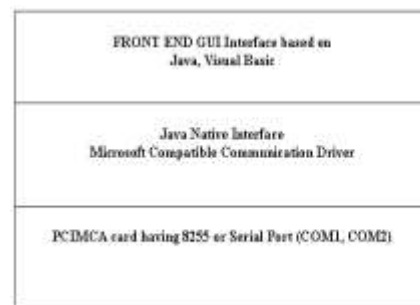


Figure 1. Present Version of communication stack and interface

D. Theoretical Interface

This section depicts what should be the logic of designing the human machine interface for the small effective PLC. This can be designed using Java, Visual Basic, Motif, C++ etc. The graphical user interface components depicted in this design section are standard components like Frame, Command Button, Check Box, Image etc. Figure 2 gives theoretical interface for this small PLC.

The “Plant Control < number> “is the container for the ladder diagram. Any number (N) of Plant Control can be run simultaneously. The ladder diagram is designed using standard drawing software tool like Microsoft Word and the cut with image file like (jpeg, gif, bmp etc.). It can be selected to run or disable from running by the left check box.

For each plant control, the output function should be written in the used programming language. Optimization of this function by Karnaugh map be used.

TABLE I : AVAILABLE COMMUNICATION STANDARD AND MECHANISM

Communication device / standard	Type	PC Interface	Remark
8255	Intel parallel port device	PCIMCIA card	Data Link Layer Layered Architecture Device driver Security in communication , Raw driver and connection to TCP/IP
RS 232	Serial communication standard, Device 8251	Windows driver exists PCIMCIA card	COM1 and COM2 available in PC
RS 485	Serial communication standard	Windows driver exists PCIMCIA card	
Radio	Device , like wireless LAN NIC Broad cast	Drivers may be available	Present Wireless LAN Driver
COM1 , COM2	Existing PC Serial port Serial devices can be used in PC		

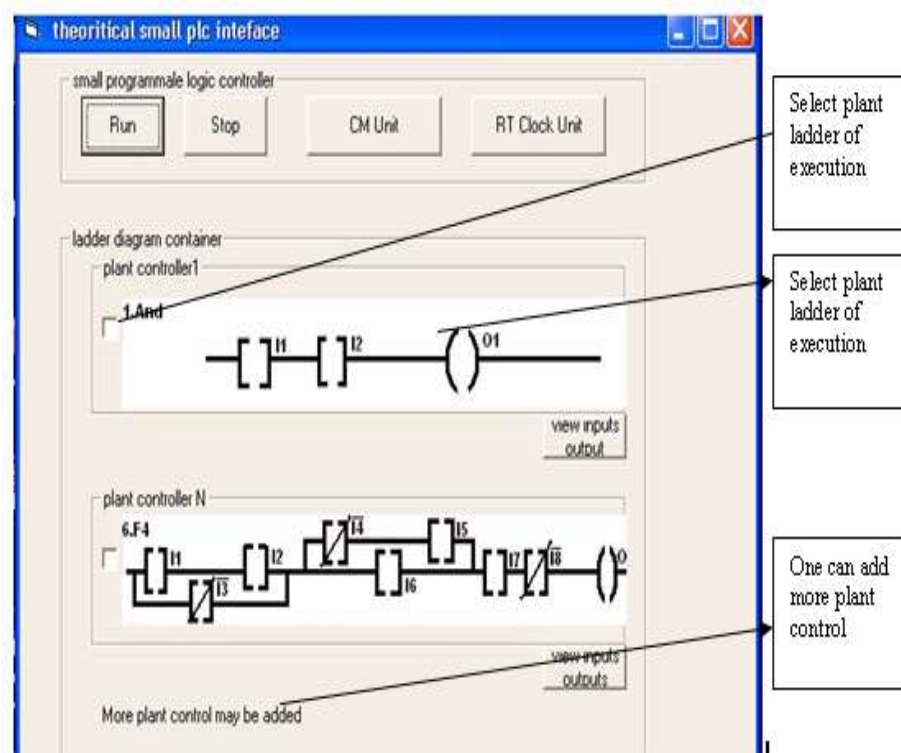


Figure 2 . Theoretical Interface GUI

The command button “Run” is for running plant control selected and “Stop” stops. The command button “CM Unit” is meant for pop up interface for communication management interfacing plant and computer side PLC interface . The command button “RT Clock Unit” is meant for real time clock and timer management for this PLC [1]. During execution one can inspect inputs and outputs of the plant in real time logic by pressing “view inputs outputs”. This small size interface may not be useful for large number of inputs and outputs in ladder of plant control.

Further integration to this theoretical interface in report generation and repository in terms of cheap available facility like Microsoft Excel (further MS Word, may be in CAD part).

V. FUTURE WORK

- CAD, A visual tool for arranging plant control diagrams, communication control, compiling and making a front end.

- Hardware interface with real plant control. Commercial PLC interface need to be studied and front end communication module need to be standardized and portable communication interface module need to be made.
- Installing the output function code in the visual tool.
- Universal make to run in all platform and operating system in all networks.
- Use and up gradation as Education Technology Kit
- Simplest possible lab illustration and complete and incomplete logic.
- Time management, internet time, time.windows.com, time related API in windows.
- Robustness and reliability for all operating system and desktop computer.
- Ethernet compatibility.
- Programming model following IEC 61131-3.

VI. CONCLUSION

In continuation of Software Engineering Research for Programmable Logic Controller, the work presented in this paper is sufficiently a primary step. The practical aspects of this work is done at Electrical Engineering Department, Veera

Surendra Sai University of Technology (www.vssut.ac.in), Orissa, India ; in under graduate project work and evaluated by experts well worthy . The work for a demo prototype is under progress.

REFERENCES

- [1] A low cost programmable logic control (PLC) trainer for use in a university agricultural electricity course. Journal of Agriculture Technology, Management and Education. March 2006, Vol 21. By Aaron Dickinson, Ronal M, Donald M Johnson.
- [2] Development of Object Oriented Modeling Tool for the design of Industrial Control logic . Kwan Hee Han; Jun Woo Park; Software Engineering Research, Management & Applications, 2007. SERA 2007. 5th ACIS International Conference on 20-22 Aug. 2007 Page(s):353 - 358
- [3] IEE Colloquium on advances in Software Engineering for PLC (Programmable Logic Controller) Systems' . Advances in Software Engineering for PLC. 14 Oct 1993
- [4] IUT X 200 to 215 recommendations on Open System Interconnection.
- [5] A low cost programmable logic control (PLC) Trainer for Use in a Agricultural Electricity Course. Journal of Agricultural Technology, management and education. Maech 2006, Vol 21.
- [6] IEC 61131-3

AUTHORS PROFILE

The author , B.K.Rana is a postgraduate in Electrical Engg from Indian Institute of Science, Bangalore, India in 1991. Subsequently worked in hardware and software in national and multinational companies. Presently a faculty in Electrical Engineering , Veer Surendra Sai University of Technology , Sambalpur , India and reseach in Industrial automation , Electrical Engineering software etc. are the fields of work.

Simulation of Packet Telephony in Mobile Adhoc Networks Using Network Simulator

Dr. P.K.Suri

Professor & Head, Dept. of Computer Sc and Applications
Kurukshetra University
Kurukshetra, India
pksuritf25@yahoo.com

Sandeep Maan

Assist. Professor, Dept. of Computer Sc.
Govt. P.G. College
Gurgaon, India.
sandeep.mann23@gmail.com

Abstract—Packet Telephony has been regarded as an alternative to existing circuit switched fixed telephony. To propagate new idea regarding Packet Telephony researchers need to test their ideas in real or simulated environment. Most of the research in mobile ad-hoc networks is based on simulation. Among all available simulation tools, Network Simulator (ns2) has been most widely used for simulation of mobile ad-hoc networks. Network Simulator does not directly support Packet Telephony. The authors are proposing a technique to simulate packet telephony over mobile ad-hoc network using network simulator, ns2.

Keywords—Network Simulator; Mobile Ad-hoc Networks; Packet Telephony; Simulator; Voice over Internet Protocol.

I. INTRODUCTION

The problem of extending the reach of fixed telephonic system over an area using mobile ad-hoc network is one of the research area that has got the attention of Computer Science research fraternity. One obvious solution to the problem comes in form of Packet Telephony, used interchangeably with Voice over Internet Protocol in this work. In packet telephony real time voice conversations are transmitted from source to destination using packet switched data networks rather than a circuit switched telephone network. With the help of Packet Telephony over mobile ad-hoc networks one can extend the reach of existing fixed telephony. This whole mechanism of extending the reach of fixed telephony is also termed as Fixed to Mobile Convergence (FMC) [1]. When this extension of fixed telephony is done over a mobile ad-hoc network, the problem becomes unique due to underlying characteristics of mobile ad-hoc network. The very nature of mobile ad-hoc networks makes the extension of telephonic call multi-hop where each intermediate node acts as potential router. The solution of extending the reach of wired telephony becomes highly beneficial with use of license free ISM band for implementing FMC. To summarize this would help forwarding telephonic call to a mobile node without any cost.

The effective extension of telephonic call over the mobile ad-hoc network is constrained by various Quality of Service requirements as recommended by United Nations Consultative

Committee for International Telephony & Telegraphy (CCITT). A number of Quality of Service, QoS parameters for implementation of fixed to mobile convergence in mobile ad-hoc networks has been suggested. These parameters include End to End Delay, Packet Delivery Rate, Packet Drop Rate, Throughput, Channel Utilization, Jitter etc. Any proposed system should follow strict QoS requirements to become practically viable. For example the End to End delay must be less than 250 ms otherwise the system may appear to be half duplex and user may complain about distortion and echo. In other words, QoS plays an important role in implementing Packet Telephony over Mobile Ad-hoc Networks.

Main deterrents in realizing the QoS based services over Mobile Ad-hoc Networks are a) Limited bandwidth of Mobile Ad-hoc Network b) Dynamic Topology of Mobile Ad-hoc Networks c) Limited Processing & Storing Capabilities of mobile nodes. Numbers of research works are in progress for ensuring QoS based Packet Telephony over Mobile Ad-hoc Networks.

It is not always feasible to develop a real time environment for conducting research. Then researchers have to resort on secondary means like simulation. In mobile ad-hoc network research, simulation techniques have been widely used. A number of simulation tools for developing mobile ad-hoc network environment are available. Most notable among these are Network Simulator (ns2), MATLAB, CSIM, OPNET, Qualinet, GoMoSlim etc. Out of these ns2 is most widely used tool for the simulation of mobile ad-hoc networks.

Network Simulator does not support VoIP or Packet Telephony directly. So a need was felt by the authors to devise a technique for the simulation of Packet Telephony with network simulator, ns2. The technique proposed should help users to test performance of the mobile ad-hoc network under different permutations and combinations of various network parameters.

II. RELATED WORK

Kurkowski et al. [2] have conducted a survey on the techniques employed by various authors for research on mobile ad-hoc networks. The authors have observed that out of 60%

authors resorting to simulation based techniques, 44% have used ns2 for drawing their conclusions.

Paolo Giacomazzi et al. [3] have worked on the issue of feasibility of fixed to mobile convergence using a mobile ad-hoc network. The authors have proposed complete system architecture for their implementation. The proposed architecture was then evaluated by the authors in terms of various quality of service (QoS) parameters like Call Drop Rate, MOS etc.

III. BACKGROUND

A Mobile Ad-hoc Network, MANET, may be defined as a collection of autonomous nodes that communicate with each other by forming multi-hop networks and maintaining connectivity in decentralized manner. All nodes in the Mobile Ad-hoc Network are of dynamic nature. This means that the topology of mobile ad-hoc networks keeps on changing.

Mobile Ad-hoc Networks do not have fixed routers. All nodes in these networks can act as routers. Apart from this mobile ad-hoc networks are characterized by a number of other salient features like range limitation, unreliable media, interference from other sources, dependency on the willingness of intermediate nodes, scarcity of power and vulnerability to security threats etc.

Mobile Ad-hoc Networks have been found to be very useful in emergency search and rescue operations. The reason behind this is their small deployment time. Moreover their deployment cost is also small.

Voice over Internet Protocol represents a set of rules and techniques to transport telephonic conversation over Internet Protocol. VoIP has proved to be one of the most admired and utilized application of internet these days. VoIP can prove to be a very beneficiary application. VoIP can help in achieving Fixed to Mobile Convergence (FMC) over mobile ad-hoc networks. The process behind this idea of achieving FMC over mobile ad-hoc network is illustrated in figure 2. In this figure various nodes are encircled representing their range. Various nodes with coinciding ranges may be termed as neighbors. In this figure node B is neighbor to nodes A and C. Different neighbors can exchange data through the channel. The extension of call from the fixed telephone to the node E can be explained as:

Initially analog voice conversations are digitized and then compressed using some suitable codec. Afterwards these compressed conversations are packetized in form of IP packets and then transported to E using underlined routing protocol. At E the packet are converted back to analog telephonic conversations. The main hurdle in implementing FMC over MANETs comes from the dynamic nature (see figure 1) and limited node range in these networks.

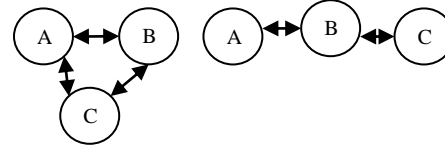


Figure 1. Dynamic Topology of MANETs

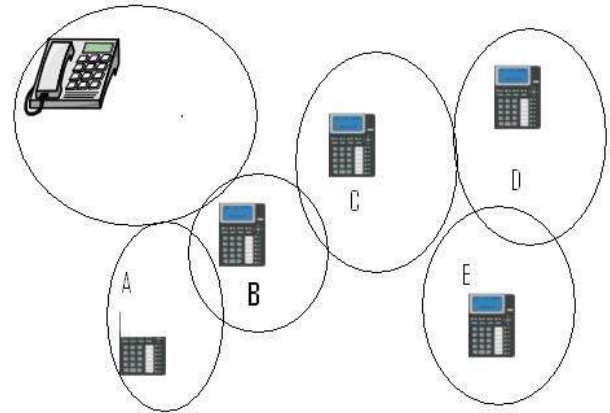


Figure 2. Fixed to Mobile Convergence over Mobile ad-hoc network

IV. SYSTEM ARCHITECTURE

System architecture represents the protocol layer used for the implementation of a network. During this work we have used a system architecture[4]-[5] composed of five network layers (see figure 3). Various responsibilities are distributed between layers as below:

A. Application Layer

The functions provided by this layer consist of digitizing & compressing the telephonic conversations in accordance with the available bandwidth. As already mentioned major constraint in implementing Packet Telephony [6] and hence FMC over the mobile ad-hoc networks comes from the limited bandwidth these networks possess. Some effective compression technique is required to overcome this limitation. A number of compression algorithms have been suggested by the International Telecommunication Union, ITU. Out of these G.729 codec [7] working at 8 kbps has been found to be most useful in scenarios where available bandwidth is small compared to the overall traffic load.

B. Transport Layer

One needs to choose between TCP and UDP for implementing transport layer. TCP is connection oriented protocol whereas UDP is comparatively unreliable connectionless protocol. The implementation of TCP would require higher bandwidth as compared to implementation of

UDP. In case of wireless mobile ad-hoc networks with limited available bandwidth UDP is the obvious choice. To overcome the limitations of UDP in relatively unreliable mobile ad-hoc network RTP (Real Time Transport Protocol) is run on the top of UDP. RTP provides services like payload identification, sequence numbering etc to the UDP. Almost every device uses a standard RTP to transmit audio and video.

C. Network Layer

Network layer in case of the mobile ad-hoc networks plays a central and pivotal role owing to the decentralized nature of mobile ad-hoc networks. All nodes participating in a mobile ad-hoc network acts as a potent router and forward the packets received from neighbors. A number of routing algorithms for mobile ad-hoc networks have been proposed. The routing algorithms for mobile ad-hoc networks have been classified [9]-[13] into two categories viz. topology based routing algorithms and position based routing algorithms. Due to various limitations most practical mobile ad-hoc networks employ topology based routing algorithms. Some major algorithms belonging to this category are DSR [14], DSDV [15], AODV [16], TORA.

D. MAC Layer

MAC layer plays a critical role in the successful implementation of mobile ad-hoc networks. Mobile ad-hoc networks have scarcity of available channel bandwidth. Moreover MAC layer not only has the responsibility of channel sharing but also hides the complexity of wireless network from upper layers.

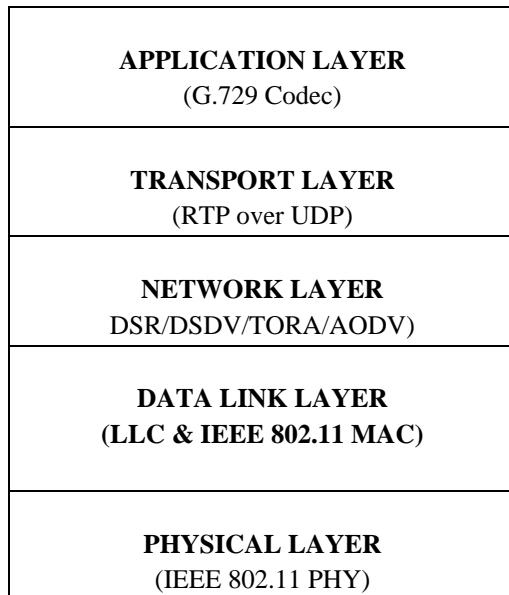


Figure 3. The Network Architecture

So, intelligent selection of MAC layer is very important. A number of MAC solutions are available these days. A good survey on these can be found in [17]-[21]. IEEE based MAC solutions have been most widely used in practical mobile ad-hoc networks.

E. Physical Layer

The responsibility with physical layer is to take data from source to destination. A number of physical layer solutions like IEEE 802.11 PHY [22]-[24], UTLA-TDD[25] are available. During this work we have implemented legacy IEEE 802.11 based physical layer.

V. SIMULATION WITH NS2

To test new ideas researchers resort to one of two underlined techniques i.e. either testing new ideas in real time environment or testing them in simulated environment. Creation of real time environment may not be always possible. In such cases authors have to depend upon the simulation tools. In case of mobile ad-hoc networks it has been observed that around 60% of work is done using simulation tools. Ns2 is most widely used tool among various available simulation tools. This can be attributed to a number of facts like:

- Ns2 is open & its source is freely available
- A full-fledged community is working on the development of this tool.
- A number of forums exist that provide for the patches to overcome the shortage in tool.
- It is easy to interpret its results with the help of easily available tools.
- Acceptability of results generated using this is very high when compared with real environment results.

Ns2 provides a number of inbuilt features while working on mobile ad-hoc networks like:

- Propagation Model: ns2 supports Friss-Space model for short distances and approximated Two Ray Ground model for long distances.
- MAC Layer: The IEEE 802.11 Distributed Coordination Function (DCF) has been implemented in ns2.
- Network Layer: ns2 supports all popular protocols like DSR, DSDV, TORA, AODV etc.
- Transport Layer: ns2 supports all popular transport layer protocols like TCP, UDP as well as RTP.

VI. PACKET TELEPHONY SIMULATION

The successful implementation of Packet Telephony is constrained with predefined range of various Quality of Service parameters as listed in table I.

Step 1: We propose following algorithm for implementing Packet Telephony over mobile ad-hoc network using

network Physical Layer Simulation: To simulate a mobile ad-hoc network using IEEE 802.11 based physical layer working at 2.4 GHz we need to set 'phyType' variable to Phy/WirelessPhy during node configuration and then initializing various ns2 variables as in table II .

Step 2: MAC Layer Simulation: We are using IEEE 802.11 based MAC, it can be setup, while configuring nodes, by initializing 'macType' variable to Mac/802_11.

Step 3: Network Layer Simulation: Main function performed by the network layer is routing. We can configure routing algorithm of our choice by setting 'adhocRouting' variable during node configuration to the desired routing algorithm and changing the values of 'ifq' & 'ifqLen' variables accordingly.

Step 4: Transport Layer Simulation: The transport layer is simulated by attaching corresponding agent with every communication between given source and destination.

Step 5: Node Configuration: Network is made of nodes. Node configuration includes creation of nodes, initialization of nodes and mobility modeling[26] of nodes. During creation of node ns2 is informed on their radius, motion type etc. During initialization a number of node related parameters are set as in table III. The node movements are created in a separate input file that is loaded when the simulator is run. The schematic representation of 'mobilenode' in ns2 is given in figure 4.

TABLE I. QOS PARAMETERS

Acceptable Range for QoS Parameters to successfully implement Packet Telephony	
Critical QoS Parameter	Acceptable Range
End to End Delay	<= 120 ms
Jitter	<= 40 ms
Packet Delivery Rate	>= 95%
Packet Drop Rate	< = 5%
Packet Loss Rate	<= 5%

TABLE II. SETTING PHYSICAL LAYER PARAMETERS

Physical Layer Parameter	Value
CPTthresh_	10.0
CSTthresh_	1.559e-11

RXThresh_	2.28289e-11
Rb_	2.1e6
Pt_	0.2818
freq_	240e+6
L_	1.0

TABLE III. NODE CONFIGURATION

Node Parameter	Explanation
AdhocRouting	Type of routing algorithm
llType	Logical Link layer setting
macType	Mac layer setting
antType	Type of antenna with node
propType	Propagation/Mobility model
phyType	Physical Layer type

Step 6: Scenario Generation: A mobile ad-hoc network is composed of nodes that are capable of moving within the premises of the network. So, next step is to create scenario that defines the positions of various nodes at a given time during simulation. The scenario data file is separately linked to the simulator during simulation run.

Step 7: Traffic Generation: In this work authors have proposed to use G .729 Codec for digitization and compression of telephonic conversations. Each compressed packet will be of 20B and packets will be transmitted at 50 packet/sec, hence an overall traffic of 8kbps will be generated. The connection between source and destination during the conversation will be maintained as two alternative pareto connections for transporting data in each directions. The detailed setting of traffic between source and destination are given in table-IV.

TABLE IV. Traffic Generation

Traffic Parameter	Explanation
Type	Application/Traffic/Pareto

rate_	8 kbps
packetSize_	20
burst_time	<Generated randomly>
idle_time_	<Generated randomly>

Traffic can be modeled in a separate input file that is loaded when simulator is run.

Step 8: Running the Simulator: Finally, the simulator created using the above steps is run and traces are collected in respective trace files.

Step 9: Analysis of output: Finally, various types of traces are analyzed using tools like NAM and MS Excel to draw conclusions. The results of trace analysis by authors are depicted in figures 5 & 6.

VII. CONCLUSIONS

Packet Telephony is one of the most attended research topic in mobile ad-hoc networks. With the help of packet telephony telephonic calls can be extended to some mobile node in the network without any additional cost. A number of wireless solutions working in ISM band around 2.4 GHz are available in market. The major problem with mobile ad-hoc networks is the limited range of its nodes, dynamic topology and scarcity of power. These limitations make study of voice over internet protocol, VoIP in mobile ad-hoc networks unique.

To establish a new idea one needs to test his idea and testing can be done either in real time environment or in simulated environment.

To test ideas on a simulated environment one must establish the authenticity of the simulation tool. For mobile ad-hoc networks network simulator, ns2 has been most extensively used for simulation of these networks.

Even the latest version of network simulator, ns-2.34 does not support VoIP directly. So, authors felt a need to propose a technique for simulation of Packet Telephony in ns2. For this purpose complete system architecture was first defined to implement packet telephony in a mobile ad-hoc network. Then a procedure was specified to perform simulation of packet telephony in network simulator.

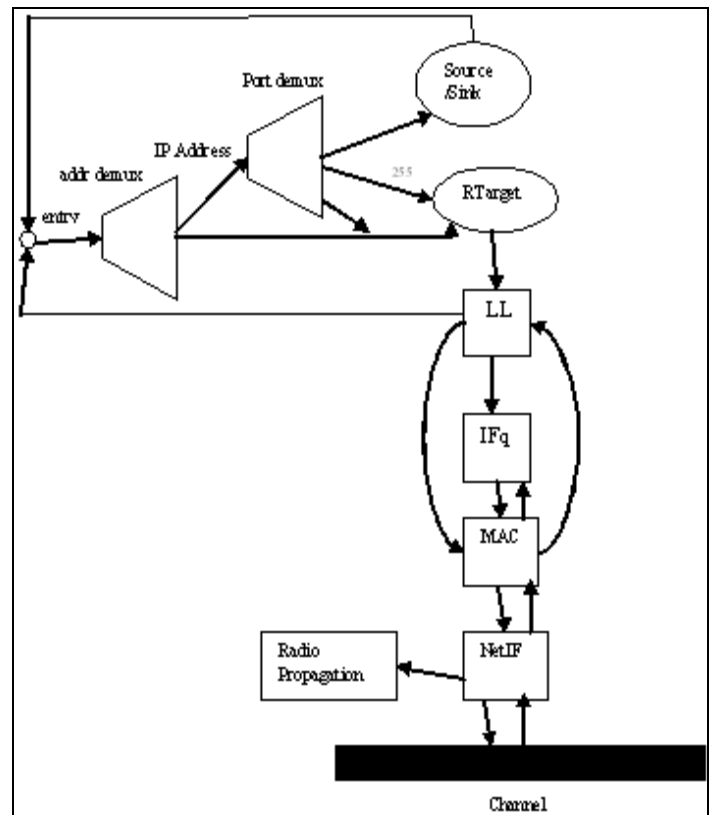


Figure 4. MobileNode in ns2 (taken from ns2 documentation)

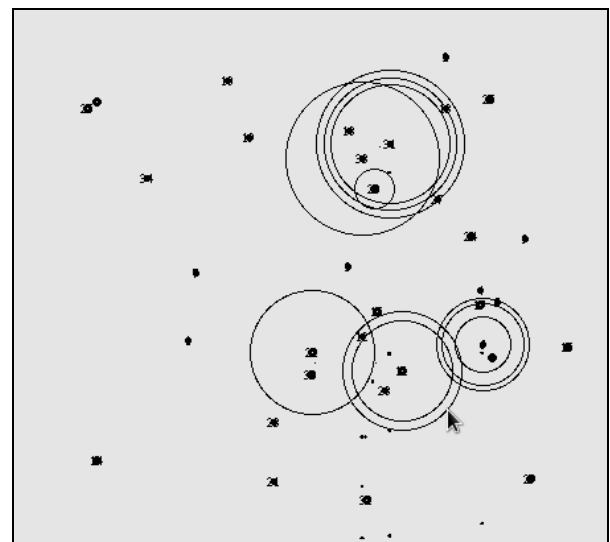


Figure 5. NAM trace output of VoIP simulation with ns2

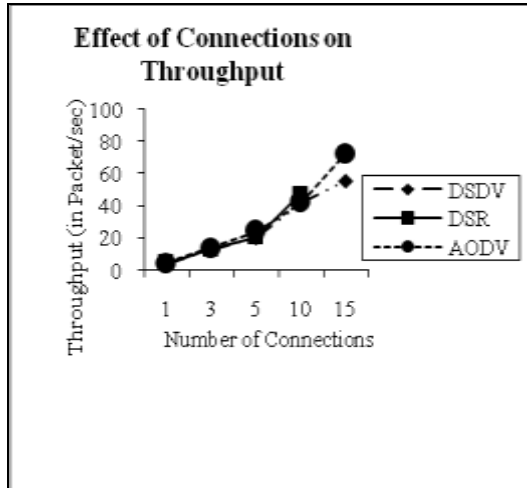


Figure 6. A Plot to compare various routing algorithms using the trace output resulting from ns2 based VoIP simulation run

VIII. FUTURE SCOPE

In this work possible extension for ns2 to implement Packet Telephony was proposed. The overall system comprises of five layers including application, transport, network, data link and physical layers. Authors encourage proposing techniques to further extend network simulator, ns2 by proposing algorithms for other possibilities in various layers. For an example authors can provide algorithms for more realistic mobility models that are not directly supported in network simulator, ns2. These extensions would help better evaluation of packet telephony based applications.

REFERENCES

- [1] J. Vong, Srivastava, and M. Finger, "Fixed to Mobile Convergence (FMC): technological convergence and the restructuring of the European telecommunications industry", SPRU 40th Anniversary B12Conference, 2006.
- [2] Stuart Kurkowski, Tracy Camp and Michael Colagrosso, "MANET Simulation Studies : The incredible", Mobile Computing and Communication Review, vol 9, no 4, 2005.
- [3] Paolo Giacomazzi, Luigi Musumeci, Giuseppe Caizzone, Giacomo Verticale, G. Liggieri, A. Proietti and S. Sabatini, "Quality of Service for Packet Telephony over Mobile Ad Hoc Network", IEEE Network, Jan/Feb 2006.
- [4] J.Schiller, "Mobile Communications", Pearson Education, Third Indian Reprint, 2004.
- [5] Andrew S. Tanenbaum, "Computer Networks", PHI.
- [6] L. Combat et. al., "Tactical Voice over Internet Protocol (VOIP) in secure wireless networks", June 2003.
- [7] M. E. Perkins et al., "Characterizing the Subjective Performance of the ITU-T 8 kb/s Speech Coding Algorithm ITU-T G.729," IEEE Commun. Mag., vol. 35, no. 9, Sep. 1997 pp. 74–81.
- [8] E. M. Royer and C.-K. Toh. "A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks," IEEE Pers. Commun., vol. 6, no. 2, Apr. 1999.
- [9] T. Camp, J. Boleng, and L. Wilcox, "Location Information Services in Mobile Ad Hoc Networks," Proc. IEEE ICC, Apr. 2002, pp. 18–22.
- [10] B. Karp et. al. "Greedy perimeter stateless routing for wireless networks," Proc. of the 6th Annual ACM/IEEE Int. Conf. on Mobile Computing and Networking (MobiCom 2000), pp 243.

- [11] B. N. Karp, "Geographic Routing for Wireless Networks.", PhD thesis, Harvard University, 2000.
- [12] C. K. Toh, "A Novel Distributed Routing Protocol To Support Ad Hoc Mobile Computing," Proc. 1996 IEEE 15th Annual Int'l, pp. 480–86.
- [13] C.C. Chiang, "Routing in Clustered Multihop, Mobile Wireless Networks with Fading Channel.", Proc. IEEE SICON '97, Apr. 1997, pp. 197–211.
- [14] D.B. Johnson and D.A. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Networks," Mobile Computing, pp. 153–81.
- [15] C.E. Perkins, and T.J. Watson, "Highly Dynamic Destination Sequenced Distance-Vector Routing (DSDV) for Mobile Computers", Comp. Commun.Rev., Oct. 1994, pp. 234–44.
- [16] C. E. Perkins and E. M. Royer, "Ad-hoc On-Demand Distance Vector Routing.", Proc. 2nd IEEE Workshop Mobile Comp. Sys. and Apps., Feb. 1999, pp. 90–100.
- [17] Sunil Kumar, Vineet S. Raghavan, Jing Deng, "Medium Access control protocols for ad hoc wireless networks: A survey", Ad Hoc Networks, 2006, pp 326-358.
- [18] A. Pal et al., "MAC layer protocols for real-traffic in ad hoc networks," Proceedings of the IEEE International Conference on Parallel Processing, 2002.
- [19] A. Muir et. al., "An efficient packet sensing MAC protocol for wireless networks," Mobile Networks, 1998, pp 221–234.
- [20] A. Nasipuri et. al., "A multichannel CSMA MAC protocol for multihop wireless networks," IEEE WCNC, September, 1999.
- [21] C.R. Lin et. al., "MACA/PR: An asynchronous multimedia multihop wireless network," Proceedings of the IEEE INFOCOM, March 1997.
- [22] "Information Technology—Telecommunications and Information Exchange Between Systems — Local and Metropolitan Area Networks-Specific Requirements — Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications", IEEE Std 802.11-1997
- [23] D.J. Deng et. al., "A priority scheme for IEEE 802.11 DCF access method," IEICE Trans. Commun., 1999, pp 96–102.
- [24] G. Bianchi. "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," IEEE JSAC, vol. 18, no. 3, Mar. 2000, pp. 535–47.-U
- [25] A. Ebner, H. Rohling, L. Wischhof, R. Halfmann, and M. Lott., "Performance of UTRA TDD Ad Hoc and IEEE 802.11b in Vehicular Environments," Proc. IEEE VTC 2003-Spring, vol. 2, Apr. 2003, pp. 18–22.
- [26] T. Camp, J. Boleng, and V. Davies, "A Survey of Mobility Models for Ad Hoc Network Research," Wiley Wireless Commun. and Mobile Comp., vol. 2, no. 5, 2002.

AUTHORS PROFILE

Dr. P.K.Suri (pkurittf25@yahoo.com) has been working in the department of Computer Science and Applications, Kurukshetra University, Kurukshetra, India for more than twenty five years. He has guided a number of PhD students. His areas of specialization includes Computer based simulation and modeling, Computer Networks etc. Presently he is acting as Head Computer Sc department and Dean Faculty of Science in the university.

Sandeep Maan (sandeep.mann23@gmail.com) is presently working as Assistant Professor in Computer Science at Govt. Post Graduate College, Sector-14, Gurgaon, India. He completed his M.Tech in Computer Science and Engineering from Kurukshetra University Kurukshetra. He is presently pursuing his PhD in Computer Science from Department of Computer Science and Applications, Kurukshetra University, Kurukshetra. His areas of interest includes Mobile Adhoc Networks, System Simulations and Artificial Intelligence.

A Novel Approach to Implement Fixed to Mobile Convergence in Mobile Adhoc Networks

Dr. P.K.Suri

Professor & Head, Dept. of Computer Sc and Applications
Kurukshetra University
Kurukshetra, India
pksuritf25@yahoo.com

Sandeep Maan

Assist. Professor, Dept. of Computer Sc.
Govt. P.G. College
Gurgaon, India
sandeep.mann23@gmail.com

Abstract— Fixed to Mobile Convergence, FMC is one of the most celebrated applications of wireless networks, where a telephonic call from some fixed telephonic infrastructure is forwarded to a mobile device. Problem of extending the reach of fixed telephony over a mobile ad-hoc network working in license free ISM band has been eluding the research fraternity. Major hindrance to FMC implementation comes from very nature of mobile ad-hoc networks. Due to the dynamic nature and limited node range in mobile ad-hoc networks, it is very difficult to realize QoS dependent applications, like FMC, over them. In this work authors are proposing complete system architecture to implement fixed to mobile convergence in a mobile ad-hoc network. The mobile ad-hoc network in the problem can hold a number of telephonic calls simultaneously. The proposed system is then implemented using network simulator, ns2. The results obtained are then compared with the predefined standards for implementation of FMC.

Keywords— Mobile Ad-hoc Networks; Packet Telephony; Voice over Internet Protocol; Fixed to Mobile Convergence; Quality of Service

I. INTRODUCTION

Mobile ad-hoc networks, MANET, are the latest member of illustrious family of wireless networks. In mobile ad-hoc networks, a number of autonomous and mobile nodes communicate with each other by forming multi-hop connections and maintaining connectivity in decentralized manner. MANETs are different from other wireless networks in terms of their distinct characteristics like limited range, unreliable media, dynamic topology, limited energy etc. In mobile ad-hoc networks each node is a potent router and forwards the packet received from other nodes using some suitable routing algorithm. The reason behind the popularity of mobile ad-hoc networks is their small setup time and deployment cost. These networks have proved their worth in emergency like situations such as natural calamity, war time etc

The area of mobile ad-hoc networks has taken another leap forward with invention of device working in license free *Industrial-Scientific-Military (ISM) band* concentrated at 2.4

GHz. Two notable technologies working in this band are WiFi and Bluetooth.

Fixed to Mobile Convergence, also referred as FMC[1]-[2] during this work, represents the process of extending reach of fixed telephony with the help of wireless techniques. One method of achieving this can be using wireless handsets and a single antenna near the wired end. The range covered by this technique cannot be large. To cover a large area one would require to purchase some frequency, which is a costly affair. During this work authors proposes a system where calls are forwarded through a mobile ad-hoc network working in ISM band. This mobile ad-hoc network under study, apart from extending the reach of fixed call, will also allow other users of the network to call each other simultaneously.

The underlying technique to be used for call forwarding will be voice over internet protocol, also termed as packet telephony during this work. In Packet Telephony, data belonging to the telephonic conversations is transported as IP packets and no permanent copper path is setup between source and destination. In order to offer real time service over the mobile ad-hoc network certain Quality of Service (QoS) parameters play vital role. These parameters include *end to end delay, packet delivery ratio and throughput etc.* Measures are required to keep the values of QoS parameters in the specified range. Quality of Service offered by the mobile ad-hoc networks depends on a number of variables like node speed, number of nodes, number of connections, area etc. Major deterrent to offering QoS based services, over the mobile ad-hoc networks, is basic nature of these networks. Due to the dynamic nature of mobile ad-hoc networks routes are also very vulnerable and re-routing is required frequently.

In this paper, we propose complete system architecture to implement FMC through a MANET carrying number of simultaneous telephonic conversations. The proposed methodology for implementing fixed to mobile convergence would be using wireless technology working in license free ISM band for deriving maximum benefits out of the technique. The proposed architecture is simulated using ns2 and the results are verified with practically acceptable values (table 1).

II. RELATED WORK

Paolo Giacomazzi et al. [3] proposed architecture for mobile ad-hoc networks carrying VoIP based telephonic conversation. The authors evaluated their architecture in terms

TABLE I. QoS PARAMETERS

Acceptable Range for QoS Parameters to successfully implement Packet Telephony		
Critical Parameter	QoS	Acceptable Range
End to End Delay		≤ 120 ms
Packet Delivery Rate		$\geq 95\%$
Packet Drop Rate		$\leq 5\%$
Packet Loss Rate		$\leq 5\%$

of various Quality of Service parameters like Call Drop Rate, MOS etc. The authors observed that the ad-hoc networks can be used for economically extending access of fixed infrastructure, termed as Fixed to Mobile Convergence (FMC). They also observed that although the size of this extension is quite limited.

III. SYSTEM ARCHITECTURE

The problem environment is sketched in figure 1. The system consists of a single fixed telephone and a number of mobile nodes forming a mobile ad-hoc network. The system not only forwards the calls at fixed telephone but various mobile nodes can themselves be engaged in telephonic conversations.

To start with, in packet telephony, analog voice conversations are to be converted to digital and then compressed. To cater the scarcity of available bandwidth with mobile ad-hoc networks, compression technique is to be carefully selected. Then the digital voice calls are converted into IP packets and forwarded over the mobile ad-hoc network. Finally, these packets are collected at destination; the digital voice is extracted from these and converted into analog voice.

During this work the authors have assumed that the area covered by the network is square in shape. As source and destination may not be neighbor to each other, the call forwarding can be multi-hop. Overall the proposed system architecture consists of five network layers [4]-[5] as explained below:

A. Physical Layer

The purpose of physical layer is to transmit the data from one node to other and hence from source to destination. A number of Physical Layer solutions for mobile ad-hoc networks are available like IEEE 802.11 PHY[6]-[9], UTRA-TDD [10]-[11]. The authors have proposed to use the most

practically used solution in IEEE 802.11 based Simple Wireless Channel working in ISM band around 2.4 GHz.

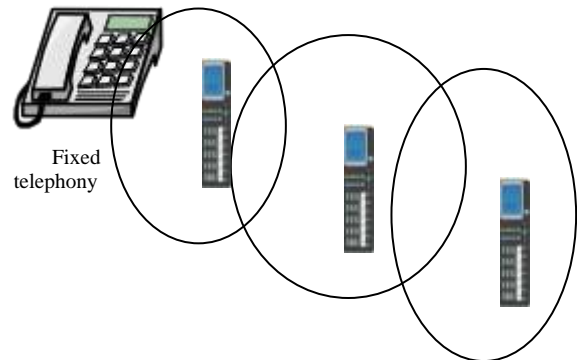


Figure 1. Fixed to Mobile Convergence over Mobile ad-hoc network

B. MAC Layer

For networks with small bandwidth at their disposal MAC layer is very critical. This layer has great impact on the performance of the Networks. There are number of MAC layer solutions like IEEE 802.11, VMAC etc. A good survey of popular MAC protocols can be found in [12]-[16]. The authors have used IEEE 802.11 based MAC layer.

C. Network Layer

The major role played by the network layer is routing. A number of routing algorithms have been proposed. These algorithms have been classified into two categories [17] viz. topology based and position based routing [18]-[24] algorithms. Out of these most practical system employ topology based routing algorithms. Topology based routing algorithms have been further classified into three categories viz. proactive, reactive and hybrid. The authors have proposed to use Dynamic Source Routing, DSR [25] algorithm in this work.

D. Transport Layer

Due to the wireless nature of network UDP is used. To realize a quality of service based application like VoIP, one cannot rely completely on unreliable UDP rather Real-Time Transport Protocol (RTP) is run on top of UDP to provide end-to-end delivery services like payload type identification, sequence numbering etc.

E. Application Layer

This layer converts telephonic conversations into the digitized form. International Telecommunication Union has standardized a number of effective speech compression algorithms; In our scenario as overall traffic load is high as compared to the available bandwidth, authors are using G.729 codec [26] working at 8 kbps(50 packets/sec with a packet of size 20B).

System proposed by the authors is sketched in figure 2.

IV. SIMULATION

Authors simulated the overall system in network simulator (ns2 ver 2.34) [27]. The system was implemented on Fedora Linux running on a Pentium computer. Number of simulation runs was made with different environmental settings. The performance of system was evaluated in terms of various Quality of Service parameters as given below:

G .729 codec at 8kbps (Application Layer)
RTP/UDP (Transport Layer)
Dynamic Source Routing (Network Layer)
LLC/IEEE 802.11 MAC (Data Link Layer)
IEEE 802.11 PHY (Physical Layer)

Figure 2. Proposed System Architecture

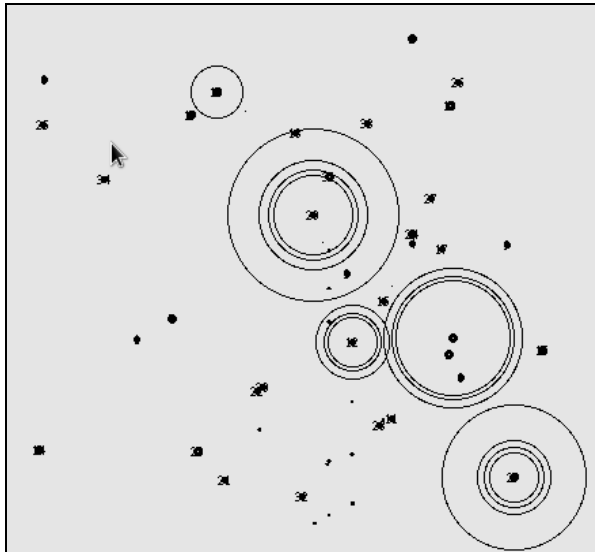


Figure 3. Snapshot of NAM trace

A. Packet Delivery Ratio

Packet delivery ratio signifies the percentage of packets that were successfully delivered from source to destination.

Failure to deliver a packet would mean loss of conversation data and hence incomplete conversation. So, a packet delivery ratio of more than 95% must be achieved for successful implementation of network.

B. End to End Delay

End to End delay represents the time elapsed between delivering of voice from speaker to listener. In quantitative terms a delay of 120 ms or less is acceptable. The problem comes if the delay is more than 250 ms as it ultimately leads to the half duplex conversation.

C. Throughput

Throughput represents number of successful packet transmissions per second. Throughput is very important evaluation parameter.

D. Packet Drop Rate

Packet Drop Rate represents percentage of packets that were dropped by intermediate nodes due to overloaded queue. Packet drop rate is significant as it identifies congestion in the network and data handling capability of a network. It is assumed that a mobile ad-hoc network would be unacceptable if packet drop rate goes beyond 5%. More drop rate means more retransmission and hence more delays.

E. Packet Loss Rate

Packet loss rate represents percentage of packets that are lost into the network. This definitely means loss of conversation and hence a packet loss rate above 5% is highly unacceptable.

V. RESULTS & OBSERVATIONS

The important observations are plotted below:

A. Parameter: End to End Delay

Observations (figures 4-6):

Authors observed that end to end delay increases with number of active connections but overall delay was within the specified experimental limits. While with increased number of participating nodes delay decreases. Both results are on expected line as with increase in number of connections traffic load on the network increases thereby increasing chances of congestion and hence delay. Whereas with increased number of participating nodes there will be lesser chances of route error and more than one route may exist simultaneously in route cache.

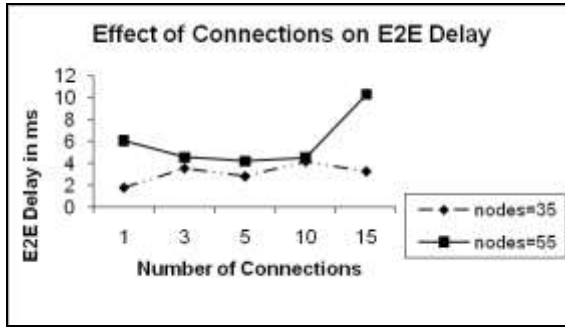


Figure 4. Effect of number of connections on End to End Delay (in ms) Other parameters are Area = 2.25 KM², Node Speed =25 KMPH & Nodes= 35/50.

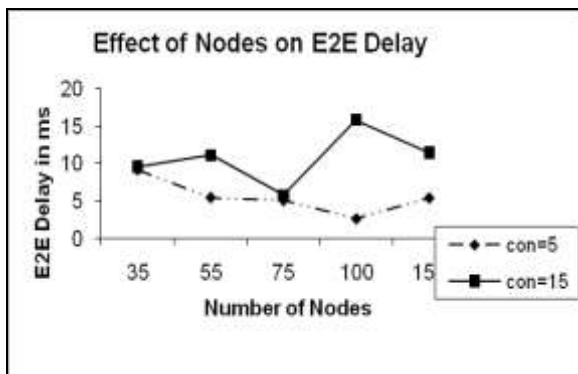


Figure 5. Effect of number of nodes on End to End Delay (in ms). The other parameters are Area = 2.25 KM², Node Speed =25 KMPH & Connections = 5/15.

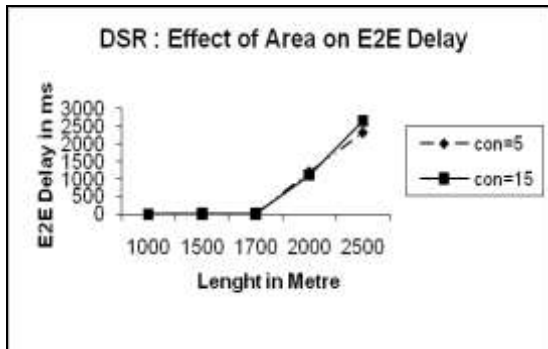


Figure 6. Effect of area (Area assumed to be square= Length* Length) on End to End Delay (in ms).

Important observation comes from the plot of E2E delay vs. area, it was observed that the delay was within the allowable limits upto smaller network size, but for larger area E2E delay increases drastically and system performance becomes unacceptable.

B. Parameter: Packet Delivery Ratio(PDR)

Observations (figures 7-9):

Authors observed that packet delivery ratio (PDR) decreases with number of active connections but overall PDR was within the experimental limits. With increasing number of participating nodes PDR remains almost constant. The result is as per expectations as with increase in number of connections traffic load on the network increases thereby increasing chances of congestion and hence lesser PDR.

Again, important observation comes from the plot of PDR vs. area; it was observed that the PDR drops drastically for large network sizes.

C. Parameter: Throughput

Observations (figures 10-12):

Authors observed that throughput increases with number of active connections. While with increase in number of participating nodes, throughput stays almost constant. The result is as per expectations, because with increase in number of connections network has more packets to deliver and hence better throughput. It was also observed that the throughput decreases with increase in size of the network.

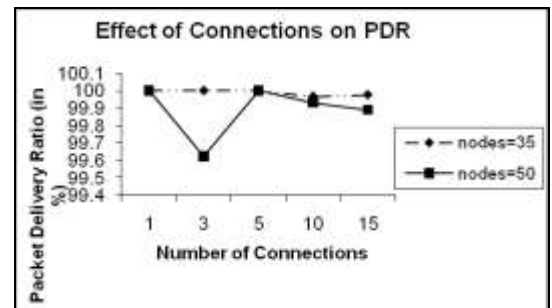


Figure 7. Effect of Number of Connections on PDR. The other parameters are Area = 2.25 KM², Node Speed =25 KMPH & Nodes= 35/50.

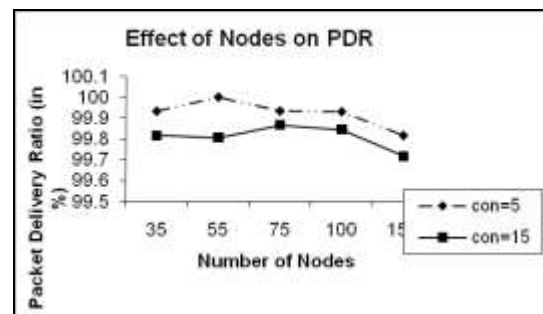


Figure 8. Effect of Number of Nodes on PDR. The other parameters are Area = 2.25 KM², Node Speed =25 KMPH & Connections = 5/15.

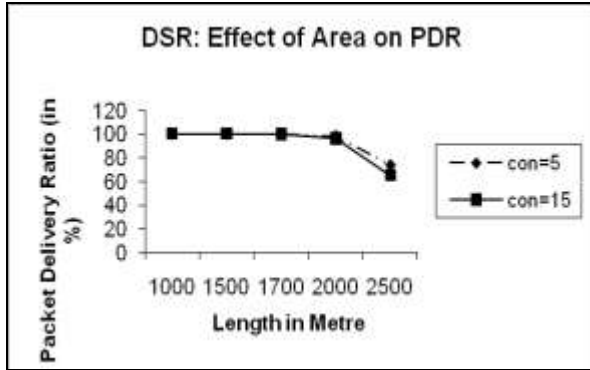


Figure 9. Effect of area (Area assumed to be square= Length* Length) on PDR.

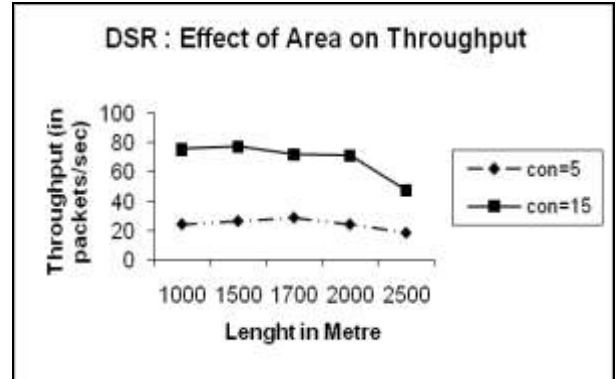


Figure 12. Effect of area (Area assumed to be square= Length* Length) on Throughput.

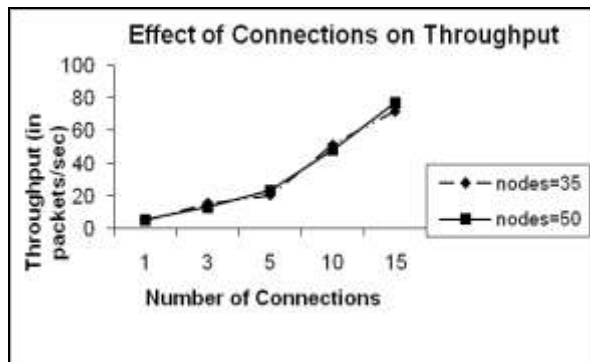


Figure 10. Effect of Number of Connections on Throughput for. The other parameters are Area = 2.25 KM², Node Speed =25 KMPH & Nodes=35/50

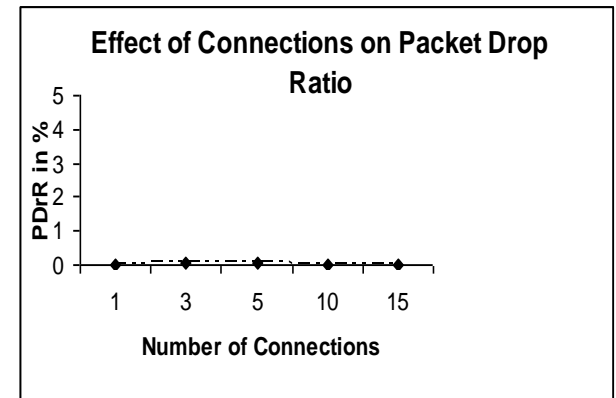


Figure 13. Effect of Number of Connections on Packet Drop Ratio for Area=2.25 KM², Node Speed =25 KMPH & Nodes=35

D. Parameter: Packet Drop Ratio(PDR)

Observations (figures 12-15):

Authors observed that Packet Drop Ratio remains in the experimental limits when plotted against number of nodes and number of connections. But it crosses allowable limit when the size of network increases.

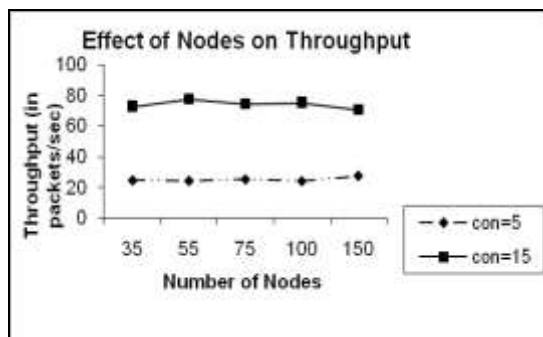


Figure 11. Effect of Number of Nodes on throughput. The other parameters are Area = 2.25 KM², Node Speed =25 KMPH & Connections = 5/15.

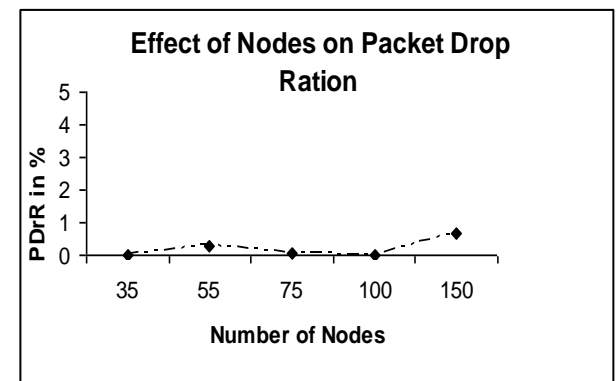


Figure 14. Effect of Number of Nodes on Packet Drop Ratio. The other parameters are Area = 2.25 KM², Node Speed =25 KMPH & Connections = 5.

VI. CONCLUSION

In this work we presented complete system architecture for implementing fixed to mobile convergence using a mobile ad-hoc network. Fixed to mobile convergence can prove to be very useful if implemented using mobile ad-hoc networks. It

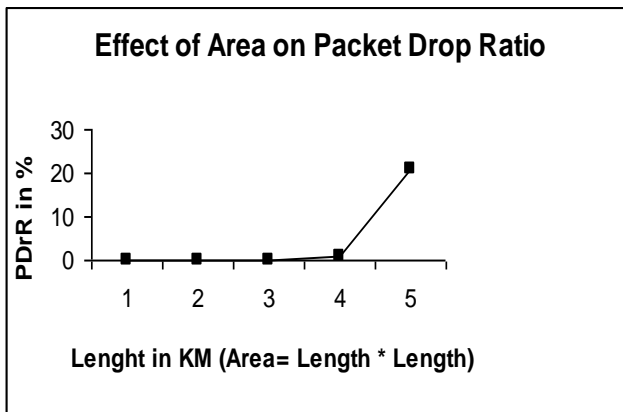


Figure 15. Effect of area (Area assumed to be square= Length* Length) on Packet Drop Ratio.

would provide a cost effective and fast solution to the problem of extending reach of existing fixed telephony. For this purpose network architecture was proposed during this work. The proposed architecture was then simulated using ns2 and observed results were compared with proposed practical limits for various network parameters like End to End Delay, Packet Delivery Ratio and throughput.

It was observed that it is possible to successfully implement FMC using a mobile ad-hoc network but this extension cannot cover a large area, it was observed that it was possible to cover a network size of around 3 KM² with the proposed system architecture.

VII. FUTURE SCOPE

During this work the feasibility of fixed to mobile convergence using a mobile ad-hoc network was evaluated using network layer parameters. Other authors are encouraged to evaluate performance of proposed architecture under different MAC level protocols.

REFERENCES

- [1] J. Vong J. Srivastava, and M. Finger, "Fixed to Mobile Convergence (FMC): technological convergence and the restructuring of the European telecommunications industry", SPRU 40th Anniversary B12Conference, 2006.
- [2] L. Combat et. al., "Tactical Voice over Internet Protocol (VOIP) in secure wireless networks", June 2003.
- [3] Paolo Giacomazzi, Luigi Musumeci, Giuseppe Caizzone, Giacomo Verticale, G. Liggieri, A. Proietti and S. Sabatini, "Quality of Service for Packet Telephony over Mobile Ad Hoc Network", IEEE Network, Jan/Feb 2006.
- [4] J.Schiller, "Mobile Communications", Pearson Education, Third Indian Reprint, 2004.
- [5] Andrew S. Tanenbaum, "Computer Networks", PHI.
- [6] F. Cali, M. Conti, and E. Gregori. "IEEE 802.11 Wireless LAN: Capacity Analysis and Protocol Enhancement," IEEE INFOCOM 1998, vol. 1, pp. 142–49.
- [7] "Information Technology—Telecommunications and Information Exchange Between Systems — Local and Metropolitan Area Networks-Specific Requirements — Part 11: Wireless LAN Medium Access

Control (MAC) and Physical Layer (PHY) Specifications", IEEE Std 802.11-1997.

- [8] D.J. Deng et. al., "A priority scheme for IEEE 802.11 DCF access method," IEICE Trans. Commun., 1999, pp 96–102.
- [9] G. Bianchi. "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," IEEE JSAC, vol. 18, no. 3, Mar. 2000, pp. 535–47.-U
- [10] M. Lott et al., "Medium Access and Radio Resource Management for Ad Hoc Networks Based on UTRA-TDD," Proc. 2nd ACM Int'l. Symp. Mobile Ad Hoc Net and Comp., 2001.
- [11] A. Ebner et al., "Performance of UTRA TDD Ad Hoc and IEEE 802.11b in Vehicular Environments," Proc. IEEE VTC 2003-Spring, vol. 2, Apr. 2003, pp. 18–22.
- [12] Sunil Kumar, Vineet S. Raghavan, Jing Deng, "Medium Access control protocols for ad hoc wireless networks: A survey", Ad Hoc Networks, 2006, pp 326-358.
- [13] A. Pal et al., "MAC layer protocols for real-traffic in ad hoc networks," Proceedings of the IEEE International Conference on Parallel Processing, 2002.
- [14] A. Muir et. al., "An efficient packet sensing MAC protocol for wireless networks," Mobile Networks, 1998, pp 221–234.
- [15] A. Nasipuri et. al., "A multichannel CSMA MAC protocol for multihop wireless networks," IEEE WCNC, September, 1999.
- [16] C.R. Lin et. al., "MACA/PR: An asynchronous multimedia multihop wireless network," Proceedings of the IEEE INFOCOM, March 1997.
- [17] E. M. Royer and C.-K. Toh. "A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks," IEEE Pers. Commun., vol. 6, no. 2, Apr. 1999.
- [18] T. Camp, J. Boleng, and L. Wilcox, "Location Information Services in Mobile Ad Hoc Networks," Proc. IEEE ICC, Apr. 2002, pp. 18–22.
- [19] B. Karp et. al. "Greedy perimeter stateless routing for wireless networks," Proc. of the 6th Annual ACM/IEEE Int. Conf. on Mobile Computing and Networking (MobiCom 2000), pp 243.
- [20] B. N. Karp, "Geographic Routing for Wireless Networks," PhD thesis, Harvard University, 2000.
- [21] C. E. Perkins and E. M. Royer, "Ad-hoc On-Demand Distance Vector Routing," Proc. 2nd IEEE Workshop Mobile Comp. Sys. and Apps., Feb. 1999, pp. 90–100.
- [22] C. K. Toh, "A Novel Distributed Routing Protocol To Support Ad Hoc Mobile Computing," Proc. 1996 IEEE 15th Annual Int'l, pp. 480–86.
- [23] C.C. Chiang, "Routing in Clustered Multihop, Mobile Wireless Networks with Fading Channel," Proc. IEEE SICON '97, Apr. 1997, pp. 197–211.
- [24] C.E.Perkins et. al., "Highly Dynamic Destination Sequenced Distance-Vector Routing (DSDV) for Mobile Computers," Comp. Commun.Rev., Oct. 1994, pp. 234–44.
- [25] D.B. Johnson and D.A. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Networks," Mobile Computing, pp. 153–81.
- [26] M. E. Perkins et al., "Characterizing the Subjective Performance of the ITU-T 8 kb/s Speech Coding Algorithm ITU-T G.729," IEEE Commun. Mag., vol. 35, no. 9, Sep. 1997 pp. 74–81.
- [27] Stuart Kurkowski, Tracy Camp and Michael Colagrosso, "MANET Simulation Studies: The incredible", Mobile Computing and Communication Review, vol 9, no 4, 2005.

AUTHORS PROFILE

Dr. P.K.Suri (pksuritt25@vahoo.com) has been working in the department of Computer Science and Applications, Kurukshetra University, Kurukshetra, India for more than twenty five years. He has guided a number of PhD students. His areas of specialization includes Computer based simulation and modeling, Computer Networks etc. Presently he is acting as Head Computer Sc department and Dean Faculty of Science & Engineering in the university.

Sandeep Maan (sandeep.mann23@gmail.com) is presently working as Assistant Professor in Computer Science at Govt. Post Graduate College, Sector-14, Gurgaon, India. He completed his M.Tech in Computer Science and Engineering from Kurukshetra University Kurukshetra. He is presently

pursuing his PhD in Computer Science from Department of Computer Science and Applications, Kurukshetra University, Kurukshetra. His areas of interest includes Mobile Adhoc Networks, System Simulations and Artificial Intelligence

IPS: A new flexible framework for image processing

Otman ABDOUN
LaRIT, IbnTofail University
Faculty of Sciences, Kenitra, Morocco
Email: otman.fsk@gmail.com

Jaafar ABOUCHABAKA
LaRIT, IbnTofail University
Faculty of Sciences, Kenitra, Morocco
Email: abouch06-etudiant@yahoo.fr

Abstract— Image processing is a discipline which is of great importance in various real applications; it encompasses many methods and many treatments. Yet, this variety of methods and treatments, though desired, stands for a serious requirement from the users of image processing software; the mastery of every single programming language is not attainable for every user. To overcome this difficulty, it was perceived that the development of a tool for image processing will help to understanding the theoretical knowledge on the digital image. Thus, the idea of designing the software platform Image Processing Software (IPS) for applying a large number of treatments affecting different themes depending on the type of analysis envisaged, becomes imperative. This software has not come to substitute the existing software, but simply a contribution in the theoretical literature in the domain of Image Processing. It is implanted in the MATLAB platform: effective and simplified software specialized in image treatments in addition to the creation of Graphical User Interfaces (GUI) [5][6]. IPS is aimed to allow a quick illustration of the concepts introduced in the theoretical part. This developed software enables users to perform several operations. It allows the application of different types of noise on images, to filter images color and intensity, to detect edges of an image and to apply image thresholding by defining a variant threshold, to cite only a few.

Keywords—Image Processing; Noise; Filter; MATLAB; Edge detection.

I. INTRODUCTION

Image processing is a set of methods that can transform images or to extract information [4][6]. This is a very wide area, which is more and more applications:

- Much of the mail is now sorted automatically, thanks to the automatic recognition of the address,
- In the military field, devices capable of detecting and recognize automatically their targets,
- In industry, automatic control by vision is increasingly common in manufacturing lines,
- Image compression is experiencing a significant expansion in recent years, particularly through the development of Internet and digital television.

In this paper, we present an image processing software IPS, is developed in MATLAB, it represents a powerful tool for scientific calculations, and creating graphical user interfaces (GUI). The program is available for use with MATLAB or as a stand-alone application that can run on Windows and UNIX systems without requiring MATLAB and its toolboxes. The home interface of the IPS platform is shown in Fig.1

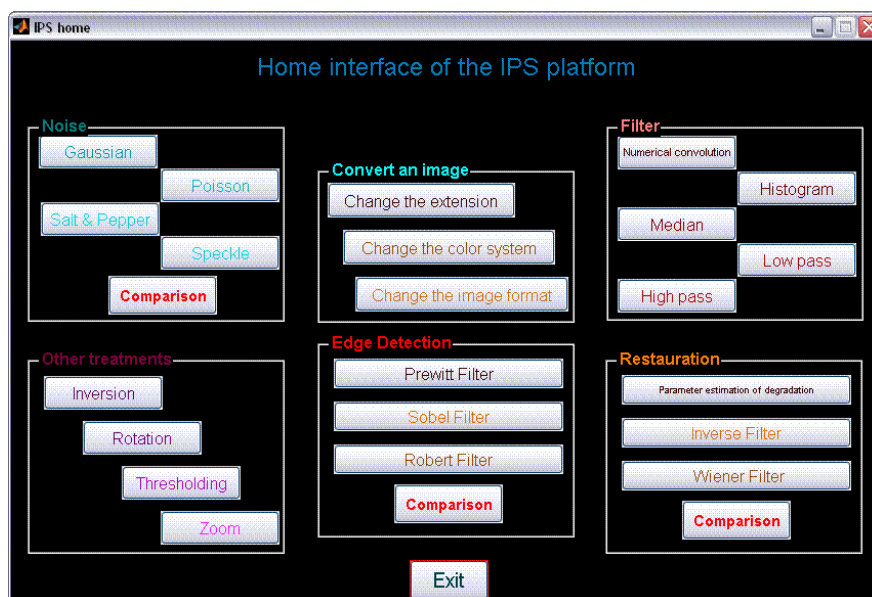


Figure. 1 : General overview of the program IPS

The **IPS** platform is simple to use and easily scalable, and allows you to do the following:

- The application of different types of noise on images ;
- Equalize the histogram of an image blurred;
- Edge detection ;
- The filter color images and intensities;
- Thresholding of images by defining a threshold ranging;
- The change in format and extension images;
- Import and export images in various locations;

II. HISTOGRAMME

A histogram is a graph to represent the statistical distribution of pixel intensities of an image, that is to say the number of pixels for each intensity levels. By convention, a histogram represents the level of intensity in the x-axis ranging from the darkest (left) to the lightest (right). In practice, for computing a histogram, it gives a number of quantization levels, and for each level, we count the number of pixels in the image corresponding to that level. MATLAB function that performs the calculation of a histogram is *imhist* [6]. It takes as parameters like the image name and the number of quantization levels desired.

A. Histogram equalization

The histogram equalization is a tool that is sometimes useful to enhance some images of poor quality (poor contrast, too dark or too bright, poor distribution of intensity levels, etc.) [4]. This is to determine a transformation of intensity levels that makes the histogram as flat as possible. If a pixel has intensity i in the original image, its intensity is smoothed image $f(i)$. In general, we chose a step function, and determine the width and height of the various steps in order to flatten the image histogram equalized. MATLAB is performed by histogram equalization *histeq* $J = (I, n)$ where I denote the original image, J equalized image, and n is the number of intensity levels in the image equalized. Avoid choosing too large n (for $n = 64$ gives good results).

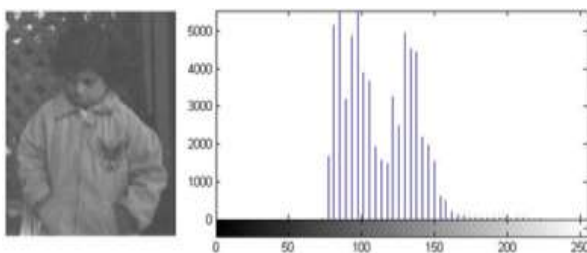


Figure 2.a : Original image blurred

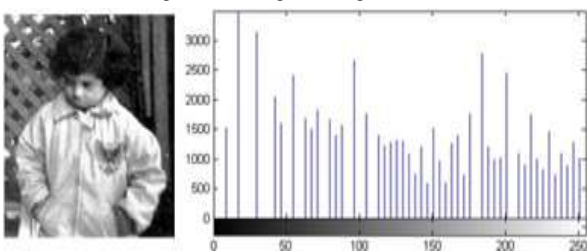


Figure 2.b. Equalized image

B. Stretching the histogram

The histogram stretching (also called "histogram linearization" or "expansion dynamics") is to allocate frequencies of occurrence of pixels on the width of the histogram. Thus it is an operation to modify the histogram so to the best allocation of intensities on the scale of values available. This amounts to extending the histogram so that the value of the lowest intensity is zero and the highest is the maximum value.

That way, if the values of the histogram are very close to each other, stretching will help to provide a better distribution to make even clearer light pixels and dark pixels close to the black.

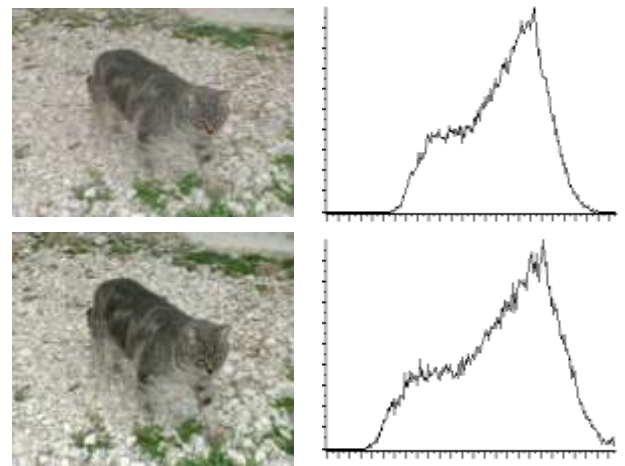


Figure 3. Stretching the histogram

It is thus possible to increase the contrast of an image. For example, a picture is too dark can become more "visible".

III. NOISE

Characterizes the noise or interference noise signal, which is to say the parts locally distorted signal. Thus, the noise of an image means the image pixels whose intensity is very different from those of neighboring pixels.

Noise can come from various causes:

- Environment during the acquisition ;
- Quality of the sensor ;
- Quality of sampling.

There are several types of noise:

A. Gaussian noise

It's a sound, whose value is randomly generated following the Gaussian:

$$B(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(\frac{-(x-m)^2}{2\sigma^2}\right)} \quad (1)$$

With:

- σ^2 : Variance

- m : median

B. Speckle noise

It is a noise (n), generated following the uniform law with an average of 0 V is a variance equal to 0 .04 default. If I is the original image, the noisy is defined by:

$$J=I+n*I \quad (2)$$

C. Salt and Pepper noise

This is a random signal generated by the Uniform Law. For the noise spectral density (D), it will be affected by noise D multiplied by the number of picture elements. Its principle is to :

- Determine the indices of its elements with a value less than half its density. Then assign 0 to the pixels corresponding to these indices in the image.
- Determine the indices of its elements with a value framed by half its density and its density. Then assign 1 to the pixels corresponding to the indices taken from the processed image

D. POISSON noise

It is an additive noise generated by the Poisson:

$$P(x = k) = \frac{a^k}{k!} e^{-a} \quad (3)$$

With a positive quantity is called the parameter of the law.

The function provided by MATLAB, which can generate noise that is IMNOISE, its syntax is:

$$\text{IMNOISE}(I, \text{TYPE}) \quad (4)$$

I: is the original image

TYPE: is the type of noise to apply, it may take the following values:

- 'GAUSSIAN': Gaussian noise to generate the function syntax IMNOISE will be as follows: IMNOISE(I, 'Gaussian', m, v), where m and v are respectively the mean and variance of the noise (Fig.4.a);
- 'POISSON': to generate the Poisson noise, the syntax is: IMNOISE(I, 'Poisson') (Fig. 4.e);
- 'SALT & PEPPER': to generate the noise with salt and pepper. The syntax is: IMNOISE(I, 'salt & pepper', D), where D is the density of noise (Fig.4.c);
- 'SPECKLE': to generate the speckle noise, the syntax is: IMNOISE(I, 'speckle', V), where V is the noise variance (Fig.4.d).

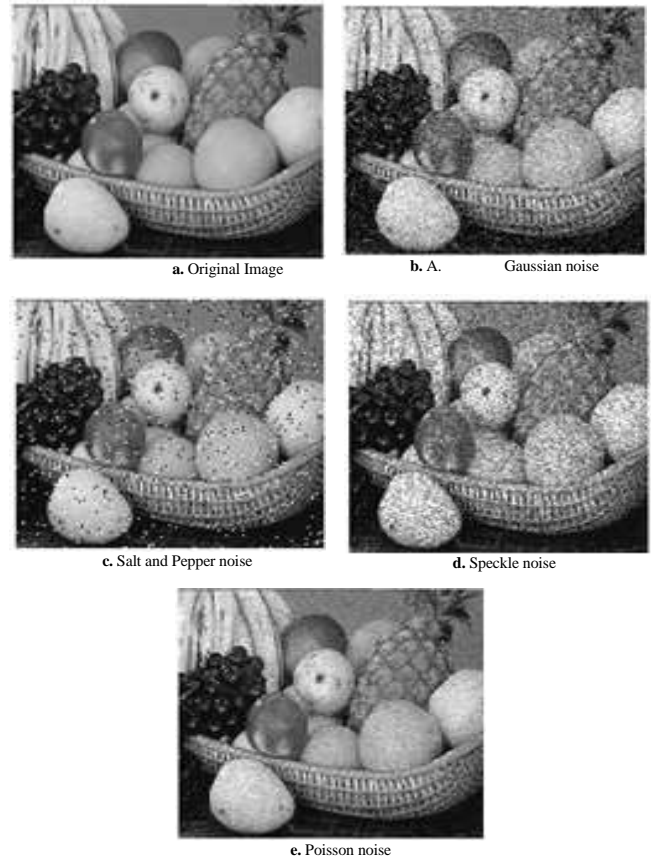


Figure 4. Noising image (a, b, c, d and e)

IV. IMAGE FILTERING

A filter is a mathematical transformation (called convolution product) for, for each pixel of the area to which it applies, to change its value based on values of surrounding pixels, multiplied by coefficients.

The filter is represented by a table (matrix), characterized by its dimensions and its coefficients, whose center is the pixel concerned. The coefficients in Table determine the filter properties. An example of filter 3 x 3 is described in Fig 5:

1	1	1
1	4	1
1	1	1

Figure 5. Filter 3x3

Thus the product matrix image, usually very large because it represents the initial image (array of pixels), the filter provides a matrix corresponding to the processed image.

We can say that filtering is to apply a transformation (called a filter) to all or part of a digital image by applying an operator. One generally distinguishes the following types of filters:

- **LOW-PASS filters**, is to mitigate the components of the image having a high frequency (dark pixels). This type of filtering is generally used to reduce the image noise is why we usually talk about smoothing. The averaging filters are a type of low-pass filters whose principle is to average the values of neighboring pixels. The result of this filter is a fuzzy picture. MATLAB has a function to apply such filtering; it is the function Filter2 (Fig.6.c);
- **Median filter** is a nonlinear filter, which consists of replacing the gray level value at each pixel by the gray level surrounded by so many values that are higher than lower values in a neighborhood of the point considered. MATLAB has a function to apply such filtering; it is the function MEDFILT2 (Fig.6.e);
- **HIGH-PASS FILTER**, unlike the low-pass, reduce the low frequency components of the image and make it possible to accentuate the detail and contrast is why the term "filter accentuation" is sometimes used (Fig.6.d);
- **Filters BANDPASS** for obtaining the difference between the original image and that obtained by applying a low-pass filter.
- **Directional Filters** applying a transformation in a given direction.

Consider an image I and a two-dimensional filter h , filtering the image I by the filter F is an image whose luminance is given by:

$$F(x, y) = \sum_{a,b} h(a,b)I(x+a, y+b) \quad (5)$$

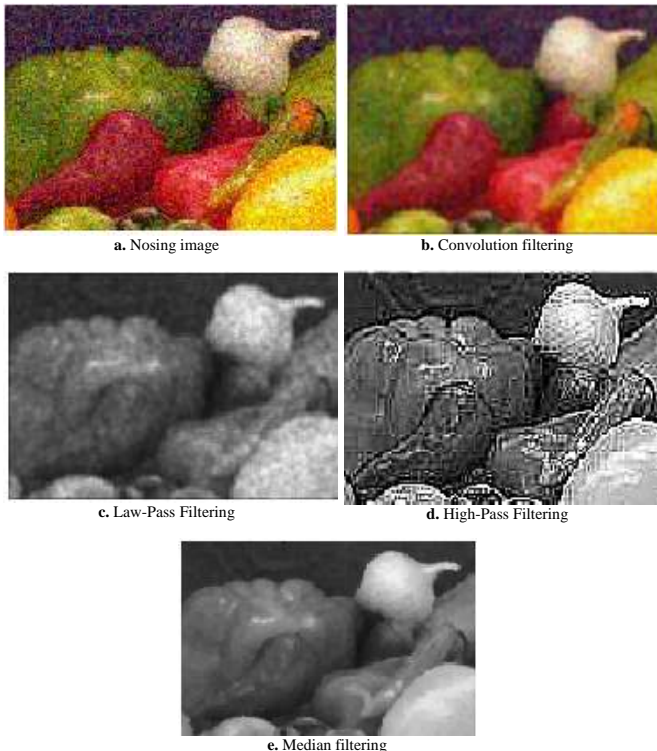


Figure 6. Image Filtering (a, b, c, d et e)

V. RESTAURATION

Restoring an image is to try to offset the damage suffered by this image [4][6]. The most common impairments are a blur or defocus shake. F image available is the result of degradation of the original image I . This degradation, when it's acts of defocus blur or camera shake, as a first approximation can be modeled by a linear filter h , followed by the addition of noise B (Fig.7). The noise can account for the actual noise at the sensor and quantization noise from the digitization of the picture, but also and above all, the difference between the adopted model and reality. In general, we assume that it is a Gaussian white noise in the frequency domain.

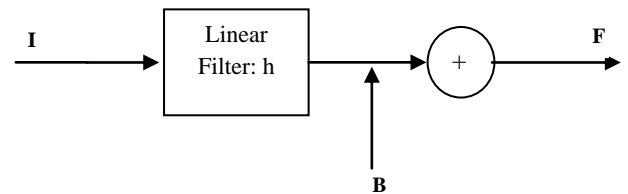


Figure 7. Principle of degradation

The restoration is calculated, from F , an image I as close to the original image I [7]. To do this, we need to know the degradation. Degradation is sometimes assumed to be known, but in practice it is generally unknown, so we have estimation from the picture deteriorated, as shown in Fig.8:

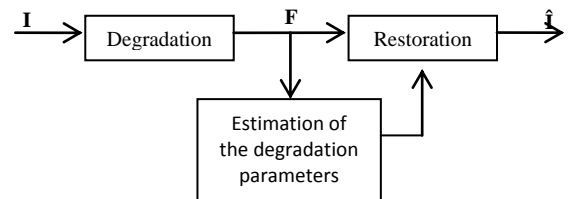


Figure 8. Principle of restoration

$$f(x, y) = \sum_{a,b} h(a,b)I(x+a, y+b) + B(x, y) \quad (6)$$

$$= (h * I)(x, y) + B(x, y)$$

By taking the Fourier transform, assuming the images I and F periodic, we obtain:

$$F(u, v) = H(u, v)I(u, v) + B(u, v) \quad (7)$$

There are several types of degradation, mention the following:

A. Defocus

Each point on the scene then gives the picture a task-shaped disk; this task is much larger than the defocusing is important [6]. Degradation can be modeled by a linear filter h whose coefficients $h(x, y)$ apply to the inside of discs and 0 outside (the value of λ is calculated so that the sum is equal to 1).

To simplify the experiments, we assume below that the degradation is performed by a filter whose impulse

response is a square (a square of $(2T+1) * (2T+1)$ pixels (where T is an integer) . We then:

$$a = \frac{1}{(2T+1)^2} \begin{cases} h(x, y) = a \text{ si } |x| \leq T \text{ ou si } |y| \leq T \\ h(x, y) = 0 \text{ sinon} \end{cases} \quad (8)$$

The parameter to determine is T .

B. Candles

If the deterioration is due to shake can be a first approximation, assuming that each point of the scene image gives a spot-shaped line segment (the orientation of this segment depends on the direction of move) [5]. This is modeled by a filter whose impulse response is in the shape of a segment.

For simplicity, assume here that this segment is horizontal. Thus, degradation is modeled by a horizontal filter

$$h = [\sqrt{a}, \sqrt{a}, \dots, \sqrt{a}] \text{ á } 2T+1 \text{ coefficients.}$$

The value of coefficients is then:

$$\sqrt{a} = 1/(2T+1) \quad (9)$$

In reality, it is clear that the value of T is not necessarily right, that orientation may not be moved horizontally, the spot defocusing provides disk-shaped and not square shaped.

The degraded image is filtered by a filter $g(x, y)$ the inverse of $h(x, y)$. The problem is that the calculation of this filter is not always trivial, because it is the opposite in the sense of convolution. That's why we pass in the frequency domain using the Fourier transform. Indeed, this transform converts convolution into multiplication. In frequency, there will therefore:

$$G(u, v) = \frac{1}{H(u, v)} \quad (10)$$

In the previous sections we saw how to estimate $h(x, y)$. Fourier transform, we deduce $H(u, v)$. We take care to observe the following precautions: Before applying the Fourier transform, fill with zeros h to reach the size of the image, otherwise the equation (10) contain extension functions different. Finally, the following equation gives us $G(u, v)$.

To restore the image, we calculate the spectrum of the restored image:

$$\hat{I}(u, v) = G(u, v) F(u, v) \quad (11)$$

This involves applying the inverse filter in the frequency domain. Finally, an inverse Fourier transform applied to $\hat{I}(u, v)$ gives the restored image $I(x, y)$.

To better understand the principle and limitations of this method, we will now express $I(u, v)$ as a function of

$I(u, v)$. Starting from the equation (11) and replacing $F(u, v)$ by its expression (10), we obtain:

$$\hat{I}(u, v) = G(u, v) H(u, v) I(u, v) + G(u, v) B(u, v) \quad (12)$$

Since, $G(u, v) H(u, v) = 1$, on a :

$$\hat{I}(u, v) = I(u, v) + G(u, v) B(u, v) \quad (13)$$

If the noise was zero, we would find exactly the original image. For nonzero noise, which will always be the case in practice, a problem arises when $H(u, v)$ becomes very low, because we will have a very high value of $G(u, v)$, which causes a strong noise amplification. A simple solution is to limit the possible values of $G(u, v)$:

If $G(u, v) > S$, then $G(u, v) = S$

If $G(u, v) < -S$, then $G(u, v) = -S$

Where, S is a positive threshold.

But because of this thresholding, G is not exactly the inverse of H , and degradation can't be totally eliminated.

VI. THRESHOLDING

The operation called "simple thresholding" is set to zero all pixels with a gray level below a certain value (called threshold, English treshold) and the maximum value pixels with a higher value. Thus the result of thresholding is a binary image containing black and white pixels (Fig.9), is the reason why the term is sometimes used for binarization. Thresholding can highlight shapes or objects in an image. However the difficulty lies in choosing the threshold to adopt.



Figure 9. Image thresholding

VII. EDGE DETECTION

The contours are essential information for certain applications of image processing. In particular, the contours of an object can generally characterize its shape. Edge detection can be achieved with filters whose coefficients were carefully chosen. We discuss in this section, three sets of special filters: filters Prewitt, Roberts and Sobel.

A set of filters edge detection (Prewitt, Roberts and Sobel) consists of a pair of linear filters that detect the contours in two orthogonal directions: vertical and horizontal.

A. Filter Prewitt

The Prewitt filters horizontal and vertical are:

$$H = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{pmatrix} \quad V = \begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix}$$

From an image I, we calculate G_h and G_v , image corresponding to the filtering of I h vs. we obtain the contour image single:

$$G = \sqrt{G_h^2 + G_v^2} \quad (14)$$

We seek a binary contour image. For this, we must choose a threshold of detection. All pixels of G whose intensity is above the threshold will increase to state 1.

B. Sobel Filtre

The Sobel filters horizontal and vertical are:

$$H = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{pmatrix} \quad V = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}$$

C. Robert Filtre

The Robert filters horizontal and vertical are:

$$H = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad V = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

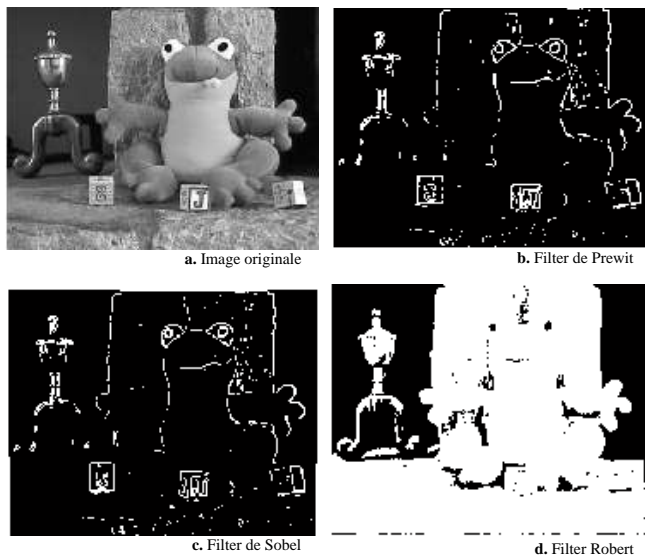


Figure 10. Edge detection of an image (a, b, c and d)

The choice of all of these treatments is justified by our desire to diversify the fields of applications enabled by our software IPS, which makes it very useful.

CONCLUSION

In this work, we developed software that we have appointed IPS and which is developed on MATLAB platform. This software is designed to allow the uninitiated to handle different programming features of MATLAB for image processing. IPS provides access to a set of operations such as: filtering, sound effects, restoration, edge detection and format change. It has a graphical interface easier to handle. We intend to improve this version by adding more functions.

REFERENCES

- [1] Agarwal S., Awan A. and D. Roth, Learning to detect objects in images via a sparse, part-based representation, IEEE Trans. Pattern Anal. Mach. Intell. 26 (11) (2004), pp. 1475–1490
- [2] Babaoğlu, Investigation of Reproduction Cells Using Digital Image Processing Techniques, Institute of Science, Selcuk University, Konya (2004).
- [3] Balasko B., J. Abonyi and B. Feil, Fuzzy Clustering and Data Analysis Toolbox for Use with MATLAB, Veszprem University, Hungary (2008)
- [4] Blanchet G. et Charbit M. : Signaux et images sous MATLAB, Edition Hermes Sciences Europe Ltd, Paris 2001.
- [5] Bres S., Jolion J.-M., LEBourgeois f. : Analyse et traitements des images numérique, Edition Lavoisier 2003.
- [6] Burel G. : Introduction au traitement d'image simulation sous MATLAB, Edition Hermes Sciences Europe Ltd, Paris et Floch, Octobre 2001.
- [7] Caroline CHAUX : Analyse en ondelettes M-bandes en arbre dual application à la restauration d'images, « <http://igm.univ-mlv.fr/LabInfo/rapportsInternes/2007/03.pdf> ».
- [8] Chien S. and H. Mortensen, Automating image processing for scientific data analysis of a large image database, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996), pp. 854–859.
- [9] Eddins S.L., R.C. Gonzalez, and R.E. Woods, Digital image processing using MATLAB, Prentice-Hall, NJ (2004).
- [10] Gonzalez, R.C., Woods, R.E., Digital Image Processing, third ed. Prentice-Hall Inc, Upper Saddle River, New Jersey, 2008.
- [11] Nouvel, A., 2002. Description de concepts par un langage visuel pour un système d'aide à la conception d'applications de traitement d'images. Ph.D. Thesis, Paul Sabatier University, Toulouse, France, September.
- [12] Russ, 2006 Russ, J.C., 2006. The Image Processing Handbook, fifth ed. (Image Processing Handbook), CRC Press Inc., Boca Raton, FL, USA.

Improved Off-Line Intrusion Detection Using A Genetic Algorithm And RMI

Ahmed AHMIM¹

Department of Computer Science,
Badji Mokhtar University
Annaba, 23000, Algeria
ahmed.ahmim@hotmail.fr

Nacira GHOUALMI²

Department of Computer Science,
Badji Mokhtar University
Annaba, 23000, Algeria
ghoualmi@yahoo.fr

Noujoud KAHYA³

Department of Computer Science,
Badji Mokhtar University
Annaba, 23000, Algeria
kahya.noudjoud@gmail.com

Abstract— This article proposes an optimization of using Genetic Algorithms for the Security Audit Trail Analysis Problem, which was proposed by L. Mé in 1995 and improved by Pedro A. Diaz-Gomez and Dean F. Hougen in 2005. This optimization consists in filtering the attacks. So, we classify attacks in “Certainly not existing attacks class”, “Certainly existing attacks class” and “Uncertainly existing attacks class”. The proposed idea is to divide the 3rd class to independent sub-problems easier to solve. We use also the remote method invocation (RMI) to reduce resolution time. The results are very significant: 0% false+, 0%false-, detection rate equal to 100%. We present also, a comparative study to confirm the given improvement.

Keywords-component; intrusion detection system; Genetic Algorithm; Off-Line Intrusion Detection; Misuse Detection;

I. INTRODUCTION

The computing networks became the paramount tool for the various sectors (social, economies, military... etc.). The phenomenal developments of networks are naturally accompanied by the increase in the number of users. These users, known or not, are not necessarily full of good intentions for these networks. They can exploit the vulnerabilities of networks and systems, to try access to sensitive information in order to read, modify or destroy them. Therefore, that these networks appear the targets of potential attacks, their securing has become an unavoidable bet.

Computer security has become in recent years a crucial problem. It rallies the methods, techniques and tools used to protect systems, data and services against the accidental or intentional threats, for ensure: Confidentiality; Availability; Integrity [1].

Nowadays, different techniques and methods have been developed to implement a security policy: authentication, cryptography, firewalls, proxies, antivirus, Virtual Private Network (VPN), Intrusion Detection System (IDS). This paper is organized after an introduction as: The second section is a state of the art on the IDS. The third section presents a formalization of the Security Audit Trail Analysis Problem (SATAP) as well as using Genetic Algorithms for the Security Audit Trail Analysis Problem proposed by Mé

[2]; the fourth section presents our contribution to optimize using Genetic Algorithms for the Security Audit Trail Analysis Problem. The fifth section presents the results obtained by our approach; the sixth section presents a comparative study between the two approaches. Finally, the conclusion presents the advantages of our approach, and the prospects work.

II. THE INTRUSION DETECTION SYSTEMS

Intrusion detection systems (IDSs) are software or hardware systems that automate the process of monitoring the events occurring in a computer system or network, analyzing them for signs of security problems [3]. The intrusion detection system was introduced by James Anderson [4], but the subject didn't have great success. After that, Denning defined the intrusion detection system models [5], where he exhibits the importance of security audit, with the aim to detect the possible violations of system security policy.

According to Intrusion Detection Working Group of IETF an intrusion detection system includes three vital functional elements: information source, analysis engine, and response component [6].

There are five concepts to classify intrusion detection Systems, which are: The detection method; The behavior on detection; The audit source location; The detection paradigm; The usage frequency [6].

The detection method is one of the principal characters of classification they describe the characteristics of the analyzer. When the intrusion detection system uses information about the normal behavior of the system it monitors, we qualify it as behavior-based. When the intrusion detection system uses information about the attacks, we qualify it as knowledge-based [6].

III. INTRUSION DETECTION BY SECURITY AUDIT TRAIL ANALYSIS

The Security Audit is as medical diagnosis, in order to determine the set of conditions, which may explain the presence of observed symptoms (in IDS: the recorded events in the audit trail). For this reason, expert uses specific knowledge (the scenarios of attack) based cause at an effect.

The expert uses its knowledge to develop assumptions that confront the reality observed. If there are still observed symptoms than the made hypothesis made is wrong. On the other hand, if there are more symptoms than those observed in the reality, a new hypothesis more relevant must be tested [2].

In this approach, the attack scenarios are modeled as a set of couples (E_i, N_i) where E_i is the type of event and N_i is the number of occurrences of this type of event in the scenario. This approach is called « the Security Audit Trail Analysis Problem».

A. Specification of the Security Audit Trail Analysis Problem [7]

Formally, the Security Audit Trail Analysis Problem can be expressed by the following statement:

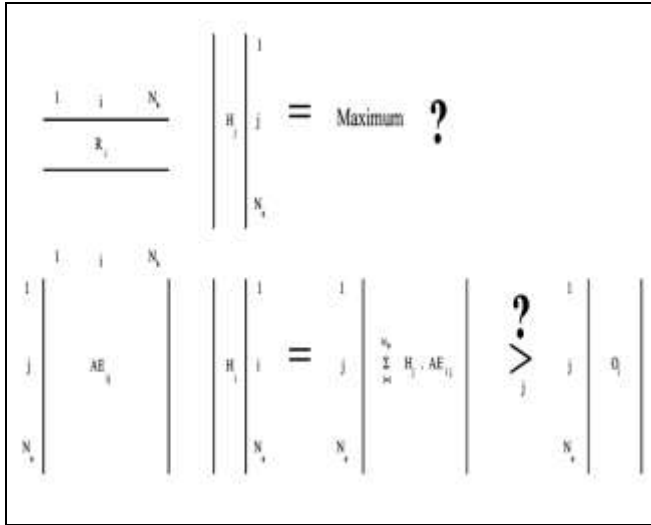


Figure 1. The Security Audit Trail Analysis Problem [7]

- N_e : the number of type of audit events.
 - N_a : the number of potential known attacks.
 - AE : is the $N_a \times N_e$ attacks-events matrix which gives the set of events generated by each attack. AE_{ij} is the number of audit events of type i generated by the scenario j
 - $AE_{ij} \geq 0$ (1)
 - R : is N_a dimensional weight vector, where:
 - $(R_i > 0)$ (2)
 - is the weight associated to the attack i (R_i is proportional to the risk inherent in the attack scenario i).
 - O : is the N_e dimensional vector where:
 - O_i counts the occurrence of events of type i present in the audit trail (O is "observed audit vector").
 - H : is N_a dimensional hypothesis vector, where:
- $$H_i = 1 \quad (3)$$

- (a) If the attack i is present according to the hypothesis and

$$H_i = 0 \quad (4)$$

- (b) Otherwise (H describes a particular attack subset).

- (c)

To explain the data contained in the audit trail (i.e. O) by the occurrence of one or more attack. We have to find the H vector which maximizes the $R \times H$ Product (it's the pessimistic approach: finding H so that the risk is the greatest) with the constraint:

$$(AE \times H)_i \leq O_i, (1 \leq i \leq N_e) \quad (5)$$

Finding H vector is NP-complete. Consequently, the application of classical algorithms is therefore, impossible where N_a equals to several hundreds.

The heuristic approach that we have chosen to solve that NP-complete problem is the following: a hypothesis is made (e.g. among the set of possible attacks, attacks i, j and k are present in the trail), the realism of the hypothesis is evaluated and, according to this evaluation, an improved hypothesis is tried, until a solution is found.

In order to evaluate a hypothesis corresponding to a particular subset of present attack, we count the occurrence of events of each type generated by all the attacks of the hypothesis. If these numbers are less than or equal to the number of events recorded in the trail, then the hypothesis is realistic.

After, we have to find an algorithm to derive a new hypothesis based on the past hypothesis: it is the role of the genetic algorithm.

B. Using Genetic Algorithms for Misuse Detection [7]

Genetic algorithms (GA) are optimum search algorithms based on the mechanism of natural selection in a population. A population is a set of artificial creatures (individuals or chromosomes). These creatures are strings of length L coding a potential solution to the problem to be solved, most often with a binary alphabet. The size L of the population is constant. The population is nothing but a set of points in a search space. The population is randomly generated and then evolves in every generation. A new set of artificial creatures is created using the fittest or pieces of the fittest individuals of the previous one. The fitness of everyone is simply the value of the function to be optimized (the fitness function) for the point corresponding to the individual. The iterative process of population creation is achieved by three basic genetic operators: selection (selects the fittest individuals), reproduction or crossover (promotes exploration of new regions of the search space by crossing over parts of individuals) and mutation (protects the population against an irrecoverable loss of information).

Two challenges arise when applying GAs to a particular problem: coding a solution for that problem with a string of bits and finding a fitness function to evaluate everyone of the population.

1) Coding a Solution with a Binary String [7]

An individual is a one length string coding a potential solution to the problem to be solved. In our case, the coding is straightforward: the length of an individual is N_a and each individual in the population corresponds to a particular H vector.

2) The Fitness Function [7]

We have to search, among all the possible attack subsets, for the one which presents the greatest risk to the system. This result in the maximization of the product $R \times H$. As GAs are optimum search algorithms, finding the maximum of a fitness function, we can easily conclude that in our case this function should be made equal to the product $R \times H$. So we have:

$$F = \sum_{i=1}^{N_a} R_i I_i \quad (6)$$

Where I is an individual.

This fitness function does not take into account the constraint feature of our problem, which implies that some individuals among the 2^{N_a} Possible are not realistic.

This is the case for some i type of events when:

$$(AE \times H)_i > O_i \quad (7)$$

As a large number of individuals do not respect the constraint. We decided to penalize them by reducing their fitness values. So we compute a penalty function (P) which increases as the realism of this individual decreases: let T_e be the number of types of events for which

$$(AE \times H)_i > O_i \quad (8)$$

The penalty function applied to such an H individual is then:

$$P = T_e^p \quad (9)$$

A quadratic penalty function (i.e. $p = 2$) allows a good discrimination among the individuals. The proposed fitness function is thus the following:

$$F(I) = \alpha + \left(\sum_{i=1}^{N_a} R_i I_i - \beta \cdot T_e^p \right) \quad (10)$$

The β parameter makes it possible to modify the slope of the penalty function and α sets a threshold making the fitness positive. If a negative fitness value is found, it is equaled to 0 and the corresponding individual cannot be selected. So the parameter allows the elimination of a too unrealistic hypothesis.

This selective function was improved by Diaz-Gomez, P. A. Hougen [8]. This improvement proved mathematically [9]

[10]. A new selective function provides less false positives and less false negative [11].

The new selective function is:

$$F(I) = N_e - T' \quad (11)$$

Where N_e corresponds to the total number of classified events. T' corresponds to the number of overestimates, i.e., the number of times $(AE \times H)_i > O_i$ for each attack H_i . That is, if a hypothesized attack H_i considered alone, would cause $(AE \times H)_i > O_i$ for some i, and another hypothesized attack H_j considered alone, would also cause $(AE \times H)_i > O_i$, then T' would have a value of 2 [12].

IV. CONTRIBUTIONS

Inspired from Ludovic Mé [7] contributions and Diaz-Gomez, P. A. Hougen [8] improvement, we classify attacks in Security Audit Trail in three classes and divide the 3rd class to independent sub-problems. Then, we apply the genetic algorithm with the proposed crossover operator in [13] and L. Mé selective function (10). The second contribution is to optimize the resolution time of the genetic algorithm. For this we apply RMI (remote method invocation) to each sub-problem.

The Figure2 represents the activity diagram that summarizes the different steps of our proposition.

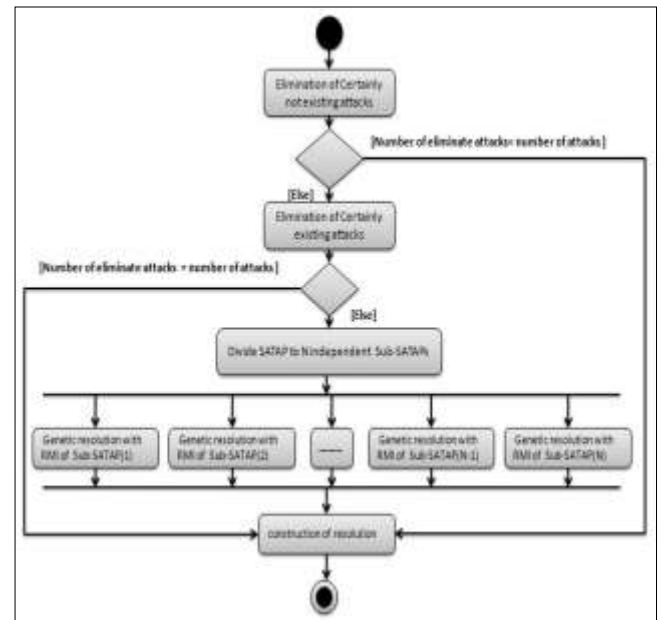


Figure 2. Activity diagram of our contribution

A. Filtration of attacks

The Filter uses Observation matrix “O” and the Matrix attack-event “AE”. The proposed idea is reducing the size of the problem in order to obtain the correct solution and to reduce the runtime.

Consequently, we classify attacks in three classes, and divide the last class to sub-problems:

1) Certainly not existing attacks' class

Eliminate attacks, which have a probability of existence equal to 0%. These attacks generate an occurrence number for one of the events greater than the occurrences number audited for this event. So, attacks i satisfy the following formula:

$$\exists j \in N_e (AE_{ij} > O_j) \quad (12)$$

To eliminate these attacks, we compare Matrix attack-event "AE" with the Observation matrix "O". The result is attacks noted N_{ap} . After that, we remove the events that have value 0 in Observation matrix "O". The number of events used for selected attacks is noted N_{ep} .

The results are matrixes with the following dimensions:

$$AE_{(N_{ap}, N_{ep})}, O_{(N_{ep})}, H_{(N_{ap})}, R_{(N_{ap})} \quad (13)$$

Figure 3 shows an example of elimination certainly not existing attacks. For example, attack A2 generates 5 events E9, while the Security Audit Trail records only 4 events E9.

AE Matrix														O Matrix	
	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13		
E1	0	0	0	0	0	0	0	0	0	0	0	4	3	E1	6
E2	1	0	5	0	0	0	0	0	0	1	0	0	0	E2	30
E3	0	0	0	0	1	0	0	3	0	0	0	0	0	E3	5
E4	0	0	0	0	0	0	0	0	0	0	3	0	0	E4	3
E5	0	1	0	0	0	0	0	0	4	0	0	0	0	E5	4
E6	0	2	0	0	2	0	0	1	0	0	0	0	0	E6	2
E7	0	0	0	0	0	1	0	0	0	4	0	0	0	E7	5
E8	0	0	4	3	0	0	0	0	2	0	1	0	0	E8	16
E9	0	5	0	0	3	1	0	0	2	0	0	0	0	E9	4
E10	0	0	0	0	0	2	3	1	0	0	0	0	0	E10	4
E11	1	0	0	4	0	3	0	0	0	0	0	0	8	E11	22
E12	1	0	0	0	0	0	0	0	0	0	0	0	0	E12	17
E13	0	0	2	0	0	0	0	0	0	0	0	0	0	E13	3
E14	0	0	3	0	0	0	0	2	0	0	3	0	0	E14	9
E15	4	0	0	0	0	0	0	0	0	1	0	0	0	E15	7
E16	0	0	0	1	0	0	0	0	0	0	3	0	0	E16	5
E17	1	0	0	1	0	1	0	0	0	0	0	7	0	E17	13
E18	0	0	0	0	0	1	0	0	1	0	0	0	0	E18	5
E19	0	1	0	0	0	0	0	1	0	0	1	0	0	E19	23

Figure 3. Example of step1

2) Certainly existing attacks' class

Eliminate attacks, which have a possibility of existence equal to 100 %. These attacks haven't a common event with other attacks. In this case, the sum of their occurrence number is less than or equal to the audited occurrences' number for this event. For eliminate these attacks we compare the $AE_{(N_{ap}, N_{ep})}$ and matrix $O_{(N_{ep})}$. So, attacks i that verify the following formula:

$$\forall j \in N_e \left((AE_{ij} > 0) \rightarrow \left(\sum_{i=0}^{i=N_a} AE_{ij} \leq O_j \right) \right) \quad (14)$$

The result is the attacks noted N_{ha} . Consequently, we resize Matrix attack-event "AE" to the size (N_{ha}, N_{ep}) . After these treatments, we eliminate the events j that verifies the formula:

$$\left(\sum_{i=0}^{i=N_a} AE_{ij} \right) \leq O_j \quad (15)$$

The number of attacks events retain is noted N_{he} .

The results are matrixes with the following dimensions:

$$AE_{(N_{ha}, N_{he})}, O_{(N_{he})}, H_{(N_{ha})}, R_{(N_{ha})} \quad (16)$$

Figure 4 shows an example of elimination certainly existing attacks. For example attack A11 haven't a common event with other attacks for event E4, and when they have a common event with other attacks (E7, E14, E19), the sum of their occurrences number is less than or equal to the audited occurrences number for this event.

AE Matrix														O Matrix	
	A1	A3	A4	A5	A7	A8	A9	A10	A11	A12	A13				
E1	0	0	0	0	0	0	0	0	0	4	3	E1	6		
E2	1	0	5	0	0	0	0	0	1	0	0	E2	30		
E3	0	0	0	0	1	0	3	0	0	0	0	E3	5		
E4	0	0	0	0	0	0	0	0	0	3	0	E4	3		
E5	0	0	0	0	0	0	0	4	0	0	0	E5	4		
E6	0	0	0	0	2	0	1	0	0	0	0	E6	2		
E7	0	0	0	0	1	0	0	0	3	0	0	E7	5		
E8	0	4	3	0	0	0	0	2	0	1	0	E8	16		
E9	0	0	0	3	0	0	2	0	0	0	0	E9	4		
E10	0	0	0	0	2	1	0	0	0	0	0	E10	4		
E11	1	0	4	0	0	0	0	0	0	0	8	E11	22		
E12	1	0	0	0	0	0	0	0	0	0	0	E12	17		
E13	0	2	0	0	0	0	0	0	0	0	0	E13	3		
E14	0	3	0	0	0	2	0	0	3	0	0	E14	9		
E15	4	0	0	0	0	0	0	1	0	0	0	E15	7		
E16	0	0	1	0	0	0	0	0	3	0	0	E16	5		
E17	1	0	1	0	1	0	0	0	0	7	0	E17	13		
E18	0	0	0	0	1	0	1	0	0	0	0	E18	5		
E19	0	0	0	0	0	1	0	0	1	0	0	E19	23		

Figure 4. Example of step 2

3) Uncertainly existing attacks' class

This last class is concerned by our contribution. These attacks that we doubt for their existence represent the real Security Audit Trail Analysis Problem (SATAP). These attacks represent the uncertainly existing attacks' class.

B. Divisions SATAP to sub-SATAP

We use the 3rd class. So, we regroup the attacks that generate the same kind event where the sum of the occurrences number exceeds the occurrences number audited for this event. This relation between attacks called "mutually exclusive".

Each attack group contain attacks "mutually exclusive" over there and we associate to each attack group the event group which they have an occurrences number higher than the audited occurrences number. We create the Sub-SATAP where each SATAP_i contains the attacks of the group i with associated events.

Figure.5 presents the associated algorithm to the process described above. Procedure add-element is called in procedure grouping.

Procedure grouping

begin

for i=1 to N_{ha} do

if (not-marked attack(i)) then

Create group();

Mark(i);

add-element(i);


```

end.
Procedure add-element(i)
begin
for  $j = 1$  to  $N_{ep}$  do
if  $AE_{ij} \neq 0$  then
for  $x = i + 1$  to  $N_{ha}$  do
if  $AE_{xj} \neq 0$  then
if (not-marked attack (x)) then
add-to-
group(x);
Mark(i);
add-
element(x);
else
fusion-group-where-
belong(x,i);
end .

```

Figure 5. Division SATAP to Sub-SATAP algorithm

Each sub-SATAP $\{AE, R, O, H\}_i$ defines the sub-SATAP_i

Figure 6 shows an example of Divisions SATAP to sub-SATAPs. For example, attack A5 is mutually exclusive with attacks A8 for the event E6 and mutually exclusive with attacks A9 for the event E9. For this reason, the attacks A5 A8 A9 with the event E6 E9 represent one of the sub-SATAP.

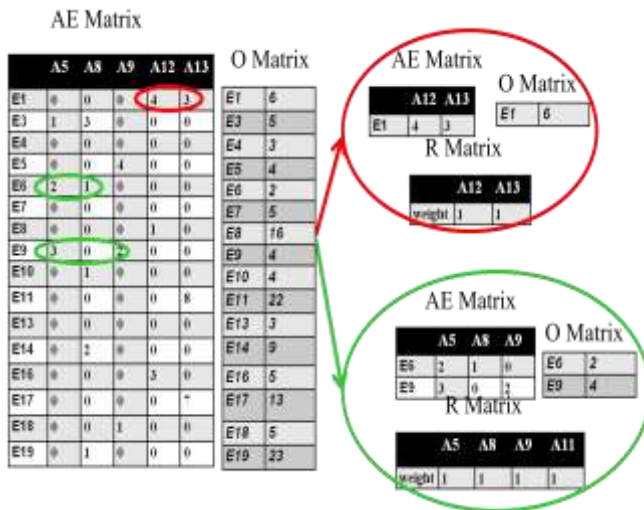


Figure 6. Example of step 3

C. The crossover

The proposed crossover operator is a crossover strongly random. All heritage possibilities are reached from the first generation in reduced time. The advantage of this crossover is the minimization of the generation number needed to generate certain individual that can be the best solution of our problem. This crossover consists, firstly, to make a cloning one of the two parents. So, the generated member inherits randomly the genes of the second parent, and we put it in the corresponding locus in the cloned parent [13] as shown in Figure 7.

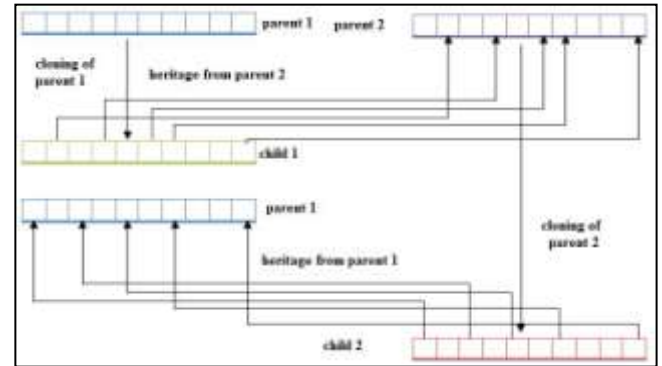


Figure 7. Crossover of our proposition

D. Resolve the sub SATAP simultaneously with RMI

This step consists to resolve the sub-SATAPs simultaneously using the remote method invocation. We associate to each sub-SATAP a thread to resolve it in the suitable computer (best performance for the biggest sub-SATAP) as shown in Figure 8.

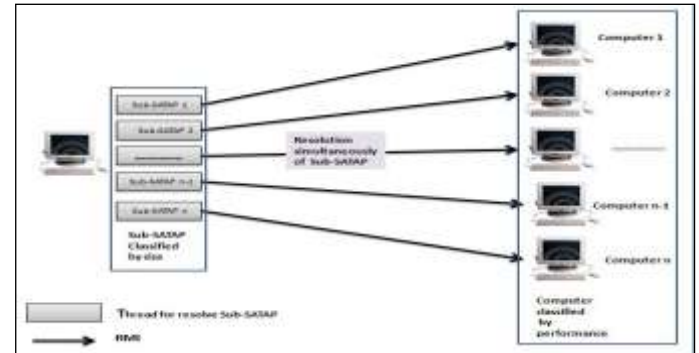


Figure 8. Simultaneous resolution mechanism of the sub-SATAP

V. EXPERIMENTATIONS

A. Used Metrics

To evaluate the performance of this contribution, several tests with several benchmarks extracted from the KDD Cup 1999 data set [14] was performed. The evaluation metrics used are the following:

- False positive: false alarms caused by legitimate changes in program behavior [15].
- False negative: missed intrusions caused by attackers who mimic benign users [15].
- Detection rate.
- Processing time.

TABLE I. IT SUMMARIZES THE DIFFERENT RESULTS OBTAINED WITH THE VARIOUS BENCHMARKS.

Benchmark	Fals e + %	Fals e - %	Dete ction rate %	Processi ng Time	Processi ng Time with RMI
benchmark(2,4)	0%	0%	100%	≈0 ms	≈0 ms
benchmark(5,9)	0%	0%	100%	≈0 ms	≈0 ms
benchmark(6,9)	0%	0%	100%	≈7 ms	≈7 ms
benchmark(9,11)	0%	0%	100%	≈0 ms	≈0 ms
benchmark(10,12)	0%	0%	100%	≈0 ms	≈0 ms
benchmark(15,19)	0%	0%	100%	≈13 ms	≈16 ms
benchmark(15,20)	0%	0%	100%	≈0 ms	≈0 ms
benchmark(17,35)	0%	0%	100%	≈0 ms	≈0 ms
benchmark(21,40)	0%	0%	100%	≈0 ms	≈0 ms
benchmark(24,50)	0%	0%	100%	≈0 ms	≈0 ms
benchmark(25,51)	0%	0%	100%	≈232 ms	≈265 ms

We remark that for all benchmarks the proportion of false positive and false negative equal to 0 % and the detection rate equal to 100 % that signify, the good quality of resultants.

We remark also that there are several benchmarks, treated in real time (0 ms). This means, that during the two first steps of attack classification we can attest about the existing attacks or not. The other benchmarks are concerned by the second step “Divisions SATAP to sub-SATAP” and the genetic algorithms must be applied to identify attacks that justify the increase of processing time. The benchmark (15, 19) and (25, 51) are treated in the more reducing time than the resolution without remote method invocation due to simultaneously treatment of the sub-SATAP.

VI. COMPARATIVE STUDY

First we compare the results of the contribution and the work of Mé [7] using the same benchmarks. The following metric are used: the number of detected attacks, the number of constraints raped during each generation, the convergence speed to the best solution, the number of generation and the necessary time for the resolution.

Results show in Figure.10 and Figure.9 and table 2 that with our work we detect the same attack's percentage that represents the real attacks.

However, there are some differences:

- Runtime: the processing time of our proposition is less than the processing time of [7] and [11] due to minimizing of problem size and dividing the problem to sub-problems and the simultaneously treatment of sub-SATAP.
- Generations number: the number of generations needed for our proposition is less than the number of generations needed for [7] and [11], because the size of the biggest sub-SATAP to be treated is less than or equal (in the worst case) the size of SATAP.
- Convergence speed: the convergence speed of our proposition is faster than [7]and[11], because the resolution of sub-SATAP is more efficient than that of SATAP.(in the worst case) the size of SATAP.
- Constraints' violation: due to the filtering operation and the dividing of SATAP to sub-SATAP, the constraint's violation of contribution is lesser (almost nonexistent) than the [7] and [11].

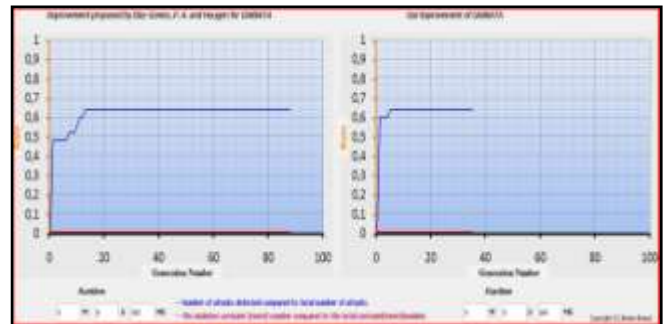


Figure 9. Comparison between our improvement and Diaz-Gomez, P. A. and Hougen for benchmark (15,19)

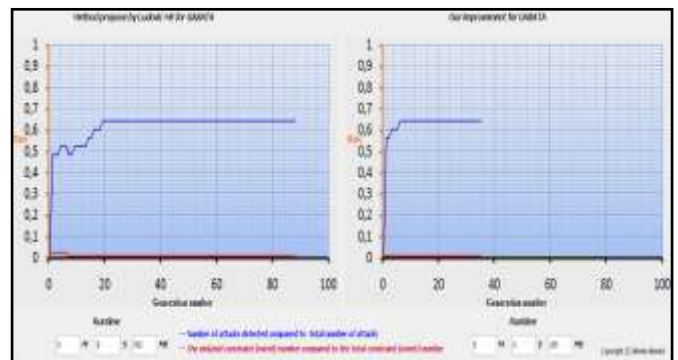


Figure 10. Comparison between improvement and L.Mé resolution for benchmark (25,51)

TABLE II. TCOMPARISON BETWEEN THE TWO RESOLUTION METHODS

	Classical resolution	Our proposition	
		Without RMI	With RMI
benchmark(2,4)	≈16 ms	≈0 ms	≈0 ms
benchmark(5,9)	≈62 ms	≈0 ms	≈0 ms
benchmark(6,9)	≈125 ms	≈7 ms	≈7 ms
benchmark(9,11)	≈234 ms	≈0 ms	≈0 ms
benchmark(10,12)	≈312 ms	≈0 ms	≈0 ms
benchmark(15,19)	≈1s 60 ms	≈16 ms	≈13 ms
benchmark(15,20)	≈1s 139 m	≈0 ms	≈0 ms
benchmark(17,35)	≈2s 777 m	≈0 ms	≈0 ms
benchmark(21,40)	≈4s 771 m	≈0 ms	≈0 ms
benchmark(24,50)	≈7s 909 m	≈0 ms	≈0 ms
benchmark(25,51)	≈8s 502 m	≈265 ms	≈232 ms

OF SATAP

VII. CONCLUSIONS AND PERSPECTIVES

Using Genetic Algorithms for the Security Audit Trail Analysis Problem has significant results. This contribution consists to classify attacks in Security Audit Trail in three classes and divide the 3rd class to sub-problems. Then, we apply the genetic algorithm with the proposed crossover operator in [13] and same selective function of Mé [7]. The second contribution is to optimize the resolution time of genetic algorithm. For this we apply RMI (remote method invocation) to each sub-problem simultaneously.

The contribution brings the following advantages:

- 0% False +.
- 0% False -.

- 100% detection rate.
- Minimizing the runtime.
- Increasing the convergence speed.
- Reducing the constraints violation.
- Reducing the generations number needed to solve this problem.

This improvement confirms the power of using Genetic Algorithms for the Security Audit Trail Analysis Problem where we detect 100% of real attacks.

Our perspective is to propose architecture of multi-agents system for real time resolution of SATAP.

REFERENCES

- [1] E. Cole, Ronald L. Krutz, J. Conley, "Network Security Bible," Wiley Publishing, Inc. January 2005 ISBN13:978-0-7645-7397-2
- [2] L. Mé, "Un algorithme génétique pour détecter des intrusions dans un système informatique," *VALGO*, 95(1):68-78, 1995.
- [3] R. Bace, P. Mell, "NIST Special Publication on Intrusion Detection Systems", 2001.
- [4] J. Anderson, "Computer Security Threat Monitoring and Surveillance", Technical report, James P. Anderson Company, Fort Washington, Pennsylvania (1980).
- [5] D. Denning, "An Intrusion-Detection Model". *IEEE transaction on Software Engineering*, 13(2):222-232 (1987).
- [6] H. Debar, M. Dacier, A. Wespi, "A Revised Taxonomy for Intrusion-Detection Systems". *Annales des Télécommunications*, 55(7-8), 2000.
- [7] L. Mé, "GASSATA, A Genetic Algorithm as an Alternative Tool for Security Audit Trails Analysis". Web proceedings of the First international workshop on the Recent Advances in Intrusion Détection 1998.
- [8] P. A. Diaz-Gomez, D. F. Hougen, "Improved Off-Line Intrusion Detection using a Genetic Algorithm" In *Proceedings of the Seventh International Conference on Enterprise Information Systems*, 2005
- [9] P. A. Diaz-Gomez, D. F. Hougen, "Analysis and Mathematical Justification of a Fitness Function used in an Intrusion Detection System" In *Proceedings of the Seventh Annual Genetic and Evolutionary Computation Conference* 2005.
- [10] P. A. Diaz-Gomez, D. F. Hougen, "Mathematical Justification of a Fitness Function used for Audit Trail Intrusion Analysis" In the *Poster Session of the Research Experience Symposium 2005*, at the University of Oklahoma.
- [11] P. A. Diaz-Gomez, D. F. Hougen, "A Case Study in Genetic Algorithms applied to Off-line Intrusion Detection Systems" In *1st Annual Computer Science Research Conference at the Stephenson Research and Technology Center* 2005.
- [12] P. A. Diaz-Gomez, D. F. Hougen, "Further Analysis of an Off-Line Intrusion Detection System": An Expanded Case Study in *Multi-Objective Genetic Algorithms SCISS'05 The South Central Information Security Symposium*
- [13] M. Rachid, N. Ghoualmi, "Crossover and Mutation Based Cloning Parent for Degree Constrained Minimum Spanning Tree Problem," *AMERICAN UNIVERSITY OF SHARJAH, UAE*, IEEE ICESMA (2010), 30-1 April, ISBN: 978-9948-427-14-8.
- [14] KDD Cup Data, <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html> October 28, 1999.
- [15] P. Roberto, M. Luigi V, "Intrusion Detection Systems," 2008, ISBN:978-0-387-77265-3.

Automatic Facial Feature Extraction and Expression Recognition based on Neural Network

S.P.Khandait
Deptt of IT, KDKCE
Nagpur, Maharashtra, India
prapti_khandait@yahoo.co.in

Dr. R.C.Thool
Deptt. of IT, SGGSIET
Nanded, Maharashtra, India
rcthool@yahoo.com

P.D.Khandait
Deptt. of Etrx, KDKCE,
Nagpur, Maharashtra, India

Abstract— In this paper, an approach to the problem of automatic facial feature extraction from a still frontal posed image and classification and recognition of facial expression and hence emotion and mood of a person is presented. Feed forward back propagation neural network is used as a classifier for classifying the expressions of supplied face into seven basic categories like surprise, neutral, sad, disgust, fear, happy and angry. For face portion segmentation and localization, morphological image processing operations are used. Permanent facial features like eyebrows, eyes, mouth and nose are extracted using SUSAN edge detection operator, facial geometry, edge projection analysis. Experiments are carried out on JAFFE facial expression database and gives better performance in terms of 100% accuracy for training set and 95.26% accuracy for test set.

Keywords- Edge projection analysis, Facial features, feature extraction, feed forward neural network, segmentation SUSAN edge detection operator.

I. INTRODUCTION

Due to technological advancements; there is an arousal of the world where human being and intelligent robots live together. Area of Human Computer Interaction (HCI) plays an important role in resolving the absences of neutral sympathy in interaction between human being and machine (computer). HCI will be much more effective and useful if computer can predict about emotional state of human being and hence mood of a person from supplied images on the basis of facial expressions. Mehrabian [1] pointed out that 7% of human communication information is communicated by linguistic language (verbal part), 38% by paralanguage (vocal part) and 55% by facial expression. Therefore facial expressions are the most important information for emotions perception in face to face communication. For classifying facial expressions into different categories, it is necessary to extract important facial features which contribute in identifying proper and particular expressions. Recognition and classification of human facial expression by computer is an important issue to develop automatic facial expression recognition system in vision community. In recent years, much research has been done on machine recognition of human facial expressions [2-6]. In last few years, use of computers for Facial expression and emotion recognition and its related information use in HCI has gained significant research interest which in turn given rise to a number of automatic methods to recognize facial

expressions in images or video [7-12]. This paper explains about an approach to the problem of facial feature extraction from a still frontal posed image and classification and recognition of facial expression and hence emotion and mood of a person. Feed forward back propagation neural network is used as a classifier for classifying the expressions of supplied face into seven basic categories like surprise, neutral, sad, disgust, fear, happy and angry. For face portion segmentation basic image processing operation like morphological dilation, erosion, reconstruction techniques with disk structuring element are used. Six permanent Facial features like eyebrows(left and right), eye (left and right), mouth and nose are extracted using facial geometry, edge projection analysis and distance measure and feature vector is formed considering height and width of left eye, height and width of left eyebrow, height and width of right eye, height and width of right eyebrow, height and width of nose and height and width of mouth along with distance between left eye and eyebrow, distance between right eye and eyebrow and distance between nose and mouth. Experiments are carried out on JAFFE facial expression database. The paper is organized as follows. Section I gives brief introduction, Section II describes about survey of existing methods, section III highlights on data collection, section IV presents methodology followed, section V gives experimental results and analysis, section VI presents conclusion and future scope and last section gives references used

II. SURVEY OF EXISTING METHODS

In recent years, the research of developing automatic facial expression recognition systems has attracted a lot of attention. A more recent, complete and detailed overview can be found in [12-14]. Accuracy of facial expression recognition is mainly based on accurate extraction of facial feature components. Facial feature contains three types of information i.e texture, shape and combination of texture and shape information. Feng et. al.[15] used LBP and AAM for finding combination of local feature information, global information and shape information to form a feature vector. They have used nearest neighborhood with weighted chi-sq statistics for expression classification. Feature point localization is done using AAM and centre of eyes and mouth is calculated based on them. Mauricio Hess and G. Martinez [16] used SUSAN algorithm

to extract facial features such as eye corners and center, mouth corners and center, chin and cheek border, and nose corner etc. Gengtao zhou et al.[17] used selective feature extraction method where expressions are roughly classified into three kinds according to the deformation of mouth as 1) sad , anger , disgust 2) happy , fear 3) surprise and again some if then rules are used to sub classify individual group expressions . Jun Ou et al.[18] used 28 facial feature points and Gabor wavelet filter for facial feature localization , PCA for feature extraction and KNN for expression classification . Md. Zia Uddin , J.J. Lee and T.S. Kim [19] used enhanced Independent component Analysis (EICA) to extract locally independent component features which are further classified by Fisher linear Discriminant Analysis (FLDA) .Then discrete HMM is used to model different facial Expressions. Feature extraction results of various conventional method (PCA , PCA-FLDA, ICA and EICA) in conjunction with same HMM scheme were compared and comparative analysis is presented in terms of recognition rate . PCA is unsupervised learning method used to extract useful features and 2nd order statistical method for deriving orthogonal bases containing the maximum variability and is also used for dimensionality reduction .V. Gamathi et al.[20] used uniform local binary pattern (LBP) histogram technique for feature extraction and MANFIS (Multiple Adaptive Neuro Fuzzy Inference system) for expression recognition. GRS Murthy and R.S. Jadon[21] proposed modified PCA (eigen spaces) for eigen face reconstruction method for expression recognition. They have divided the training set of Cohn kanade and JAFFE databases into 6 different partitions and eigen space is constructed for each class , then image is reconstructed. Mean square error is used as a similarity measure for comparing original and reconstructed image. Hadi Seyedarabi et al.[22] developed facial expression recognition system for recognizing basic expression . They have used cross correlation based optical flow method for extracting facial feature vectors. RBF neural network and fuzzy inference system is used for recognizing facial expressions. Zhengyou Zhang et al. [23] presented a FER system where they have compared the use of two type of features extracted from face images for recognizing facial expression .Geometric positions of set of fiducial point and multiscale & multi orientation gabor wavelet coefficient extracted from the face image at the fiducial points are the two approaches used for feature extraction. These are given to neural network classifier separately or jointly and results were compared.

Comparison of the recognition performance with different types of features shows that Gabor wavelet coefficients are more powerful than geometric positions. Junhua Li and Li Peng [24] used feature difference matrix for feature extraction and QNN (Quantum Neural Network) for expression recognition From the survey, it is observed that various approaches have been used to detect facial features [25] and classified as holistic and feature based methods to extract facial feature from images or video sequences of faces. These are geometry based, appearance based, template based and skin color segmentation based approaches. Recently large

amount of contributions were proposed in recognizing expressions using dynamic textures features using both LBP and gabor wavelet approach and appearance features and increases complexity. Moreover one cannot show features located with the help of bounding box. Hence, the proposed facial expression recognition system aimed to use image preprocessing and geometry based techniques for feature extraction and feed forward neural network for expression recognition for the frontal view face images.

III. DATA COLLECTION

Data required for experimentation is collected from JAFFE database for neural network training and testing. JAFFE stands for The Japanese Female Facial Expression (JAFFE) Database. The database contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by ten different Japanese female models. Sixty Japanese subjects have rated each image on 6 emotion adjectives. The database was planned and assembled by Miyuki Kamachi, Michael Lyons, and Jiro Gyoba with the help of Reiko Kubota as a research assistant. The photos were taken at the Psychology Department in Kyushu University. Few samples are shown in Fig. 1



Fig. 1 Few samples of facial expressions of person YM

IV. METHODOLOGY

Fig. 2 shows the proposed pseudo code for Automatic Facial Expression Recognition System.

1. Read input image from database and localize face using morphological image processing operations
2. Crop the face image.
3. Extract features from cropped face.
4. Find facial feature vectors.
5. Train neural network.
6. Recognize expression

Fig. 2 Pseudo code for AFERS

A. Face Portion Localization and Feature Extraction

Face area and facial feature plays an important role in facial expression recognition. Better the feature extraction rate more is the accuracy of facial expression recognition. Precise localization of the face plays an important role in feature extraction, and expression recognition. But in actual application, because of the difference in facial shape and the quality of the image, it is difficult to locate the facial feature precisely. Images from JAFFE database are taken as input. This database contains low contrast images therefore images

are first pre-processed using contrast limited adaptive histogram equalization operation and is used for enhancing contrast of an image. Face area is segmented using morphological image processing operations like dilation, erosion reconstruction, complementation, regional max and clear border(to get Region of Interest).

In order to extract facial features, segmented face image (RoI) is then resized to larger size to make facial components more prominent. SUSAN edge detection operator [26] along with noise filtering operation is applied to locate the edges of various face feature segment components. SUSAN operator places a circular mask around the pixel in question. It then calculates the number of pixels within the circular mask which have similar brightness to the nucleus and refers it as USAN and then subtract USAN size from geometric threshold to produce edge strength image.

Following steps are utilized for facial feature segment localization-

1. Apply Morphological operations to remove smaller segments having all connected components (objects) that have fewer than P pixels where P is some threshold value.
2. Trace the exterior boundaries of segments and draw bounding box by taking into account x,y coordinates and height and width of each segment.
3. Image is partitioned into two regions i.e upper and lower portion on the basis of centre of cropped face assuming the fact that eyes and eyebrows are present in the upper part of face and mouth and nose is present in the lower part. Smaller segments within the region are eliminated by applying appropriate threshold value and remaining number of segments are stored as upper left index, upper right index and lower index. Following criteria is used for selecting appropriate upper left index, upper right index and lower index-
 - a) A portion is an upper part if x and y values are less than centx and centy where centx and centy are x- and y-coordinates of center of cropped image. Eyes and eyebrows are present in this area. For left eye and eyebrow portion certain threshold for values of x and y is considered for eliminating outer segments. For right eye and eyebrow also specific threshold value is chosen for eliminating outer segments
 - b) A portion is a lower portion if its value is greater than centx and centy where centx and centy are x- and y-coordinates of center of an image. Nose and mouth are present in this area. For nose and mouth area segments, x lies in the specific range and y also lies in certain range is considered as region of interest for eliminating outer segments. Here number of segments for each portion are stored. If number of segments are > 2 then following procedure for combining the segments is called (step 4)

4. Segments are checked in vertical direction. If there is overlapping then the segments are combined. Again if segments are >2 then distance is obtained and the segments which are closer are combined. This process is repeated until we get two segments for each part and in all total six segments(Fig. 4 (j)).

This gives the bounding box for total six segments which will be left and right eyes, left and right eyebrows, nose and mouth features of the supplied face (Fig. 3).

B. Formation of Feature Vector

Bounding box location of feature segments obtained in the above step are used to calculate the height and width of left eyebrow, height and width of left eye, height and width of right eyebrow, height and width of right eye, height and width of nose and height and width of mouth. Distance between centre of left eye and eyebrow, right eye and eyebrow and mouth and nose is also calculated. Thus total 15 parameters are obtained and considered as feature vector (Fig.3). Thus-

$$F_v = \{H_1, W_1, H_2, W_2, H_3, W_3, H_4, W_4, H_n, W_n, H_m, W_m, D_1, D_2, D_3\} \quad (1)$$

Where,

H_1 =height of left eyebrow, W_1 = width of left eyebrow
 H_2 = height of left eye, W_2 = width of left eye
 H_3 =height of right eyebrow, W_3 = width of right eyebrow
 H_4 = height of right eye, W_4 = width of right eye
 H_n = height of nose, W_n = width of nose,
 H_m = height of mouth, W_m = width of mouth
 D_1 = distance between centre of left eyebrow and left eye,
 D_2 = distance between centre of right eyebrow and right eye,
 D_3 = distance between centre of nose and mouth

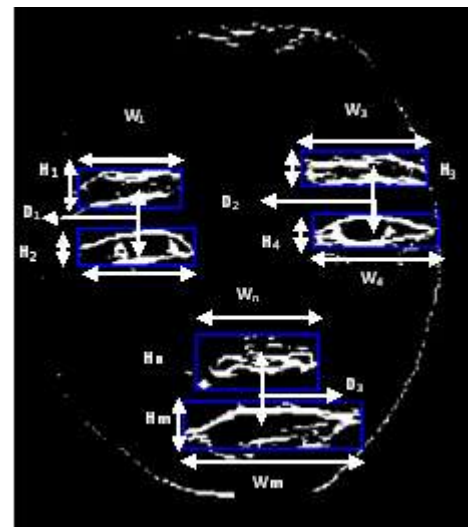


Fig. 3 Feature vector for Expression recognition

C. Expression Classification using Neural Network

Neural computing has re-emerged as an important programming paradigm that attempts to mimic the functionality of the human brain. This area has been developed to solve demanding pattern processing problems, like speech and image processing. These networks have demonstrated their ability to deliver simple and powerful solutions in areas that for many years have challenged conventional computing approaches. A neural network is represented by weighted interconnections between processing elements (PEs). These weights are the parameters that actually define the non-linear function performed by the neural network. Back-Propagation Networks is most widely used neural network algorithm than other algorithms due to its simplicity, together with its universal approximation capacity. The back-propagation algorithm defines a systematic way to update the synaptic weights of multi-layer perceptron (MLP) networks. The supervised learning is based on the gradient descent method, minimizing the global error on the output layer. The learning algorithm is performed in two stages [27]: feed-forward and feed-backward. In the first phase the inputs are propagated through the layers of processing elements, generating an output pattern in response to the input pattern presented. In the second phase, the errors calculated in the output layer are then back propagated to the hidden layers where the synaptic weights are updated to reduce the error. This learning process is repeated until the output error value, for all patterns in the training set, are below a specified value. The Back Propagation, however, has two major limitations: a very long training process, with problems such as local minima and network paralysis; and the restriction of learning only static input-output mappings [27].

Fifteen values so obtained in the section IV B) are given as an input to the neural network. Model uses an input layer, two hidden layer with 15 and 7 neurons and an output layer.

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this paper Neural Network model is constructed for JAFFE Face Database for frontal view facial images. Fig. 4 shows the results of facial feature extraction. Initially Face portion segmentation is done using morphological image processing operation and hence face localization is achieved. Region of interest is cropped using localized face and then this image is resized to larger size so that facial feature components should appear prominent. SUSAN edge detection operator along with noise filtering operation is applied to locate the edges of various face feature segment components. Multiple facial feature candidates after applying our algorithm steps 1 and 2 are shown in Fig.4 b).

Cropped Facial image is divided into two regions based on the centre of an image and location of permanent facial feature. The step 3 and 4 are applied to facial segments and results are shown in Fig. 4 (c-i). Fig 4 j) shows localized permanent facial features which are used as input to neural networks.

Training Phase: In this work, supervised learning is used to train the back propagation neural network. The training samples are taken from the JAFFE database. This work has considered 120 training samples for all expressions. After getting the samples, supervised learning is used to train the network. It is trained three times and shown good response in reduction of error signal.

Testing Phase: This proposed system is tested with JAFFE database. It was taken totally 30 sample images for all of the facial expressions.

The Fig. 5 shows the GUI for displaying the results of face localization, extracted permanent facial features with bounding box and its recognition and classification. In this figure, the sample image exhibits sad(SA) expression. Performance plot of neural network is shown in Fig. 6. Plot shows that the network learns gradually and reaches towards the goal.

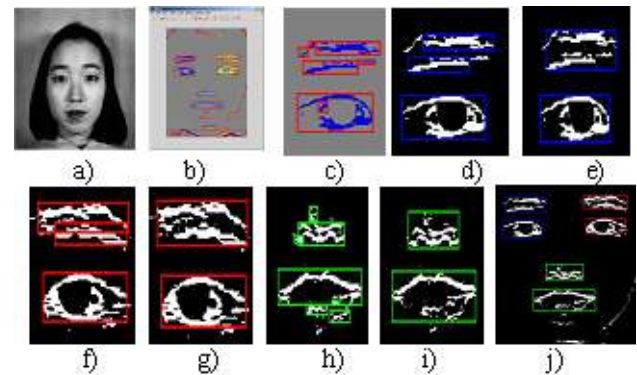


Fig. 4 Results of facial feature extraction a) Original Image b) Multiple facial feature candidates after applying our algorithm step1,2,3 and 4 c) Possible Feature candidates of upper left portion with overlapped eyebrow segments. d) Eye and eyebrow segments after applying step 6 e) Required feature segments of upper left portion f) Possible Feature candidates of upper right portion with overlapped eyebrow segments. g) Required Eye and eyebrow segments after applying step 6 h) Possible Feature candidates of lower portion with overlapped nose and mouth segments. i) Required nose and mouth segments after applying step 6 j) Located Facial features



Fig. 5 GUI for classification and recognition of facial

VI. CONCLUSION AND FUTURE SCOPE

In this paper, automatic facial expression recognition (AFER) system is proposed. Machine recognition of facial expression is a big challenge even if human being recognizes it without any significant delay. The combination of SUSAN edge detector, edge projection analysis and facial geometry distance measure is best combination to locate and extract the facial feature for gray scale images in constrained environments and feed forward back-propagation neural network is used to recognize the facial expression. 100% accuracy is achieved for training sets and 95.26% accuracy is achieved for test sets of JAFFE database which is promising. Table 1 presents the % recognition accuracy of facial expression which appears in literature [28] and our approach. Proposed combination method for feature extraction does not extract exactly six features parameters properly if there are hairs on face area. Therefore in future an attempt can be made to develop hybrid approach for facial feature extraction and recognition accuracy can be further improved using same NN approach and hybrid approach such as ANFIS. An attempt can also be made for recognition of other database images or images captured from camera.

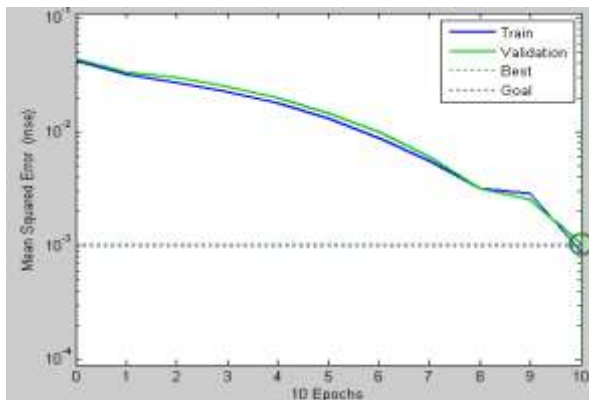


Fig. 4 Performance plot of neural network

TABLE I
% RECOGNITION ACCURACY

Authors	No. of subjects Used	Images Tested	% accuracy
Kobayashi and Hara[28]	15	90	85
Zhang[12]	10	213	90.1
Lyons et. al.[28]	10	193	92
Sebe et. al.[28]	-	-	85-95
Kulkarni SS et. al.[28]	62	282	90.4
Chang JY,Chen JL [29]	08	38	92.1(for 3 expressions)
Our approach	10	30	96.42

REFERENCES

- [1] Mehrabian.A, 1968. "Communication without Words", Psychology Today, Vol.2, No.4, pp 53-56.
- [2] Daw-tung lin. "Facial Expression Classification Using PCA and Hierarchical Radial Basis Function Network", in Journal of Information Science and Engineering, Vol . 22, pp 1033-1046, 2006.
- [3] Hong-Bo Deng, Lian-Wen Jin, Li-Xin Zhen, Jian-Cheng Huang,. "A New Facial Expression Recognition Method Based on Local Gabor Filter Bank and PCA plus LDA", in International Journal of Information Technology Vol. 11 No. 11, 2005.
- [4] Aleksic.P.S. and A. K. Katsaggelos. "Automatic Facial Expression Recognition Using Facial Animation Parameters and Multi-Stream HMMs," 6th European Workshop on Image Analysis for Multimedia Interactive Services , Montreux, Switzerland, 2005.
- [5] Limin Ma, David Chelberg and Mehmet Celenk. "Spatio-Temporal Modeling of Facial Expressions using Gabor-Wavelets and Hierarchical Hidden Markov Models" in the Proc.of ICIP 2005, pp-57-60, 2005.
- [6] Pantic.M. and Ioannis Patras. "Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments from Face Profile Image Sequences", IEEE transactions on Systems, Man, and Cybernetics—Part B: cybernetics, vol. 36, no. 2, 2006.
- [7] Ruicong Zhi, Qiuqi Ruan, "A Comparative Study on Region-Based Moments for Facial Expression Recognition," in Congress on Image and Signal Processing, Vol. 2, pp.600-604, 2008.
- [8] Irene Kotsia, Ioannis Pitas, " Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines" in IEEE Transactions on Image Processing 16(1): pp. 172-187, 2007.
- [9] Kakumanu.P., Nikolaos G. Bourbakis, "A Local-Global Graph Approach for Facial Expression recognition", ICTAI, pp 685-692,2006.
- [10] Aleksic. P.S., Aggelos K. Katsaggelos. "Automatic facial expression recognition using facial animation parameters and multistream HMMs".IEEE Transactions on Information Forensics and Security 1(1): pp. 3-11,2006.
- [11] Spiros Ioannou, George Caridakis, Kostas Karpouzis, and Stefanos Kollias, "Robust Feature Detection for Facial Expression Recognition",EURASIP Journal on Image and Video Processing, vol. 2007, Article ID 29081, 2007.
- [12] Pantic.M and L. J. M. Rothkrantz. "Automatic analysis of facial expressions: the state of the art," IEEE Trans. on PAMI, vol. 22, no. 12,pp. 1424-1445, 2000
- [13] Fasel.B and J. Luetttin. "Automatic facial expression analysis: A survey", Pattern Recognition, vol. 36, pp. 259-275, 2003.
- [14] <http://www.columbia.edu/~vb2266/files/FaceExpressionSurvey-Vinay.pdf>
- [15] Xiaoyi Feng, Baohua Lv, Zhen Li, Jiling Zhang, "A Novel Feature Extraction Method for Facial expression Recognition."
- [16] Mauricio Hess, Geovanni Martinez, "Facial Feature Extraction based on Smallest Univalued Assimilating Nucleus (SUSAN) Algorithm"
- [17] Gengtao Zhou, Yongzhao Zhan, Jianming Zhang, "Facial Expression Recognition Based on Selective Feature Extraction", Proceedings of the sixth International Conference on Intelligent System Design and Applications (ISDA'06) 2006 IEEE
- [18] Jun Ou, Xiao-Bo Bai, Yun Pei, Liang Ma, Wei Liu, "Automatic facial expression recognition using gabor filter and expression analysis.", 2010 Second International Conference on Computer Modeling and Simulation , 2010 IEEE, pp 215-218
- [19] Md. Zia Uddin, J. J. Lee, and Y. -S. Kim, "An enhanced independent component-based human facial expression recognition from video.", IEEE Transactions on Consumer Electronics, Vol 55, No. 4. November 2009, pp. 2216-2224.

- [20] V.Gomathi, Dr. K. Ramar, and A. Santhiyaku Jeevakumar, “ A Neuro fuzzy approach for facial expression recognition using LBP Histograms”, International Journal of Computer Theory and Engineering, Vol 2, No. 2 ,April 2010, 1793-8201, pp 245-249.
- [21] G. R. S. Murthy, R. S. Jadon, “Effectiveness of Eigenspaces for facial expression recognition”, International Journal of Computer Theory and Engineering Vol. 1, No. 5, December 2009 ,1793-8201, pp. 638-642
- [22] Hadi Seyedarabi, Ali Aghagolzadeh , Soharb Khanmihammadi, “Recognition of six basic facial expression By feature-points tracking using RBF Neural network and fuzzy inference system”, 2004 IEEE International Conference on Multimedia and Expo (ICME), pp 1219-1222
- [23] Nectarios Rose, “Facial Expression Classification using Gabor and Log-Gabor Filters”, Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR’06) IEEE computer society
- [24] Junhua Li , Li Peng, “Feature Difference Matrix and QNNs for Facial Expression Recognition.”, 2008 Chinese Control and Decision Conference (CCDC 2008), pp 3445-3449
- [25] S.P. Khandait, P.D. Khandait and Dr.R.C.Thool, “An Efficient approach to Facial Feature Detection for Expression Recognition”,International Journal of Recent Trends in Engineering, Vol 2, No. 1, November 2009,PP. 179-182
- [26] Hua Gu Guangda Su Cheng Du, “ Feature Points Extraction from Faces Image and Vision Computing NZ, Palmerston North, November 2003 , pp.154-158
- [27] HAYKIN , S., “*Neural Networks: A Comprehensive Foundation*,” Prentice Hall, Upper Saddle River, NJ,1999.
- [28] Kulkarni SS, Reddy NP, Hariharan SI, “ Facial Expression (Mood) recognition from facial images using Committee Neural Networks”, Biomedical Engineering online 2009,8:16
- [29] Chang J.Y, Chen J.L , “Automated Facial Expression Recognition System Using Neural Networks”, Journal of Chinese Institute of Engineers, vol. 24,No. 3 , pp. 345-356(2001)

AUTHORS PROFILE

Dr. R.C. Thool is Professor in Deptt. of Information Technology, SGGSIET, SRTMNU, Nanded. His area of interest is computer vision, robot vision and image processing and pattern recognition.

Mrs. S.P. Khandait is a research scholar and Assistant professor in the department of Information Technology, KDKCE, RSTMNU, Nagpur. Presently she is pursuing her PhD in CSE. Her research interest is Image processing and computer vision.

P. D. Khandait is an Assistant professor in the department of Electronics Engineering , KDKCE, RSTMNU, Nagpur. His area of research interest is Signal and Image processing, soft computing etc..

Coalesced Quality Management System

A Collaborative Service Delivery Model

A. Pathanjali Sastri

Lecturer, Department of Computer Application,
V.R.Siddhartha Engineering College,
Kanuru Vijayawada – 520 007, Andhra Pradesh, India.
akellapatanjali@yahoo.com

K. Nageswara Rao

Professor & Head, Department of Computer Science and
Engineering,
P.V.P.Siddhartha Institute of Technology,
Kanuru, Vijayawada – 520 007, Andhra Pradesh, India
drknrao@ieee.org

Abstract— Developing software in a stipulated time frame and within a budget is not good enough if the product developed is full of bugs and today end users are demanding higher quality software than ever before. Project lifecycle starts with pre-Project work all the way through to post-Project. Projects need to be set up correctly from the beginning to ensure success. As the software market matures, users want to be assured of quality. In order to check such unpleasant incidents or potential problems lurking around the corner for software development teams, we need a quality framework, that not only to assess the common challenges that are likely to come in the way while executing the projects but also to focus on winning the deal during proposal stage.

Our research paper is an honest appraisal of the reasons behind the failure of projects and an attempt to address valuable pointers for the successes of future projects. “Coalesced Quality Management Framework (CQMF)” is a theoretical model to bring the best of Quality to the work products developed and to gain the firsthand knowledge of all the projects, defects, and quality metrics and report to the Management so that missed deadlines and enhancement of budget are avoided providing an opportunity to deliver the end product to the satisfaction of the customer. With this framework the project stakeholders and the management constantly validate what is built and verify how it is being built.

Keywords- *Quality Assurance, Operational Excellence, Coalesced Quality Management System, Business Analyst, phase gate reviews*

I. INTRODUCTION

There are many studies attempting to quantify the cost of software failures. They don't agree on percentages but they generally agree that the number is at least 50 to 80 billion dollar range annually [1]. Implementing a complex new product may be the most difficult and risky effort an organization is facing today. An experienced project manager is identified for a heavy budget project and is made responsible for delivering on the contractual commitments. But they have to deliver a solution that meets customer's objectives and for this, the project team should have experience and capability to ensure that the project will succeed. If the staff has limited experience in the required technologies, methodologies and in management, failure is not an option. Poor quality management

can lead to rework, customer dissatisfaction, higher costs, and missed deadlines. There is a need to apply an effective framework that must be prudently applied in order to identify, anticipate and address these risks before they cripple the project. With all that we have on the line, there must be always an unbiased advice on the potential risks to the project and to the organization. The organizations need an effective governance to standardize on a framework comprising process, tools, and resources (experts) that would help them save time and reduce product/process failures in order to bring in the results that may likely be in the best interests of the organization.

This paper proposes a CQMF model that recommends that by implementing CQMS model and adapting effective people practices i.e. involving appropriate stakeholders in each stage of the life cycle, a project will definitely be executed towards the success. For such, the paper is organized as follows. Section II briefly surveys the reasons for failure of projects; Section III describes the existing quality processes in organizations; Section IV discusses the limitations of the present quality structures in organizations; Section V describes about the importance of effective people practices; and Sections VI, VII, VIII, and IX describe the proposed approaches.

II. A SURVEY OF REASONS FOR FAILURE OF PROJECTS

Our research work started with good amount of literature survey through Internet and journal articles on project failure [1][4][10][11][14]. Besides, some informal interviews were also conducted with few program managers and project managers working in top notch companies. Data were collected by interviewing project managers and consultants using a semi-structured interview schedule. The errors are not actual code defects but are considered to be the errors or failures of governance, committed by a governing body like a steering committee, errors and failures of project management and errors and failures of software engineering. The samples of seven large IT projects were examined from four perspectives: Governance, Business case, Organizational capacity, and Project management as shown in table 1.

TABLE 1 ANALYSIS OF FAILURES TO DETERMINE ROOT CAUSES OF POOR OUTCOME

S.No	Parameter	Explanation of the issues	No. of projects (Out of seven projects)
1	Governance	<ol style="list-style-type: none"> Varied widely from project to project Governance responsibilities were not carried out adequately because the processes used to approve and manage large IT projects did not increase the project's likelihood of success. All the projects experienced lack of scrutiny at project conception and initiation, and was eventually proven to be fundamentally unwise. 	Seven
2	Business Case	Projects looked at were allowed to proceed with a business case that was incomplete and the projects and the steering team did not clearly define the business needs it expected the project to meet.	Six
3	Organizational capacity	Projects undertaken lacked the appropriate skills and experience to manage the large IT projects.	Five
4	Project Management	<ol style="list-style-type: none"> Quality of project management ranged from good to poor but in two cases, poor project management led to long delays and large cost overruns. The PM did not follow accepted best practices in managing the project. 	Five

Worth noting is that most organizations that experienced the software failures have not attempted to study the root causes of the failures. Unfortunately, most organizations don't see preventing failure as an urgent matter, even though that view risks harming and maybe even destroying the organization [5]. Understanding why this attitude persists is not

just an academic exercise; it has tremendous implications for business and society [9]. The ways in which an organization develops software can be viewed as a system within which the organization prevents defects in it, output through different methods of prevention, detection and removal. Figure 1 depicts a defect prevention model for any organization.

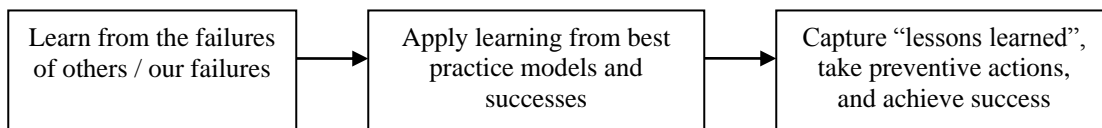


Figure 1 Defect Prevention Model

III. EXISTING QUALITY ASSURANCE PROCESS

Software Quality Assurance is the planned and systematic set of activities that ensures that software life cycle processes and products conform to requirements, standards, and procedures [IEEE 610.12]. Full life cycle Software Assurance activities provide independent and objective assessments of the processes and quality of the product. Figure 2 shows the activities performed by quality group throughout project life cycle. The Quality Management System (QMS) consists of detailed checklists, standards, templates and guidelines exist within the processes to bring in rigor and predictability into every aspect of project planning and execution. These processes and templates are maintained in a centralized repository and are made available across various types of projects (testing, conversion, maintenance, development, package implementation, etc) within the organization which are used in every aspect of the project (requirements analysis, design, change/ configuration management, tailoring, defect or schedule estimation, etc). Software Quality Assurance (SQA) team maintains and enhances the Quality Management System repository and Knowledge Database, based on the experiences gained from the project implementation and bench marking

against international practices. The knowledge database consists of project metrics database, process metrics database and a process-capability baseline. Project leaders use these to estimate effort, schedule tasks and predict defect levels during the project-planning phase. The process database is based on data from past projects and ensures that project plans are realistic. Further, project monitoring based on these metrics increases its effectiveness. The role of SQA team is to create the process-oriented mind-set within the organization and always stick to its commitment and help/facilitate projects to consistently deliver quality software solutions on time by conforming to existing quality standards and monitoring the work products for conformance to standards and processes.

IV. LIMITATIONS WITH THE PRESENT QUALITY MANAGEMENT STRUCTURE

Quality Assurance activity will be considered as a no value-added function if it does not focus on opportunities for early error detection, problem prevention, and risk identification and mitigation and earlier detection and identification yields fewer costs to fix and less schedule impact. By adhering to comprehensive quality management system and quality assurance processes, it could be possible to leverage project

experiences and learning to bring about predictability in processes and continuity & sustainability in quality if this system is controlled and run by the right people. Most of the times the SQA/SQC activities end up in corrective action by analyzing the origin of the defects, examining the SDLC to

determine where the defect was introduced (Requirements or Design or Coding) and reviewing these with the project managers and other associates for possible improvements. This type of approach will not work for large projects or critical projects.

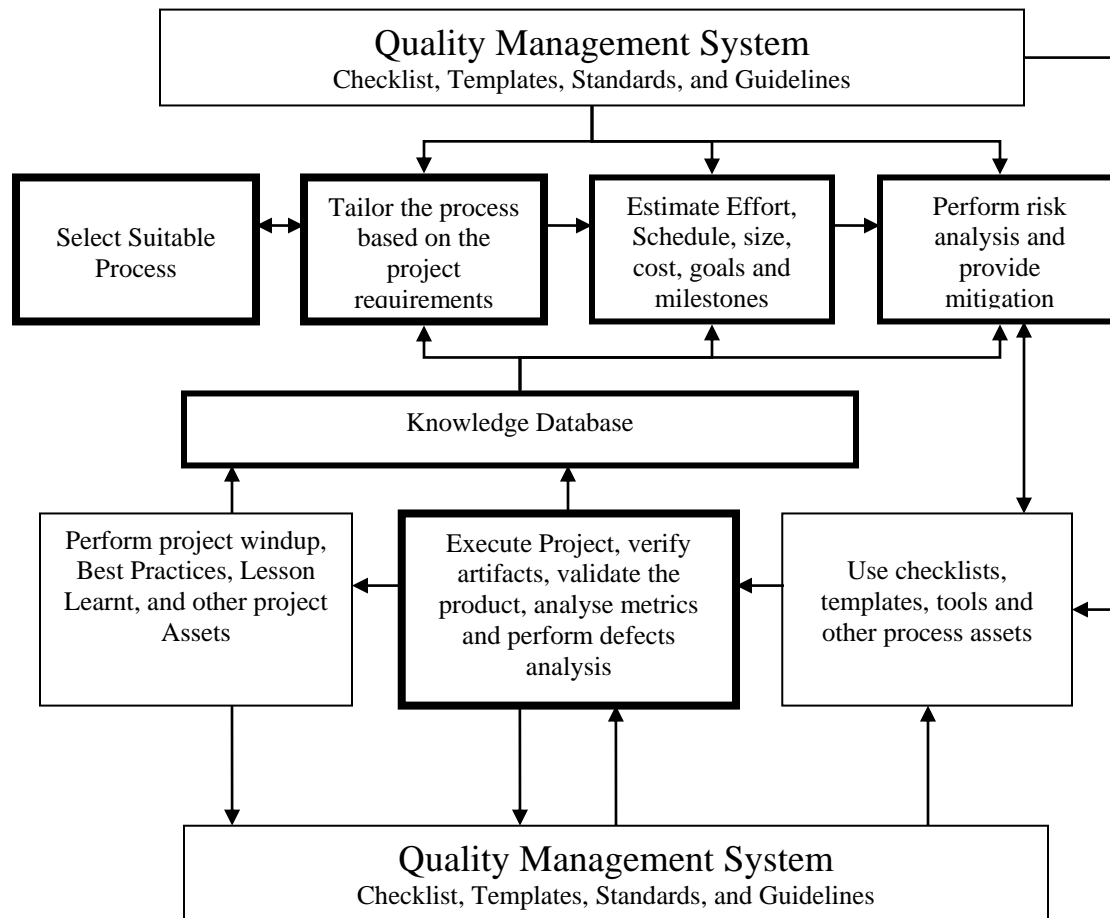


Figure 2 Typical QMS and performed activities by the system

Suitable development processes have a considerable influence or in improving software quality and models such as ISO or CMMI have been deployed for a long time for improving development processes. But it seems that these efforts remain almost fruitless when we look into the reasons of failure of projects. However, it must also be noted that, good processes may well ensure better products, but good processes alone are by far no guarantee for perfect products. It is thus absolutely essential to take the time to work on a suitable development process in an iterative way together with those involved and accompanied by experienced people.

V. OPERATIONAL EXCELLENCE THROUGH EFFECTIVE PEOPLE PRACTICES

Figure 3 shows the success of achieving the goals largely depends on the people in the organization and key to constant focus on operational excellence emerges from the assignment of right set of people with required competencies [8].

Quality Result = Functional Quality + Quality of Reliability + Quality of Cost + Quality of Delivery Schedule

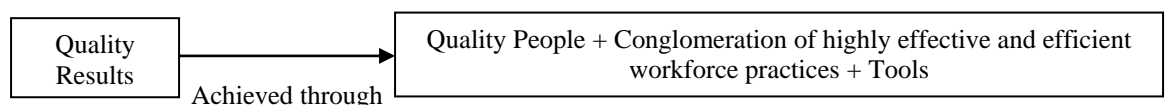


Figure 3 Framework to drive operational excellence

VI. COALESCED QUALITY MANAGEMENT SYSTEM (CQMS) – A COLLABORATIVE SERVICE DELIVERY MODEL

Many systems were developed with functionality in mind rather than with operations or usability in mind. According to the man who invented management, Peter F. Drucker, “Efficiency is doing things right, Effectiveness is doing right things”. The proposed framework shown in figure 4 is a Coalesced Quality Management System (CQMS) that defines a Collaborative Service Delivery Model and serves as organisation’s “eyes and ears” into the inner workings of the project which is independent of the project team and the customer. It provide insight into all aspects of the project: requirements management, adherence to the schedule and budget, project governance, technical architectures and change management with the involvement of IV&V pool, subject matter experts, groups like portfolio management office, quality group, and process/product/tool repository.

The model shown in figure 4 will address the issues shown in table 1, with earlier detection & Prevention of errors either in the product or in process thus bringing down the total cost of quality. This model assures that reviews are conducted by experienced, qualified, and dispassionate experts and projects are on track before proceeding to the next gate as shown in the figure 5. Viewing defects in released product is not the desired

one and there should be defenses in depth to prevent defects. So the activities of QMS shown as dark boxes in figure 2 need to be performed and assessed by the people from expert pool (shown as a dark box in figure 4).

VII. JUSTIFICATION OF THE CQMS MODEL

According to new research, success in 68 percent of technology projects is “improbable”. Poor requirements analysis causes many of these failures, meaning projects are doomed right from the start [2]. In many organizations, requirements are not detailed enough to enable project to make needed changes and get to the end goal reliably. Key to the success of the design and development of a software system like bio-informatics software or an e-security solution mostly depends on the obtaining requirements from different people with expertise in different competencies. E.g., developing bio-informatics software requires the services of a Biology and/or bio-chemistry expert to provide domain related requirements and test the software, a mathematician to provide guidance towards optimal algorithms, few computer scientists to develop the code etc. Using this model we intend to reduce the number of iterations in the process of creating the final product by involving expert pool in conception, initiation, requirements, design and testing phases of the life-cycle.

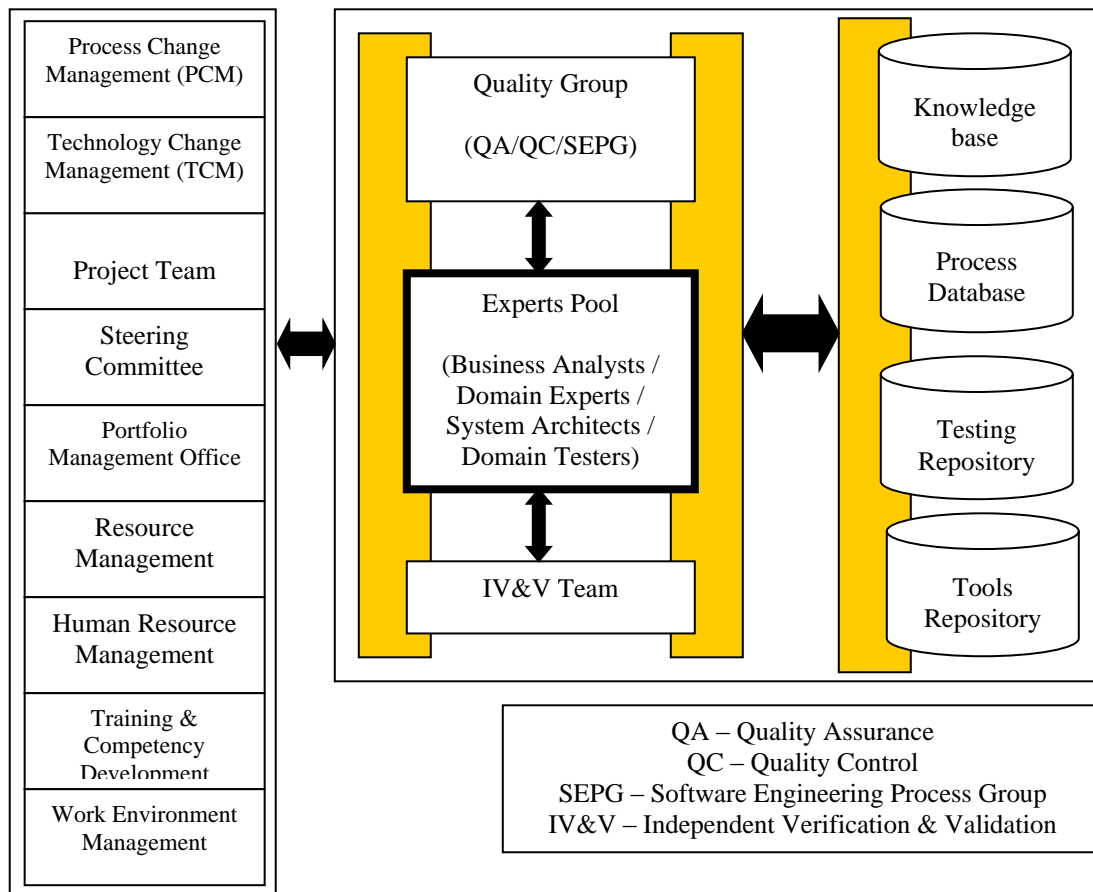


Figure 4 Coalesced Quality Management System (CQMS)

As the operational excellence is defined as a goal of executing projects in a manner that improves timely delivery and quality of deliverables while reducing the rework, the CQMS addresses the problems like the project conception that results in unwise approaches because of (a) inadequate analyses of business issues because the participants lack the necessary qualifications or experience and (b) ineffective review programs performed in unsupportive project environments (barriers to success of the projects) that contribute more towards the failure of the large projects by primarily focusing on the customer needs and optimization of people, assets, and processes.

VIII. INVOLVEMENT OF EXPERTS DURING PROPOSAL STAGE, REQUIREMENTS AND TESTING PHASES

It has been observed industry wide that in many cases, the root cause of high cost and schedule overruns and poor quality of the deliverables lie in the proposal stage [7]. Also the projects surveyed seem to be not utilizing sufficient business analysis skill to consistently bring projects in on time and budget [6]. New study from IAG consulting finds companies with poor requirements spend on average \$2.24m more per project [14]. Many project teams contain domain generalists who learn just enough domain information required for the project or sometimes developers are moved to business analysis role and this would prevent them from working based on anticipation. By understanding the business we don't mean having an in-depth knowledge of how business operates and the projects need to involve a deep domain expert/Business Analyst to perform business process analysis, requirements specification and outline design, acceptance testing and system implementation work. With inadequate, inappropriate or inaccurate requirements as a major contributor to project overruns and failure, the role of a skilled Business Analyst in a project team has become more critical than ever [3].

The expected skill set of Business Analyst [9]:

Hard Skills: (a) Requirements Elicitation (Investigate, Analyse, and Specify)
(b) Business Systems Modeling (Process, Data, and Business rules)
Soft Skills: Analysis, Creative thinking, Interviewing, Presentation and Negotiation.

Curve 1 in figure 5 shows the underestimation of the complexity of the project with inadequate, inappropriate or inaccurate requirements which end up in project overruns and failure. Since a broad experience of business is required during the conception, initiation, requirements and testing phases, the involvement of professional business analyst is more critical than ever during these phases. Curve 2 in figure 5 depicts the decreasing complexity of the project with the involvement of business analyst which compromises on the three major elements on-time delivery, within cost budget and quality of product. The focus of the CQMS is to identify defects as early as possible, when they are easier and more cost effective to

correct. This plan provides a framework of activities that are performed within phases of the project life cycle (figure 5).

The key practices help significantly in this regard are:

- (i) Involvement of Experts (Business Analysts / Domain Experts or Subject Matter Experts) during proposal stage and requirements phase.
- (ii) Involvement of Experts (Business Analysts / Domain Experts or Subject Matter Experts / System Architects) during design phase.
- (iii) Involvement of Experts (Business Analysts / Domain Experts or Subject Matter Experts / Domain Testers) during testing phase.

IX. INVOLVEMENT OF EXPERTS IN INDEPENDENT VERIFICATION VALIDATION (IVV) TEAM TO CONDUCT PHASE GATE REVIEWS

As observed from the failure projects from the literature survey, the organizations struggle to implement an effective review program. The survey analysis[12][13] also shows that there is no visible and sustained commitment to reviews from project manager or quality manager in most of the projects (management problem) and it was also found that holding reviews would be too unpleasant in an environment which is not supportive of quality practices. But we see project managers want to deliver quality products and at the same time they also feel pressure to release products quickly and turn resistance towards inspections due to time shortage for conducting reviews before product delivery. It is also demanded by them for code reviews whenever a project is in trouble. Until now, there has been no set methodology for doing project reviews, no requirement that reviews be conducted under defined circumstances or at specified points, and no precise qualifications required for project reviewers. Having this methodology, phase gate reviews, shown in figure 5 and a pool of reviewers shown in figure 4 (box darkened) provide added value to projects planned or under way. This also helps projects conduct reviews quickly on any aspect of their project provide reducing the iterations of project management reviews during project life cycles.

SQA provides the objective evidence that all life cycle processes have been properly and adequately performed. The IV&V team conducts the phase gate reviews, provides the objective evidence of product compliance with the system's functional requirements and the users' needs. IV&V is conducted parallel to, but separate from, the system software development activities. IV&V apart from doing assessment activities like reviewing, analyzing, testing, and monitoring should act as technical advisors to help a project manager oversee a project by handling the project's schedule, budget, deliveries, etc. IV&V group is composed of experienced domain experts, in-depth technical expertise with strong communication, management and planning skills for assessment, analysis, evaluation, review, inspection, and testing of software products and processes. In other words IV&V

maintains an overall system perspective and analyzes each activity to ensure that progress continues toward the completion of the system. This analysis includes ensuring that the configuration baseline is established and maintained and this approach will help to maintain system integrity. Both SQA and IV&V will utilize the services of the expert pool, and PCM and TCM teams. It will help projects implement the best practices for ensuring the success of their projects and avoid failure of the projects. This ensures continuous improvement based on the industry's best practices and lessons learned from the reviews performed.

It is suggested that the organization has to focus on:

- Developing a set of criteria for the use of independent project reviews (IVV) — critical assessments of a project conducted by people who are at arm's length from it.

- Selecting independent reviewers from the established pools of qualified reviewers and these experienced reviewers are independent of the oversight functions in project. Phase Gate Review Program.
- The complexity monitoring during the phases of SDLC as shown in figure 5 to achieve a better control and a more robust estimate of the next phase, thus minimizing the project overruns.
- Using Software Reliability Growth Models by designers/developers to estimate the remaining bugs in the system which can be used to take a call on fit to release the product at phase gate 5 as the amount of testing needed to find and eliminate all the bugs is time and cost prohibitive.

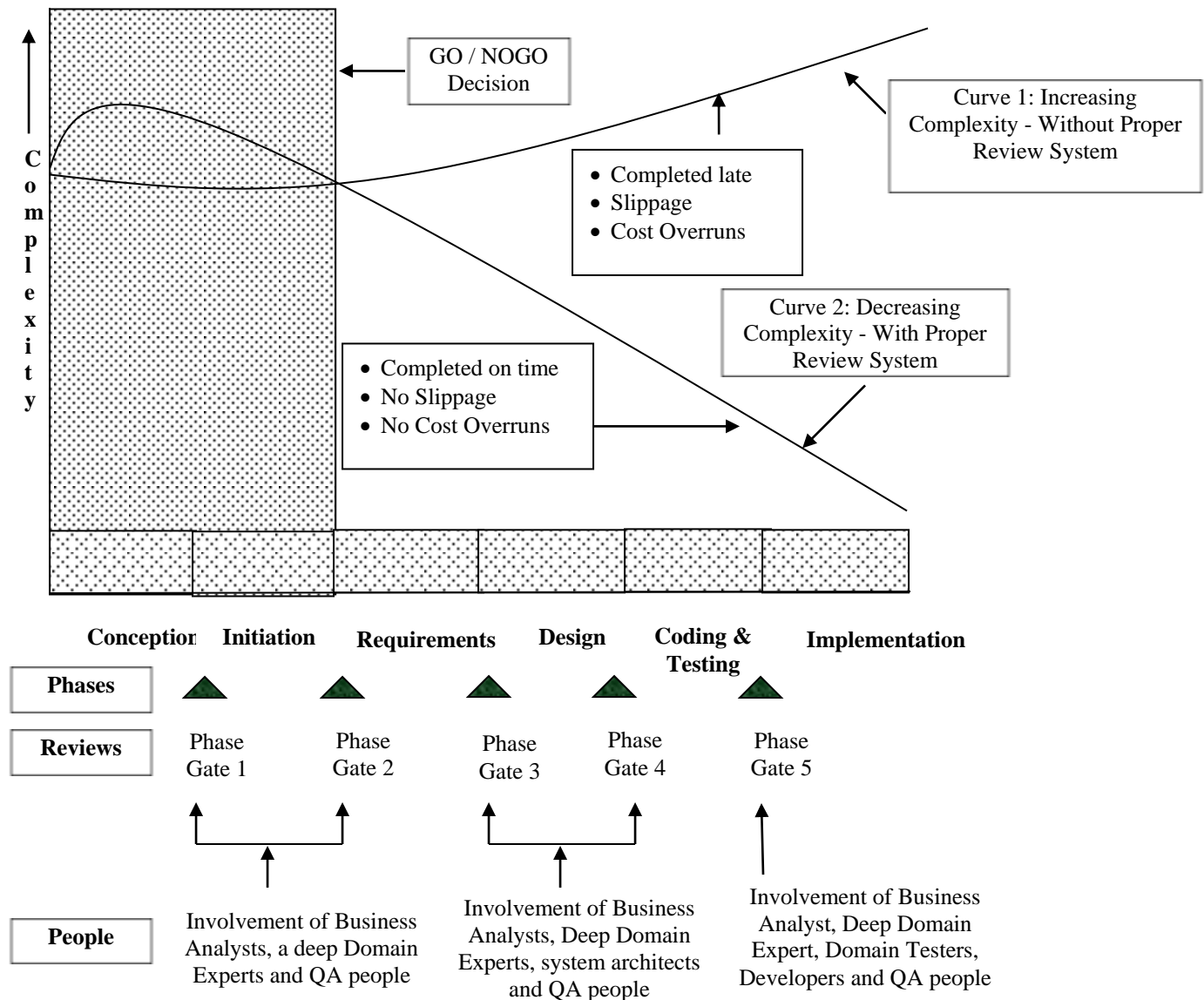


Figure 5: System Complexity and involvement of key resources in various phases and reviews

CONCLUSION

In this paper we have proposed a theoretical model CQMS that stresses on establishing an expert pool to address the risk of failure due to an incorrect understanding of the project's goals. More and more organizations have to realize that they must be able to wire solutions together more quickly to get to market rapidly and be competitive. In a custom development projects, the needs of the customer are relatively static whereas, in a product development situation, the needs of the "marketplace" are constantly changing. To ensure this, the industry must adopt better practices for software development. For this the organizations need better governance and focus more on to standardizing a framework comprising process, tools, and assets that would help them save time and reduce project failures.

By making people aware of the above discussed practices / methods may significantly increase the effectiveness of the implementing process in the project life cycle. The future investigations need to happen in the direction that the existing methodologies (waterfall, spiral, agile etc.) may be critically examined by implementing the proposed framework to suggest improvements to the direction of implementing operational excellence through effective people practices.

REFERENCES

- [1] Vanaja Arvind, Executive Director, Thinksoft Global Services Ltd, Chennai, <http://bit.ly/F4TThinksoft>, Business Line <http://www.thehindu.com/business/Industry/article507513.ece>
- [2] Michael Krigsman, <http://www.zdnetasia.com/study-68-percent-of-it-projects-fail-62049950.htm>
- [3] http://www.irm.com.au/papers/what_is_a_business_analyst.pdf
- [4] Who Killed the Virtual Case File?, <http://www.spectrum.ieee.org/sep05/1455>
- [5] <http://www.newswise.com/articles/view/513919?print-article,10/14/2009>
- [6] http://www.irm.com.au/papers/Golden_Rules_for_Business_analysts.pdf
- [7] Sudha Gopalakrishnan, Vice President and Head-Business Process Excellence Group, Corporate Quality, Polaris Software Labs Ltd, Chennai, "Key Best Practices @ Polaris Software driving Operational Excellence, CSI Communications, Jan-2010, pp12-14
- [8] Sankaran Venkatramani, Associate Director, KPMG, Chennai, "Operational Excellence through effective People Practices", CSI Communications, Jan-2010, pp 20-21.
- [9] Derrick Brown, Jan Kusaik, IRM Training Pty Ltd, CSI Communications, May-2010, pp 7-10
- [10] Patrick Malone, <http://www.tagonline.org/articles.php?id=259>, Project Failure, Jun 9, 2008.
- [11] <http://nclarity.net/data/Documentation/pmo.pdf>
- [12] Labuschagne C, Brent A.C, "Sustainable Project Life Cycle Management: Aligning Project Management Methodologies With The Principles Of Sustainable Development"
- [13] Dave Breda, "Getting Far Beyond Stages and Gates", Pittiglio, Rabin, Todd & McGrath, Inc., 2006
- [14] <http://www.iag.biz>

AUTHORS PROFILE

Mr. A.Pathanjali Sastri is currently pursuing Ph.D Computer Science from Raylaseema University, Kurnool. He is working as a Lecturer in Velagapudi Siddhartha Engineering college since 2008 and has 10 years of Industrial experience prior to this. He has published papers in reputed international conferences recently. He has 12 years of industrial experience and 2 years of teaching experience. His area of interest includes Software Engineering, Quality Assurance, Artificial Intelligence and RDBMS.

Dr. K. Nageswara Rao is currently working as Professor & Head in the Department of Computer Science Engineering, Prasad V. Potluri Siddhartha Institute of Technology, Kanuru, Vijayawada-7. He has an excellent academic and research experience. He has contributed various research papers in the journals, conferences of International/national repute. His area of interest includes Artificial Intelligence, Software Engineering, Robotics, & Datamining.

Detection of Routing Misbehavior in MANETs with 2ACK scheme

Chinmaya Kumar Nayak¹, G K Abani Kumar Dash², Kharabela parida³ and Satyabrata Das⁴

^{1,2,3,4} Department of Computer Science and Engineering, College of Engineering Bhubaneswar,
BPUT, Odisha, INDIA

¹Chinmaya.confidentone@gmail.com, ²abanidash1982@gmail.com, ³kharabelaparida@gmail.com,

⁴satya.das73@gmail.com

Abstract—The Routing misbehavior in MANETs (Mobile Ad Hoc Networks) is considered in this paper. Commonly routing protocols for MANETs [1] are designed based on the assumption that all participating nodes are fully cooperative. Routing protocols for MANETs are based on the assumption which are, all participating nodes are fully cooperative. Node misbehaviors may take place, due to the open structure and scarcely available battery-based energy. One such routing misbehavior is that some nodes will take part in the route discovery and maintenance processes but refuse to forward data packets. In this, we propose the 2ACK [2] scheme that serves as an add-on technique for routing schemes to detect routing misbehavior and to mitigate their effect. The basic idea of the 2ACK scheme is to send two-hop acknowledgment packets in the opposite direction of the routing path. To reduce extra routing overhead, only a few of the received data packets are acknowledged in the 2ACK scheme.

Keywords- MANET; routing in MANETS; misbehavior of nodes in MANETS; credit based scheme; reputation based scheme; the 2ACK scheme; network security.

I. INTRODUCTION

A. MOBILE ADHOC NETWORK

Mobile Ad-hoc networks (MANET) are self-configuring and self-organizing multi hop wireless networks where, the network structure changes dynamically. In a MANET nodes (hosts) communicate with each other via wireless links either directly or relying on other nodes as routers [3]. The nodes in the network not only acts as hosts but also as routers that route data to/from other nodes in network. The operation of MANETs does not depend on preexisting infrastructure or base stations. Network nodes in MANETs can move freely and randomly.

An Example is shown in figure 1. Node A can communicate directly (single hop) [4] with node C, node D and node B. If A wants to communicate with node E, node C must work as an intermediate node for communication between them. That's why the communication between nodes A and E is multi-hop. The operation of MANETs does not depend on preexisting infrastructure or base stations. Network nodes in MANETs can move freely and randomly.

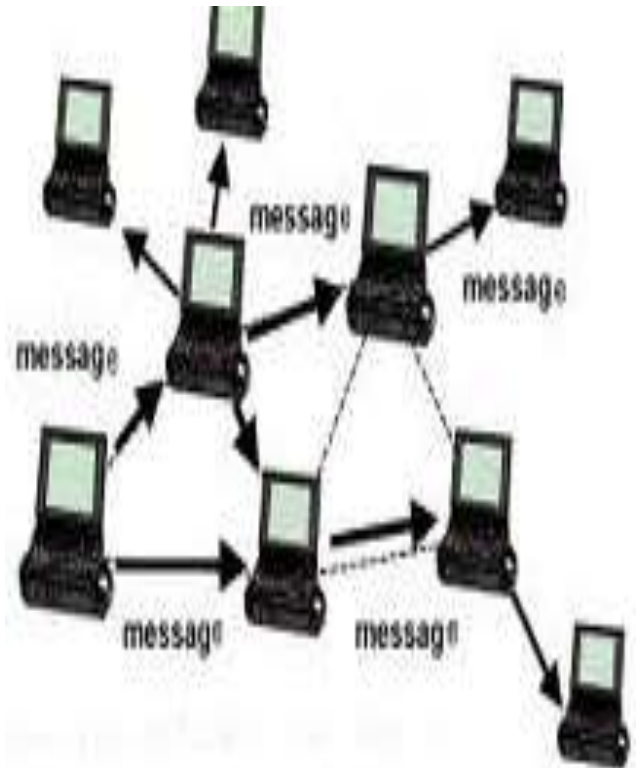


Figure 1: A Mobile ad hoc network

B. CHARACTERISTICS OF MANETS:

- It having the dynamic topology, which links formed and broken with mobility.
- Possibly uni-directional links [4].
- Constrained resources like battery power and wireless transmitter range.
- Network partitions.

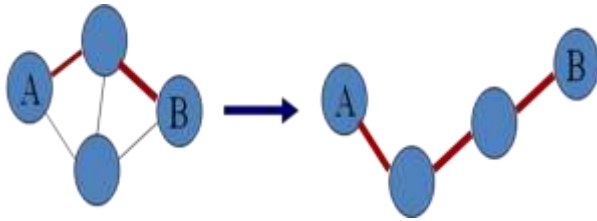


Figure 2: Representation of dynamic topology

C. MANET ROUTING

To find and maintain routes between dynamic topology with possibly uni-directional links, using minimum resources. The use of conventional routing protocols in a dynamic network is not possible because they place a heavy burden on mobile computers and they present convergence characteristics that do not suit well enough the needs of dynamic networks [5]. For Example, any routing scheme in a dynamic environment for instance ad hoc networks must consider that the topology of the network can change while the packet is being routed and that the quality of wireless links is highly variable. The network structure is mostly static in wired networks that are why link failure is not frequent. Therefore, routes in MANET must be calculated much more frequently in order to have the same response level of wired networks. Routing schemes in MANET are classified in four major groups, namely, proactive routing, flooding, reactive routing, and hybrid routing [6].

D. MISBEHAVIOUR OF NODES IN MANET:

Ad hoc networks increase total network throughput by using all available nodes for forwarding and routing. Therefore, the more nodes that take part in packet routing, the greater is the overall bandwidth, the shorter is the routing paths, and the smaller the possibility of a network partition. But, a node may misbehave by agreeing to forward packets and then failing to do so, because it is selfish, overloaded, broken, or malicious [7].

An overloaded node lacks the buffer space, CPU cycles or available network bandwidth to forward packets. A selfish node is unwilling to spend CPU cycles, battery life or available network bandwidth to forward packets not of direct interest to it, even though it expects others to forward packets on its behalf. A malicious node creates a denial of service (DOS) [7] attack by dropping packets. A broken node might have a software problem which prevents it from forwarding packets.

II. PROPOSED MODEL

A. THE 2ACK SCHEME

The main idea of the 2ACK scheme is to send two-hop acknowledgment packets in the opposite direction of the routing path. In order to reduce additional routing overhead, only a fraction of the received data packets are acknowledged

in the 2ACK scheme. Thus it detects the misbehaving nodes, eliminate them and choose the other path for transmitting the data. The watchdog detection mechanism has a very low overhead. Unfortunately, the watchdog technique suffers from several problems such as ambiguous collisions, receiver collisions, and limited transmission power [8]. The main issue is that the event of successful packet reception can only be accurately determined at the receiver of the next-hop link, but the watchdog technique only monitors the transmission from the sender of the next-hop link.

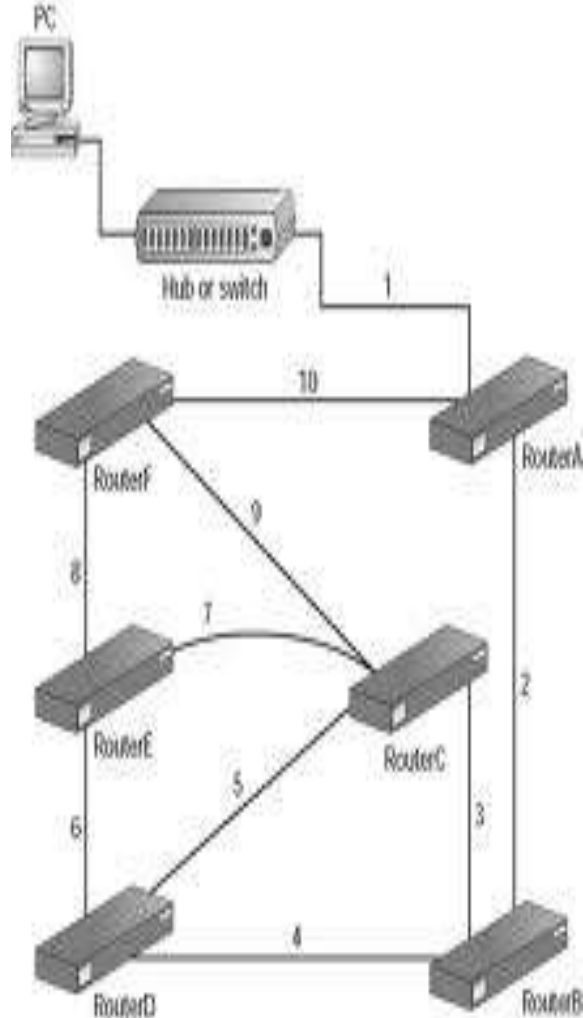


Figure 3: Scenario for packet dropping and misrouting

Noting that a misbehaving node can either be the sender or the receiver of the next-hop link, we focus on the problem of detecting misbehaving links instead of misbehaving nodes. In the next-hop link, a misbehaving sender or a misbehaving receiver has a similar adverse effect on the data packet [8]: It will not be forwarded further. The result is that this link will be tagged. 2ACK scheme significantly simplifies the detection mechanism.

B. DETAILS OF THE 2ACK SCHEME

The 2ACK scheme is a network-layer technique to detect misbehaving links and to mitigate their effects. It can be implemented as an add-on to existing routing protocols for MANETs, such as DSR. The 2ACK scheme detects misbehavior through the use of a new type of acknowledgment packet, termed 2ACK. A 2ACK packet is assigned a fixed route of two hops (three nodes) in the opposite direction of the data traffic route.

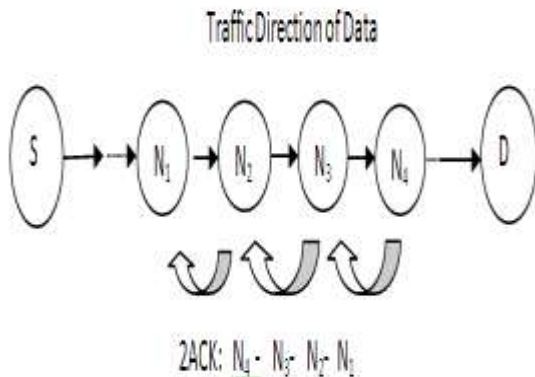


Figure 4: The 2ACK Scheme

Figure 4 illustrates the operation of the 2ACK scheme. Suppose that N1, N2, N3 and N4 are three consecutive nodes (tetra) along a route [9]. The route from a source node, S, to a destination node, D, is generated in the Route Discovery phase of the DSR protocol. When N1 sends a data packet to N2 and N2 forwards it to N3 and so on, it is unclear to N1 whether N3 or N4 receives the data packet successfully or not. Such an ambiguity exists even when there are no misbehaving nodes. The problem becomes much more severe in open MANETs with potential misbehaving nodes.

The 2ACK scheme requires an explicit acknowledgment to be sent by N3 and N4 to notify N1 of its successful reception of a data packet: When node N3 receives the data packet successfully, it sends out a 2ACK packet over two hops to N1 (i.e., the opposite direction of the routing path as shown), with the ID of the corresponding data packet. The triplet $N1 \rightarrow N2 \rightarrow N3 \rightarrow N4$ is derived from the route of the original data traffic.

Such a tetra is used by N1 to monitor the link $N2 \rightarrow N3 \rightarrow N4$. For convenience of presentation, we term N1 in the tetra $N1 \rightarrow N2 \rightarrow N3 \rightarrow N4$ the 2ACK packet receiver or the observing node and N4 the 2ACK packet sender. Such a 2ACK transmission takes place for every set of tetra along the route. Therefore, only the first router from the source will not serve as a 2ACK packet sender. The last router just before the destination and the destination will not serve as 2ACK receivers.

III. APPLICATION

Ad-hoc networks are suited for use in situations where an infrastructure is unavailable or to deploy one is not cost effective.

A mobile ad-hoc network can also be used to provide crisis management services applications, such as in disaster recovery, where the entire communication infrastructure is destroyed and resorting communication quickly is crucial. By using a mobile ad-hoc network, an infrastructure could be set up in hours instead of weeks, as is required in the case of wired line communication. Another application example of a mobile ad-hoc network is Bluetooth, which is designed to support a personal area network by eliminating the need of wires between various devices, such as printers and personal digital assistants. The famous IEEE 802.11 or Wi-Fi protocol also supports an ad-hoc network system in the absence of a wireless access point [9]. Another application example of a mobile ad-hoc network is Bluetooth, which is designed to support a personal area network by eliminating the need of wires between various devices, such as printers and personal digital assistants [10].

IV. ADVANTAGES

As compared to the watchdog, the 2ACK scheme has the following advantages:

1) **Flexibility [9]:** One advantage of the 2ACK scheme is its flexibility to Control overhead with the use of the Rack parameter.

2) **Reliable data Transmission:** It deals with the reliable transfer of file from source to destination. The file needs to be stored at source for certain amount of time even if it has been transmitted. This will help to resend the file if it gets lost during transmission from source to destination.

3) **Reliable route discovery [10]:** Reliable Route Discovery deals with discovering multi-hop route for wireless transmission. Routing in a wireless ad-hoc network is complex. This depends on many factors including finding the routing path, selection of routers, topology, protocol etc.

4) **Limited Overhearing Range [10]:** A well-behaved N3 may use low transmission power to send data toward N4. Due to N1's limited overhearing range, it will not overhear the transmission successfully and will thus infer that N2 is misbehaving, causing a false alarm. Both this problem occurs due to the potential asymmetry between the communication links. The 2ACK scheme is not affected by limited overhearing range problem.

5) **Limited Transmission Power:** A misbehaving N2 may maneuver its transmission power such that N1 can overhear its transmission but N4 cannot. This problem matches with the Receiver Collisions problem. It becomes a threat only when the distance between N1 and N2 is less than that between N2 and N3 and so on. The 2ACK scheme does not suffer from limited transmission power problem.

V. CONCLUSION

The proposed system is a simulation of the algorithm that detects misbehaving links in Mobile Ad Hoc Networks. The 2ACK scheme identifies misbehavior in routing by using a new acknowledgment packet, called 2ACK packet. A 2ACK packet is assigned a fixed route of two hops (four nodes N1, N2, N3, N4), in the opposite direction of the data traffic route. The system implements the 2ACK scheme which helps detect misbehavior by a 3 hop acknowledgement. The 2ACK scheme for detecting routing misbehavior is considered to be network-layer technique for mitigating the routing effects.

REFERENCES

- [1] J. J. Garcia-Luna-Aceves et al: Source Tree Adaptive Routing (STAR) protocol, draft-ietf-manet-star-00.txt, 1998, IETF Internet Draft.
- [2] D. Johnson, D. Maltz, Y.C. Hu, and J. Jetcheva. The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR) in 10th IEEE International Conference, 27-30 Aug 2002, Year of Publication :2002, ICON 2002
- [3] Kong, P. Zerfos, H. Luo, S. Lu, and L. Zhang, "Providing Robust and Ubiquitous Security Support for Mobile Ad-Hoc Networks," Proc. IEEE Int'l Conf. Network Protocols (ICNP '01), 2001.
- [4] L.M. Feeney and M. Nilsson. Investigating the Energy Consumption of a Wireless Network Interface in an Ad Hoc Networking Environment, Proc. IEEE INFOCOM, 2001: Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies, Volume 3 (2001), Pages. 1548-1557, Year of Publication: 2007
- [5] Elizabeth Royer and C-K Toh: A Review of Current Routing Protocols for Ad-Hoc Mobile Wireless Networks. IEEE Personal Communications Magazine, pages 46-55, April 1999.
- [6] Josh Broch , David A. Maltz , David B. Johnson , Yih-Chun Hu , Jorjeta Jetcheva , A performance comparison of multi-hop wireless ad hoc network routing protocols, Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking, p.85-97, October 25-30, 1998, Dallas, Texas, United States [doi>10.1145/288235.288256].
- [7] K. Balakrishnan, J. Deng, and P.K. Varshney. TWOACK: Preventing Selfishness in Mobile Ad Hoc Networks , Proc. IEEE Wireless Comm. and Networking Conf. (WCNC '05), Mar. 2005, Volume 4, Pages 2137-2142, IEEE Press 2005, Year of Publication:2005
- [8] Charles E. Perkins , Pravin Bhagwat, Highly dynamic Destination-Sequenced Distance-Vector routing (DSDV) for mobile computers, ACM SIGCOMM Computer Communication Review, v.24 n.4, p.234-244, Oct. 1994.
- [9] Charles E. Perkins , Elizabeth M. Royer, Ad-hoc On-Demand Distance Vector Routing, Proceedings of the Second IEEE Workshop on Mobile Computer Systems and Applications, p.90, February 25-26, 1999.
- [10] David B. Johnson, David A. Maltz: Dynamic Source Routing (DSR) in Ad Hoc Wireless Networks. In Mobile Computing, edited by Tomasz Imielinski and Hank Korth, chapter 5, pages 153-181. Kluwer Academic Publishers, 1996.

AUTHORS PROFILE



Chinmaya Kumar Nayak is a scholar of M.Tech (CSE) at College of Engineering, Biju Pattanaik University, Bhubaneswar, Odisha, INDIA. He is an author of the book Data structure using 'C'. He published many papers in national seminars. His research areas include Image processing, Image transformation techniques, Adhoc-network etc.

G K Abani Kumar Dash is a scholar of M.Tech (CSE) at College of Engineering, Biju Pattanaik University, Bhubaneswar, Odisha, INDIA. His research areas includes Image processing, Adhoc-network etc.

Kharabela Parida is a scholar of M.Tech (CSE) at College of Engineering, Biju Pattanaik University, Bhubaneswar, Odisha, INDIA. His research areas includes Datamining, Adhoc-network etc.



Satyabrata Das is as Assistant Professor and Head in the department of Computer Sc. & Engineering, College of Engineering Bhubaneswar (CEB) and He is a research scholar in the department of ICT Under F.M University, Balasore. He received his Masters degree from Siksha 'O' Anusandhan University, Bhubaneswar. His research area includes DSP, Soft Computing, Data Mining, Adhoc-network etc. Many publications are there to his credit in many

International and National level journal and proceedings.

Framework for Automatic Development of Type 2 Fuzzy, Neuro and Neuro-Fuzzy Systems

Mr. Jeegar A Trivedi

Department of Computer Science & Technology
Sardar Patel University
Vallabh Vidyanagar, India
jeegar.trivedi@yahoo.com

Dr. Priti Srinivas Sajja

Department of Computer Science & Technology
Sardar Patel University
Vallabh Vidyanagar, India
priti_sajja@yahoo.com

Abstract—This paper presents the design and development of generic framework which aids creation of fuzzy, neural network and neuro fuzzy systems to provide expert advice in various fields. The proposed framework is based on neuro fuzzy hybridization. Artificial neural network of the framework aids learning and fuzzy part helps in providing logical reasoning for making proper decisions based on inference of domain expert's knowledge. Hence by hybridizing neural network and fuzzy logic we obtain advantages of both the fields. Further the framework considered type 2 fuzzy logic for more human like approach. Developing a neuro fuzzy advisory system is tedious and complex task. Much of the time is wasted in developing computational logic and hybridizing the two methodologies. In order to generate a neuro fuzzy advisory system quickly and efficiently, we have designed a generic framework that will generate the advisory system. The resulting advisory system for the given domain is interactive with its user and asks question to generate fuzzy rules. The system also allows provision of training sets for neural network by its users in order to train the neural network. The paper also describes a working prototype implemented based on the designed framework; which can create a fuzzy system, a neural network system or a hybrid neuro fuzzy system according to information provided. The working of the prototype is also discussed with outputs in order to develop a fuzzy system, a neural network system and a hybrid neuro fuzzy system for a domain of course selection advisory. The generated systems through this prototype can be used on web or on desktop as per the user requirement.

Keywords—Artificial Neural Network; Fuzzy Logic; Type 2 Fuzzy Logic; Neuro Fuzzy Hybridization.

I. INTRODUCTION

Neuro fuzzy is an emerging branch of artificial intelligence. Various types of expert and intelligent systems are being developed using this methodology. Neuro fuzzy methodology involves proper hybridization of neural network with fuzzy logic. The fuzzy logic can be further fuzzified into type 2 fuzzy logic; this provides much better approximation and reasoning capability for the derived solution. However developing neuro fuzzy system is complicated and tedious task, because to develop a neuro fuzzy system, one has to develop neural network along with fuzzy rules and hybridize

the two components, also while hybridizing the developer has to take care that the input/output of the neural network are in accordance with the fuzzy rules [2]. Hence developing neuro fuzzy system consumes much time and efforts for training such system appropriately [6]. Hence there is a need to develop a framework which automates development of neuro fuzzy systems. This paper presents the generic design model for development of such neuro fuzzy system. Currently no such system exists and few are there which aids in development of individual areas of neural network and fuzzy logic separately. Software like ANFIS (Adaptive Neuro Fuzzy Inference System) or DENFIS (Dynamic Evolving Neuro Fuzzy Inference System) which are developed using MATLAB uses given input/output data set, the toolbox function `anfis` constructs a fuzzy inference system (FIS) whose membership function parameters are tuned (adjusted) using either a back propagation algorithm alone or in combination with a least squares type of method [1, 16]. This adjustment allows fuzzy systems to learn from the data which are to be modeled. However these two software's lack's the usage of type 2 fuzzy membership function which can be used during fuzzy inference as described by [12, 13]. Hence they are limited to type 1 fuzzy system which is merged with the neural network; however the result or prediction of such system is just derivation of mathematical formulation as type1 fuzzy logic is generalization of crisp logic. Hence this kind of systems lack proper reasoning and produce more mathematical result than human understandable results. To develop a system which takes input in form of layman's reasoning and understanding ability and give output in the same manner as expected by human being, the developer has to take help of type 2 fuzzy logic. Type 2 fuzzy logic interpretation reflects to human intelligence and logical reasoning ability this is demonstrated by [3, 8]. To generate type 2 fuzzy sets and separate rules the developer has to deal with development of type 2 fuzzy member ship function which are also complex mathematical computational tasks. Another important concept about the neuro fuzzy system is to train and keep the neural network in balanced state along with inferring fuzzy rules properly with the trained neural network. Hence while developing neuro fuzzy system manpower is consumed in

developing neuro fuzzy systems, rather than its training and its practical implementation. To avoid such wastage of time in development of neural network and fuzzy membership function, it is advisable to have generic neuro fuzzy development framework which will suit to develop different neuro fuzzy advisory systems in different subject areas and fields. The framework is developed using Microsoft dot net technology. Much of the code is developed using C#.net programming language. The development methodology addresses our problem domain to generate framework for automatic development of neuro fuzzy system by achieving following two objectives:

- The users of the framework have to specify what task to perform on basis of their broad objectives and need not specify how to carry out task.
- The output of the framework is an interactive advisory system which can be used by any novice user.

Chapter II discusses the three methodologies used in the framework. Chapter III discusses neural network system, chapter IV discusses Type 2 fuzzy system and chapter V discusses neuro fuzzy system methodologies with an application example of each of them.

II. METHODOLOGY

The proposed framework allows generation of neuro fuzzy systems; hence to generate system with only one of the two components namely artificial neural network and fuzzy logic is also possible. Hence with our frame work it is possible to generate three kinds of systems which are listed as follows:

- Artificial Neural Network Systems: The system where only training is required to reach the result along with mathematical activation functions are developed under this module.
- Fuzzy Systems: The systems where only reasoning with proper logical condition is required are developed under this module.

- Neuro-Fuzzy System: The hybridization of artificial neural network with fuzzy logic in proper manner, where both training and logical reasoning are required is developed under this module.

The framework is designed with different learning algorithms for artificial neural network which also allows control over different respective learning parameters. The main interface will provide three options to choose development in one of these three categories. The interface of framework is displayed in Fig. 1. There are three buttons respectively for three different systems. Pressing any one of these button will open an interface of the selected button's development approach.

III. NEURAL NETWORK SYSTEM

This module of the framework allows developer to generate and train artificial neural network according to developers own requirements. Artificial neural network is a computational simulation of the biological neurons inside a human brain [14]. It mimics the human mind by using artificial neurons, input hidden & output layers and learning strategy in terms of learning algorithms. The developer of the neural network system has to specify parameter such as input/output variable, number of hidden layers and number of hidden neurons in each layer, selected training methodology namely supervised learning or unsupervised learning or reinforcement learning, specify learning algorithm along with its parameters & finally training set data for the artificial neural network. The developer has to train the network by specifying its learning rate, training loops (iterations), bias values if any and select the activation function as in [7, 18] to transfer input to proper output. Fig. 2 shows sample screen for the neural network module. The trained neural network configuration can be saved and its output can be saved in excel style sheet file.

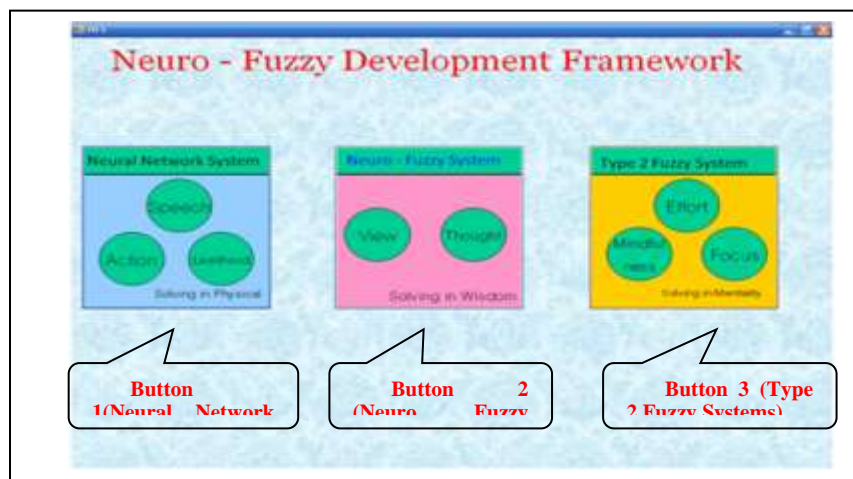


Figure 1. Interface of the Proposed Framework

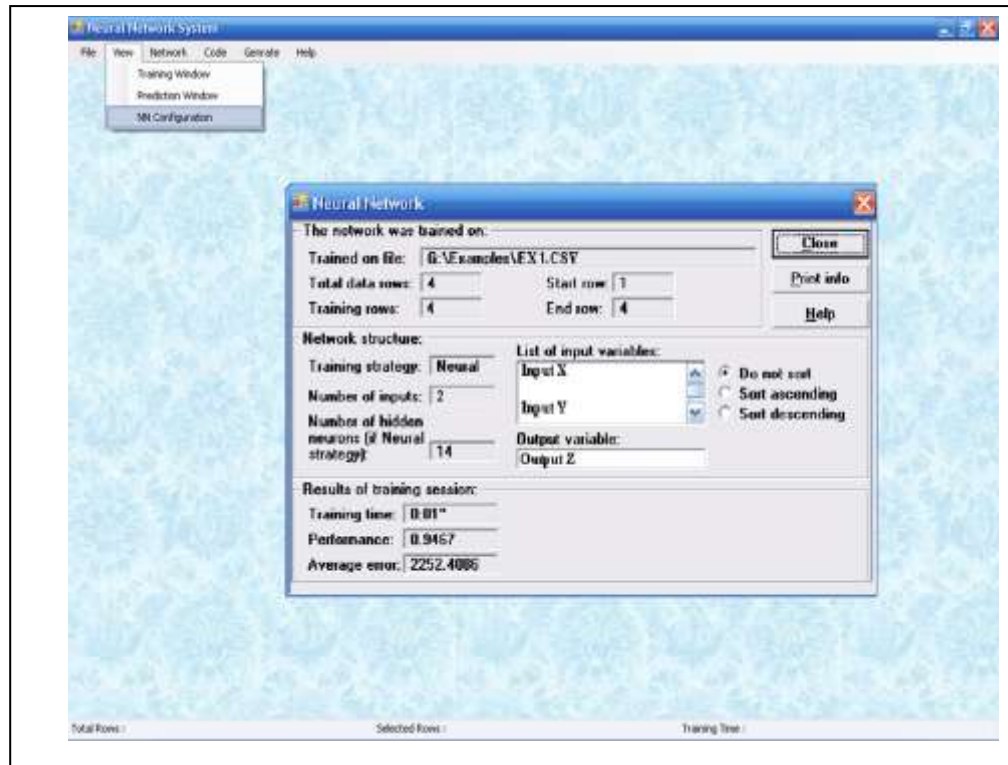


Figure 2. Artificial Neural Network (Configuration).

A. Example 1

An artificial neural network with **eight** input broad categories and with **six** output broad categories with **three** hidden layers of **seven** neurons each having threshold value at 0.6 and learning rate using back propagation algorithm is considered. The training set data are collected through

questionnaires from Department of Computer Science & Technology, Sardar Patel University, which offers different courses at post graduate level and conducts PhD programmes. The department comprise of more than 500 students whose data are collected individually for training neural network. The sample test case data for input and output are shown in Table I & Table II respectively.

TABLE I. DATA VALUES SUPPLIED FOR INPUT BROAD CATEGORIES OF ARTIFICIAL NEURAL NETWORK

Field Name	Percentage	Strategic	Area of Interest	Scholarship	Family Background	Technical Skills	Knowledge Domain	Category
Input Data 1	0.7	0.6	0.3	0.4	0.6	0.6	0.4	0.4
Input Data 2	0.8	0.8	0.4	0.5	0.2	0.6	0.4	0.7

TABLE II. DATA VALUES GENERATED BY OUTPUT BROAD CATEGORIES OF ARTIFICIAL NEURAL NETWORK

Field Name	Further Studies	Foreign Opportunities	Technical	Management	Artistic	Job/Business
Output Data 1	0.46	0.30	0.48	0.41	0.42	0.60
Output Data 2	0.70	0.51	0.48	0.42	0.37	0.75

Selected Categories for Data 1: Job/Business

Selected Categories for Data 2: Job/Business, Further Studies

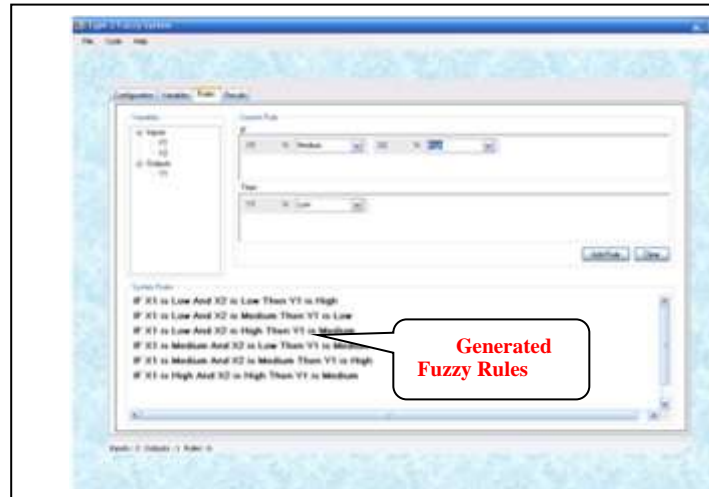


Figure 3. Type 2 Fuzzy System (Rule Addition/Deletion).

IV. TYPE 2 FUZZY SYSTEM

This module of the framework allows developer to generate Fuzzy Inference System with the use of type 2 fuzzy logic. Fuzzy logic presents the concept of linguistic variable as shown by [5]. In order to infer the reasoning behind the fuzzy logic one has to use fuzzy set operations [15]. However the drawback of fuzzy logic is that, it's a generalization of crisp sets. In the process of removing the limitation of fuzzy logic sets, type 2 fuzzy logic sets were introduced. Type 2 fuzzy logic sets provide better understanding and support to human logical reasoning. To use type 2 fuzzy logic sets, various set operations and membership functions are used on type 2 fuzzy logic as shown by [4, 9]. This will infer almost accurate logic behind the current problem domain like a human expert. Type 2 fuzzy logic covers the region called footprint of uncertainty under it [19]. The developer of the type 2 fuzzy system has to specify fuzzy variables and condition. The membership function is chosen according to the application of the system to be developed. The chosen membership function is applied

to generate fuzzy inference for the given problem domain. The architecture of type 2 fuzzy system is shown by [3]. Type 2 fuzzy systems additionally includes type reducer that converts type 2 fuzzy logic to type 1 fuzzy logic, rest of the components in both type 1 and type 2 fuzzy systems work in similar manner. Fig. 3 shows sample screen of type 2 fuzzy systems module.

Inferred rule base is applied to the problem domain to obtain desired result with human understandable logic and reasoning abilities. More the number of rules more accurate are the predictions. If all possible rules are covered then the system will work as human expert generating advice to the respective problem domain.

A. Example 2

Consider a fuzzy system in academic field, then considering percentage & human intelligence as fuzzy variable the conditions generated for student in the situation for deciding to pursue further studies are mentioned in Fig. 4.

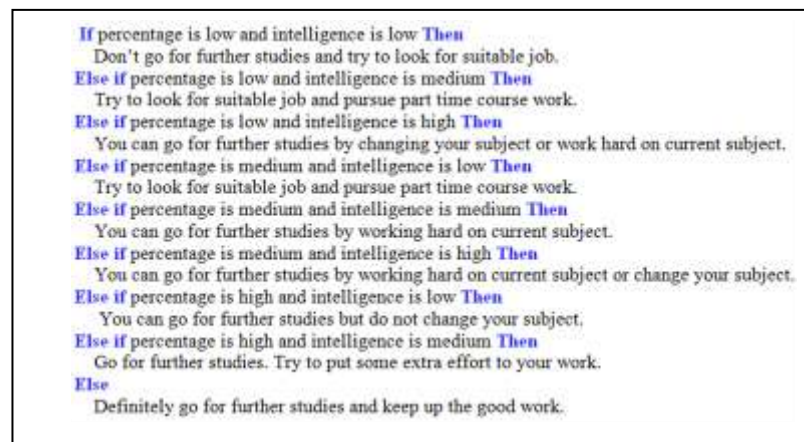


Figure 4. Genrated Fuzzy Rules.

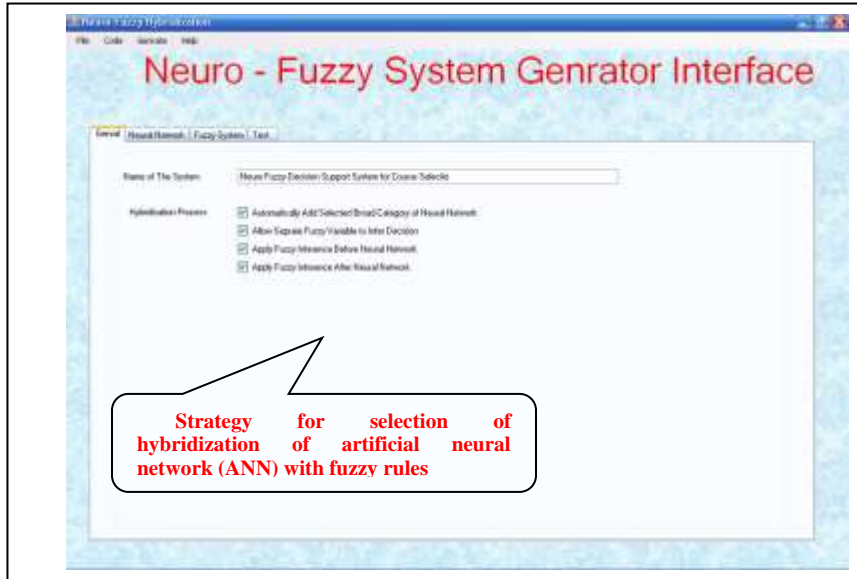


Figure 5. Neuro-Fuzzy System (Hybridization Process).

V. NEURO - FUZZY SYSTEM

This module of the framework allows developer to generate neuro-fuzzy systems. Neuro-fuzzy systems are expert systems based on hybridization of artificial neural network with fuzzy logic. The fuzzy rules are mapped with input/output broad categories of the artificial neural network. The power of machine learning that mimics learning of human brain and rule base generated from various domain experts make neuro-fuzzy system an expert intelligent system that can make decision on its own to guide and solve various day to day problems. Neuro-fuzzy systems are used widely when confusion and dilemma prevails in human mind, they not only

support to overcome the dilemmas but also provide their decision with a strong support of logical reasoning. Thus behaving like an intelligent expert of the given problem domain. Fig. 5 shows sample figure of neuro-fuzzy system development module.

To facilitate the hybridization process, the developer has to select the options which will allow generating fuzzy variables and fuzzy inference before and after the neural network processing. This process will help to decide the architecture of system to be generated. The resulting system will be neuro-fuzzy based decision support system as in [10, 11].

S.No	Percentage	Passing Month	Passing Year	Board/Inst.	Stream
1	80.14	April	2008	GOB	Science
2	82	April	2008	GOB	Science
3	83.33	June	2008	MSU	B.C.A.
4	86	June	2009	DPU	M.C.A.

Figure 6. Educational Detail of User.



Figure 7. Advice Generated by the System.

A. Example 3

Consider a neuro-fuzzy advisory system on course selection that will allow students and their parents to choose the carriers which are most suitable for the student's growth and development. By hybridization of Example 1 & 2, we get a neuro fuzzy decision support system for course selection. The resulting system when displayed in web is seen in Fig. 6 & Fig. 7. In Fig.6 we see an interface to display educational detail of the system user. The user of the system has to fill in appropriate educational detail from the options provided by the system. On the basis of educational details provided by the user as fuzzy input to the system; the system will combine neural network broad categories as mentioned in Example 1 and apply fuzzy rules as shown in Example 2 to generate neuro fuzzy advice to user, which is displayed in Fig.7.

Advantages of the Proposed Framework

- Generates Web based and Workstation based system in Neural Network, Fuzzy and Neuro-fuzzy area with respective problem domain.
- Reusability of the code for new development and modification in current systems.
- Faster development of neuro-fuzzy system without going into computational details of programming.
- The developer will just specify what to use and apply for development of the system, there is no need to specify how to apply.
- Availability of generic library for development. The library of functions and algorithms will also help to develop other major system in the field of computer science.
- It will save time and effort for development of neuro fuzzy systems.

- The developer will be able to concentrate on working of the system rather than making of the system.
- The resulting neuro-fuzzy system generated from the framework will be used by layman to solve their daily dilemma and help own self with proper guidance.
- Many experts knowledge combined with their wisdom and experience will be at finger tips of the user of generated system.
- Expert's knowledge is documented by the system and can be applied even years after it was actually fed to system.
- There is no retirement date of the system, once the system is in stable state the developer can continuously monitor and improve its performance by changing parameters and adding rules to the system.
- This framework is developed using Microsoft's Dot Net technology (Visual Studio 2010) and can be updated easily to future release of versions.

Future Scope of the Proposed Framework

The framework is developed to create and hybridize artificial neural network with fuzzy system for which class libraries have been already developed. In future it is possible to introduce more concept of artificial intelligence like Genetics, ACO, Chaos Theory, Swarm Intelligence etc as in [17, 20] to hybridize with existing methodology to generate wide range of application areas of artificial intelligence.

The source code libraries of the framework will aid in developing various kind of neuro fuzzy application like time series analysis, neuro fuzzy controller for different appliances and many such respective areas of neuro fuzzy development other than neuro fuzzy advisory systems.

VI. CONCLUSION

In this paper we have proposed design of generic framework for automatic development of neuro-fuzzy advisory system. The framework shows development of three kinds of systems namely Neural Network Systems, Fuzzy Systems and Neuro-Fuzzy Systems. We have presented design, development strategy and a application example for each of these systems. We have proved that generic framework for development of neuro-fuzzy system is very important as it has many advantages. The derived solution of the proposed problem domain is explained clearly in Example 3. The resulting advisory system generated from the example is ongoing UGC based major research project for developing neuro fuzzy decision support system for course selection for students and their parents. The generated system acts as an expert having domain in the field of academics and guides students as well as their parents that which courses are most suitable for respective students and justifies the advice generated with proper reasoning and concepts. The proposed generic framework will boost development of neuro fuzzy advisory systems and hence provide common man to take advice of field expert of the respective problem domain at finger tips at any place and at any time.

ACKNOWLEDGMENT

Authors are grateful to the University Grants Commission, New Delhi, India for funding this research work [File No. 36-203/2008(SR)].

REFERENCES

- [1]. Ching Long Su, Chuen Jyh Chen, and Shih Ming Yang, "A self-organized neuro-fuzzy system for stock market dynamics modeling and forecasting", WSEAS Transactions on Information Science and Applications, Vol.7, No.9, pp.1137-1149, 2010.
- [2]. J S R Jang, C T Sun and E Mizutani, Neuro-fuzzy Soft Computing, Prentice Hall of India Ltd, pp. 27-33, 1997.
- [3]. Jerry M. Mendel, "Type-2 fuzzy sets and systems: an Overview", IEEE Computational Intelligence Magazine, Vol. 2, pp. 20-29, 2007.
- [4]. JOHN, R.I., and COUPLAND, S, "Type-2 fuzzy logic: A historical view", IEEE Computational Intelligence Magazine, Vol. 2, pp. 57-62, 2007.
- [5]. L.A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning", Information Sciences, vol. 8, pp. 43-80, 1975.
- [6]. M.Tim Jones, Artificial Intelligence Application Programming, dreamtech press, pp. 232-239, 2003.
- [7]. Mehmet KARAKOSE and Erhan AKIN, "Type 2 Fuzzy activation functions for multilayer feed forward neural network", IEEE International Conference on Systems, Man and Cybernetics, pp. 3762-3767, 2004.
- [8]. N. Karnik, J. Mendel, and Q. Liang, "Type-2 Fuzzy Logic Systems", IEEE Trans. On Fuzzy Systems, Vol. 7, No. 6, pp. 643-658, 1999.
- [9]. Oscar Castillo and Patricia Melin, Type 2 Fuzzy Logic: Theory and Application, Springer, pp. 5-28, 2008.
- [10]. Priti Srinivas Sajja, "Fuzzy artificial neural network decision support system for course selection", Journal of Engineering and Technology, vol.19, pp.99-102, 2006.
- [11]. Priti Srinivas Sajja, "Type-2 Fuzzy User Interface for Artificial Neural Network based Decision Support System for Course Selection". International Journal of Computing and ICT Research, Vol. 2, No. 2, pp. 96-102, 2008.
- [12]. Priti Srinivas Sajja, Anbumani K, and Nedunchezian R (Eds.), Soft Computing Applications for Database Technologies: Techniques and Issues, IGI Global Book Publishing, Hershey, PA, USA, pp.72-92, 2010.
- [13]. Priti Srinivas Sajja, and Jeegar A Trivedi, "Using Type-2 Hyperbolic Tangent Activation Function in Artificial Neural Network", Research Lines, Vol. 3, No. 2, pp. 51-57, 2010.
- [14]. Rich and Knight, Artificial Intelligence, Tata McGraw Hill Publishing Co. Ltd. 21st Indian Reprint, pp. 492-495, 2001.
- [15]. S N Sivanandam and S N Deepa, Principles Of Soft Computing. Wiley, pp. 318-322, 2007.
- [16]. S. M. Seyedhoseini, J. Jassbi, and N. Pilevari, "Application of adaptive neuro fuzzy inference system in measurement of supply chain agility: Real case study of a manufacturing company", African Journal of Business Management Vol.4, No.1, pp.83-96, 2010.
- [17]. Shuxiang Xu, "Data Mining Using Higher Order Neural Network Models With Adaptive Neuron Activation Functions", IJACT : International Journal of Advancements in Computing Technology, Vol. 2, No. 4, pp. 168-177, 2010
- [18]. T.Subbulakshmi, S. Mercy Shalinie and A. Ramamoorthi, "Implementation of Artificial Neural Network with Faster Training Strategies for Classification of Intrusion", International Journal of Artificial Intelligence and Computational Research, Vol. 2, No. 1, pp.47-51, 2010.
- [19]. WU, H. AND MENDAL, J.M., "Uncertainty bounds and their use in the design of interval type-2 fuzzy logic system", IEEE Transactions on fuzzy systems, Vol. 10, No. 5, pp. 622-639, 2002.
- [20]. Zahra Khanmirzaei, and Mohammad Teshnehlab, "Prediction Using Recurrent Neural Network Based Fuzzy Inference system by the Modified Bees Algorithm", IJACT : International Journal of Advancements in Computing Technology, Vol. 2, No. 2, pp. 42-55, 2010.

AUTHORS PROFILE

- [1]. Mr. Jeegar A Trivedi is working as a Project Fellow in the Department of Computer Science & Technology at Sardar Patel University, India. He is also a full time research student and carrying out research work in the fields of neuro fuzzy systems and type 2 fuzzy systems.
- [2]. Dr. Priti Srinivas Sajja is working as an Associate Professor at the Department of Computer Science, Sardar Patel University, India. Her research interests include knowledge-based systems, soft computing, multiagent systems, and software engineering. She has 80 publications in books, book chapters, journals, and in the proceedings of national and international conferences. Three of her publications have won best research paper awards. She is co-author of 'Knowledge-Based Systems' published by Jones & Bartlett Publishers, USA. She is serving as a member in editorial board of many international science journals and served as program committee member for various international conferences.

A study on Feature Selection Techniques in Bio-Informatics

S.Nirmala Devi

Department of Master of Computer Applications
Guru Nanak College
Chennai, India
csnirmala77@yahoo.co.in

Dr. S.P Rajagopalan

Department of Master of Computer Applications
Dr.M.G.R Educational and Research Institute
Chennai, India
sasirekaraj@yahoo.co.in

Abstract— The availability of massive amounts of experimental data based on genome-wide studies has given impetus in recent years to a large effort in developing mathematical, statistical and computational techniques to infer biological models from data. In many bioinformatics problems the number of features is significantly larger than the number of samples (high feature to sample ratio datasets) and feature selection techniques have become an apparent need in many bioinformatics applications. This article provides the reader aware of the possibilities of feature selection, providing a basic taxonomy of feature selection techniques, discussing its uses, common and upcoming bioinformatics applications.

Keywords- *Bio-Informatics; Feature Selection; Text Mining; Literature Mining; Wrapper; Filter Embedded Methods.*

I. INTRODUCTION

During the last ten years, the desire and determination for applying feature selection techniques in bioinformatics has shifted from being an illustrative example to becoming a real prerequisite for model building. The high dimensional nature of the modeling tasks in bioinformatics, going from sequence analysis over microarray analysis to spectral analyses and literature mining has given rise to a wealth of feature selection techniques are presented in the field.

The application of feature selection techniques is focused in this article. While comparing with other dimensionality reduction techniques like projection and compression, feature selection techniques do not alter the original representation of the variables, but merely select a subset of the representation. Thus, it preserves the original semantics of the variables and Feature selection is also known as variable selection, feature reduction, attribute selection or variable subset selection.

Feature selection helps to acquire better understanding about the data by telling which the important features are and how they are related with each other and it can be applied to both supervised and unsupervised learning. The interesting topic of feature selection for unsupervised learning (clustering) is a more complex issue, and research into this field is recently getting more attention in several communities and the problem of supervised learning is focused here, where the class labels are known already.

The main aim of this study is to make aware of the necessity and benefits of applying feature selection techniques. It provides an overview of the different feature selection techniques for classification by reviewing the most important application fields in the bioinformatics domain, and the efforts done by the bioinformatics community in developing procedures is highlighted. Finally, this study point to some useful data mining and bioinformatics software packages that can be used for feature selection.

II. FEATURE SELECTION TECHNIQUES

Feature selection is the process of removing features from the data set that are irrelevant with respect to the task that is to be performed. Feature selection can be extremely useful in reducing the dimensionality of the data to be processed by the classifier, reducing execution time and improving predictive accuracy (inclusion of irrelevant features can introduce noise into the data, thus obscuring relevant features). It is worth noting that even though some machine learning algorithms perform some degree of feature selection themselves (such as classification trees); feature space reduction can be useful even for these algorithms. Reducing the dimensionality of the data reduces the size of the hypothesis space and thus results in faster execution time.

As many pattern recognition techniques were originally not designed to cope with large amounts of irrelevant features, combining them with FS techniques has become a necessity in many applications. The objectives of feature selection are

- (a) to avoid over fitting and improve model performance, i.e. prediction performance in the case of supervised classification and better cluster detection in the case of clustering
- (b) to provide faster and more cost-effective models
- (c) to gain a deeper insight into the underlying processes that generated the data.

Instead of just optimizing the parameters of the model for the full feature subset, we now need to find the optimal model parameters for the optimal feature subset [1], as there is no guarantee that the optimal parameters for the full feature set are equally optimal for the optimal feature subset.

There are three types of feature subset selection approaches: depending on how they combine the feature selection search with the construction of the classification model: filters, wrappers and embedded methods which perform the features selection process as an integral part of a machine learning (ML) algorithm. Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. Wrappers can be computationally expensive and have a risk of over fitting to the model. Filters are similar to Wrappers in the search approach, but instead of evaluating against a model, a simpler filter is evaluated. Embedded techniques are embedded in and specific to a model.

A. Filter Methods

These methods do not require the use of a classifier to select the best subset of features. They use general characteristics of the data to evaluate features. Filter techniques use the intrinsic properties of the data to assess the relevance of features. In many cases the low-scoring features are removed and feature relevance score is calculated, then this subset is given as input to the classification algorithm.

They are pre-processing methods. They attempt to assess the merits of features from the data, ignoring the effects of the selected feature subset on the performance of the learning algorithm. Examples are methods that select variables by ranking them through compression techniques or by computing correlation with the output.

Advantages of filter techniques are that they are independent of the classification algorithm, computationally simple and fast and easily scale to very high-dimensional datasets. Feature selection needs to be performed only once, and then different classifiers can be evaluated.

Disadvantages of filter methods is that they ignore the interaction with the classifier i.e., the search in the feature subset space is separated from the search in the hypothesis space. Each feature is considered separately and compared to other types of feature selection techniques it lead to worse classification performance thereby ignoring feature dependencies. A number of multivariate filter techniques were introduced in order to overcome the problem of ignoring feature dependencies.

B. Wrapper methods

These methods assess subsets of variables according to their usefulness to a given predictor. The method conducts a search for a good subset using the learning algorithm itself as part of the evaluation function. The problem boils down to a problem of stochastic state space search. Examples are the stepwise methods proposed in linear regression analysis. This method embeds the model hypothesis search within the feature subset search. A search procedure of possible feature subsets is defined and various subsets of features are generated and evaluated. The training and testing a specific classification model evaluation produces a specific subset of features. A search algorithm is then 'wrapped' around the classification model to search the space of all feature subsets. These search

methods can be divided in two classes deterministic and randomized search algorithms.

Advantages of Wrapper Method include the interaction between feature subset search and model selection, and the ability to take into account feature dependencies. Disadvantages are that they have a higher risk of over fitting than filter techniques.

III. APPLICATIONS IN BIOINFORMATICS

A. Feature Selection for Sequence Analysis

A multistage process that includes the determination of a sequence (protein, carbohydrate, etc.), its fragmentation and analysis, and the interpretation of the resulting sequence information. This information is useful in that it: (a) reveals the similarities of homologous genes, thereby providing insight into the possible regulation and functions of these genes; and (b) leads to a better understanding of disease states related to genetic variation. New sequencing methodologies, fully automated instrumentation, and improvements in sequencing-related computational resources contribute to the potential for genome-size sequencing projects.

In the context of feature selection, two types of problems can be distinguished: signal and content analysis. Signal analysis focuses on identifying the important motifs in the sequence, such as gene regulatory elements or structural elements. On the other hand content analysis focuses on the broad characteristics of a sequence, such as tendency to code for proteins or fulfillment of a certain biological function and feature selection techniques are then applied to focus on the subset of relevant variables.

1) Content Analysis

In early days of bioinformatics the prediction of subsequence's that code for proteins has been focused. Many versions of Markov models were developed because many features are extracted from a sequence, and most dependencies occur between adjacent positions. Interpolated Markov model was introduced to deal with limited amount of samples [2], and the high amount of possible features. This method used filter method to select only relevant features and interpolation between different orders of the Markov model to deal with small sample sizes. Later Interpolated Markov Model was extended to deal with non-adjacent feature dependencies, resulting in the interpolated context model (ICM), which crosses a Bayesian decision tree with a filter method (λ_2) to assess feature relevance. Recognition of promoter regions and the prediction [3], of microRNA targets are the use of FS techniques in the domain of sequence analysis.

2) Signal Analysis

For the recognition of short, more or less conserved signals in the sequence many sequence analysis methods are used and also to represent the binding sites for various proteins or protein complexes. Regression Approach is the common approach to find regulatory motifs and to relate motifs to gene expression levels to search for the motifs that maximize the fit to the regression model [4], Feature selection is used .In 2003

to find discriminative motifs a classification approach is chosen. This method uses the threshold number of misclassification (TNoM) to score genes for relevance to tissue classification. From the TNoM score, t represents the significance of each motif a P-value is calculated and according to their P-value Motifs are then sorted.

Another line of research is performed in the context of the gene prediction setting, where structural elements such as the translation initiation site (TIS) and splice sites are modeled as specific classification problems. In future research, FS techniques can be expected to be useful for a number of challenging prediction tasks, such as identifying relevant features related to alternative TIS and alternative splice sites.

B. Feature Selection for Microarray Analysis

The human genome contains approximately 20,000 genes. At any given moment, each of our cells has some combination of these genes turned on, and others are turned off. Scientists can answer this question for any cell sample or tissue by gene expression profiling, using a technique called microarray analysis. Microarray analysis involves breaking open a cell, isolating its genetic contents, identifying all the genes that are turned on in that particular cell and generating a list of those genes.

During the last decade, the introduction of microarray datasets stimulated a new line of research in bioinformatics. Microarray data pose a great challenge for computational techniques, because of their small sample sizes and their large dimensionality. Furthermore, additional experimental complications like noise and variability render the analysis of microarray data an exciting domain. A dimension reduction technique was realized in order to deal with these particular characteristics of microarray data and soon their application became a de facto standard in the field. Whereas in 2001, the field of microarray analysis was still claimed to be in its infancy a considerable and valuable effort has since been done to contribute new and adapt known FS methodologies.

1) The Univariate Filter Paradigm

This Method is simple yet efficient because of the high dimensionality of most microarray analyses, fast and efficient FS techniques such as univariate filter methods have attracted most attention. The prevalence of these techniques has dominated the field and now comparative evaluations of different FS techniques and classification over DNA microarray datasets focused on the univariate. This domination of the this approach can be explained by a number of reasons:

- (a) The univariate feature rankings output is intuitive and easy to understand;
- (b) the objectives and expectations that bio-domain experts have when wanting to subsequently validate the result by laboratory techniques or in order to explore literature searches is fulfilled by the output of the gene ranking. The experts could not feel the need for selection techniques that take into account gene interactions;

(c) multivariate gene selection techniques the needs extra computation time.

(d) the possible unawareness of subgroups of gene expression domain experts about the existence of data analysis techniques to select genes in a multivariate way;

The detection of the threshold point in each gene that reduces the number of training sample misclassification and setting a threshold on the observed fold-change differences in gene expression between the states under study are some of the simplest heuristic rule for the identification of differentially expressed genes. A wide range of new univariate feature ranking techniques has since then been developed. These techniques can be divided into two classes: parametric and model-free methods.

Parametric methods assume a given distribution from which the observations (samples) have been generated. t-test and ANOVA are the two samples among the most widely used techniques in microarray studies, although the usage of their basic form, possibly without justification of their main assumptions, is not advisable [5]. To deal with the small sample size and inherent noise of gene expression datasets include a number of t- or t-test like statistics (differing primarily in the way the variance is estimated) and a number of Bayesian frameworks are the modifications of the standard t-test. Regression modeling approaches and Gamma distribution models are the other types of parametrical approaches found in the literature.

Due to the uncertainty about the true underlying distribution of many gene expression scenarios, and the difficulties to validate distributional assumptions because of small sample sizes, non-parametric or model-free methods have been widely proposed as an attractive alternative to make less stringent distributional assumptions. The Wilcoxon rank-sum test [6], between-within classes sum of squares (BSS/WSS) [7], and the rank products method [8]. Are the model-free metrics of statistics field have demonstrated their usefulness in many gene expression studies.

These model-free methods uses random permutations of the data to estimate the reference distribution of the statistics allowing the computation of a model-free version of the associated parametric tests. These techniques deal with the specificities of DNA microarray data, and do not depend on strong parametric assumptions. Their permutation principle partly alleviates the problem of small sample sizes in microarray studies and enhancing the robustness against outliers.

2) The multivariate paradigm for filter, wrapper and embedded techniques

Univariate selection methods have certain restrictions and it leads to less accurate classifiers by, e.g. not taking into account gene-gene interactions. Thus, researchers have proposed techniques that try to capture these correlations between genes. Correlation-based feature selection (CFS) [9], and several variants of the Markov blanket filter method are the application of multivariate filter methods ranges from

simple bivariate interactions towards more advanced solutions exploring higher order interactions. The two other solid multivariate filter procedures are Minimum Redundancy-Maximum Relevance (MRMR) [10], and Uncorrelated Shrunken Centroid (USC) [11], algorithms highlighting the advantage of using multivariate methods over univariate procedures in the gene expression domain.

Feature selection uses an alternative way to perform a multivariate gene subset selection, incorporating the classifier's bias into the search and thus offering an opportunity to construct more accurate classifiers. The scoring function is another characteristic of any wrapper procedure and is used to evaluate each gene subset found. As the 0-1 accuracy measure allows for comparison with previous works, the vast majority of papers use this measure. However, recent proposals advocate the use of methods for the approximation of the area under the ROC curve [12], or the optimization of the LASSO (Least Absolute Shrinkage and Selection Operator) model [13]. For screening different types of errors in many biomedical scenarios ROC curves certainly provide an interesting evaluation measure.

The embedded capacity of several classifiers to discard input features and thus propose a subset of discriminative genes has been exploited by several authors. A random forest (a classifier that combines many single decision trees) is an example to calculate the importance of each gene. The weights of each feature in linear classifiers, such as SVMs and logistic regression are used by embedded FS techniques and these weights are used to reflect the relevance of each gene in a multivariate way, and thus allow for the removal of genes with very small weights.

Due to the lesser degree embedded approaches and higher computational complexity of wrapper, these techniques have not received as much interest as filter proposals. However univariate filter method is an advisable practice to pre-reduce the search space, and only then apply wrapper or embedded methods, hence fitting the computation time to the available resources.

C. Mass Spectra Analysis

For disease diagnosis and protein-based biomarker profiling the emerging new and attractive framework is the Mass spectrometry technology (MS). A mass spectrum sample is characterized by thousands of different mass/charge (m/z) ratios on the x-axis, each with their corresponding signal intensity value on the y-axis. A typical MALDI-TOF low-resolution proteomic profile can contain up to 15,500 data points in the spectrum between 500 and 20,000 m/z , and the number of points even grows using higher resolution instruments.

For data mining and bioinformatics purposes, it can initially be assumed that each m/z ratio represents a distinct variable whose value is the intensity. The data analysis step is severely constrained by both high-dimensional input spaces and their inherent sparseness, just as it is the case with gene expression datasets. Although the amount of publications on

mass spectrometry based data mining is not comparable to the level of maturity reached in the microarray analysis domain, an interesting collection of methods has been presented in the last 4-5 years.

The following crucial steps is to extract the variables that will constitute the initial pool of candidate discriminative features and starting from the raw data, and after an initial step to reduce noise and normalize the spectra from different samples. Some studies employ the simplest approach of considering every measured value as a predictive feature, thus applying FS techniques over initial huge pools of about 15,000 variables, up to around 1,00,000 variables. The elaborated peak detection and alignment techniques are the great deal of current studies performs aggressive feature extraction procedures. These procedures tend to seed the dimensionality from which supervised FS techniques will start their work in less than 500 variables. To set the computational costs of many FS techniques to a feasible size the feature extraction step is thus advisable in these MS scenarios. Univariate filter techniques seem to be the most common techniques used which is similar to the domain of microarray analysis, even though the use of embedded techniques is certainly emerging as an alternative. The other parametric measures such as notable variety of non-parametric scores and F-Test have also been used in several MS studies. Although the t-test maintains a high level of popularity. Multivariate filter techniques on the other hand, are still somewhat underrepresented.

In MS studies Wrapper approaches have demonstrated their usefulness by a group of influential works. In the major part of these papers different types of population-based randomized heuristics are used as search engines: genetic algorithms [14], particle swarm optimization (Ressom et al., 2005) and ant colony procedures [15]. To discard input features an increasing number of papers uses the embedded capacity of several classifiers. Variations of the popular method originally proposed for gene expression domains using the weights of the variables in the SVM-formulation to discard features with small weights, have been broadly and successfully applied in the MS domain. Based on a similar framework, to rank the features by the weights of the input masses in a neural network classifier. The alternative embedded FS strategy is the embedded capacity of random forests and other types of decision tree-based algorithms.

IV. DEALING WITH SMALL SAMPLE DOMAINS

Small sample sizes and their over fitting and inherent risk contain a great challenge for many modeling problems in bioinformatics. Two initiatives have emerged in the context of feature selection (i.e.) the use of adequate evaluation criteria, and the use of stable and robust feature selection models in response to this novel experimental situation.

A. Adequate evaluation criteria

Several papers have warned about the substantial number of applications not performing an independent and honest validation of the reported accuracy percentages. In such cases, a discriminative subset of features is often selected by the

users using the whole dataset. This subset is used to estimate the accuracy of the final classification model thus testing the discrimination rule on samples that were already used to propose the final subset of features. The need for an external feature selection process in training the classification rule at each stage of the accuracy estimation procedure is gaining space in the bioinformatics community practices. Furthermore, novel predictive accuracy estimation methods with promising characteristics, such as bolstered error estimation have emerged to deal with the specificities of small sample domains.

B. Ensemble feature selection approaches

An ensemble system, on the other hand is composed of a set of multiple classifiers and performs classification by selecting from the predictions made by each of the classifiers. Since wide research has shown that ensemble systems are often more accurate than any of the individual classifiers of the system alone and it is only natural that ensemble systems and feature selection would be combined at some point.

Instead of choosing one particular FS method different FS methods can be combined using ensemble FS approaches and accepting its outcome as the final subset. Based on the evidence that there is often not a single universally optimal feature selection technique and due to the possible existence of more than one subset of features that discriminates the data equally well [11], model combination approaches such as boosting have been adapted to improve the robustness and stability of final, discriminative methods [16]. To assess the relevance of each feature in an ensemble FS the methods based on a collection of decision trees (e.g. random forests) can be used. Although the use of ensemble approaches requires additional computational resources, we would like to point out that they offer an advisable framework to deal with small sample domains, provided the extra computational resources are affordable.

V. FEATURE SELECTION IN UPCOMING DOMAINS

A. Single nucleotide polymorphism analysis

A **single-nucleotide polymorphism** (SNP, pronounced *snip*) is a DNA sequence variation occurring when a single nucleotide — A, T, C, or G — in the genome (or other shared sequence) differs between members of a species or paired chromosomes in an individual. Single nucleotide polymorphisms (SNPs) are mutations at a single nucleotide position that occurred during evolution and were passed on through heredity, accounting for most of the genetic variation among different individuals. SNPs are number being estimated at about 7 million in the human genome and it is the forefront of many disease-gene association studies. The important step towards disease-gene association is selecting a subset of SNPs that is sufficiently informative but still small enough to reduce the genotyping overhead. Typically, the number of SNPs considered is not higher than tens of thousands with sample sizes of about 100.

In the past few years several computational methods for htSNP selection (haplotype SNPs; a set of SNPs located on one chromosome) have been proposed. One approach is based on the hypothesis that the human genome can be viewed as a set of discrete blocks that only share a very small set of common haplotypes. The aim of this approach is to identify a subset of SNPs that can either explain a certain percentage of haplotypes or at least distinguish all the common haplotypes. Another common htSNP selection approach is based on pairwise associations of SNPs, and tries to select a set of htSNPs such that each of the SNPs on a haplotype is highly associated with one of the htSNPs [17]. The remaining SNPs can be reconstructed and it is the third approach considering htSNPs as a subset of all SNPs. The idea is to select htSNPs based on how well they predict the remaining set of the unselected SNPs.

B. Text and literature mining

It is emerging as a promising area for data mining in biology. Text mining or text data mining, or text analytics, refers to the process of deriving high-quality information from text. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. Bag-of-Words (BOW) representation is one important representation of text and documents where the variable represents each word in the text representation of the text may lead to very high dimensional datasets, pointing out the need for feature selection techniques.

In the field of text classification the application of feature selection techniques is common and the application in the biomedical domain is still in its infancy. A large number of feature selection techniques that were already developed in the text mining community for tasks such as biomedical document clustering and classification and it will be of practical use for researchers in biomedical literature mining.

VI. FS SOFTWARE PACKAGES

Table I shows an overview of existing software. In order to provide the interested reader with some pointers to existing software packages implementing a variety of feature selection methods. The software is organized into four sections: general purpose FS techniques, techniques tailored to the domain of microarray analysis, techniques specific to the domain of mass spectra analysis and techniques to handle SNP selection and all software packages mentioned are free for academic use. For each software package, the main reference, implementation language and website is shown.

For each software package, the main reference, implementation language and website is shown.

TABLE I SOFTWARE FOR FEATURE SELECTION

General Purpose FS software			
WEKA	Java	Witten and Frank(2005)	http://www.cs.waikato.ac.nz/ml/weka
Fast Correlation Based Filter	Java	Yu and Liu(2004)	http://www.public.asu.edu/~huanliu/FCBF/FCBFsoftware.html
MLC++	C++	Kohavi et al.(1996)	http://www.sgi.com/tech/mlc
Feature selection Book	Ansi C	Liu and Motoda(1998)	http://public.asu.edu/~huanliu/FSbook
Microarray analysis FS software			
SAM	R,Excel	Tusher et al.(2001)	http://www-stat.stanford.edu/~tibs/SAM/
PCP	C,C++	Buturovic(2005)	http://pcp.sourceforge.net
GALGO	R	Trevino & Falciani(2006)	http://www.bip.bham.ac.uk/bioinf/galgo.html
GA-KNN	C	Li et al(2001)	http://dir.niehs.nih.gov/microarray/datamining/
Nudge(Bioconductor)	R	Dean & Raftery(2005)	http://www.bioconductor.org/
Qvalue(Bioconductor)	R	Storey(2002)	http://www.bioconductor.org/
DEDS(Bioconductor)	R	Yang et.al(2005)	http://www.bioconductor.org/
Mass Spectra analysis FS software			
GA-KNN	C	Li et al(2004)	http://dir.niehs.nih.gov/microarray/datamining/
R-SVM	R,C,C++	Zhang et al.(2006)	http://www.hsph.harvard.edu/bioinfocore/RSVMhome/R-SVM.html
SNP analysis FS software			
CHOISS	C++, Perl	Lee and Kang(2004)	http://biochem.kaist.ac.kr/choiss.htm
WCLUSTAG	Java	Sham et al.(2007)	http://bioinfo.hku.hk/wclustag

VII. CONCLUSIONS AND FUTURE PERSPECTIVES

In this article, it is reviewed the main contributions of feature selection research in a set of well-known bioinformatics applications. The large input dimensionality and the small sample sizes are the two main issues emerge as common problems in the bioinformatics domain. Researchers designed FS techniques to deal with these problems in bioinformatics, machine learning and data mining.

During the last years a large and fruitful effort has been performed in the adaptation and proposal of univariate filter FS techniques. In general, it is observed that many researchers in the field still think that filter FS approaches are only restricted to univariate approaches. The proposal of multivariate selection algorithms can be considered as one of the most promising future lines of work for the bioinformatics community.

A second line of future research is The development of especially fitted ensemble FS approaches to enhance the robustness of the finally selected feature subsets is the second line of future research. In order to alleviate the actual small sample sizes of the majority of bioinformatics applications, the further development of such techniques, combined with appropriate evaluation criteria, constitutes an interesting direction for future FS research.

SNPs, text and literature mining, and the combination of heterogeneous data sources are the other interesting opportunities for future FS research will be the extension towards upcoming bioinformatics domains. While in these domains, the FS component is not yet as central as, e.g. in gene expression or MS areas, I believe that its application will become essential in dealing with the high-dimensional character of these applications.

ACKNOWLEDGMENT

I would like to thank the anonymous reviewers for their constructive comments, which significantly improved the quality of this review.

REFERENCES

- [1] Daelemans, W., et al. (2003) Combined optimization of feature selection and algorithm parameter interaction in machine learning of language. In Proceedings of the 14th European Conference on Machine Learning (ECML – 2003), pp. 84-95.
- [2] Salzberg, et al. (1998) Microbial gene identification using interpolated markov models. Nucleic Acids Res., 26, 544–548.
- [3] Saeys, Y., et al. (2007) In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi, and protists. Bioinformatics, 23, 414–420.
- [4] Keles, S., et al. (2002) Identification of regulatory elements using a feature selection method. Bioinformatics, 18, 1167–1175.
- [5] Jafari, P. and Azuaje, F. (2006) An assessment of recently published gene

- expression data analyses: reporting experimental design and statistical factors, *BMC Med. Inform. Decis. Mak.*, 6, 27.
- [6] Thomas, J., et al. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.*, 11, 1227–1236.
- [7] Dudoit, S., et al. (2002) Comparison of discriminant methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, 97, 77–87.
- [8] Breitling, R., et al. (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, 573, 83–92.
- [9] Wang, Y., et al. (2006) Tumor classification based on DNA copy number aberrations determined using SNPS arrays. *Oncol. Rep.*, 5, 1057–1059.
- [10] Ding, C. and Peng, H. (2003) Minimum redundancy feature selection from microarray gene expression data. In *Proceedings of the IEEE Conference on Computational Systems Bioinformatics*, pp. 523–528
- [11] Yeung, K. and Bumgarner, R. (2003) Multiclass classification of microarray data with repeated measurements: application to cancer. *Genome Biol.*, 4, R83.
- [12] Ma, S. and Huang, J. (2005) Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics*, 21, 4356–4362.
- [13] Ghosh, D. and Chinnaiyan, M. (2005) Classification and selection of biomarkers in genomic data using LASSO. *J. Biomed. Biotechnol.*, 2005, 147–154.
- [14] Li, T., et al. (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20, 2429–2437.
- [15] Resson, H., et al. (2007) Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics*, 23, 619–626.
- [16] Ben-Dor, A., et al. (2000) Tissue classification with gene expression profiles. *J. Comput. Biol.*, 7, 559–584
- [17] Carlson, C., et al. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, 74, 106–120..
- [18] Margaret H. Dunham S. Sridhar (2008) *Data Mining Introductory and Advanced Topics*.
- [19] Efron, B., et al. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, 96, 1151–1160.
- [20] Kohavi, R., et al. (1996) Data mining using MLC++: a machine learning library in C++. In *Tools with Artificial Intelligence*, IEEE Computer Society Press, Washington, DC, pp. 234–245.
- [21] Inza, I., et al. (2000) Feature subset selection by Bayesian networks based optimization. *Artif. Intell.*, 123, 157–184.
- [22] Sofie Van Landeghem (2008), *Extracting Protein –Protein Interactions from Text using Rich Feature Vectors and Feature Selection*. 77-84.
- [23] Michael Gutkin (2009) , *A method for feature selection in gene expression-based disease classification*.
- [24] Varshavsky, R., et al. (2006) Novel unsupervised feature filtering of biological data. *Bioinformatics*, 22, e507–e513.
- [25] Jake Y. Chen , Stefano Lonardi (2009), *Biological Data Mining*.
- [26] Pawel Smialowski, *Bioinformatics (2010)*, Pitfalls of supervised feature selection, *oxford journals* 26(3): 440-443.

Software Effort Prediction using Statistical and Machine Learning Methods

Ruchika Malhotra

Department of Software Engineering,
Delhi Technological University
Bawana, Delhi 110042
ruchikamalhotra2004@yahoo.com

Ankita Jain

Department of Computer Engineering,
Delhi Technological University
Bawana, Delhi 110042
ankita4813@yahoo.com

Abstract - Accurate software effort estimation is an important part of software process. Effort is measured in terms of person months and duration. Both overestimation and underestimation of software effort may lead to risky consequences. Also, software project managers have to make estimates of how much a software development is going to cost. The dominant cost for any software is the cost of calculating effort. Thus, effort estimation is very crucial and there is always a need to improve its accuracy as much as possible. There are various effort estimation models, but it is difficult to determine which model gives more accurate estimation on which dataset. This paper empirically evaluates and compares the potential of Linear Regression, Artificial Neural Network, Decision Tree, Support Vector Machine and Bagging on software project dataset. The dataset is obtained from 499 projects. The results show that Mean Magnitude Relative error of decision tree method is only 17.06%. Thus, the performance of decision tree method is better than all the other compared methods.

Keywords— Software effort estimation, machine learning, decision tree, linear regression

I. INTRODUCTION

For any software organization, accurate estimation of effort is crucial for successful management and control of software project. In other words, in any software effort estimation, making an estimate of the person-months and the duration required to complete the project, is very important. Software effort estimation also plays very important role in determining cost of the software. Thus, effort estimation is crucial for the quality of the software.

Software Effort estimation techniques fall under following categories: Expert judgment, Algorithmic estimation, Machine Learning, Empirical techniques, Regression techniques, and Theory-based techniques. It is difficult to determine which model gives more accurate result on which dataset. Thus, there is a need for predicting effort and making a comparative analysis of various machine learning methods.

In this paper, we have done empirical study and comparison of some of the models on well-known China dataset [21]. The models which we are dealing with are developed using statistical and machine learning methods in order to verify which model performs the best. Linear Regression, Artificial Neural Network, Support Vector machine, Decision Tree, and bagging are the methods which are used in this work. These methods have seen an explosion

of interest over years and hence it is important to analyse the performance of these methods. We have analysed these methods on large datasets collected from 499 projects.

The paper is organized as follows: Section 2 summarizes the related work. Section 3 explains the research background, i.e. describes the dataset used for the prediction of effort and also explains various performance evaluation measures. Section 4 presents the research methodology followed in this paper. The results of the models predicted for software development effort estimation and the comparative analysis are given in section 5. Finally, the paper is concluded in section 6.

II. RELATED WORK

Software effort estimation is a key consideration to software cost estimation [5]. There are numerous Software Effort Estimation Methods such as Algorithmic effort estimation, machine learning, empirical techniques, regression techniques and theory based techniques. Various models have been discussed in previous researches. An important task in software project management is to understand and control critical variables that influence software effort [5]. The paper by K.Smith, et.al. [17] has discussed the influence of four task assignment factors, team size, concurrency, intensity, and fragmentation on the software effort. These four task assignment factors are not taken into consideration by COCOMO I and COCOMO II in predicting software development effort. The paper [17] has proposed the Augmented and Parsimonious models which consider the task assignment factors to calculate effort and thus has proved that estimates are improved significantly by adding these factors while determining effort. Besides these task assignment factors which influence the effort estimation, the paper by Girish H. Subramanian, et.al.[5], concluded that the adjustment variables i.e. software complexity, computer platform, and program type have a significant effect on software effort. COCOMO I, COCOMO II, Function Points [1] and its various extensions all use adjustment variables, such as software complexity and reliability among others, to arrive at an adjusted estimate of software effort and cost. Also there is significant interaction between the adjustment variables which indicate that these adjustment variables influence each other and their interactions also have a significant effect on effort.

Some recent study is also done in the field of “Analogy based Estimations”. Analogy based estimations compare the

similarities between the projects whose effort is to be estimated with all the historical projects. In other words, it tries to identify that historical project which is most similar to the project being estimated. To measure the similarity between pairs of projects, distance metrics are used. Euclidean (Jeffery et al., 2000), Manhattan (Emam et al., 2001) and Minkowski distances (Stamelos et al., 2003) are the widely used distance metrics in analogy-based estimations, [14]. Various researches have been done to improve the estimation accuracy in analogy – based estimations. The author Chiu, et.al. [14], proposed an adjusted analogy-based software effort estimation model by adopting the GA method to adjust the effort based on the similarity distances. In other words, the effort of the closest project is not used directly, but it is adjusted to improve the accuracy. Another method of improving the estimation accuracy is proposed by Tosun, et.al [3]. In the traditional formula for Euclidean distance, the features are either unweighted or same weight is assigned to each of the features. The problem in the unweighted case is that importance of each feature is not taken into account. In the paper [3], the authors have proposed a novel method for assigning weights to features by taking their particular importance on cost into consideration. Two weight assignment heuristics are implemented which are inspired by a widely used statistical technique called PCA.

A lot of research has also been done in Machine learning techniques of estimation. The paper by Finnie and Wittig [7], has examined the potential of two artificial intelligence approaches i.e. artificial neural networks (ANN) and case-based reasoning (CBR) for creating development effort estimation models using the same dataset which is ASMA (Australian Software Metrics Association). Also, the potential of artificial neural networks (ANN) and case-based reasoning (CBR), for providing the basis for development effort estimation models in contrast to regression models is examined by the same authors in their paper [6]. The authors concluded that artificial intelligence models are capable of providing adequate estimation models. Their performance is to a large degree dependent on the data on which they are trained, and the extent to which suitable project data is available will determine the extent to which adequate effort estimation models can be developed. CBR allows the development of a dynamic case base with new project data being automatically incorporated into the case base as it becomes available while ANNs will require retraining to incorporate new data.

Besides ANN and CBR, other important machine learning techniques is CART (Classification and regression trees). Recently, MART (Multiple additive regression trees) has been proposed that extends and improves the CART model using stochastic gradient model. The paper by Elish [12] empirically evaluates the potential and accuracy of MART as a novel software effort estimation model when compared with recently published models, i.e. radial basis function (RBF) neural networks, linear regression, and support vector regression models with linear and RBF kernels. The comparison is based on a well-known and respected NASA software project dataset. The paper [2] has compared the results of Support vector regression with both linear regression and RBF kernels.

Genetic Algorithms are also widely used for accurate effort estimation. The paper by Burgess and Lefley [4], evaluates the potential of genetic programming (GP) in software effort estimation and comparison is made with the Linear LSR, ANN etc. The comparison is made on the Desharnais data set of 81 software projects. The results obtained depend on the fitness function used.

As we have seen, software repositories or datasets are widely used to obtain data on which effort estimation is done. But software repositories contain data from heterogeneous projects. Traditional application of regression equations to derive a single mathematical model results in poor performance [8]. The paper by Gallogo [8] has used Data clustering to solve this problem.

In this research, the models are predicted and validated using both statistical and machine learning methods. The comparative analysis with previous researches has also been done. The results showed that the Decision Tree was the best among all the other models used with MMRE of 17 %.

III. RESEARCH BACKGROUND

A. Feature Sub Selection Method

The data we have used is obtained from Promise data repository. The dataset comprises of 19 features, one dependent and eighteen independent variables. But, some of the independent variables are removed as they are not much important to predict the effort, thus making the model much simpler and efficient. There are various techniques used for reducing data dimensionality. We have used Feature sub selection technique which is provided in the WEKA tool [21] to reduce the number of independent variables. After applying Correlation Based Feature Subselection (CFS), the 19 variables were reduced to 10 variables (one dependent and nine independent variables). Correlation based feature selection technique (CFS) is applied to select the best predictors out of independent variables in the datasets [11], [18]. The best combinations of independent variable were searched through all possible combinations of variables. CFS evaluates the best of a subset of variables by considering the individual predictive ability of each feature along with the degree of redundancy between them. The dependent variable is Effort. Software development effort is defined as the work carried out by the software supplier from specification until delivery measured in terms of hours [18].

The independent variables are Output, Enquiry, Interface, Added, PDR_AFP, PDR_UFP, NPDR_AFP, NPDU_UFP and Resource. All the independent variables correspond to function point method [1].

B. Empirical Data Collection

The dataset which we have used consists of 19 drivers of effort for predicting effort estimation model. The descriptive statistics of nine independent variables chosen by CFS method is shown in table 1. The mean value of the effort (dependent variable) is found to be 3921.

TABLE I. DATASET STATISTICS

Variables	Mean	Min.	Max.	Median	Standard Deviation
Output	114	0	2455	42	221
Enquiry	62	0	952	24	105
Interface	24	0	1572	0	85
Added	360	0	1358	135	829
PDR_AFP	12	0.3	83.8	8.1	12
PDR_UFP	12	0.3	96.6	8	13
NPDR_AFP	13	0.4	101	8.8	14
NPDR_UFP	14	0.4	108.3	8.9	15
Resource	1	1	4	1	1
Effort	3921	26	50620	1829	6474

C. Performance Measures

We have used the following evaluation criterion to evaluate the estimate capability. Among all the mentioned measures, the most commonly used are PRED(A) and MMRE. Hence, we will use these two measures to compare our results with the results of the previous researches.

1. *Mean Magnitude of relative error (MMRE)* (or mean absolute relative error) [9], [6]

$$MMRE = \frac{1}{n} \sum_{i=1}^n \frac{|P_i - A_i|}{A_i} \quad (1)$$

Where P_i is the predicted value for datapoint i ;

A_i is the actual value for datapoint i ;

n is the total number of datapoints

2. *PRED(A)*

It is calculated from the relative error. It is defined as the ratio of datapoints with error (MRE) less than equal to A to the total number of datapoints. Thus, higher the value of PRED(A), the better it is considered.

$$PRED(A) = d/n \quad (2)$$

Where d is the value of MRE where datapoints have less than or equal to A error. Commonly used value of A is 25% in the literature.

3. *Root Mean Squared Error (RMSE)*

The root mean squared error is defined as:

$$E = \sqrt{\frac{1}{n} \sum_{i=1}^n (P_i - A_i)^2} \quad (3)$$

Where P_i is the predicted value for datapoint i ;

A_i is the actual value for datapoint i ;

n is the total number of datapoints

If $P_i = A_i, \forall i = 1, 2, \dots, n$; then $E=0$ (ideal case)

Thus, range of E is from 0 to infinity. RMSE gives high importance to large errors because the errors are squared before they are averaged. Thus, RMSE is used the most when large errors are undesirable.

4. *Relative absolute Error (RAE)*

The relative absolute error of individual dataset j is defined as:

$$E_j = \frac{\sum_{i=1}^n |P_{ij} - A_i|}{\sum_{i=1}^n |A_i - A_m|} \quad (4)$$

Where P_{ij} is the value predicted by the individual dataset j for datapoint i ;

A_i is the actual value for datapoint i ;

n is the total number of datapoints;

A_m is the mean of all A_i

For ideal case, the numerator is equal to 0 and $E_j = 0$. Thus, the E_j ranges from 0 to infinity.

5. *Root Relative Squared Error*

The root relative squared error of individual dataset j is defined as:

$$E_j = \sqrt{\frac{\sum_{i=1}^n (P_{ij} - A_i)^2}{\sum_{i=1}^n (A_i - A_m)^2}} \quad (5)$$

Where P_{ij} is the value predicted by the individual dataset j for datapoint i ;

A_i is the actual value for datapoint i ;

n is the total number of datapoints;

A_m is the mean of all A_i

For ideal case, the numerator is equal to 0 and $E_j = 0$. Thus, the E_j ranges from 0 to infinity.

6. *Mean Absolute error*

The mean absolute error measures of how far the estimates are from actual values. It could be applied to any two pairs of numbers, where one set is "actual" and the other is an estimate, prediction.

7. *Correlation Coefficient*

Correlation measures of the strength of a relationship between two variables. The strength of the relationship is indicated by the correlation coefficient. The larger the value of correlation coefficient, the stronger the relationship.

D. *Validation Measures*

There are three validation techniques namely hold-out, leave-one-out and K-cross validation [13]. As our dataset is large consists of 499 data points, the hold out method is used where the dataset is divided into two parts, i.e. the training and validation set.

IV. RESEARCH METHODOLOGY

In this paper, we are using the machine learning techniques in order to predict effort. We have used one regression and four machine learning methods in order to predict effort. Support vector machines, Artificial Neural

Network, Decision tree and Bagging methods have seen an explosion of interest over the years, and have successfully been applied in various areas.

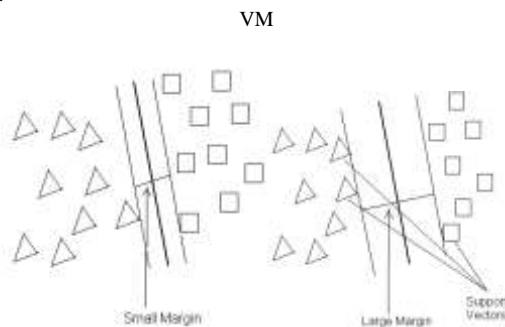
A. Linear Regression

Linear regression analyses the relationship between two variables, X and Y. One variable is the dependent variable and the other is the independent variable. For doing this, it finds a line which minimizes the sum of the squares of the vertical distances of the points from the line. In other words, it is method of estimating the conditional expected value of one variable y given the values of some other variable or variables x.

B. Support Vector Machine

Support Vector Machine (SVM) is a learning technique which is used for classifying unseen data correctly. For doing this, SVM builds a hyperplane which separates the data into different categories. The dataset may or may not be linearly separable. By 'linearly separable' we mean that the cases can be completely separated i.e. the cases with one category are on the one side of the hyperplane and the cases with the other category are on the other side. For example Figure 1 shows the dataset where examples belong to two different categories – triangles and squares. Since these points are represented on a 2 – dimensional plane, they can be separated by a 1-dimensional line. To separate these points into 2 different categories, there is infinite number of lines possible. Two possible candidate lines are shown in the figure 1. However, only one of the lines gives maximum separation/margin and that line is selected. 'Margin' is defined as distance between the dashed lines (as shown in figure) drawn parallel to the separating lines. These dashed lines give the distance between the separating line and closest vectors to the line. These vectors are called as support vectors. SVM can also be extended to the non-linear boundaries using kernel trick. The kernel function transforms the data into higher dimensional space to make the separation easy. We have used SVM for estimating continuous variable, i.e. effort. [19].

Figure 1.



C. Artificial Neural network

Artificial Neural Network (ANN) comprises a network of simple interconnected units called "neurons" or "processing units". The ANN has three layers, i.e. the input layer, hidden layer and the output layer. The first layer has input neurons which send data via connections called weights to the second layer of neurons and then again via more weight to the third layer of output neurons. More complex systems have more than one hidden layers. But it has been proved in literature that more than one hidden layer may not be acceptable [20].

The most common algorithm for training or learning is known as error back-propagation algorithm.

Error back-propagation learning consists of two passes: a forward pass and a backward pass. In the forward pass, an input is presented to the neural network, and its effect is propagated through the network layer by layer. During the forward pass the weights of the network are all fixed. During the backward pass the weights are all updated and adjusted according to the error computed. An error is composed from the difference between the desired response and the system output. This error information is fed back to the system and adjusts the system parameters in a systematic fashion (the learning rule). The process is repeated until the performance is acceptable. The model predicted in this study consists of 9 input layers (independent variables chosen in our study) and one output layer.

D. Decision Tree

Decision tree is a methodology used for classification and regression. It provides a modelling technique that is easy for human to comprehend and simplifies the classification process. Its advantage lies in the fact that it is easy to understand; also, it can be used to predict patterns with missing values and categorical attributes. Decision tree algorithm is a data mining induction techniques that recursively partitions a data set of records using depth-first greedy approach or breadth-first approach until all the data items belong to a particular class.

A decision tree structure is made of (root, internal and leaf) nodes and the arcs. The tree structure is used in classifying unknown data records. At each internal node of the tree, a decision of best split is made using impurity measures. We have used M5P implemented in WEKA Tool [21]. This method generated M5 model rules and trees. The details of this method can be found in [16].

E. Bagging

Bagging which is also known as bootstrap aggregating is a technique that repeatedly samples (with replacement) from a data set according to a uniform probability distribution [10]. Each bootstrap sample has the same size as the original data. Because the sampling is done with replacement, some instances may appear several times in the same training set, while others may be omitted from the training set. On average, a bootstrap sample D_i contains approximately 63% of the original training data because each sample has a probability $1 - (1 - 1/N)^N$ of being selected in each D_i . If N is sufficiently large, this probability converges to $1 - 1/e = 0.632$. After training the k classifiers, a test instance is assigned to the class that receives the highest number of votes.

V. ANALYSIS RESULTS

A. Model Prediction Results

China Dataset [15] was used to carry out the prediction of effort estimation model. A holdout technique of cross validation was used to estimate the accuracy of effort estimation model. The dataset was divided into two parts i.e. training and validation set in ratio of 7:3. Thus 70% was used for training the model and 30% was used for validating accuracy of the model. Four machine learning methods and one regression method was used to analyse the results.

The model with the lower MMRE, RMSE, RAE, RRSE, MAE and the higher correlation coefficient and PRED(25) is considered to be the best among others. As shown in table, the decision tree results are found to be best with the MMRE value 17.06%, RAE value 32.02, RRSE value 38.40, correlation coefficient 0.93, and PRED(25) value

52%. Hence decision tree method is found to be effective in predicting effort. Also, the results of decision tree are competent with the traditional linear regression model.

TABLE II. OUR RESULTS

Performance Measures	Linear regression	Support Vector Machine	Artificial Neural Network	Decision tree	Bagging
Mean Magnitude Relative Error (MMRE) %	17.97	25.63	143.79	17.06	74.23
Root mean squared error (RMSE) %	4008.11	3726.71	5501.18	2390.47	3656.75
Relative absolute error (RAE) %	54.16	48.49	71.50	32.02	45.79
Root relative squared error (RRSE) %	64.74	60.13	91.17	38.40	59.11
Correlation coefficient	0.79	0.81	0.75	0.93	0.83
Mean absolute error (MAE) %	1981.48	1774.36	2561.00	1173.43	1668.03
PRED(25) %	36	38.66	11.33	52	34.66

The graphs as shown in figures 2-6, for the actual and the values as predicted by the particular model are shown on Y-axis and they correspond to the 499 projects. The 'black' curve presents the curve for the actual values, whereas the

'red' curve presents the curve for the predicted values. The closer the actual and predicted curves, the lesser are the error and better are the model. The graphs show that the actual and the predicted values are very close to each other. As shown in figure 5, the decision tree shows the best result.

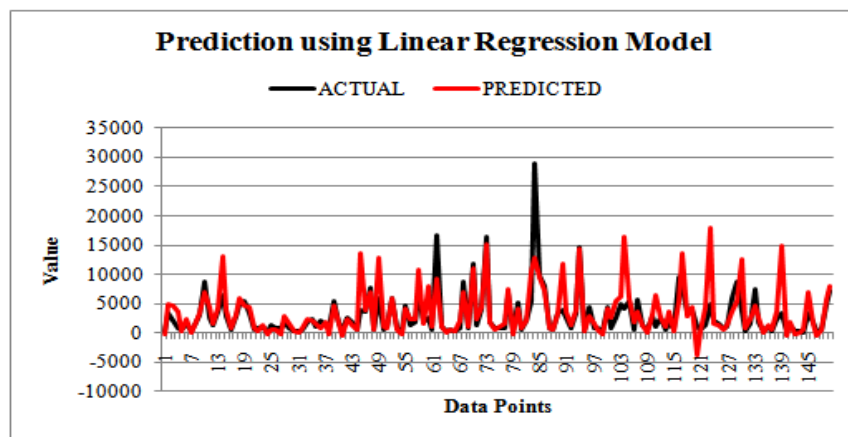


Figure 2. Results Using Linear Regression

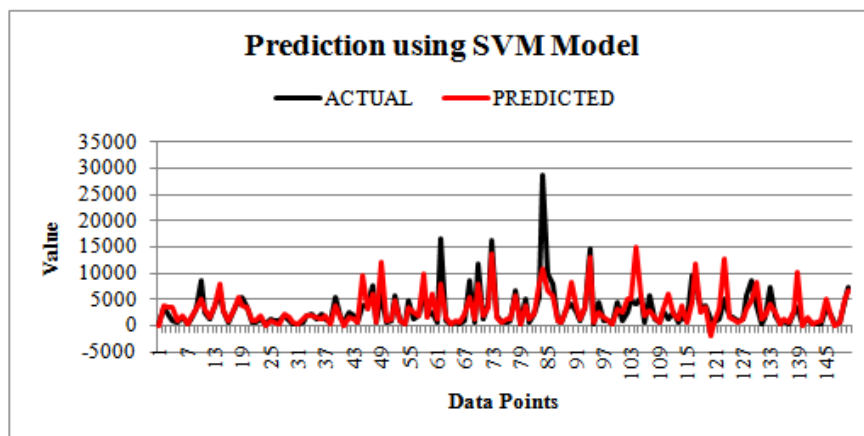


Figure 3. Results Using Support Vector Machine

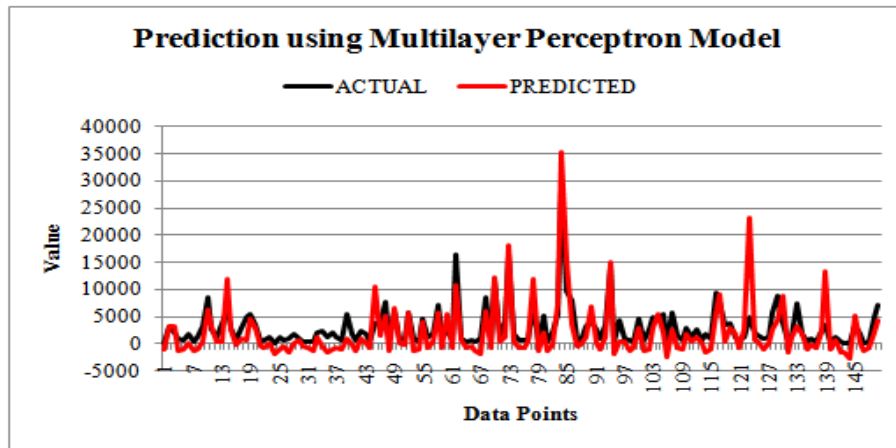


Figure 4. Results Using Artificial Neural Network

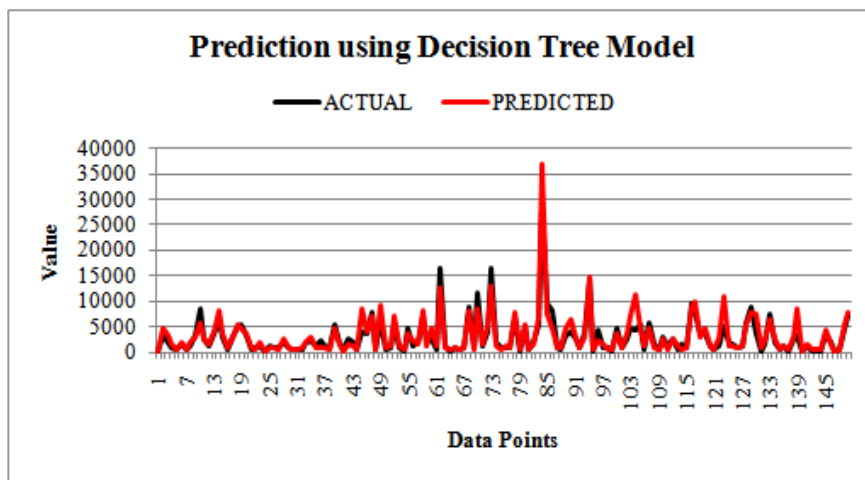


Figure 5. Results Using Decision Tree

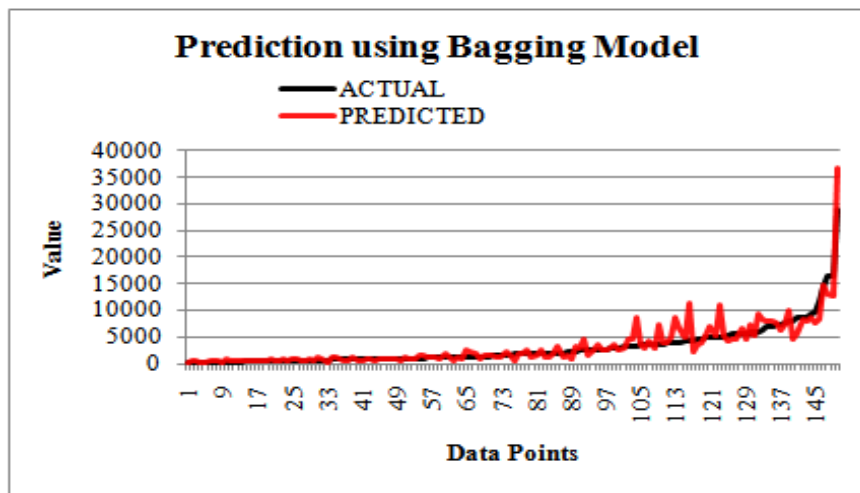


Figure 6. Results Using Bagging

B. Comparative analysis with previous studies

Our results of effort estimation model predicted using decision tree method were better than all the other four methods used in our study. We have compared our results with eight other previous studies. The previous studies under comparison have same dependent variable, although the

independent variables vary for each study. As shown in table 3, out of 24 models of literature, our decision tree model has outperformed the MMRE values of 21 models. The results of PRED(25) are also very good and higher as compared to most of the models. Thus, our effort estimation model using decision tree method is widely acceptable.

TABLE III. COMPARISON ANALYSIS

Papers	Methods Used	Pred(25)	MMRE
Our Results	Decision Tree	52	17.06
	Bagging	34.66	74.23
	Linear Regression	36	17.97
	SVM	38.66	25.63
	ANN	11.33	143.79
[14]	ANN	22	90
	Classification And Regression Trees	26	77
	Ordinary Least Square Regression	33	72
	Adjusted analogy-based estimation using Euclidean distance	57	38
	Adjusted analogy-based estimation using Manhattan distance	52	36
	Adjusted analogy-based estimation using Minkowski distance	61	43
[17]	Augmented COCOMO	Pred(20)31.67	65
	Parsimonious COCOMO	30.4	64
[8]	Clustering	Pred(30) 35.6	1.03
[6]	Regressive	-	62.3
	ANN	-	35.2
	Case Based Reasoning	-	36.2
[12]	Multiple Additive Regression Trees	88.89	8.97
	Radial Basis Function	72.22	19.07
	SVR Linear	88.89	17.4
	SVR RBF	83.33	17.8
	Linear Regression	72.22	23.3
[4]	Genetic Programming	23.5	44.55
	ANN	56	60.63
[7]	ANN	-	17
	Case Based Reasoning	-	48.2
[21]	SVR	88.89	16.5
	Linear Regression	72.22	23.3
	Radial Basis Function	72.22	19.07

VI. CONCLUSION

In this research we have made a comparative analysis of one regression with four machine learning methods for predicting effort. We have obtained results using the data obtained from Promise data repository. The dataset consists of 19 features which we have reduced to 10 features using CFS method. The results show that the decision tree was the best method for predicting effort with MMRE value 17% and PRED(25) value 52%. The software practitioners and researchers may apply decision tree method for effort estimation. Hence, machine learning methods selected in this study have shown their ability to provide an adequate model for predicting maintenance effort.

The future work can further replicate this study for industrial software. We plan to replicate our study to predict effort prediction models based on other machine learning algorithms such as genetic algorithms. We may carry out cost benefit analysis of models that will help to determine whether a given effort prediction model would be economically viable.

REFERENCES

- [1] A. Albert and J.E. Gaffney, "Software Function Source Lines of Code and Development Effort Prediction: A Software Science Validation," IEEE Trans. Software Engineering, vol. 9, pp. 639-648, 1983.
- [2] A.L.I. Oliveira, " Estimation of software effort with support vector regression," Neurocomputing, vol. 69, pp. 1749-1753, Aug. 2006.

- [3] A.Tosun, B.Turhan and A.B. Bener, "Feature weighting heuristics for analogy-based effort estimation models," *Expert Systems with Applications*, vol. 36, pp. 10325-10333, 2009.
- [4] C.J. Burgess and M.Lefley, "Can genetic programming improve software effort estimation? A comparative evaluation," *Information and Software Technology*, vol. 43, pp. 863-873, 2001.
- [5] G.H. Subramanian, P.C. Pendharkar and M.Wallace, "An empirical study of the effect of complexity, platform, and program type on software development effort of business applications," *Empirical Software Engineering*, vol. 11, pp. 541-553, Dec. 2006.
- [6] G.R. Finnie and G.E. Wittig, "A Comparison of Software Effort Estimation Techniques: Using Function Points with Neural Networks, Case-Based Reasoning and Regression Models," *Journal of Systems and Software*, vol. 39, pp. 281-289, 1997.
- [7] G.R.Finnie and G.E. Wittig, "AI Tools for Software Development Effort Estimation," in *Proc. SEEP '96*, 1996, International Conference on Software Engineering: Education and Practice (SE:EP '96).
- [8] J.J.C Gallego, D.Rodriguez, M.A.Sicilia, M.G.Rubio and A.G. Crespo, "Software Project Effort Estimation Based on Multiple Parametric Models Generated Through Data Clustering," *Journal of Computer Science and Technology*, vol. 22, pp. 371-378, May 2007.
- [9] K. Srinivasan and D. Fisher, "Machine Learning Approaches to Estimating Software Development Effort," *IEEE Transactions on Software Engineering*, vol. 21, Feb. 1995.
- [10] L.Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123-140, Aug. 1996.
- [11] M.Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proceedings of the 17th International Conference on Machine Learning*, pp.359-366.
- [12] M.O. Elish, "Improved estimation of software project effort using multiple additive regression trees," *Expert Systems with Applications*, vol. 36, pp. 10774-10778, 2009.
- [13] M.Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal Royal Stat. Soc.*, vol. 36, pp. 111-147, 1974.
- [14] N.H. Chiu and S.J. Huang, "The adjusted analogy-based software effort estimation based on similarity distances," *The Journal of Systems and Software*, vol. 80, pp. 628-640, 2007.
- [15] Promise. Available: <http://promisedata.org/repository/>.
- [16] R.J. Quinlan, "Learning with Continuous Classes," in *5th Australian Joint Conference on Artificial Intelligence*, Singapore, pp. 343-348, 1992.
- [17] R.K. Smith, J.E.Hale and A.S.Parrish, "An Empirical Study Using Task Assignment Patterns to Improve the Accuracy of Software Effort Estimation," *IEEE Transactions on Software Engineering*, vol. 27, pp. 264-271, March 2001.
- [18] R.Malhotra, A.Kaur and Y.singh, "Application of Machine Learning Methods for Software Effort Prediction," in *Newsletter ACM SIGSOFT Software Engineering Notes*, vol. 35, May 2010.
- [19] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy, "Improvements to the SMO Algorithm for SVM Regression," *IEEE Transactions on Neural Networks*, vol. 13, March 2001.
- [20] T.M. Khoshgafaar, E.D. Allen, J.P. Hudepohl, S.J. Aud, "Application of neural networks to software quality modeling of a very large telecommunications system," *IEEE Transactions on Neural Networks*, vol. 8, pp. 902-909, July 1997.
- [21] Weka. Available: <http://www.cs.waikato.ac.nz/ml/weka/>

AUTHORS PROFILE

Ruchika Malhotra She is an assistant professor at the Department of Software Engineering, Delhi Technological University (formerly Delhi College of Engineering), Delhi, India. She was an assistant professor at the University School of Information Technology, Guru Gobind Singh Indraprastha University, Delhi, India. Prior to joining the school, she worked as full-time research scholar and received a doctoral research fellowship from the University School of Information Technology, Guru Gobind Singh Indraprastha Delhi, India. She received her master's and doctorate degree in software engineering from the University School of Information Technology, Guru Gobind Singh Indraprastha University, Delhi, India. Her research interests are in software testing, improving software quality, statistical and adaptive prediction models, software metrics, neural nets modeling, and the definition and validation of software metrics. She has published more than 40 research papers in international journals and conferences. Malhotra can be contacted by e-mail at ruchikamalhotra2004@yahoo.com

Ankita Jain She is at the Department of Computer Engineering, Delhi Technological University (formerly Delhi College of Engineering), Delhi, India. She received her bachelor's degree in computer engineering from the Guru Gobind Singh Indraprastha University, Delhi, India. Her research interests are statistical and adaptive prediction models and improving software quality. She can be contacted by e-mail at ankita4813@yahoo.com