IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 2 Issue 5

May 2011

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

SAI

www.ijacsa.thesai.org

IJACSA

WHERE WISDOM SHARES

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS

A Publication of
The Science and Information Organization

IJACSA

WHERE WISDOM SHARES

INTERNATIONAL JOURNAL OF
ADVANCED COMPUTER SCIENCE AND APPLICATIONS

SAI

A Publication of
The Science and Information Organization

IJACSA Editorial

From the Desk of Managing Editor…

It is a pleasure to present our readers with the May 2011 Issue of International Journal of Advanced Computer Science and Applications (IJACSA).

With monthly feature peer-reviewed articles and technical contributions, the Journal's content is dynamic, innovative, thought-provoking and directly beneficial to the readers in their work.

The number of submissions have 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.

Some of the papers have an introductory character, some of them access highly desired extensions for a particular method, and some of them even introduce completely new approaches to computer science research in a very efficient manner. This diversity was strongly desired and should contribute to evoke a picture of this field at large. As a consequence only 29% of the received articles have been finally accepted for publication.

With respect to all the contributions, we are happy to have assembled researchers whose names are linked to the particular manuscript they are discussing. Therefore, this issue may not just be used by the reader to get an introduction to the methods but also to the people behind that have been pivotal in the promotion of the respective research.

By having in mind such future issues, we hope to establish a regular outlet for contributions and new findings in the field of Computer science and applications. Therefore, IJACSA in general, could serve as a reliable resource for everybody loosely or tightly attached to this field of science.

And if only a single young researcher is inspired by this issue to contribute in the future to solve some of the problems sketched here or contribute to exiting methodologies and research work, the effort of all the contributors will be rewarded. In that sense we would like to thank all the authors and reviewers that contributed to this issue for their efforts and their collaboration in this project.

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
Volume 2 Issue 5, May 2011
editorijacsa@thesai.org
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

IJACSA Associate Editors

IJACSA Reviewer Board

CONTENTS

Paper 1: Using Merkle Tree to Mitigate Cooperative Black-hole Attack in Wireless Mesh Networks

Authors: Shree Om, Mohammad Talib

PAGE 1 – 6

Paper 2: QVT transformation by modelling - From UML Model to MD Model

Authors: I.Arrassen, A.Meziane, R.Sbai, M.Erramdani

PAGE 7 – 14

Paper 3: Fuzzy Particle Swarm Optimization with Simulated Annealing and Neighborhood Information Communication for Solving TSP

Authors: Rehab F. Abdel-Kader

PAGE 15 – 21

Paper 4: Empirical Validation of Web Metrics for Improving the Quality of Web Page

Authors: Yogesh Singh, Ruchika Malhotra, Poonam Gupta

PAGE 22 – 28

Paper 5: Systematic and Integrative Analysis of Proteomic Data using Bioinformatics Tools

Authors: Rashmi Rameshwari, Dr. T. V. Prasad

PAGE 29 – 35

Paper 6: A Conceptual Framework for an Ontology-Based Examination System

Authors: Adekoya Adebayo Felix, Akinwale Adio Taofiki, Sofoluwe Adetokunbo

PAGE 36 – 42

Paper 7: Design and Analysis of a Novel Low-Power SRAM Bit-Cell Structure at Deep-Sub-Micron CMOS Technology for Mobile Multimedia Applications

Authors: Neeraj Kr. Shukla, R.K.Singh, Manisha Pattanaik

PAGE 43 – 49

Paper 8: Generating Performance Analysis of GPU compared to Single-core and Multi-core CPU for Natural Language Applications

Authors: Shubham Gupta, Prof. M.Rajasekhara Babu

PAGE 50 – 53

Paper 9: Main Stream Temperature Control System Based on Smith-PID Scheduling Network Control

Authors: Jianqiu Deng, Haijun Li, Zhengxia Zhang

PAGE 54 – 58

Paper 10: FPGA-Based Design of High-Speed CIC Decimator for Wireless Applications

Authors: Rajesh Mehra, Rashmi Arora

PAGE 59 – 62

Paper 11: Implementation and Performance Analysis of Video Edge Detection System on Multiprocessor Platform

Authors: Mandeep Kaur, Kulbir Singh

PAGE 63 – 70

Paper 12: A robust multi color lane marking detection approach for Indian scenario

Authors: L N P Boggavarapu, R S Vaddi, K R Anne, H D Vankayalapati, J K Munagala

PAGE 71 – 75

Paper 13: A Comprehensive Analysis of Materialized Views in a Data Warehouse Environment

Authors: Garima Thakur, Anjana Gosain

PAGE 76 – 82

Paper 14: A Routing Scheme for a New Irregular Baseline Multistage Interconnection Network

Authors: Mamta Ghai

PAGE 83 – 86

Paper 15: Application of Fuzzy Logic Approach to Software Effort Estimation

Authors: Prasad Reddy P.V.G.D, Sudha K. R, Rama Sree P

PAGE 87 – 92

Paper 16: A Load Balancing Policy for Heterogeneous Computational Grids

Authors: Said Fathy El-Zoghdy

PAGE 93 – 100

Paper 17: Generating PNS for Secret Key Cryptography Using Cellular Automaton

Authors: Bijayalaxmi Kar, D.Chandrasekhra Rao, Dr. Amiya Kumar Rath

PAGE 101 – 105

Paper 18: Computerised Speech Processing in Hearing Aids using FPGA Architecture

Authors: V. Hanuman Kumar, Prof. P. Seetha Ramaiah

PAGE 106 – 112

Paper 19: A Neural Approach for Reliable and Fault Tolerant Wireless Sensor Networks

Authors: Vijay Kumar, R. B. Patel, Manpreet Singh, Rohit Vaid

PAGE 113 – 118

Paper 20: Deploying an Application on the Cloud

Authors: N. Ram Ganga Charan, S. Tirupati Rao, Dr .P.V.S Srinivas

PAGE 119 – 125

Paper 21: Radial Basis Function For Handwritten Devanagari Numeral Recognition

Authors: Prerna Singh, Nidhi Tyagi

PAGE 126 – 129

Paper 22: The Multi-Liaison Algorithm

Authors: Ms. Anjali Ganesh Jivani, Ms.Amisha Hetal Shingala, Dr. Paresh. V. Virparia

PAGE 130 – 134

Paper 23: A Performance Study of Some Sophisticated Partitioning Algorithms

Authors: D.Abhyankar, M.Ingle

PAGE 135 – 137

Using Merkle Tree to Mitigate Cooperative Black-hole Attack in Wireless Mesh Networks

Shree Om
Department of Computer Science
University of Botswana
Gaborone, Botswana

Mohammad Talib
Department of Computer Science
University of Botswana
Gaborone, Botswana

***Abstract*— Security is always a major concern and a topic of hot discussion to users of Wireless Mesh Networks (WMNs). The open architecture of WMNs makes it very easy for malicious attackers to exploit the loopholes in the routing protocol. Cooperative Black-hole attack is a type of denial-of-service attack that sabotages the routing functions of the network layer in WMNs. In this paper we have focused on improving the security of one of the popular routing protocols among WMNs, Ad-hoc on demand distance vector (AODV) routing protocol and present a probable solution to this attack using Merkle hash tree.**

Keywords- WMN, MANET; Cooperative black-hole attack; AODV; Merkle tree; malicious; OWHF.

I. INTRODUCTION

A black-hole attack is a network layer denial-of-service (DoS) attack that exploits the route discovery process of on-demand routing protocols. The network layer of WMN defines how interconnected networks (inter-networks) function. The network layer is the one that is concerned with actually getting data from one computer to another even if it is on a remote network. It is at this layer that the transition really begins from the more abstract functions of the higher layers – which do not concern themselves as much with data delivery – into specific tasks required to get its data to the destination. Its job is to provide a best-efforts (i.e., not guaranteed) way to transport data-grams from source to destination, without regard to whether these machines are on the same network or whether there are other networks in between them. This whole communication is made possible through the use of Internet Protocol (IP). IP is the primary network protocol used on the internet. IP is a connectionless protocol. IP supports unique addressing for computers on a network. Most networks use the IP version 4 (IPv4) standard that features IP addresses four bytes (32 bits) in length. The newer IP version 6 (IPv6) standard features addresses 16 bytes (128 bits) in length. All in all, the point that we are trying to make here is that IP is the network layer protocol that holds the whole internet together and intruders love to interrupt the functions of this layer. You can imagine the seriousness of the damage caused if an intruder is able to sabotage the functions of this layer.

The aim of this paper is to reflect light on Cooperative Black-hole attack, a serious form of Black-hole attack and the challenges with its security mechanisms. Section II presents threats and attacks at network. Section III takes a look at Cooperative Black-hole attack, section IV presents related works against the attack and challenges, section V gives the problem statement, section VI, gives background on the hashing tree, section VI discusses a probable solution, section VIII highlights expected results and section IX discusses future work.

II. THREATS AND ATTACKS AT NETWORK LAYER

We expect a secured WMN to have accomplished objectives such as confidentiality, integrity, availability, authenticity, non-repudiation, authorization and anonymity. In this section, some of the most critical threats and attacks present at network layer are discussed.

A. Black-hole attack:

In this attack, the malicious node always replies positively to a route request from a source node although it may not have a valid route to the destination and will always be the first to reply to the route request message. Therefore, all the traffic from the source node will be directed toward the malicious node, which may drop all the packets, resulting in DoS [1].

B. Wormhole attack:

To launch this attack, an attacker connects two distant points in the network using a direct low latency communication link called the wormhole link. Once the wormhole-link is established, the attacker captures wireless transmission on one end, sends then through the wormhole link, and replays them at the other end [8]. Then the attacker starts dropping packets and cause network disruption. The attacker can also spy on the packets going through, use the information gained to launch new attacks, and thus compromise the security of the network.

C. Sink-hole attack:

In this attack, a malicious node can be made very attractive through the use of powerful transmitters and high-gain antennas to the surrounding nodes with respect to the routing algorithm [15].

D. Sybil Attack:

This attack is defined as a “malicious device illegitimately taking on multiple identities” [4]. This attack abuses the path diversity in the network used to increase the available bandwidth and reliability. The malicious node creates multiple identities in the network. The legitimate nodes, assuming these identities to be distinct network nodes, will add these identities in the list of distinct paths available to a particular destination thus including the malicious node on path of a data, which can affect packet transfer as well as drop them. But, Even if the malicious node does not launch any attack, the advantage of path diversity is diminished, resulting in degraded performance [15].

A summary table has been drawn below that show the comparison of the above mentioned attacks based on the following properties:

1) *Type of attack:* Four types: Masquerade, Replay, Modify, and DoS attack [12].
2) *Type of attacker:* internal attacker or external attacker.
3) *Required knowledge:* The amount of information needed to be gathered or collected from the network in order to effectively perform the attack.
4) *Cost:* The cost of running an attack, not necessarily economic, but also measured in terms of resources or time requirements.
5) *Detectability:* An attack on the network layer or routing protocols is desired to be as less detectable as possible.

TABLE I. COMPARISON OF ATTACKS AT NETWORK LAYER

Attack	**Type of Attacker**	**Type of Attack**	**Required Knowledge**	**Cost**	**Detectability**
Black-hole	Insider	DoS	Low	Low	High
Wormhole	Insider & Outsider	Modify & Dos & Replay	High	High	Low
Sink-hole	Insider	Modify & DoS	Medium	Medium	Low
Sybil	Insider	Masquerade & DoS	Low	Medium	Low

As we see in the table, a black-hole attack will be favoured by most attackers because any attacker whose intentions are to bring down the whole network communication at a low cost with least amount of information about the network can carry out this attack. The detectability is surely higher than other attacks and that is why a more complex form of Black-hole attack called Cooperative Black-hole attack which is hard to detect, is being carried out by attackers. The next section takes a look in to this form of Black-hole attack.

III. COOPERATIVE BLACK-HOLE ATTACK

Since, WMNs share common features with the wireless ad-hoc networks, the routing protocols developed for MANETs can be applied to WMNs. In this paper, we focus on AODV and we explain operation of black-hole and cooperative black-hole attack by using AODV as an example protocol.

Black-hole attack is a type of active attack. These attacks involve some modification of the data stream or the creation of a false stream [12]. Figure 1 below show a simple scenario of this attack with one malicious node.

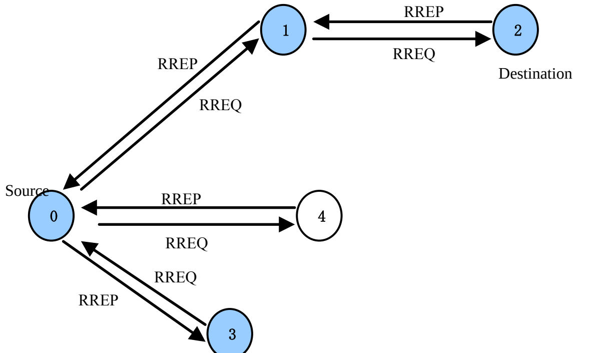
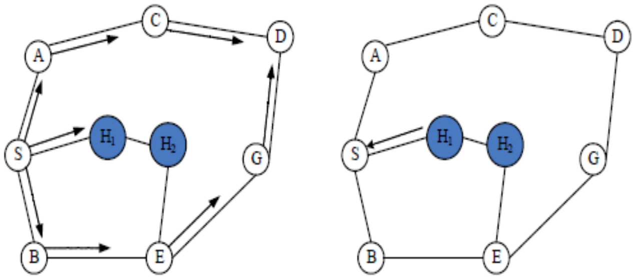


Figure 1. Black-hole attack in progress

The core functionality of WMNs is the routing capability and attackers take advantage of the shortcomings as the routing protocol has some loop holes. The AODV protocol is vulnerable to the well-known black hole attack. AODV uses sequence numbers to determine the freshness of routing information and to guarantee loop-free routes. In case of multiple routes, a node selects the route with the highest sequence number. If multiple routes have the same sequence number, then the node chooses the route with the shortest hop count. A malicious node sends Route Reply (RREP) messages without checking its routing table for a fresh route to a destination. As shown in fig. 1 above, source node 0 broadcasts a Route Request (RREQ) message to discover a route for sending packets to destination node 2. A RREQ broadcast from node 0 is received by neighbouring nodes 1, 3 and 4. However, malicious node 4 sends a RREP message immediately without even having a route to destination node 2. A RREP message from a malicious node is the first to arrive at a source node. Hence, a source node updates its routing table for the new route to the particular destination node and discards any RREP message from other neighbouring nodes even from an actual destination node. Once a source node saves a route, it starts sending buffered data packets to a malicious node hoping they will be forwarded to a destination node. Nevertheless, a malicious node (performing a black hole attack) drops all data packets rather than forwarding them on.

A more complex form of the attack is a Co-operative Black Hole Attack where multiple malicious nodes collude together resulting in complete disruption of the routing and packet forwarding functionality of the network. For example, in figure 2, when multiple black hole nodes are acting in coordination with each other, the first black hole node H1 refers to one of its team-mates H2 as the next hop. According to the proposed methods in [3], the source node S sends a further request message to ask H2 if it has a routing to node H1 and a routing to destination node D. Because H2 is cooperating with H1, its further reply is “yes” to answer both the questions. So source node S starts passing the date packets. Unfortunately, in reality, the packets are abstracted by node H1 and the security of the network is compromised [10].



(a) Network Flooding of RREQ (b) Propagation of RREP message

Figure 2. Figure 2. Co-operative Black-hole attack [16]

IV. Related Works and Challenges

AODV does not incorporate any specific security mechanism, such as strong authentication. Therefore, there is no straightforward method to prevent mischievous behaviour of a node such as media access control (MAC) spoofing, IP spoofing, dropping packets, or altering the contents of the control packets.

Solutions have been proposed to mitigate black-hole nodes in [2, 3, 10]. However, the solution, which are designed for MANETs consider malicious nodes that work alone, i.e., each node is an attacker, and do not target attackers working in groups. For example, method proposed in [3] can help mitigate individual node attack because it requires the intermediate node to include information about the next hop to destination in the RREP packet. Then after the source node has received this packet, it sends a further route request (FREQ) to the next hop node asking if the node has route to the destination. Now, if this next hop node has been working together with the malicious node, then it will reply “yes” to the FREQ and the source node will transmit the packet to the malicious node that sent the first reply which is a black-hole node. A solution to defending cooperative black-hole attacks was proposed in [10] but no simulations or performance evaluations had been done. The methodology uses the concept of Data Routing Information (DRI) table and cross-checking further request (FREQ) and further reply (FREP). [14] have used the algorithm proposed by [10] and modified it slightly to improve the accuracy of preventing the attack and efficiency of the process and simulated the new modified algorithm. The solution has been proposed for MANETs which are usually mobile devices powered by battery. The maintenance of DRI increases overhead and cross-checking delays the communication process which in-turn drains more battery power. However [14] have compared their results with [3] and proved that their method is more efficient and accurate. Two authentication mechanism for identifying multiple black hole nodes cooperating as a group in MANETs is proposed by [16]. The mechanism is based on the assumption that no other authentication mechanism such as a Public Key Infrastructure (PKI) is present which is usually not practical in MANETs. The source node checks the RREP messages to determine the data packets to pass with the authentication mechanisms proposed in [16]. However, the question that arises is, how will this authentication mechanism be protected from malicious nodes that might forge the reply if the hash key of any node is to be disclosed to all nodes. In [13], authors propose an enhancement of the basic AODV routing protocol to combat the cooperative black hole attack in MANET. They use a structure, which they call fidelity table wherein every participating node will be assigned a fidelity level that acts as a measure of reliability of that node. In case the level of any node drops to 0, it is considered to be a black hole node and is eliminated. In their approach, they assume that nodes are already authenticated which is a little strong assumption. [2] present a solution to avoid single node and co-operative Black-hole attacks in a MANET based on the principle of Merkle tree. However, factors such as network density, nodes mobility and the number of black hole nodes which are determining factors in a solutions performance, in term of end to end delay and network load, were not considered.

V. Problem Statement

The state-of-the-art work is still insufficient for deploying sizeable WMNs because important aspects such as security still remain open problems. Cooperative black-hole attack is a severe denial-of-service attack routing protocol threat, accomplished by dropping packets, which can be easily employed against routing in Wireless Mesh Networks, and has the effect of making the destination node unreachable or downgrade communications in the network. The black holes are invisible and can only be detected by monitoring lost traffic. The emergence of new applications of WMNs necessitates the need for strong privacy protection and security mechanisms of WMNs. The AODV, our case study protocol, does not have any security mechanisms and malicious nodes can perform many attacks by taking advantage of the loopholes in the protocol. The next section proposes a solution to prevent Cooperative black-hole attack in hybrid WMNs.

A solution is proposed by [2] to black-hole and cooperative black-hole attack in MANETs based on the principle of Merkle tree but has challenges. Our solution uses its fundamentals and makes modifications to address these challenges and helps mitigate Cooperative black-hole attack in hybrid WMNs. Before we get in to the description of the solution, we would like to give a brief background of Merkle Tree.

VI. MERKLE TREE

Also called Merkle hash tree (MHT) is a binary tree relies on the properties of one way hash functions (OWHFs) [7]. A sample MHT is shown in figure 3.

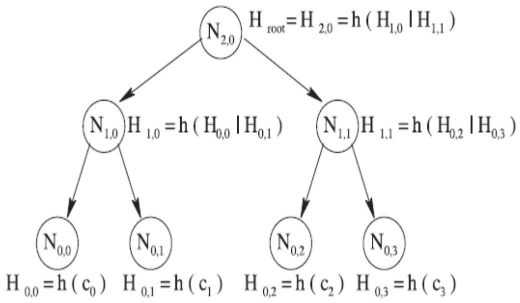

$N_{2,0}$ $H_{root} = H_{2,0} = h(H_{1,0} | H_{1,1})$
$N_{1,0}$ $H_{1,0} = h(H_{0,0} | H_{0,1})$
$N_{1,1}$ $H_{1,1} = h(H_{0,2} | H_{0,3})$
$N_{0,0}$
$N_{0,1}$
$N_{0,2}$
$N_{0,3}$
$H_{0,0} = h(c_0)$
$H_{0,1} = h(c_1)$
$H_{0,2} = h(c_2)$
$H_{0,3} = h(c_3)$

Figure 3. A sample MHT [7]

- N_{ij} denotes the nodes within the MHT where *i* and *j* represent, respectively, the *i*-th level and the *j*-th node.
- H_{ij} denotes the cryptographic variable.
- *h* denotes a one way hash function e.g. the function SHA-1 [6].
- | is the concatenation operator.
- Nodes at level 0 are called "leaves".
- Each leaf carries a given value e.g. $h(C_0)$, $h(C_1)$, $h(C_2)$ and $h(C_3)$ in Fig. 3.
- The value of an interior node (including the root) is a one-way hash function of the node's children values e.g. value of interior node $N_{1,0}$ is: $h(\ H_{0,0} \mid H_{0,1})$ which is the hashing result of the concatenation of values of children $N_{0,0}$ and $N_{0,1}$.

VII. The Proposed Solution

Table 3 contains the notations used to describe the solution.

TABLE II. NOTATIONS

Notation	Significance
IDi	Identity of node i.
Si	Secret generated by node i.
h	OWHF
|	Concatenation operator

In figure 4, we consider a piece of network made up of 4 nodes A, B, C and D. On this last, a Merkle tree is juxtaposed. We point out that our goal is to check that B and C conveys well, towards D, the traffic sent by A.

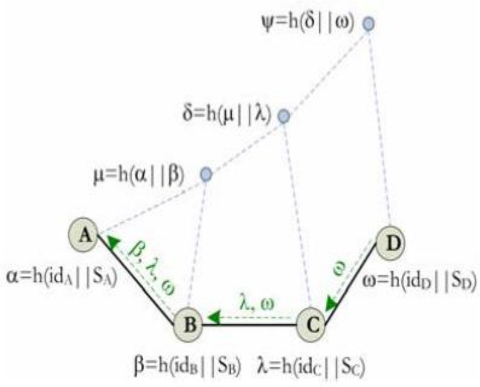


Figure 4. Basic principle of the solution [2].

Node A is source node and has the value ψ (value of the root of the Merkle tree). Each node *i* holds the value $h(id_i \mid S_i)$. So as per method proposed by [2], if A has to send data to D through B and C, in order to make sure that B and C are not cooperating as black hole nodes D sends ω (value held by D) to C, then C sends λ and ω to B which in turn sends β, λ and ω to A. A then recalculates ψ from α, β, λ and ω, then compares the result with the value ψ of already held, if equality, the route (A,B,C,D) is secured, otherwise, the route contains a black hole node.

Nodes B and C can cooperate to conduct black hole attack, this is easy if D communicates to C its secret S_D based on trust and since C is cooperating with B, it will pass the secret S_D to B so that it can calculate ω. [2] have not addressed this problem as how to protect the secret S_i from being compromised. This could create problems in dense network. Our solution adds to [2]. When A requests for a route to D, B being an attacker replies with the highest sequence number to A. According to AODV, A would discard other RREPs. Our solution looks to modify AODV such that it records the second best RREP from the node claiming to have a route to D. We assume that this node is safe. We call this node X. Node X has the value θ which is equal to $h(id_x \mid S_x)$. Since B already has β, λ and ω, it forwards all these values to A without any further communication with C (assumption). We introduce change of secret on the source and destination node whenever there is a request for the hash value. That means that when D sends ω to node X, it is a completely different value from the value of B. But B did not even communicate with C or D. Similarly, A would hold a new ψ and new α. Now, when A recalculates ψ from new α, β, λ and ω, then compares the result with the new value of ψ, it would be different but when A recalculates ψ from new α, θ and new ω and compares the result with the value ψ, they will be same. This would mark node B to be a malicious node and it will be black listed from future communication. At the same time a update will be sent with the packet to D through node X informing it of the malicious behavior of B. D will black list node C because it never received any RREQ from it because B never communicated with C which should not have been the case if both nodes were trusted.

The steps below give a rough idea of how the solution will work assuming node D has already shared its secret with node C and node C has forwarded is secret to node B along with secret of node D.

Step 1: Source node A sends RREQ for destination D.
Step 2: Source node A updates its value of ψ and generates new secret for itself.
Step 3: Intermediate node B sends RREP with highest sequence number.
Step 4: Node A stores this information.
Step 5: RREP from node X is received after RREP from node B.
Step 6: Instead of discarding this RREP, node A temporarily stores this information
Step 7: In order to prove legitimacy, node B and node X have to send the hash values including that of destination D.
Step 8: Node X requests for ω from node D.
Step 9: Node D generates new secret, recalculates new ω and passes it to node X.
Step 10: Node X passes θ and new ω to node A.
Step 11: Node B passes β, λ and old ω (calculated on the basis of secret of D sent by node C).
Step 12: Node A recalculates two values of ψ, ψ1 based on values from node B and ψ2 based on values from node X.

Step 13: Node A compares $\psi 1$ and $\psi 2$ to already held new value ψ.
Step 14: Node A discovers $\psi 1$ is not the same as new ψ.
Step 15: Node A black lists node B.
Step 16: Node A sends packet to node X to be delivered to node D with attached information about node B.
Step 17: Node D receives packet.
Step 18: Node D black lists node B.
Step 19: Node D black lists node C based on assumption that node B was able to calculate ω because node C must have shared secret with node B, hence making node C an untrusted node.
Step 20: Node D sends acknowledgement (ACK) packet to node A including information about untrustful behaviour of node C.
Step 21: Node A updates its list of black listed nodes and adds node C to it.

VIII. Expected Results

The method would successfully identify the colluding malicious nodes and when compared with other proposed methods would have

- Better packet delivery ratio (PDR) – the number of packets generated by the sources vs. the number of packets received at the destination.
- Reduced detection time - This is the time to detect the network, which has a black hole attack, measured by the attack detection time minus the traffic start time [14].
- Better average end-to-end delay - this is the average delay between the sending of the data packet by the source and its receipt at the corresponding receiver and includes all the delays caused during route acquisition, buffering and processing at intermediate nodes, retransmission delays at the MAC layer, etc [9]. It is measured in milliseconds.
- Reduced routing overhead – ratio of number of control packets generated to the data packets transmitted.

IX. Future Work

In this paper we have studied the routing security issues of WMNs, described the Cooperative black hole attack that can be mounted against a WMN and proposed a possible solution for it in the AODV protocol. The proposed solution can be applied to identify multiple black hole nodes cooperating with each other in a hybrid WMN. As future work, we intend to develop concrete algorithms and simulations to analyze the performance of the proposed solution based on network density, nodes mobility and the number of black hole nodes.

References

[1] I. Aad, P. J. Hubaux, W. E. Knightly, “Impact of denial-of-service attacks on ad-hoc networks,” IEEE/ACM Trans. Net. USA, vol. 16, iss. 4, pp. 791-802, August 2008.

[2] A. Baddache, A. Belmehdi, “Avoiding black hole and cooperative black hole attacks in wireless ad hoc networks,” Int. J. of Comp. Sc. and Info. Sec. (IJCSIS). USA, vol. 7, iss. 1, pp. 10-16, January 2010.

[3] H. Deng, W. Li, P. D. Agarwal, “Routing security in wireless ad-hoc networks,” IEEE Comm. Mag. USA, vol. 40, iss. 10, pp. 70-75, December 2002

[4] M. Imani, M. E. Rajabi, M. Taheri, M. Naderi, “Vulnerabilities in network layer at WMN,” Int. Conf. on Ed. and Net. Tech. China, pp. 487-492, June 2010.

[5] J. Yin, S. Madria, “A hierarchical secure routing protocol against black hole,” IEEE Int. Conf. on Sensor Net., Ubiquitous, and Trustworthy Computing. Taiwan, vol. 1, pp. 376-383, June 2006.

[6] R. C. Merkle, “A certified digital signature,” Advances in Crypt. (CRYPTO89) USA, pp 218-238, August 1989.

[7] L. J. Munoz, J. Forne, O. Esparaza, M. Soriano, “Certificate revocation system implementation based on the Merkle hash tree,” Int. J. of Info. Sec. Heidelberg, vol. 2, iss. 2, pp. 110-124, January 2004.

[8] F. Nait-Abdesselam, B. Bensaou, T. Taleb, “Detecting and avoiding wormhole attacks in wireless ad-hoc networks,” IEEE Comm. Mag. Canada, vol. 4, iss. 64, pp. 127-133, April 2008.

[9] S. S. Ramaswami, S Upadhyaya, “Smart handling of colluding black hole attacks in MANETs and wireless sensor networks using multipath routing,” IEEE workshop on Info. Assurance. USA, pp. 253-260, July 2006.

[10] S. Ramaswamy, H. Fu, M. Sreekantaradhya, J. Dixon, K. Nygard, “Prevention of cooperative black hole attack in wireless ad hoc networks,” Int. Conf. on Wireless Net. USA, vol. 1, pp. 570-575, June 2003.

[11] S. M. Siddiqui, S. C. Hong, “Security issues in wireless mesh networks,” IEEE Inter. Conf. on Multimedia and Ubiquitous Eng. (MUE'07). South Korea, vol. 1, pp. 717-722, April 2007.

[12] W. Stallings, Network security essentials: Applications and Standards. New Jersey, USA: Prentice Hall, 2003

[13] L. Tamilselvan, V. Sankarnarayanan, “Prevention of blackhole attack in MANET,” Int. Conf. on Wireless Broadband and Ultra Wideband Comms. Australia, vol. 2, pp. 21, April 2007.

[14] H. Weerasinghe, H. Fu, “Preventing cooperative black hole attacks in mobile ad hoc networks: simulation implementation and evaluation,” Future Gen. Comm. and Networking (FGCN). South Korea, pp. 362-367, December 2007.

[15] V. Zhang, J. Zheng, H. Hu, Security in wireless mesh networks. Florida, USA: Auerbach Publications, 2009

[16] M. Zhao, J. Zhou, “Cooperative black hole attack prevention for mobile ad hoc networks,” IEEC '09 Proceedings of the 2009 Int. Symp. on Info. Engg. and Elec. Commerce. USA, vol. 1, pp. 26-30, May 2009

[17] Bhakthavathsalam, R., Shashikumar, R., Kiran, V., & Manjunath, Y. R. (2010). Analysis and Enhancement of BWR Mechanism in MAC 802 . 16 for WiMAX Networks. International Journal of Advanced Computer Science and Applications - IJACSA, 1(5), 35-42.

[18] Prasad, D. (2011). A Reliable Security Model Irrespective of Energy Constraints in Wireless Sensor Networks. International Journal of Advanced Computer Science and Applications - IJACSA, 2(4), 20-29.

[19] Jaigirdar, F. T. (2011). Grid Approximation Based Inductive Charger Deployment Technique in Wireless Sensor Networks. International Journal of Advanced Computer Science and Applications - IJACSA, 2(1).

AUTHORS PROFILE

Shree Om received his B.Sc. degree in computer engineering from University of South Alabama, USA in 2007. He is currently pursuing M.Sc. in Information Systems at University of Botswana, Botswana. His research interests are in the networking field particularly in mesh networking. He has attended several workshops conferences locally and internationally.

Professor Mohammad Talib has, presently, been associated with the Computer Science Department of the University of Botswana and has also been an adjunct professor at a couple of Universities in the United States. 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 Science. He has an excellent industrial relevance and has worked as Software Engineer in the Silicon Valley at 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 & Engineering 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, M.Sc. in Statistics and a 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 about eighty 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, conferences etc. in various capacities. He is conferred upon an honorary doctorate in computer science and engineering by the Yorker International University, USA

QVT transformation by modeling

From UML Model to MD Model

I.Arrassen
Laboratory for Computer Science Research
Faculty of Sciences
Mohammed First University
524, Oujda, Morocco

A.Meziane
Laboratory for Computer Science Research
Faculty of Sciences
Mohammed First University
524, Oujda, Morocco

R.Sbai
Laboratory of Applied Mathematics and Computer Signal Processing
Superior School of Technology
Mohammed First University
524, Oujda, Morocco

M.Erramdani
Laboratory of Applied Mathematics and Computer Signal Processing
Superior School of Technology
Mohammed First University
524, Oujda, Morocco

***Abstract* – To provide a complete analysis of the organization, its business and its needs, it is necessary for leaders to have data that help decision making. Data warehouses are designed to meet such needs; they are an analysis and data management technology. This article describes an MDA (Model Driven Architecture) process that we have used to automatically generate the multidimensional schema of data warehouse. This process uses model transformation using several standards such as Unified Modeling Language, Meta-Object Facility, Query View Transformation, Object Constraint Language, ... From the UML model, especially the class diagram, a multidimensional model is generated as an XML file, the transformation is carried out by the QVT (Query View Transformation) language and the OCL (Object Constraint Language) Language. To validate our approach a case study is presented at the end of this work.**

Key-Words:* *Datawarehouse; Model Driven Architecture; Multidimensional Modeling; Meta Model; Transformation rules; Query View Transformation.

I. Introduction

To support the process of making management decisions, development of data warehouses is important for organizations. According to the definition given by Bill Inmon (1996), data warehouse is a collection of data that is subject-oriented, integrated, time-varying and non-volatile. His ultimate goal is integrating data from all corners of the enterprise in a single directory, from which users can easily find answers to queries, generate reports and perform analysis.

A data warehouse is a management and data analysis technology. On this basis, the establishment of a process of building data warehouse is very important. Through this work, we use a UML class diagram summarizing the activities: requirements expression, analysis and design of information system of the organization. From this diagram, we will generate future objects decision diagram such as facts and dimensions. The decisional diagram will be in the form of a multidimensional schema, in fact, multidimensional modeling is the one that best represents the data warehouse schema.

The approach used in this work is the MDA. A models transformation process is used to transform a UML model (Class Diagram) into a multidimensional model, the QVT and the OCL languages was chosen as implementation language processing.

In Section 2, we will discuss the works that are related to our theme. In section 3 we explain the concepts of multidimensional modeling. In Section 4 we outline the main concepts of the MDA architecture (Model Driven Architecture), which is at the heart of the approach followed in our work. Then in Section 5 we present the source and target Meta models used in the transformation program. In Section 6, we present the work of generating the multidimensional model. A case study is presented in Section 7. We conclude this work in Section 8, with suggestions and possible extensions.

II. Related Works

In recent years, several approaches to developing data warehouse have been proposed. In this section, we present a brief overview of some best known approaches.

In [1] several DWH case studies are presented. DWH design is based on using the star schema and its variants (snowflake schema and fact in constellation) using a relational approach: tables, columns, foreign keys and so on. However this work is considered as a benchmark in the field of DWH, the authors are only interested in the representation of relational DWH, and they regarded any other technology.

In [2], the authors propose a model Fact-Dimension (Dimensional-Fact Model DFM), in which they define a special notation for the conceptual model of the DWH. They also show how to derive a DWH schema from the data sources described by the entity-relationship diagram.

A goal-oriented approach has been added to DFM in [3]. This approach does not consider important aspects such as the ETL (Extract Transform Load) process. In addition, the authors consider that relational schemas of data exist, which is not always true, in addition, the use of special notations makes it difficult to apply this approach.

In [4], a model for the design of the DWH was proposed: YAM2. It is an object oriented model that uses UML notation to represent the multidimensional structure of data. However, no method that shows how to get the conceptual model is described in the paper.

In [5] the authors propose a multidimensional meta-model extended by generalizing the model heart, based on medical data, this work was done under the MAP project (Personalized Medicine Anticipation). In our opinion this type model is specific to medical data; in addition, the authors do not specify the modeling approach.

In [6] the authors created a decision support system called BIRD integrated data grid Decrypthon which is designed to process data in functional genomics. This system is a specific platform that does not meet all requirements of any DWH.

III. Multidimensional modeling

It is a logical design technique that aims to present data in a standard, intuitive way, which allows high performance access.

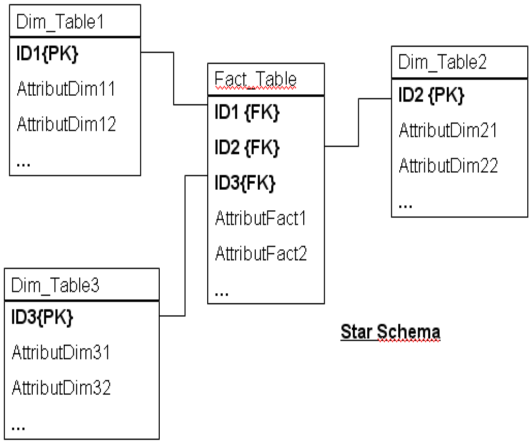


Figure 1. Dimensionnel Model Schema

IV. MDA Architecture

In November 2000, the OMG, a consortium of over 1,000 companies, initiates the MDA approach [7]. The purpose of this standard is to separate the business logic of the enterprise, from any technical platform. It is a new way to design applications. Indeed, the technical architecture is unstable and undergoes many changes over time, unlike the business logic. It is therefore easy to separate the two to face the increasing complexity of information systems and high costs of technology migration. This separation then allows the capitalization of software knowledge and the know-how of the company.

Figure 2 below shows schematically the architecture of MDA [7]. It is divided into four layers. The OMG was based on several standards. In the center are the standard UML (Unified Modeling Language) [8], MOF (Meta-Object Facility) [9] and CWM (Common Warehouse Meta-model) [10].

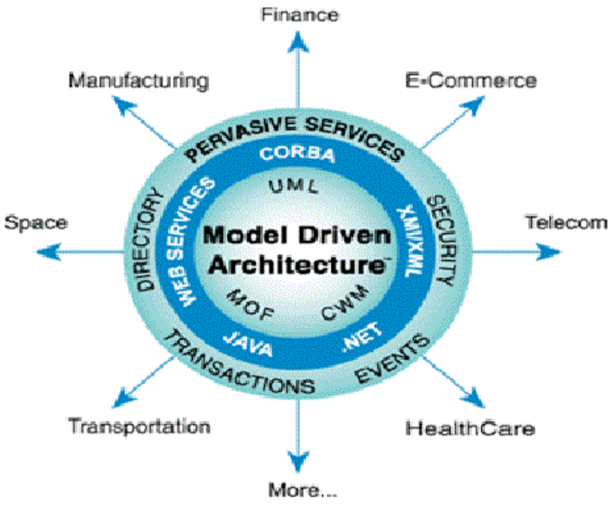


Figure 2. MDA Architecture [7]

In the next layer, it is also a standard XMI (XML Metadata Interchange), which allows communication between the middlewares (Java, CORBA, .NET and Web Services). The third layer contains the services that manage events, security, directories, and transactions. The final layer offers specific frameworks in fields (Finance, Telecommunications, Transportation, Space, Medicine, Commerce, Manufacturing ...)

A. CIM (Computation Independent Model)

CIM stands for Computation Independent Model [12]. In UML, a requirements model can be summarized as a use case diagram. Because they contain the functionality provided by the application and the various entities that interact with them (actors) without providing information on the operation of the application. The role of requirements models in an MDA approach is the first models to be perennial. The modeled requirements provide a contractual basis validated by the customer and vary little.

B. PIM (Platform Independent Model)

The PIM represents the business logic specific to the system or the design model. It represents the operating entities and services. It must be perennial and last over time. It describes the system, but shows no details of its use on the platform. At this level, the formalism used to express a PIM is a class diagram in UML, which can be coupled with a constraint language like OCL (Object Constraint Language). Analysis models and design are independent of platforms where they are implemented J2EE, .NET, PHP, etc.. [12].

C. PSM (Platform Specific Model)

The PSM is the work of code generation after performing the analysis models and design. This phase, the most delicate of MDA, must also use templates. It includes the application of design patterns techniques [12]. PSM code models facilitate the generation of code from a model analysis and design. They contain all information necessary to operate an execution platform, such as information systems to manipulate file systems or authentification systems.

D. OCL (Object Constraint Language)

OCL [11] was developed in 1997 by Jos Warmer (IBM), on the basis of language IBEL (Integrated Business Engineering Language). It was formally incorporated into UML 1.1 in 1999. It is a formal language that is based on the notion of constraint. A constraint is a boolean expression that can be attached to any UML element. It usually indicates a restriction or gives information on a model. The constraints are used in particular to describe the semantics of UML and its various extensions, participating in the definition of profiles.

E. MOF2.0 (MétaObject Facility) QVT

MOF [9] defines the structure that should have any meta-model. A meta-model is a set of meta-classes with meta-associations. MOF2.0 meta-meta-model is unique, and UML 2.0 meta-model is dedicated to the modeling of object-oriented applications. Among the objectives of MOF2.0: capitalize on existing commonalities between UML and MOF-level class diagrams and to explain the differences. The advantage of the MOF is to structure all meta-models in the same way.

QVT (Query / View / Transformation) in the MDA architecture is a standard for model transformations defined by the OMG (Object Management Group) in 2007. It is central to any proposed MDA. It defines a way to transform source models to target models. These source and target models must conform to the MOF meta-model. Specifically, this means that the abstract syntax of QVT must conform to the MOF 2.0 meta-model. QVT defines three specific languages named: Relations, Core Operational/Mapping. These languages are organized in a layered architecture. Relations and Core are declarative languages to two different levels of abstraction. QVT Core is to relational, what is the Java byte code to Java source code. The QVT Operational / Mapping is an imperative language, it provides common constructions in imperative languages (loops, conditions ...).

F. Kinds of MDA transformation.

The MDA identifies several transformations during the development cycle [12]. It is possible to make four different types of transformations:

1) CIM to CIM
2) CIM to PIM
3) PIM to PIM
4) PIM to PSM

There are three approaches in MDA to perform these transformations:

- Approach by programming: using the object-oriented programming languages such as Java, to write computer programs that are unique to manipulate models. This approach was used in [13] and [14].The authors automatically generate a web application from a simplified class diagram.
- Approach by template: Consists of taking a "template model", canvas of configured target models, these settings will be replaced by the information contained in the source model. This approach requires a special language for defining model template.
- Approach by Modeling: The objective is to model the transformations of models, and make transformation models sustainable and productive and to express their independence of execution platforms. The standard MOF2.0 QVT [9] is used to define the meta model for the development of models for the models transformations.

In this paper we chose to transform PIM to PSM, with an approach by modeling. This type of transformation will allow us to automatically generate the multidimensional data warehouse schema from UML schema. Indeed, as shown in Figure 2, the definition of model transformation is a model structured according to the meta-model MOF2.0 QVT. Models instances of meta-model MOF2.0 QVT express structural mapping rules between meta-model source and meta-model target of transformation. This model is a sustainable and productive, it must be transformed to allow the execution of processing on an execution platform. The following figure illustrates the approach by modeling MOF2.0 QVT [12].

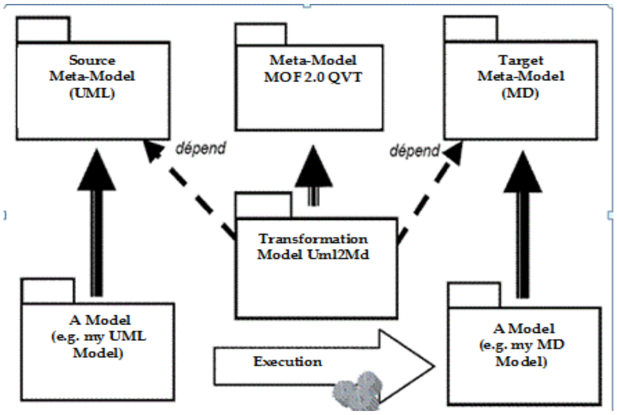


Figure 3. Modeling approach (MOF2.0 QVT)

V. Source Meta Model - Target Meta Model

In our MDA approach, we opted for the modeling approach to generate the data warehouse multidimensional schema. As mentioned above, this approach requires a source meta-model and a target meta-model.

A. Source Meta-Model

We present in this section the different meta-classes which form the source UML meta model used in [13], and our target meta-model MD to achieve the transformation between the source and target model. This transformation is based on that proposed by [13]. The source meta-model is a simplified UML model based on packages containing Class and Datatype. These classes contain typed properties and are characterized by multiplicities (upper and lower). The classes contain operations with typed parameters. The following figure illustrates the source meta-model:

- UmlPackage: expresses the notion of UML package. This meta-class is related to the meta-class Classifier.

- Classifier: it is a meta abstract class that represents both the concept of UML class and the concept of data type.
- Class: represents the concept of UML class.
- DataType: represents the data type of UML.
- Operation: expresses the concept of operations of a UML class
- Parameter: parameter expresses the concept of an operation. They can be of two types or Class Datatype. This explains the link between meta-Parameter class and meta-class Classifier.
- Property: expresses the concept of Property of a UML class. These properties are represented by the multiplicities and meta-attributes upper and lower. A UML class is composed of properties, which explains the link between the meta-class Property and meta-class Class. These properties can be primitive type or class type. This explains the link between the meta-class and the Classifier meta-class Property.

B. *Target Meta-Model*

The target is a simplified dimensional meta-model based on packages containing Facts and Dimensions. A Fact contains attributes related to dimensions. Figure 5 shows the Target Meta Model:

- FactPackage: expresses the concept of package and we consider that each class contains a FactPackage.
- Fact: expresses the concept of Fact in a multidimensional schema. Each fact is contained in FactPackage, which explains the link between FactPackage and Fact.
- FactAttribute: expresses the concept of attributes for the element Fact and Fact contains FactAttribute
- Dimension: expresses the concept of dimension in a multidimensional schema.
- DimensionAttribute: expresses the concept of dimensions attributes. A dimension contains attributes, which explains the link between a DimensionAttribute and Dimension.
- Datatype: represents the type of data in a multidimensional schema.

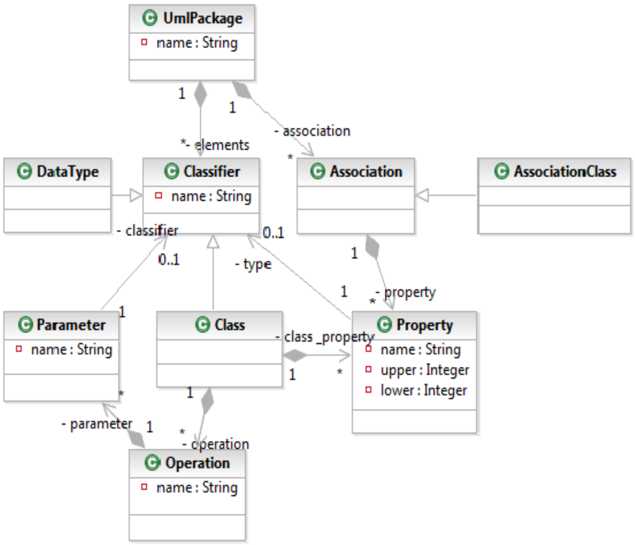


Figure 4. Simplified Meta-Model class diagram of UML.

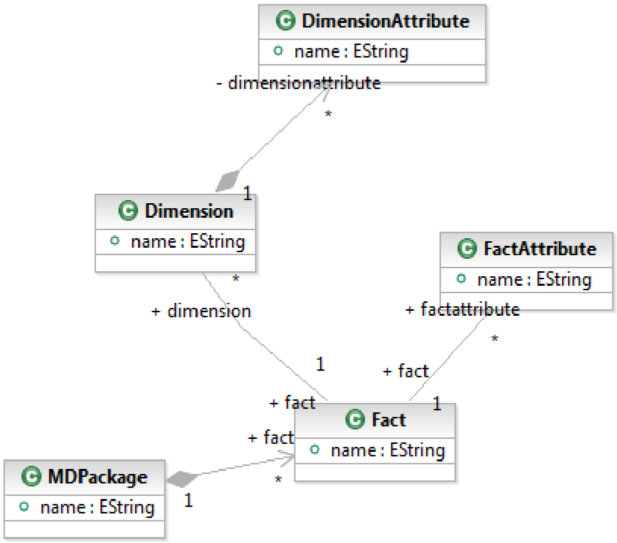


Figure 5. Simplified Meta-Model of multidimensional star schema.

VI. QVT TRANSFORMATION OF THE UML META MODEL TO THE MULTIDIMENSIONAL META MODEL.

A. *SmartQvt*

To achieve the transformation cited above, we opted for the tool SmartQvt. This tool appeared in 2007 [17] which is also an implementation of the QVT-Operational language standard that allows the transformation of models. This tool compiles model transformations specified in QVT to produce Java code used to performs these transformations. This tool is provided as Eclipse plug-ins based on the meta-modeling framework, EMF, and is composed of three elements:

- QVT Editor: helps end users to write QVT specifications.
- QVT Parser: converts the textual concrete syntax on its corresponding representation in terms of meta-model QVT.
- QVT Compiler: produces, from the QVT model, a Java API on top of EMF generated for the implementation of transformations. The input format is a QVT specification provides in XMI 2.0 format in accordance with the QVT meta-model.

B. The QVT transformation

The QVT transformation is a function of two arguments: the first is the source model, the second is the target model as shown in the figure below:

```
transformation uml2md(in srcModel:UML,out dest:MD);
main(){
srcModel.objects()[UmlPackage] ->
map UmlPackage2MDPackage(); }
```

Figure 6. Transformation function uml2md()

In this figure, we see the transformation function called uml2md() that takes as input source model, the simplified UML model shown in Figure 4, and as output the target model, the simplified multidimensional model shown in Figure 5.

As explained in the previous section, we consider a simplified UML model that consists of a package called UmlPackage. This package will be transformed into a multidimensional package called MDPackage, using the transformation rule sets in Figure 7. The MDPackage generated will be named MD followed by with the name of UmlPackage. The Association class will be transformed into a Fact by using the function associationclass2fait ().

```
mapping UmlPackage::UmlPackage2MDPackage () :
MDPackage { name := 'MD' + self.name;
fact:=srcModel.objects()[AssociationClass]->map
associationclass2fact();}
```

Figure 7. Transformation rule of UMLPackage to MDPackage

```
mapping AssociationClass::associationclass2fact () : Fact {
name := 'Fait'+self.name;
factattribute:=self.assoproperty->map
property2factattribute();
dimension= self.assoproperty->select (upper=-1)->type-
>map type2dimension(); }
```

Figure 8. Transformation rule of Associationclass to a Fact

Figure 8 shows the transformation rule of an association class in a UML class diagram to a Fact.

```
mapping Classifier::type2dimension () : Dimension {
srcModel.objects()[Class]->forEach(c){
    if(self=c){ name:='Dimension'+self.name ;
    base:= c._property->select (upper=-1)->type-> map
type2base();        } } }
```

Figure 9. Transformation rule of a Classifier to a Dimension

Figure 9 shows the transformation rule of Classifier to an element Dimension of Dimensional Model.

```
mapping Property::property2factattribute(): FactAttribute
    {   name:=self.name; }
```

Figure 10. Transformation rule of Property to FactAttribut

Figure 10 shows the transformation rule of UML class property to a Fact entity attribute, this rule uses the function property2factattribute ().

```
mapping Parameter::property2factattribute () :
FactAttribute {
    name:=self.name; }
```

Figure 11. Transformation rule of Parameter to FactAttribut

```
mapping Property::property2dimensionAttribute () :
DimensionAttribute
    {   name:=self.name;}
```

Figure 12. Transformation rule of Property to DimensionAttribut

VII. CASE STUDY

In our case study, we use the UML model representing a business, among its activities, the sale of products to its customers, these products are classified under categories, and several sub categories form a category. The UML model representing part of the Information System of the company is represented in a format file XMI2.0. This file is composed of the element UmlPackage, which contains the following elements:

```
<?xml version="1.0" encoding="ASCII"?>
<xmi:XMI xmi:version="2.0" xmlns:xmi="http://www.omg.org/XMI"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:UmlMdMM="http:///UmlMdMM.ecore">
<UmlMdMM:UmlPackage name="Factory">
<elements xsi:type="UmlMdMM:DataType" name="String"/>
<elements xsi:type="UmlMdMM:Class" name="Custommer">
<_property upper="1" lower="1" name="CustomerID"
type="//@elements.0"/>
<_property upper="1" lower="1" name="AccountNumber"
type="//@elements.0"/>
<_property upper="1" lower="1" name="CustomerType"
type="//@elements.0"/>
</elements>
<elements xsi:type="UmlMdMM:Class" name="Product">
<_property upper="1" lower="1" name="ProductID"
type="//@elements.0"/>
<_property upper="1" lower="1" name="Name"
type="//@elements.0"/>
<_property upper="1" lower="1" name="StandardCost"
type="//@elements.0"/>
<_property upper="1" lower="1" name="ListPrice"
type="//@elements.0"/>
<_property upper="1" lower="1" name="Weight"
type="//@elements.0"/>
<_property upper="1" lower="1" name="Style"
type="//@elements.0"/>
<_property upper="1" lower="1" name="Color"
type="//@elements.0"/>
<_property upper="1" lower="1" name="DayToManufacture"
type="//@elements.0"/>
</elements>
 <association xsi:type="UmlMdMM:AssociationClass"
name="SalesOrder">
 <assoproperty upper="1" lower="1" name="SalesOrderID"
type="//@elements.0"/>
 <assoproperty upper="1" lower="1" name="OrderQty"
type="//@elements.0"/>
<assoproperty upper="1" lower="1" name="UnitPrice"
type="//@elements.0"/>
<assoproperty upper="1" lower="1" name="OrderDate"
type="//@elements.0"/>
 <assoproperty upper="-1" lower="1" name="CustomerID"
type="//@elements.1"/>
<assoproperty upper="-1" lower="1" name="ProductID"
type="//@elements.3"/>
  </association>
  </UmlMdMM:UmlPackage>
</xmi:XMI>
```

Figure 13. Overview of the instance of the UML model of our case study

An element named Customer, type of class, which has the following properties:

- CustomerID : Custormer Identifier.
- CustomerName : Customer Name.
- AccountNumber:Numbre of the Customer Account.
- Customer Type: type of Customer, I=Individual, S=Store.
- AddressID: the Adress Identifier.

An element named Address, type of class, which has the following properties:

- AddressID: Identifier of customer address.
- AdressLine1: Line 1 of the address.
- AdressLine2: Line 2 of the address.
- City: defines the city of customer.
- State: defines the state of customer.
- PostalCode: defines the postal code.

An element named Product, type of class, which has the following properties:

- ProductID: Product Identifier.
- ProductName: Product Name.
- StandardCost: Product Standard Cost.
- ListPrice : sale price
- Weight : Product Weight
- Style : W=Women, M=Male, U=Both
- Color: Product Color.
- DayToManufacture: Number of days to manufacture the product.
- ProductSubcategoryID : he product belongs to this sub category.

An element named ProductSubcategory, type of class, which has the following properties:

- ProductSubcategoryID: Product Sub category Identifier.
- ProductSubcategoryName: Product Sub category Name.
- ProductCategoryID: the sub category belongs to this category.

An element named ProductCategory, type of class, which has the following properties:

- ProductCategoryID: Product Category Identifier.
- ProductCategoryName : Product Category Name.

An element named SalesOrderDetail, type of class, which has the following properties:

- SalesOrderDetailID: Identifier of sales order.
- OrderQty: Quantity of Sales.
- UnitPrice: the selling price of a single product.
- OrderDate: Creation date of the sales order.
- CustomerID: Customer Identifier.
- ProductID: Product Identifier.

This file in XMI format is provided as input of QVT transformation, achieved under Eclipse Europa. As output of the QVT transformation, file in XMI format is generated automatically. It represents the multidimensional schema of company Datawarehouse in our case study. This file is composed of a Fact named FactSalesOrderDetail surrounded by Dimensions: Customer, Adress, Product, Category, SubCategory.

The generated XMI file contains the element MDPackage composed of the following:

```
<?xml version="1.0" encoding="ASCII"?>
<xmi:XMI xmi:version="2.0" xmlns:xmi="http://www.omg.org/XMI"
xmlns:BdMdMM="http:///BdMdMM.ecore">
 <BdMdMM:MDPackage name="MDFactory">
  <fact name="FactSalesOrder" factattribute="/1 /2 /3 /4 /5 /6"/>
 </BdMdMM:MDPackage>
 <BdMdMM:FactAttributename="SalesOrderID" fact="/0/@fact.0"/>
 <BdMdMM:FactAttribute name="OrderQty" fact="/0/@fact.0"/>
 <BdMdMM:FactAttribute name="UnitPrice" fact="/0/@fact.0"/>
 <BdMdMM:FactAttribute name="OrderDate" fact="/0/@fact.0"/>
 <BdMdMM:FactAttribute name="CustomerID" fact="/0/@fact.0"/>
 <BdMdMM:FactAttribute name="ProductID" fact="/0/@fact.0"/>
 <BdMdMM:Dimension name="DimensionCustommer">
  <dimensionattribute name="CustomerID"/>
  <dimensionattribute name="AccountNumber"/>
  <dimensionattribute name="CustomerType"/>
  <dimensionattribute name="AddressID"/>
 </BdMdMM:Dimension>
 <BdMdMM:Dimension name="DimensionProduct">
  <dimensionattribute name="ProductID"/>
  <dimensionattribute name="Name"/>
  <dimensionattribute name="StandardCost"/>
  <dimensionattribute name="ListPrice"/>
  <dimensionattribute name="Weight"/>
  <dimensionattribute name="Style"/>
  <dimensionattribute name="Color"/>
  <dimensionattribute name="DayToManufacture"/>
  <dimensionattribute name="ProductSubcategoryID"/>
 </BdMdMM:Dimension>
</xmi:XMI>
```

Figure 14. Overview of the instance of Multidimensional Model generated after execution of the transformation

VIII. Conclusion and Future Works

We applied the MDA approach for the engineering of SID. The objective is to define a model transformation. It is based on a UML source model and a Multidimensional target model. This transformation takes as input the source model and as output the target model. Once done, we can generate from a simplified UML model instance, a simplified multidimensional model.

Looking ahead, we plan to develop a graphical plug-in integrated to Eclipse, which automatically generates from a UML class diagram, a multidimensional diagram. In addition, we can extend this work to more generalized meta-models.

References

[1] R. Kimball The Data Warehouse Toolkit, Second Edition, Wiley Computer Publishing 2002.

[2] M. Golfarelli, D. Maio, S. Rizzi, The dimensional fact model: a conceptual model for data warehouses, International Journal of Cooperative Information Systems 7 (2-3) (1998) 215–247.

[3] P. Giorgini, S. Rizzi, M. Garzetti, Goal-oriented requirement analysis for data warehouse design, DOLAP, 2005, pp. 47–56

[4] A. Abelló, J. Samos, F. Saltor, YAM2: a multidimensional conceptual model extending UML, Information Systems 31 (6) (September 2006) 541–567

[5] Dj. Midouni, J.Darmont, F.Bentayeb, Approach to modeling complex multidimensional data: Application to medical data, « 5ème journée francophones sur les entrepôts de données et l'analyse en ligne » Montpellier, France (EDA 2009).

[6] A. Nguyen, A. Friedrich, G. Berthommier, L. Poidevin, L. Moulinier, R. Ripp,O. Poch, Introduction du nouveau Centre de Données Biomédicales Décrypthon, CORIA - Conférence en Recherche d'Information et Applications-, 2008

[7] Object Management Group (OMG), MDA Guide 1.0.1. http://www.omg.org/cgi-bin/doc?omg/03-06-01.

[8] Object Management Group (OMG) Unified Modeling Language Specification 2.0. http://www.omg.org/cgi-bin/doc?formal/05-07-04.

[9] Object Management Group (OMG), MOF 2.0 Query/View/Transformation.http://www.omg.org/spec/MOF/2.0/PDF/

[10] Object Management Group (OMG), Common WarehouseMetamodel (CWM) Specification 1.1. http://www.omg.org/spec/cwm/1.1/PDF/

[11] Object Management Group (OMG), Object Constraint Language (OCL) Specification 2.0. http://www.omg.org/spec/OCL/2.2/PDF/

[12] MDA en action Ingénierie logicielle guidée par les modèles, Xavier Blanc, édition Eyrolles 2005.

[13] S. Mbarki, M. Erramdani, Model-Driven Transformations: From Analysis to MVC 2 Web Model. (I.RE.CO.S.), Vol 4 N 5 Septembre 2009.

[14] S. Mbarki, M. Erramdani, Toward automatic generation of mvc2 web applications InfoComp, Journal of Computer Science, Vol.7 n.4, pp. 84-91, December 2008, ISSN: 1807-4545

[15] J. Trujillo, S. Lujan-Mora, , I.Song , a UML profile for multidimensional modeling in data warehouses, Data & knowledge Engineering, (2006) 725-769

[16] JN. Mazón , J. Trujillo, An MDA approach for the development of data warehouses, Decision Support Systems 45 (2008) 41–58

[17] SmartQVT, http://smartqvt.elibel.tm.fr/

[18] Ganapathy, G., & Sagayaraj, S. (2011). Extracting Code Resource from OWL by Matching Method Signatures using UML Design Document. International Journal of Advanced Computer Science and Applications - IJACSA, 2(2), 90-96.

[19] Dhindsa, K. S. (2011). Modelling & Designing Land Record Information System Using Unified Modelling Language. International Journal of Advanced Computer Science and Applications - IJACSA, 2(2), 26-30.

[20] Acharya, A. A. (2010). Model Based Test Case Prioritization for Testing Component Dependency in CBSD Using UML Sequence Diagram. International Journal of Advanced Computer Science and Applications - IJACSA, 1(6).

AUTHORS PROFILE

I.Arrassen Graduate as Computer Science Enginner from the INPT(National Institut of Poste and Telecommunication) and Ph-D-Student at Faculty of Sciences, Laboratory for Computer Science Research, Mohammed First University, Oujda, Morocco.

A.Meziane is a Professor of Computer Sciences, Laboratory for Computer Science Research , Mohammed First University, Oujda, Morocco.

R.Esbai Ph-D-Student at Faculty of Sciences, Laboratory AMCSP: Applied Mathematics and Computer Signal Processing, Mohammed First University, Oujda, Morocco.

M.Erramdani is a Professor of Computer Sciences at Superior School of TechnologyLaboratory AMCSP: Applied Mathematics and Computer Signal Processing Mohammed First University, Oujda, Morocco.

Fuzzy Particle Swarm Optimization with Simulated Annealing and Neighborhood Information Communication for Solving TSP

Rehab F. Abdel-Kader
Electrical Engineering Department
Faculty of Engineering, Port-Said University
Port Fouad 42523, Port-Said, Egypt

Abstract— **In this paper, an effective hybrid algorithm based on Particle Swarm Optimization (PSO) is proposed for solving the Traveling Salesman Problem (TSP), which is a well-known NP-complete problem. The hybrid algorithm combines the high global search efficiency of fuzzy PSO with the powerful ability to avoid being trapped in local minimum. In the fuzzy PSO system, fuzzy matrices were used to represent the position and velocity of the particles in PSO and the operators in the original PSO position and velocity formulas were redefined. Two strategies were employed in the hybrid algorithm to strengthen the diversity of the particles and to speed up the convergence process. The first strategy is based on Neighborhood Information Communication (NIC) among the particles where a particle absorbs better historical experience of the neighboring particles. This strategy does not depend on the individual experience of the particles only, but also the neighbor sharing information of the current state. The second strategy is the use of Simulated Annealing (SA) which randomizes the search algorithm in a way that allows occasional alterations that worsen the solution in an attempt to increase the probability of escaping local optima. SA is used to slow down the degeneration of the PSO swarm and increase the swarm's diversity. In SA, a new solution in the neighborhood of the original one is generated by using a designed λ search method. A new solution with fitness worse than the original solution is accepted with a probability that gradually decreases at the late stages of the search process. The hybrid algorithm is examined using a set of benchmark problems from the TSPLIB with various sizes and levels of hardness. Comparative experiments were made between the proposed algorithm and regular fuzzy PSO, SA, and basic ACO. The computational results demonstrate the effectiveness of the proposed algorithm for TSP in terms of the obtained solution quality and convergence speed.**

Keywords- Information Communication; Particle Swarm Optimization; Simulated Annealing; TSP.

I. INTRODUCTION

Traveling Salesman Problem (TSP) is a well-known NP-complete problem that has important practical applications as many complicated problems in various fields can be abstracted and changed to TSP [1-3]. The problem can be described as a single salesman who wants to visit a number of cities. The main objective of TSP is to find a minimal length closed tour that visits each city exactly once. TSP has been studied extensively over the past several decades. Although it is simple to model the TSP mathematically, there is no definite algorithm that can be used to solve TSP in polynomial time. It is evident that the computational cost of TSP by exhaustive permutations is O(n!). Today many bio-inspired intelligence techniques are rapidly developing such as Genetic Algorithms (GA), Ant Colony Optimization (ACO), and Particle Swarm Optimization (PSO) are used to solve combinational optimization problems such as TSP.

PSO is a computation paradigm based on group intelligent global optimization methods initially introduced by Kennedy and Eberhart in 1995 [4 - 6] to simulate the process of birds forage for food. PSO is different from other evolutionary algorithms, as it convergences quickly, has less parameters, encodes with real numbers and can directly deal with the problem domain, without conversion. PSO is simple, easy-to implement, and is widely used, in continuous function optimization problems where it yields excellent optimization performance. Recently, some researches apply this algorithm to problems of discrete quantities. However, the basic PSO algorithm suffers a serious problem that all particles are prone to be trapped into the local minimum in the later phase of convergence. The optimal value found is often a local minimum instead of a global minimum.

Aiming at solving the shortcoming of the basic PSO algorithm, many variations, such as Fuzzy PSO [3, 6], Hybrid PSO [7], Intelligent PSO [8], Niching PSO [9] and Guarantee Locally Convergent PSO [10] have been proposed to increase the diversity of particles and improve the convergence performance. In this paper, a new algorithm that combines the fuzzy PSO algorithm with Neighborhood Information Communication (NIC) strategy and Simulated Annealing (SA) was proposed and applied to solve the TSP. The NIS strategy incorporates the shared information provided by the individual's neighborhood into the PSO update equations. SA is a kind of stochastic method that is well known for its effective capability of escaping local optima. By integrating NIC and SA to the fuzzy PSO, the new algorithm, which we call it PSO-NIC-SA can not only escape from local minimum trap in the later phases of convergence, but also simplify the implementation of the algorithm. In the experiments, three additional algorithms: fuzzy PSO, SA, and ACO have been implemented and the results have been compared to the results of the proposed algorithm. Test results demonstrate that the PSO-NIC-SA algorithm outperforms the other algorithms in solving TSP. The rest of this paper is organized as follows: The basic particle swarm optimization algorithm is presented in Section II. The mathematical model for TSP and the Fuzzy PSO algorithm for solving the TSP are described in Section III. In Section IV, the proposed PSO-NIC-SA algorithm for TSP is explained. The experimental results are reported in Section V. Finally, we summarize the paper with some concluding remarks in Section VI.

II. PARTICLE SWARM OPTIMIZATION

PSO proposed by Dr. Eberhart and Dr. Kennedy in 1995 is a computational technique based on the idea of collaborative behavior and swarming in biological populations inspired by the social behavior of bird flocking or fish schooling [4 - 6].

The algorithm, which is based on a metaphor of social interaction, searches a space by adjusting the trajectories of individual vectors, called “particles” as they are conceptualized as moving points in multidimensional space. The individual particles are drawn stochastically toward the position of their own previous best performance and the best global performance among its neighbors.

The PSO algorithm is simple, easy to implement, robust to control parameters, and computationally efficient compared to other heuristic optimization techniques. The original PSO has been applied to a learning problem of neural networks and function optimization problems, and the efficiency of the method has been confirmed.

When PSO is used to solve an optimization problem, a swarm of particles, is used to explore the solution space for an optimum solution. Each particle represents a candidate solution and is identified with specific coordinates in the *D*-dimensional search space. The position of the *i*'th particle is represented as $X_i = (x_{i1}, x_{i2}, \ldots\ldots, x_{iD})$. The velocity of a particle is denoted as $V_i = (v_{i1}, v_{i2}, \ldots\ldots\ldots, v_{iD})$. The fitness function is evaluated for each particle in the swarm and is compared to the fitness of the best previous result for that particle and to the fitness of the best particle among all particles in the swarm. After finding the two best values, the particles evolve by updating their velocities and positions according to the following equations:

$$V_i^{t+1} = \omega * V_i^t + c_1 * rand_1 * (p_{i\_best} - X_i^t) + c_2 * rand_2 * (g_{best} - X_i^t) \tag{1}$$

$$\boldsymbol{X}_i^{t+1} = \boldsymbol{X}_i^t + \boldsymbol{V}_i^{t+1} \tag{2}$$

Where $i = (1, 2, \ldots\ldots., Pop\_Size)$ and *Pop_Size* is the size of the swarm; p_{i_best} is the particle best reached solution and g_{best} is the global best solution in the swarm. c_1 and c_2 are cognitive and social parameters that are bounded between 0 and 2. $rand_1$ and $rand_2$ are two random numbers, with uniform distribution *U[0,1]*. $-V_{max} \leq V_i^{t+1} \leq V_{max}$ (V_{max} is the maximum velocity).

The inertia weight *ω* is a factor used to control the balance of the search algorithm between exploration and exploitation. The recursive steps will go on until we reach the termination condition.

III. TRAVELING SALESMAN PROBLEM

A. Mathematical Model of TSP

The TSP can be described as follows: In the graph $G = (X, E)$, *X* is the set of nodes, or cities to be visited, *E* is the set of edges, $E= \{(x_i, x_j): x_i, x_j \in X\}$. The objective of TSP is to find a minimal length closed tour that visits a number of cities '*n*' such that each city is visited exactly once. This closed tour is called the *Hamiltonian cycle*. When the total distance traveled is the main metric for evaluating the cost then the problem of finding the best path *S* is the same as the minimization of the target fitness function *F(S)* defined as follows:

$$F(S) = \sum_{i=1}^{n-1} d(x_i, x_{i+1}) + d(x_n, x_1) \tag{3}$$

Where $d(x_i, x_{i+1})$, $0 \leq i \leq n-1$ is the *Euclidean* distance from city x_i to city x_{i+1}. If for all pairs of nodes x_i, x_j, the distances $d(x_i, x_j)$ and $d(x_j, x_i)$ are equal then the problem is said to be *symmetric*, otherwise it is said to be *asymmetric*. In the TSP the solution space increases rapidly as the total number of cities increases. For example, with number of cities *n=100*, the total number of possible solution paths will be *4.67*10 155*. Tremendous research has focused on this research problem due to its significance both in theory and applications.

B. Fuzzy Matrix to Represent TSP Solution

The Fuzzy Discrete PSO (FD-PSO) was first introduced by Wei Pang et al. [1, 2] to solve the TSP, and subsequently used and improved in [11] and was found to achieve satisfactory results.

1) Construction of Fuzzy Matrix

Assume the solution of TSP is $T= \{(x_1, x_2), (x_2, x_3), \ldots\ldots, (x_{n-1}, x_n), (x_n, x_1)\}$, where *n* is the number of cities, x_i $(i =1, \ldots\ldots,n)$ is the *i*-th visited city (node) in this solution, and (x_1, x_2), $(x_2, x_3), \ldots\ldots, (x_{n-1}, x_n)$, (x_n, x_1) are the visited directed edges in turn. Then fuzzy position matrix can be represented as an *nxn* matrix *P* as follows:

$$P= \begin{pmatrix} p_{11} \cdots\cdots\cdots p_{1n} \\ . \qquad\qquad . \\ p_{n1} \cdots\cdots\cdots p_{nn} \end{pmatrix} \tag{4}$$

Where $p_{ij} \in [0, 1]$ means the possibility of choosing directed edge (x_i, x_j), after city x_i has been chosen in the TSP solution. In order to avoid directed edge (x_i, x_i) *for* $(i =1, \ldots\ldots,n)$ appearing in TSP solution, we need to set the elements in diagonal of fuzzy matrix to very small values, thus let p_{ii} $(i= 1, \ldots\ldots ,n) =-Max$.

The velocity of the particle is defined as:

$$V= \begin{pmatrix} v_{11} \cdots\cdots\cdots v_{1n} \\ . \qquad\qquad . \\ v_{n1} \cdots\cdots\cdots v_{nn} \end{pmatrix} \tag{5}$$

2) Fuzzy PSO operators

The original PSO position and velocity equations (1) and (2) should be redefined to the form of matrices. The symbol "⊗" was used to denote the modified multiplication operator which represents a scalar matrix multiplication where all elements in the matrix are multiplied by the scalar. The symbols "⊕" and " Θ " denote the addition and subtraction between matrices respectively. The modified position and velocity vectors can be rewritten as follows:

$$V_i^{t+1} = \omega * V_i^t \oplus (c_1 * rand_1) \otimes (p_{i\_best} \Theta X_i^t) \oplus (c_2 * rand_2) \otimes (g_{best} \Theta X_i^t) \quad (6)$$

$$\boldsymbol{X_i^{t+1} = X_i^t \oplus V_i^{t+1}} \quad (7)$$

3) Initialization

The elements of the position matrix *P given* in (4) are randomly generated subject to the following constraints:

$$\sum_{j=1}^{n} p_{ij} = 1, \text{ for } i=\{1,\ 2,\ \ldots\ldots,n\} \quad (8)$$

$$P_{ij} \in [0,1] \quad (9)$$

Similarly, the elements of the velocity matrix *V* given in equation (5) are randomly generated subject to the following constraint:

$$\sum_{j=1}^{n} v_{ij} = 0 \quad \text{for} \quad i=\{1,\ 2,\ \ldots\ldots,n\} \quad (10)$$

4) Normalization of the Position Matrix

The position matrix *P* may violate the constraint given in (9) after the update of the position matrix in the forthcoming generations. Thus, it is necessary to normalize the position matrix. First, all negative elements in *P* are converted to *0*. The other elements in *P* are recalculated according to the following transformation:

$$P = \begin{pmatrix} p_{11} / \sum_{i=1}^{n} p_{1i} & \ldots\ldots\ldots & p_{1n} / \sum_{i=1}^{n} p_{1i} \\ . & & . \\ p_{n1} / \sum_{i=1}^{n} p_{ni} & \ldots\ldots\ldots & p_{nn} / \sum_{i=1}^{n} p_{ni} \end{pmatrix} \quad (11)$$

5) Defuzzication

The fuzzy position matrix *P* represents the potential solution for the TSP. However, matrix *P* has to be decoded in order to obtain the feasible solution representing the best route (the ordered sequence of nodes or cities visited). This procedure is called "*Defuzzication*". In this paper the global maximum method is used for the *Defuzzication* of the position matrix. In this method we have a flag array to record whether we have selected the columns of the matrix and a route array to record the route solution, first all the columns are not selected, then for each row of the matrix, we choose the element which is not selected by the previous rows and has the max value, then we mark the column of the max element “selected”, and the column number are recorded to the route array. After all the rows have been processed, we get the route solution from the route array and the cost of the TSP route is calculated according to (3).

6) Fuzzy PSO Algorithm for TSP

The pseudo code of the fuzzy PSO algorithm for solving the TSP is presented in Fig. 1.

IV. Hybrid Algorithm for TSP

PSO algorithm is problem-independent, which means little specific knowledge relevant to the given problem is required. This advantage makes PSO more robust than many other search algorithms. However, as a stochastic search algorithm, PSO is prone to lack global search ability at the end of its run. PSO may fail to find the required optima as it can easily get trapped into local optima in complex optimization problems. Two strategies were proposed in the hybrid algorithm to strengthen the diversity of the particles and to speed up the convergence process. The first strategy is based on NIC among the particles where a particle absorbs better historical experience of the neighboring particles. The second strategy is the use of SA which randomizes the search algorithm in a way that allows occasional alterations that worsen the solution in an attempt to increase the probability of escaping local optima. The NIC and the SA strategies are explained in more detail in the following Sections.

A. Neighborhood Information Communication

Biological results suggest that information sharing among neighboring individuals contributes to evolution as the current state of neighbors significantly impact on the decision process of group members. However, PSO, as a simulation of group foraging behavior, does not include any neighborhood information sharing into its evolutionary equations. In traditional PSO, the global best solution g_{best} is the only information shared among the particles of the swarm. In this paper, an information sharing strategy among the particles is utilized in the proposed hybrid algorithm. In the renewing process of the position and velocity matrices, a particle absorbs better historical experience of the neighboring particles with better fitness values than its own. Better particles will guide the other particles to improve their fitness. This results in a very small probability to be trapped by local optima. All particles in the swarm will be ranked according to their fitness values. In minimization problems, the particle with the smallest fitness value will be ranked 1 and similarly for all other particles. Each particle shows interest in other particles according to their rank. The modified velocity matrix can be rewritten as follows:

$$V_i^{t+1} = \omega * V_i^t \oplus (\frac{1}{rank(i)} * rand_1) \otimes (p_{i\_best} \Theta X_i^t) + rand_2 \otimes (g_{best} \Theta X_i^t) + social\_\mathrm{info}(i) \quad (12)$$

Where *social_info(i)* is calculated as shown in Fig. 2. The *social_info()* module gives the direction of the swarm by sharing information with all other individuals that have better fitness values. V_{max} has been set to small values to prevent abrupt variations in the solution domain.

```
Step1: initialization
   1.1.  Initialize the swarm size to Pop Size and the maximum number of
         iterations Max iterations
   1.2. Initialize the position and velocity matrices for all particles in the
         swarm.
   1.3. Evaluate the fitness of each particle
   1.4. Initialize the local best for every particle $P_i = X_i^0$ and the global best
         $g_{best}$ is the best among all $P_i$
Step 2: if Current iteration equals Max iterations goto step 5.
Step 3: for i:=0 to Pop Size -1
   3.1. Current iteration= Current iteration +1
   3.2. Calculate the new velocity matrices according to equation (6)
   3.3. Calculate the new position matrices according to equation (7)
   3.4. Normalize the position matrices according to equation (11)
   3.5. Defuzzy the new position matrices and calculate the cost for each
        position matrix using(3).
   3.6. If the cost of the new position of a particle is less than that of the
        local best of the particle then update the local best position with
        new position
Step 4: If the cost of the local best of some particles is less than the $g_{best}$ then
         update the $g_{best}$ with the local best particle. Goto step 2
Step 5: output $g_{best}$ route and its corresponding cost
```

Figure 1. Fuzzy PSO Algorithm

```
social_info(i)
{
  social effect =0;
     for k=1: Pop_Size
         if F(k)<F(i)
              social_effect= social_effect +(1/rank(k)*rand$_3$) $\otimes$ ( $X_k^t \Theta X_i^t$ ))
      if (social effect > $V_{max}$ )
           return $V_{max}$
      else
         return social effect }
```

Figure 2. Neighborhood Information Communication Algorithm.

B. SA-based Local Search for TSP

SA is an intelligent stochastic strategy used in solving optimization problems. It was successfully applied to the optimization problems by Kirkpatrick [3, 12]. SA employs certain probability to avoid becoming trapped in a local optimum by allowing occasional alterations that increases the diversity of the particles in the swarm. In the search process, the SA accepts not only better but also worse neighboring solutions with a certain probability. Such mechanism can be regarded as a trial to explore new space for new solutions, either better or worse. The probability of accepting a worse solution is larger at higher initial temperature. As the temperature decreases, the probability of accepting worse solutions gradually approaches zero. More specifically, starting from an initial state, the system is perturbed at random to a new state in the neighborhood of the original one. Then the change ΔE of the fitness function value is calculated. For minimization problems, the new state is accepted with probability $min\{1, exp(-\Delta E\ /T)\}$, where T is a control parameter corresponding to the temperature in the analogy. The SA algorithm generally starts from a high temperature, and then the temperature is gradually lowered. At each temperature, a search is carried out for a certain number of iterations. The above technique can increase the diversity in the particles and enable PSO to accept a bad solution with a probability that will gradually decrease to zero as the temperature decreases. In this paper a simple λ search method is designed for generating the SA neighborhood solutions, where λ is a parameter representing the depth of the local search. The λ search method includes two steps:

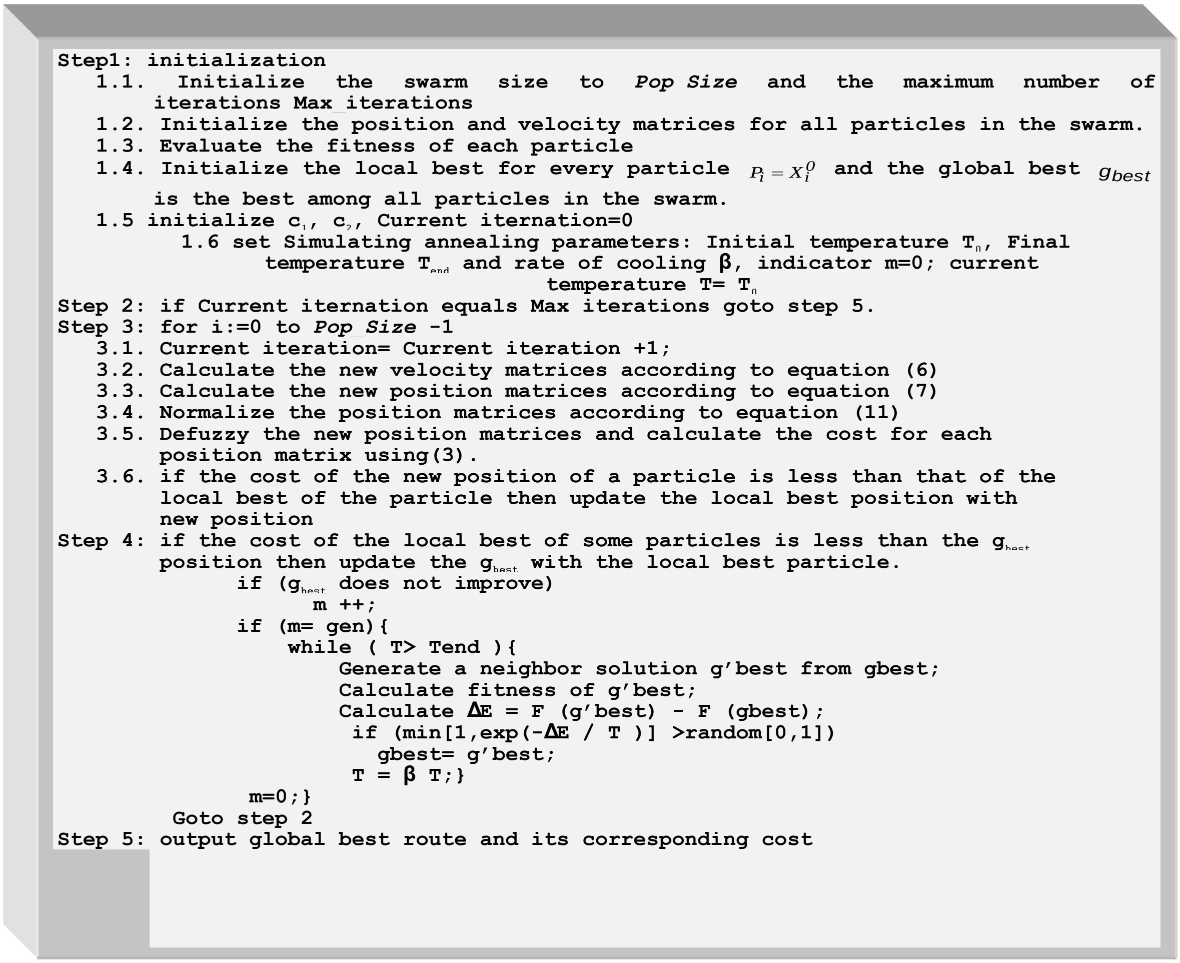
1. Swap the order of a pair of randomly selected rows in the position matrix

2. Perform the matrix normalization transformation according to (11)

This process is repeated for λ randomly selected pairs of rows and λ new solutions are produced. The best solution among the λ generated solutions is selected as the newly produced solution.

C. Hybrid PSO-NIC-SA Algorithm

The objective of TSP is to minimize the fitness function given in (3) that represents the cost of a particular route. Combining the fast optimal search ability of fuzzy PSO with the Neighboring Information Communication model and the probability jump property of SA, we design a new algorithm



```
Step1: initialization
   1.1.  Initialize the swarm size to Pop Size and the maximum number of
        iterations Max_iterations
   1.2. Initialize the position and velocity matrices for all particles in the swarm.
   1.3. Evaluate the fitness of each particle
   1.4. Initialize the local best for every particle $P_i = X_i^0$ and the global best $g_{best}$
        is the best among all particles in the swarm.
   1.5 initialize $c_1$, $c_2$, Current iternation=0
          1.6 set Simulating annealing parameters: Initial temperature $T_0$, Final
                temperature $T_{end}$ and rate of cooling $\beta$, indicator m=0; current
                                  temperature T= $T_0$
Step 2: if Current iternation equals Max iterations goto step 5.
Step 3: for i:=0 to Pop_Size -1
   3.1. Current iteration= Current iteration +1;
   3.2. Calculate the new velocity matrices according to equation (6)
   3.3. Calculate the new position matrices according to equation (7)
   3.4. Normalize the position matrices according to equation (11)
   3.5. Defuzzy the new position matrices and calculate the cost for each
        position matrix using(3).
   3.6. if the cost of the new position of a particle is less than that of the
        local best of the particle then update the local best position with
        new position
Step 4: if the cost of the local best of some particles is less than the $g_{best}$
        position then update the $g_{best}$ with the local best particle.
              if ($g_{best}$ does not improve)
                    m ++;
              if (m= gen){
                  while ( T> Tend ){
                      Generate a neighbor solution g'best from gbest;
                      Calculate fitness of g'best;
                      Calculate $\Delta$E = F (g'best) - F (gbest);
                       if (min[1,exp(-$\Delta$E / T )] >random[0,1])
                         gbest= g'best;
                       T = $\beta$ T;}
               m=0;}
         Goto step 2
Step 5: output global best route and its corresponding cost
```

Figure 3. Hybrid PSO-NIC-SA Algorithm

PSO-NIC-SA to solve TSP problem. The pseudo code of the PSO-NIC-SA algorithm for solving the TSP is shown in Fig. 3.

In the proposed algorithm, fuzzy PSO is performed first using the position and velocity update equations that incorporates the neighborhood information communication strategy. If the best particle g_{best} does not improve for a specified number of generations *gen*, then SA is used. The best individual from the fuzzy PSO algorithm provides the initial solution for SA during the hybrid search process. The hybrid algorithm implements easily and reserves the generality of PSO and SA. Moreover, it can be applied to many combinatorial optimization problems by simple modification.

V. Experimental Results

The proposed algorithm has been implemented using MATLAB 7.1 and executed on a Pentium IV, 2.8 GHz computer. To evaluate the efficiency of the proposed algorithm and for comparison purpose, three other artificial algorithms including fuzzy PSO, basic SA, and basic ACO have been implemented, and applied to the same TSP problems. Three TSP benchmark problems (Burma14, Berlin52, Eil75) selected from the TSPLIB [13] were tested. The parameters used in PSO, NIC and SA were determined through the preliminary experiments. The following parameter setting was used in the proposed algorithm: the swarm size *Pop_Size* = 50, *Max_iterations*= 1000, initial temperature T_0=1000, final temperature $T_{end}=1$, rate of cooling β=0.99 and the SA search parameter $\lambda=5$. All algorithms were run 20 times for all the TSP problems. The results shown in Table I represent the mean and best solutions found by running the various trails of each method.

From the results of the four algorithms, it is clear that our algorithm is significantly better than the other algorithms. For example, in Burma14 problem, the best fitness value 30.87 achieved by the hybrid PSO-NIC-SA algorithm is not only the smallest value among the four algorithms but also the optimal solution for the problem. Meanwhile the mean value case, performance is also greatly improved by the new algorithm. It is shown that the new algorithm is a better and more effective means to solve TSP problem. Fig. 4 presents the convergence speed for the various algorithms in the Burma14 benchmark. The Figure shows the mean fitness found versus the number of iterations in the four implemented algorithms. As can be seen from Fig. 4, the proposed PSO-NIC-SA algorithm was able to reach good solutions faster than other methods in the early stages of the search process, and reach better solutions than others at the end of the search process.

TABLE I. COMPARISONS OF FUZZY PSO, SA, ACO AND PSO-NIC-SA FOR BURMA14, BERLIN52, EIL75 BENCHMARKS

TSP Problem	Optimal Solution	Fuzzy PSO			SA			ACO			Hybrid PSO-NIC-SA		
		Mean	Best	Time (ms)	Mean	Best	Time (ms)	Mean	Best	Time (ms)	Mean	Best	Time (ms)
Burma14	30.87	34.6	30.87	444	34.98	33.56	199	30.95	30.87	107	30.87	30.87	354
Berlin52	7542	7645	7543	8976	7331	7389	5890	7601.5	7542	5432	7593.8	7542	6766
Eil75	542.37	593.83	587.65	13543	598.76	580	10250	544.7	542.37	10876	543. 9	542.37	11240

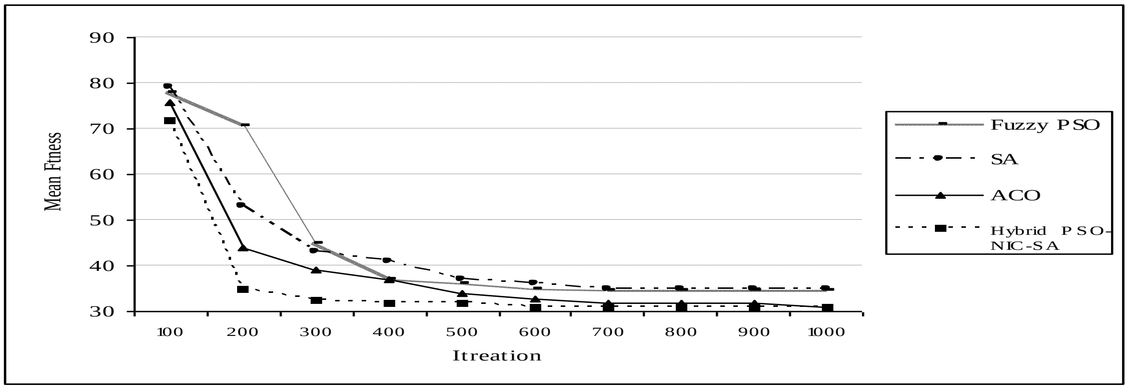


Figure 4. Convergence Curves for the Different Algorithms in Burma14.

VI. CONCLUSIONS

In this paper, an effective hybrid algorithm based on fuzzy PSO is proposed for solving the TSP, which is a well-known NP-complete problem. The proposed algorithm combines the high global search efficiency of fuzzy PSO with the powerful ability to avoid being trapped in local minimum. In the fuzzy PSO system, fuzzy matrices were used to represent the position and velocity of the particles in PSO and the operators in the original PSO position and velocity formulas were redefined.

Two strategies were employed in the hybrid algorithm to strengthen the diversity of the particles and to speed up the convergence process. The first strategy is based on neighborhood information communication among the particles where a particle absorbs better historical experience of the neighboring particles. This strategy integrates the individual experience of the particles with the neighbor sharing information of the current state.

The second strategy is the use of simulated annealing which randomizes the search algorithm in a way that allows occasional alterations that worsen the solution in an attempt to increase the probability of escaping local optima. SA is used to slow down the degeneration of the PSO swarm and increase the swarm's diversity. In SA, a new solution in the neighborhood of the original one is generated by using a designed λ search method. A new solution with fitness worse than the original solution is accepted with a probability that gradually decreases at the late stages of the search process. The hybrid algorithm is examined using a set of benchmark problems from TSPLIB with various sizes and levels of hardness.

Comparative experiments were made between the proposed algorithm and PSO, SA, and basic ACO. The computational results validate the effectiveness of the proposed approach.

REFERENCES

[1] W. Pang, K.-P. Wang, C.-G Zhou, and L.-J Dong, "Fuzzy Discrete Particle Swarm Optimization for Solving Traveling Salesman Problem," In Proceedings of the Fourth International Conference on Computer and Information Technology, pp. 796-800, 2004.

[2] W. Pang, K.-P. Wang, C.-G Zhou, and L.-J Dong "Modified Particle Swarm Optimization Based on Space Transformation for Solving Traveling Salesman Problem," In Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, pp. 2342-2346, 2004.

[3] L. Fang, P. Chen, and S. Liu, "Particle Swarm Optimization with Simulated Annealing for TSP," In Proceedings of 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Databases, pp. 206-209, 2007.

[4] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," In Proceedings of IEEE International Conf. on Neural Networks, Piscataway, NJ, Vol. 4: pp. 1942-1948, 1995.

[5] R. Eberhart and Y. Shi, "Comparison Between Genetic Algorithms and Particle Swarm Optimization," In e. a. V. William Porto, editor, Evolutionary Programming, Lecture Notes in Computer Science, Vol. 1447, pp. 611–616. Springer, 1998.

[6] Y. Shi and R. Eberhart, "Fuzzy Adaptive Particle Swarm Optimization," In Proceedings of Congress on Evolutionary Computation, pp.101-106, 2001.

[7] M. Lovbjerg, T. Rasmussen, and T. Krink, "Hybrid Particle Swarm Optimizer with Breeding and Subpopulation," In Proceedings of Evolutionary Computation Conference, 2001.

[8] G. Ciuprina, D. Ioan, and I. Munteanu, "Use of Intelligent-Particle Swarm Optimization in Electromagnetic," IEEE Transactions on Magnetics, Vol. 38, No. 2, pp. 1037-1040, 2002.

[9] R. Brits, A. Engelbrecht, and A. Niching, "Particle Swarm Optimizer," In Proceedings of 4th Asia-Pacific Conference on Simulated Evolution and Learning, 2002.

[10] E. Bergh and A. Engelbrecht, "A New Locally Convergent Particle Swarm Optimizer," In Proceedings of IEEE Conference on Systems, Man, and Cybernetics, 2002.

[11] B. Shen, M. Yao, and W. Yi, "Heuristic Information Based Improved Fuzzy Discrete PSO Method for Solving TSP," Lecture Notes in Artificial Intelligence, Vol. 4099, Springer, Berlin, pp. 859-863, 2006.

[12] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimizaion by Simulated Annealing," Science Journal, Vol. 220, No. 4598, pp. 671-680, 1983.

[13] TSPLIB: http://elib.zib.de/pub/mp-testdata/tsp/tsplib/tsplib.html

[14] Mohsen, F. M. A. (2011). A new Optimization-Based Image Segmentation method By Particle Swarm Optimization. International Journal of Advanced Computer Science and Applications - IJACSA, (Special Issue), 10-18.

[15] Trivedi, J. A. (2011). Framework for Automatic Development of Type 2 Fuzzy , Neuro and Neuro-Fuzzy Systems. International Journal of Advanced Computer Science and Applications - IJACSA, 2(1), 131-137.

[16] Khan, A. R., Rehman, Z. U., & Amin, H. U. (2011). Application of Expert System with Fuzzy Logic in Teachers ' Performance Evaluation. International Journal of Advanced Computer Science and Applications - IJACSA, 2(2), 51-57.

AUTHOR PROFILE

Rehab Farouk Abdel-Kader, attended Suez Canal University, Port-Said, Egypt majoring in Computer Engineering, earning the BS degree in 1996. She graduated from Tuskegee University, Tuskegee, Alabama with a MS degree in Electrical Engineering in 1999. She joined the Ph.D. program at Auburn University and earned her Doctor of Philosophy degree in 2003. She worked as an assistant Professor in the Engineering Studies Program in Georgia Southern University, Statesboro, Georgia from 2003 to 2005. She is currently an assistant professor in the Electrical Engineering department, Faculty of Engineering at Port-Said, Port-Said University, Port-Said, Egypt. Her current research interests include Signal Processing, Artificial Intelligence, Pattern Recognition, and Computer Vision

Empirical Validation of Web Metrics for Improving the Quality of Web Page

Yogesh Singh
University School of Information Technology
Guru Gobind Singh Indraprastha University
Delhi-110006 India

Ruchika Malhotra
Software Engineering
Delhi Technological University
Delhi-110042 India

Poonam Gupta
Computer science
Maharaja Agrasen Institute of Management Studies
Rohini Sec -22

***Abstract*— Web page metrics is one of the key elements in measuring various attributes of web site. Metrics gives the concrete values to the attributes of web sites which may be used to compare different web pages .The web pages can be compared based on the page size, information quality ,screen coverage, content coverage etc. Internet and website are emerging media and service avenue requiring improvements in their quality for better customer services for wider user base and for the betterment of human kind. E-business is emerging and websites are not just medium for communication, but they are also the products for providing services. Measurement is the key issue for survival of any organization Therefore to measure and evaluate the websites for quality and for better understanding, the key issues related to website engineering is very important.**

In this paper we collect data from webby awards data (2007-2010) and classify the websites into good sites and bad sites on the basics of the assessed metrics. To achieve this aim we investigate 15 metrics proposed by various researchers. We present the findings of quantitative analysis of web page attributes and how these attributes are calculated. The result of this paper can be used in quantitative studies in web site designing. The metrics captured in the predicted model can be used to predict the goodness of website design.

Keywords- Metrics; Web page; Website; Web page quality; Internet; Page composition Metrics; Page formatting Metrics.

I. Introduction

A key element of any web site engineering process is metrics. Web metrics are used to better understand the attributes of the web page we create. But, most important, we use web metrics to assess the quality of the web engineered product or the process to build it. Since metrics are crucial source of information for decision making, a large number of web metrics have been proposed in the last decade to compare the structural quality of a web page [1].

Software metrics are applicable to all the phases of software development life cycle from beginning, when cost must be estimated, to monitoring the reliability of the products and sub products and end of the product, even after the product is operational.

Study of websites is relatively new convention as compared to quality management. It makes the task of measuring web sites quality very important. Since metrics are crucial source of information for decision making, a large number of web metrics have been proposed in the last decade to compare the structural quality of a web page.

Web site engineering metrics are mainly derived from software metrics, hyper media and Human computer interaction. The intersection of all the three metrics will give the website engineering metrics [2]. Realizing the importance of web metrics, number of metrics has been defined for web sites. These metrics try to capture different aspects of web sites. Some of the metrics also try to capture the same aspect of web sites e.g., there are number of metrics to measure the formatting of page. Also a number of metrics are there for page composition.

Web site engineers need to explicitly state the relation between the different metrics measuring the same aspect of software. In web site designing, we need to identify the necessary metrics that provide useful information, otherwise the website engineers will be lost into so many numbers and the purpose of metrics would be lost.

As the number of web metrics available in the literature is large, it become tedious process to understand the computation of these metrics and draw conclusion and inference from them. Thus, properly defined metrics is used for predictions in various phases of web development process. For proper designing of websites, we need to understand the subset of metrics on which the goodness of website design metrics depends. In this paper we present some attributes related to web page metrics and calculate the values of web attributes with the help of an automated tool. This tool is developed in JSP and calculates about 15 web page metrics with great accuracy.

To meet the above objective following steps are taken:

- Set of 15 metrics is first identified and their values are computed for 514 different web sites (2007-2010) webby awards data.
- The interpretations are drawn to find the subset of attributes which are related to goodness of website

design. Further, these attributes can be used to assess the data into good sites and bad sites.

The goal of this paper is to find the subset of metrics out of 15 metrics to capture the criteria of goodness of web sites. The paper is organized as follows: In section II of this paper the web page metrics which we use in our research is tabulated. Section III describes the research methodology, data collection and description of the tool which we use for calculating the attributes of the web page. In section IV we describe the methodology used to analyze the data .Section V presents the result .Conclusion is discussed in section VI. Future work is given in section VII

II. Description of Web Page Metrics Selected for the Study

Although, several researcher proposed many metrics [3-7] for web page, out of those we identified only 15 metrics for our study that are tabulated in table 1.There are 42 web page metrics and classification of those is given below:-

- Page composition metrics:-The example of this metrics are No. of words, Body Text words, Words in page title, Total number of links etc.
- Page formatting metrics:- They comprise of Font size, Font style, Screen coverage etc
- Overall page quality or assessment metrics;- Example of these metrics are Information quality, Image quality, Link Quality etc.

The overall page quality metrics cannot be easily evaluated as they require human intervention. So in our study, we only use Page formatting metrics and page composition metrics which can be easily calculated.

TABLE I WEBSITE ENGINEERING METRICS

Number Of Words	Total Words On A Page
Body text words	Words that are body Vs. display text
Number of links	Links on a page
Embedded links	Links embedded in text on a page
Wrapped links	Links spanning multiple lines
Within page links	Links to other areas of the same page
Number of!'s	Exclamation points on a page
Page title length	Words in page title
Page size	Total bytes for the page and images
Number of graphics	Total images on a page
Text emphasis	Total emphasized text
Number of list	List on a page
Frames	Use of frames
Number of tables	Number of tables present on a web page
Emphasized body text	Total emphasized body text

The description of the parameters used in this study is given below:-

1) *Number of words*

Total number of words on a page is taken. This attribute is calculated by counting total number of words on the page. Special characters such as & / are also considered as words.

2) *Body text words*

This metrics counts the number of words in the body Vs display text (i.e. Headers). In this, we calculate the words that are part of body and the words that are part of display text that is header separately. The words can be calculated by simply counting the number of words falling in body and number of words falling in header.

3) *Number of links*

These are the total number of links on a web page and can be calculated by counting the number of links present on the web page.

4) *Embedded links*

Links embedded in text on a page. These are the links embedded in the running text on the web page.

5) *Wrapped links*

Links that spans in multiple lines. These are the links which take more than one lines and can be calculated by counting the number of links that spans in multiple lines.

6) *Within page links*

These are the links to other area of the same page. This can be calculated by counting the number of links that links to other area of the same page. Example in some sites have top bottom.

7) *Number of !'s*

Exclamations points on a page can calculated by counting total number of ! marks on a page.

8) *Page title length*

These refer to the words in the page title and can be calculated by counting the total no of words in the page title.

9) *Number of graphics*

These refer to the total number of images on a page. And can be calculated by counting the total number of images present on the page.

10) *Page size*

It refers to the total size of the web page and can be found in properties option of the web page.

11) *Number of list*

This metrics can be calculated by counting total number of ordered and unordered list present on a web page.

12) *Number of tables*

This metrics gives the answer of the question .How many number of tables is used in making a web page?

13) *Frames*

This metrics can be calculated by analyzing whether a web page contains frames or not.

14) Text emphasis

This metric can be calculated by analyzing the web page and counting the total number of words which are in bold, italics and capital.

III. Research Methodology

This study calculates quantitative web page metrics for example number of words, body text words, number of graphics, emphasized body text, number of links etc from the web pages that was evaluated for 2007-2010 webby awards.

The organizers of webby awards places the sites in 70 categories example travel, sports, science, fashion, student, youth, education, School University etc.

The Webby Awards is the leading international award honoring excellence on the Internet. Established in 1996 during the Web's infancy, the Webbys are presented by The International Academy of Digital Arts and Sciences, which includes an Executive 750-member body of leading Web experts, business figures, luminaries, visionaries and creative celebrities, and Associate Members who are former Webby Award Winners and Nominees and other Internet professionals.

The Webby Awards presents two honors in every category -- The Webby Award and The People's Voice Award -- in each of its four entry types: Websites, Interactive Advertising, Online Film & Video and Mobile Web. Members of The International Academy of Digital Arts and Sciences select the nominees for both awards in each category, as well as the winners of the Webby Awards. However the online community, determine the winners of The People's Voice by voting for the nominated work that believe to be the best in each category[8]

For our study we take all the 70 categories for example travel, sports, science, fashion, student ,youth, education, school university etc. these categories contain about 514 sites . Mainly we want to determine the subset of metrics which we use to classify the sites into good and bad sites.

A. Data Collection

The web sites are taken from webby awards sites. We collected about 514 sites from various categories; only the home page is collected for evaluating different web pages. There are three levels in the site as level 1, level 2 and level 3 pages. The level 1 page is the home pages. The level 2 consist of pages that are accessible directly from level 1 that is home page and the level 3 pages that are accessible from level 2 but not from the home page. In this paper we only consider level 1 page.

The data collection process in explained in the block diagram of figure 1.

The data points for each year are tabulated in table II.

TABLE II DESCRIPTION OF DATA POINTS USED IN THE STUDY

YEAR	GOOD DATA POINTS	BAD DATA POINTS	TOTAL DATA POINTS
2007	54	75	129
2008	41	90	131
2009	69	54	123
2010	60	71	131

From the above table we can conclude that the total number of data points is 514.

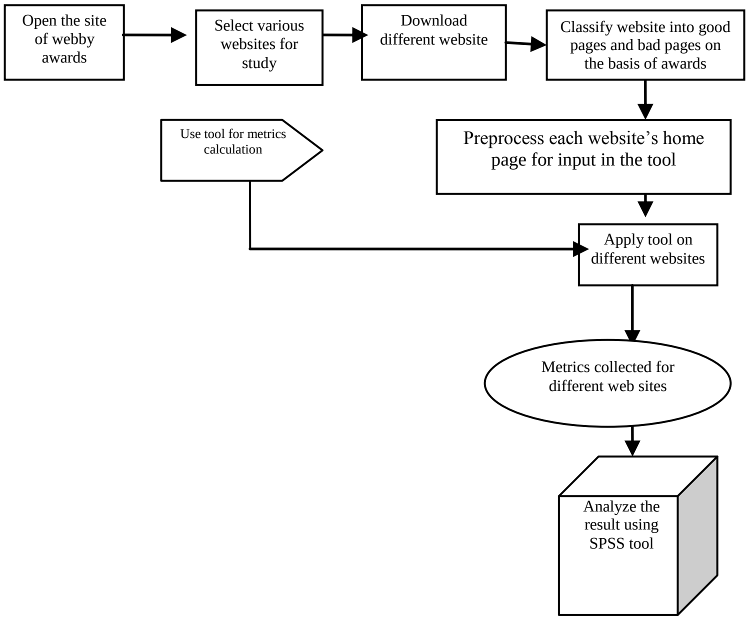
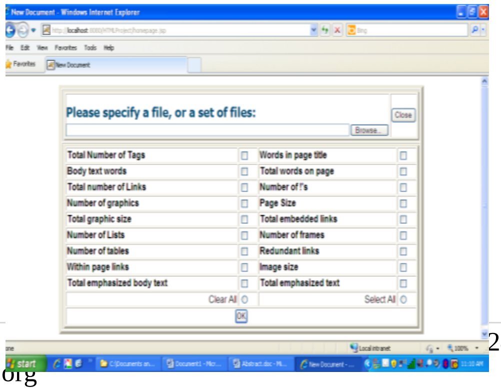

Open the site of webby awards
Select various websites for study
Download different website
Classify website into good pages and bad pages on the basis of awards
Use tool for metrics calculation
Preprocess each website's home page for input in the tool
Apply tool on different websites
Metrics collected for different web sites
Analyze the result using SPSS tool

Figure 1: Block Diagram of Data Collection Process

B. Description of tool

To automate the study of web page metrics we develop a tool for calculating 15 web page attributes. We use JSP for this purpose. JSP technology is one of the most powerful, easy to use and fundamental tools in a Web site developer's toolbox. JSP technology combines HTML and XML with Java servlet (server application extension) and JavaBeans technologies to create a highly productive environment for developing and deploying reliable, interactive, high performance, platform-independent web sites. JSP technology facilitates creation of dynamic content on the server. It is part of the Java platform's integrated solution for server-side programming which provides a portable alternative to other server-side technologies, such as CGI. JSP technology integrates numerous Java application technologies, such as Java servlet, JavaBeans, JDBC, and Enterprise JavaBeans. It also separates information presentation from application logic and fosters a reusable-component model of programming [9].

From the above mentioned tool we can calculate different web attributes. We can select all the attributes or select some of the above list. We can also save the result for further use. The interface of the tool is shown in figure 2.



Please specify a file, or a set of files:

Total Number of Tags	□	Words in page title	□
Body text words	□	Total words on page	□
Total number of Links	□	Number of !'s	□
Number of graphics	□	Page Size	□
Total graphic size	□	Total embedded links	□
Number of Lists	□	Number of frames	□
Number of tables	□	Redundant links	□
Within page links	□	Image size	□
Total emphasized body text	□	Total emphasized text	□
Clear All	○	Select All	○

Figure 2: Tool Interface for calculating web metrics

IV. Data Analysis Methodology

In this section we describe the methodology used to analyze the metrics data computed for 514 web sites. We use Logistic Regression to analyze the data.

Logistic Regression:-LR is the common technique that is widely used to analyze data. It is used to predict dependent variable from a set of independent variables. In our study the dependent variable is good/bad and the independent variables are web metrics. LR is of two types (1) Univariate LR and (2) Multivariate LR.

Univariate LR is a statistical method that formulates a mathematical model depicting relationship between the dependent variable and each independent variable.

Multivariate LR is used to construct a prediction model for goodness of design of web sites. The multivariate LR formula can be defined as follows:-

$$\text{Prob}(X1, X2 \ldots Xn) = \frac{e^{(A_0+A_1X_1+\ldots+A_nX_n)}}{1+e^{(A_0+A_1X_1+\ldots+A_nX_n)}}$$

In LR, two stepwise selection methods, forward selection and backward elimination can be used [10]. Stepwise variable entry examines the variable that is selected one at a time for entry at each step. This is a forward stepwise procedure. The backward elimination method includes all independent variables in the model. Variables are deleted one at a time from the model until stopping criteria are fulfilled.

We used forward selection method to analyze 2007-2009 webby awards data and backward elimination method for 2010 webby awards data.

V. Analysis Results

We employed statistical techniques to describe the nature of the data of the year 2007-2010 webby awards. We also apply Logistic Regression for the prediction of different models to examine differences between good and bad design. This section presents the analysis results, following the procedure described in section IV. Descriptive statistics are for every year data is presented in section A and model prediction is presented in section B

A. Descriptive Statistics

Each table [III-VI] presented in the following subsection show min, max, mean and SD for all metrics considered in this study.

TABLE III-Descriptive Statistics of year 2007 webby awards data

Descriptive Statistics				
Metrics	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Words in page title	0	23	6.22	4.01
Body text words	0	16302	2787.94	3338.00
Total words on page	27	17041	3078.74	3435.61
Total number of Links	0	921	130.66	175.42
Number of!’s	2	264	34.21	40.017
Number of graphics	0	147	25.30	31.98
Page Size	312	204372	41522.69	46382.42
Total embedded links	0	213	13.43	33.39
Number of Lists	0	117	13.12	21.77
Number of frames	0	1	0.01	0.12
Number of tables	0	89	2.33	8.21
Within page links	0	46	2.19	5.91
Total emphasized body text	0	129	6.67	16.73
Total emphasized text	0	129	6.8	16.71
Total Display Word On The Page	17	11443	2112.57	2347.33

TABLE IV:-Descriptive Statistics of year 2008 webby awards data

Descriptive Statistics				
Metrics	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Words in page title	0	22	5.83	4.05
Body text words	0	17312	2141.63	3183.18
Total words on page	27	17571	2406.36	3239.02
Total number of Links	0	838	93.23	149.69
Number of!’s	2	163	26.36	34.21

Number of graphics	0	172	22.19	32.42
Page Size	321	255663	32228.66	45002.67
Total embedded links	0	157	9.27	21.95
Number of Lists	0	95	9.42	18.58
Number of frames	0	0	0	0
Number of tables	0	109	3.31	12.39
Within page links	0	69	2.64	9.81
Total emphasized body text	0	65	4	9.064
Total emphasized text	0	65	4.23	9.17
Total Display Word On The Page	17	11813	1689.91	2239.08

TABLE V:-Descriptive Statistics of year 2009 webby awards data

Descriptive Statistics				
Metrics	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Words in page title	0	24	6.77	4.34
Body text words	1	21940	4250.93	4774.21
Total words on page	96	22110	4715.31	4727.05
Total number of Links	0	1031	185.41	216.29
Number of!’s	2	167	48.91	43.62
Number of graphics	0	201	33.01	36.95
Page Size	1054	255665	60398.7	61056.79
Total embedded links	0	334	28.16	53.63
Number of Lists	0	95	18.16	24.58
Number of frames	0	1	0.01	0.09
Number of tables	0	78	3.76	9.79
Within page links	0	69	4.43	11.48
Total emphasized body text	0	118	7.81	17.34
Total emphasized text	0	118	8.37	17.41
Total Display Word On The Page	73	18745	3402.15	3582.72

TABLE VI:-Descriptive Statistics of year 2010 webby awards data

Descriptive Statistics				
Metrics	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Words in page title	0	17	6.72	3.86
Body text words	1	34523	2942.06	4757.21
Total words on page	85	35443	3343.40	4861.28
Total number of Links	0	844	117.01	178.52
Number of!’s	2	417	43.04	67.36
Number of graphics	0	230	25.97	38.78
Page Size	930	429870	44581.33	63574.64
Total embedded links	0	239	11.80	30.63
Number of Lists	0	197	11.16	23.60
Number of frames	0	0	0	0
Number of tables	0	66	2.15	7.58

Within page links	0	70	3.10	9.41
Total emphasized body text	0	42	4.06	8.01
Total emphasized text	0	42	4.49	8.34
Total Display Word On The Page	64	28989	2400.75	3690.57

B. Model Prediction

We used Logistic Regression to discriminate good from bad pages. This technique is suitable where we have one dependent variable and many independent variables. As in our study, we have one dependent variable named good/bad and independent variables are the whole web metrics of webby awards. We built four predictive models for identifying good pages. Model 1 is with respect to data 2007, Model 2 is with respect to 2008, Model 3 is with respect to data 2009 and Model 4 is with respect to data 2010.These model predict the goodness of the website design based on the subset of metrics which we get from the Logistic regression technique. Table VII summarizes the attributes of the web page. If theses selected metrics have the higher values then we say that, these attributes contributes to a bad design of a web site.

TABLE VII:-subset of Metrics selected in each model using Logistic Regression

Data	Metrics selected
2007	Total embedded links, Number of list
2008	Words in page title
2009	Number of lists
2010	Body text words, Number of!'s, Page size, Number of lists ,Number of tables, Within page link

The description of each model is described below:-

MODEL 1:-This model is based on 2007 webby awards and the attributes which contributes towards the bad design are total embedded links and number of lists. If we have higher values of these attributes then we predict that we can have a bad design.

MODEL 2:-This model is based on 2008 webby awards. The only attributes which contributes to the bad design is words in page title means if we have higher value of this metrics then we could have a bad design.

MODEL 3:-This model is w.r.t 2009 webby awards. In this model we also get only one attribute named number of list which could lead to a bad design of a website.

MODEL 4:-Model 4 is based on webby awards of the year 2010. In this model we get many metrics which leads to a bad design like body text words, number of!'s, Page size, Number of tables and within page links. If we have higher values of these metrics we will get a bad design. For predicting model 4 we use backward elimination method of Logistic Regression.

The results of table 7 shows that we can create profile of good pages that is attributes which can be used to make a good design.

VI. CONCLUSION

The goal of this research is to capture quality of web sites. As E-business is emerging and websites are not just medium for communication, but they are also a product of providing services. Therefore imparting quality, security and reliability to web sites are very important. We empirically validate the relationship of web metrics and quality of websites using logistic regression technique the results are based on webby awards data obtained 2007-2010.

The webby awards data set is possibly the largest human – rated corpus of web sites available. Any site that is submitted for the award is examined by three judges on six criteria. It is unclear and unknown how the experts rates the website but hope we present a way towards a new methodology for creating empirically justified recommendation for designing a good web sites. In this paper we present the attributes which, if have higher value can lead to a bad design. From the above attributes we also find profile of good pages.

The type of metrics explored here are only one piece of the web site design puzzle; this work is part of a larger project whose goal are to develop techniques to empirically investigate all aspect of web site design and to develop tools to help designers of the web site to improve the quality of the web page.

VII. FUTURE WORK

In future, we replicate this work on the larger data set and we will explore the tools and methods in all dimensions with the help of that work the web site engineers will simplify and improve the quality of the web sites. Also in future we will take level 1 and level 2 web pages because the home page has different characteristics from other levels of the page .In future we will propose some guidelines to make effective web sites which are easily downloaded and have good scanability.

REFERENCES

[1] K.K. Aggarwal, Yogesh Singh, Arvinder Kaur, Ruchika Malhotra, "Empirical Study of Object- oriented Metrics", *Journal of Object Technology*, vol 5. No 8, pp. 149-173, November –December 2006.

[2] K.K Aggarwal, Yogesh Singh, "Software Engineering", 3rd edition, New Age Publication, India , 2008.

[3] Jakob Nielsen, "Designing Web Usability the Practice of Simplicity", New Riders Publishing Indianapolis, IN, 2000.

[4] Karen A. Shriver, "Dynamics in Document Design", Wiley Computer Publishing, John Wiley & Sons, Inc., New York, 1997.

[5] Lincoln D. Stein, "The rating game", http://stein .cshl.org/lstein/rater/,1997.

[6] George W. Furans, "Effective view navigation", *in proceedings of ACM CHI 97 conference on human factors in computing systems*, volume 1 of PAPERS: information structures, pp. 367-374, 1997.

[7] Kevin Larson and Mary Czerwinski., "Web page design: Implications of memory, structure and scent for information retrieval", *In proceedings of*

ACM CHI 98 Conference on human Factors in Computing Systems, volume 1 of Web Page Design , pp. 25-32, 1998.

[8] http://WWW.webby awards.com/

[9] Vincent Flanders and Michael Willis, "Web Pages That Suck: Learn Good design by Looking at bad Design" , SYBEX, San Francisco, 1998.

[10] Mayers, J.H and Forgy E.W. (1963). The Development of numerical credit evaluation systems. Journal of the American Statistical Association, Vol.58 Issue 303 (Sept) pp 799–806.

AUTHORS PROFILE

Yogesh Singh

He is a professor at the University School of Information Technology, Guru Gobind Singh Indraprastha University, Delhi, India. He is also Controller of Examinations at the Guru Gobind Singh Indraprastha University, Delhi, India. He was founder Head (1999-2001) and Dean (2001-2006) of the University School of Information Technology, Guru Gobind Singh Indraprastha University, Delhi, India. He received his master's degree and doctorate from the National Institute of Technology, Kurukshetra, India. His research interests include software engineering focusing on planning, testing, metrics, and neural networks. He is coauthor of a book on software engineering, and is a Fellow of IETE and member of IEEE. He has more than 200 publications in international and national journals and conferences. Singh can be contacted by e-mail at ys66@rediffmail.com.

Ruchika Malhotra

She is an assistant professor at the Department of Software Engineering, Delhi Technological University (formerly known as 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 for than 50 research papers in international journals and conferences. Malhotra can be contacted by e-mail at ruchikamalhotra2004@yahoo.com.

Poonam Gupta

She is an assistant professor at the Maharaja Agrasen Institute of Management Studies Department of Computer science, Delhi, India,She is pursuing M.Tech(I.T) from University School of Information Technology, Guru Gobind Singh Indraprastha Delhi, India and received her master's in software engineering from the University School of Information Technology, Guru Gobind Singh Indraprastha University, Delhi, India. Her research interests are in software metrics,software testing, improving software quality, Gupta can be contacted by e-mail at poonam.goel@gmail.com

Systematic and Integrative Analysis of Proteomic Data using Bioinformatics Tools

Rashmi Rameshwari
Asst. Professor, Dept. of Biotechnology,
Manav Rachna International University,
Faridabad, India

Dr. T. V. Prasad
Dean (R&D), Lingaya's University,
Faridabad, India

***Abstract*— The analysis and interpretation of relationships between biological molecules is done with the help of networks. Networks are used ubiquitously throughout biology to represent the relationships between genes and gene products. Network models have facilitated a shift from the study of evolutionary conservation between individual gene and gene products towards the study of conservation at the level of pathways and complexes. Recent work has revealed much about chemical reactions inside hundreds of organisms as well as universal characteristics of metabolic networks, which shed light on the evolution of the networks. However, characteristics of individual metabolites have been neglected in this network. The current paper provides an overview of bioinformatics software used in visualization of biological networks using proteomic data, their main functions and limitations of the software.**

Keywords- Metabolic network; protein interaction network; visualization tools.

I. Introduction

Molecular interaction network visualization is one of the most user-friendly features developed for the simulation process of biological interactions [29]. Drawing of any molecule for example, protein may seem to be easy but generating the same protein with all the types of conformation that it can attain during any interactions and to simulate this process is quite difficult. In this context one of the greatest manually produced molecular structures of its time was done by Kurt Kohn's 1999 map of cell cycle control. Protein interactions network visualization deals with territory that is very similar to that of protein-protein interaction prediction, but differs in several key ways. proteomics data are often associated with pathways or protein interactions, and both of these are easily visualized as networks[22]. Even types of data not normally viewed as networks (*e.g.* microarray results) are often painted onto signaling, metabolic, or other pathways or protein interaction networks for visualization and analysis.

Visualization and analysis tools are commonly used to interact with proteomic data. Most, visualization tools were developed simply for illustrating the big picture represented by protein-protein interaction data or expression data, for qualitative assessment, not necessarily for quantitative analysis or prediction. Expression and interaction experiments tend to be on such a large scale that it is difficult to analyze them, or indeed grasp the meaning of the results of any analysis. Visual representation of such large and scattered quantities of data allows trends that are difficult to pinpoint numerically to stand out and provide insight into specific avenues of molecular functions and interactions that may be worth exploring first out of the bunch, either through confirmation or rejection and then later of significance or insignificance to the research problem at hand. With a few recent exceptions, visualization tools were not designed with the intent of being used for analysis so much as to show the workings of a molecular system more clearly.

Visualization tools also do not actually predict molecular interactions themselves or their characteristics. On the contrary, visualization tools only create a graphical representation of what is already "known" in literature and in molecular interaction repositories such as the Gene Ontology (GO) [7]. A side-effect of displaying interaction networks by treating some proteins as members of more general family groupings or going by interactions in different tissues or as analogues in other tissues or organisms is the apparent display of a protein-protein or other molecular interaction that is inferred but may or may not have actually been observed and documented, something that could be misconstrued as a predicted interaction.

The most significant difference between molecular interaction network visualization and molecular interaction prediction is the nature of the information it provides [11]. Protein-protein interaction prediction is characterized by its concern with how proteins will interact, where they will interact, under what conditions they were interact, and what parts are necessary for their interaction. These characteristics are governed by physical and chemical properties of the proteins or other molecules involved, which may be actual molecules that have been described through extensive proteomics experiments or hypothetical, *in silico* generated species that are being investigated for pharmaceutical and other applications. Interaction network visualization tools are given no knowledge of physical or chemical properties of the proteins that, why they interact. As a result, the information they inadvertently impart concerns only whether or not certain proteins putatively interact with certain other proteins, not how they interact, when they interact, or why they interact.

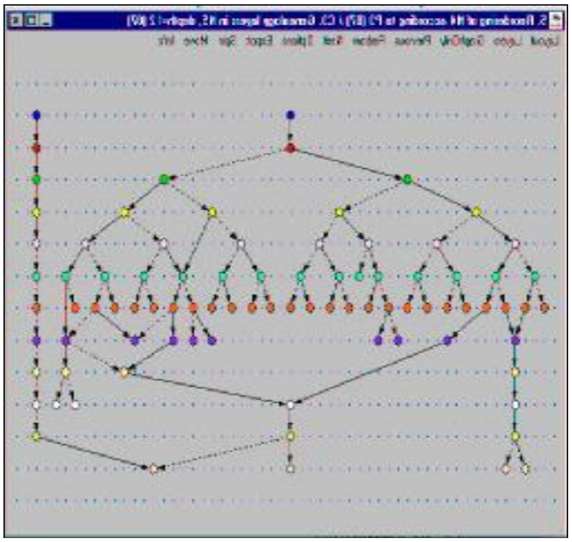


Figure:(A)

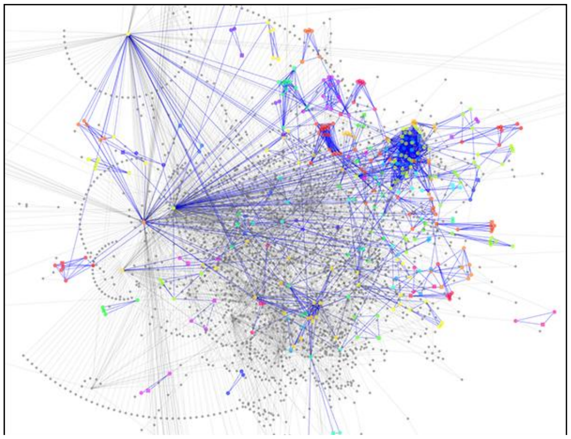


Figure: (B)
Figure1: Comparative visualization of Protein-Network drawn by different tools like Pajek (Fig.A), and Cytoscape (Fig. B). Here genes are represented as nodes and interaction as edges.

II. Background and Related Work

With the recent advances in high-throughput technologies, software tools have been developed to visualize and analyze large-scale data. This paper deals with various visualization techniques for proteomic data. Major emphasis is on network graph generated during protein-protein interaction. Many tools are being used for this purpose which is based on different algorithm. For example, Pajek [6] and Cytoscape [8] use force directed layout algorithm which produces graph by computing force between pairs of nodes in each iteration of the optimization process. The networks can be visualized as indicated in Fig 1A and Fig 1B.

As protein interactions data also helps in study related to evolutionary analysis. In a new era, it is necessary to understand how the components which involve in biological systems from the various biological data and knowledge of components at molecular level. It reveals the structure of biological systems and lead to “ontological” comprehension of biological systems. Comprehensive data of protein interactions is also suitable for systems level evolutionary analysis.

There are many commercial software such as Ingenuity Pathway Analysis (Figure 2), MetaCore and Pathway studio, designed to visualize high-throughput data in the context of biological networks. Biological networks have a scale-free and modular organization. In a scale free networks the degree of distribution follows a power-law, which means that only a small number of nodes, called hubs are highly connected [12]. Hubs usually play essential roles in biological systems [13]. On the other hand, groups of proteins with similar functions tend to form clusters or modules in the network architecture. Many commercial software for network visualization follow this law.

Metacore is an integrated knowledge database and software suite for pathway analysis of experimental data and gene lists. It is based on manually curetted databases of human protein-protein, Protein-DNA and protein compound interactions. This package includes easy to use, intuitive tools for search, data visualization, mapping and exchange, biological networks and interactomes [32].

The other software tool known as Pathway studio, is based on a mammalian database, named: ResNet 5 mammalian, which is generated by text mining of the PubMed database and 43 full text journals [14]. The advantages of using this tool are that it increases the depth of analysis of high-throughput data generation experiments like microarray gene expression, proteomics, metabolomics. Enables data sharing in a common analysis environment. This tool simplifies keeping up to date with the literature and brings this knowledge into an analysis environment. This also enables visualization of gene expression values and status in the context of protein interaction networks and pathways. However, free software like Pathway Voyager (Figure 5), GenMapp and Cytoscape are also available. Pathway Voyager applies flexible approach that uses the KEGG database [27] to pathway mapping [17]. GenMapp (Figure 6) is also designed to visualize gene expression data on maps representing biological pathways and gene groupings. GenMapp has more option which can modify or design new pathways and apply complex criteria for viewing gene expression data on pathways [14].

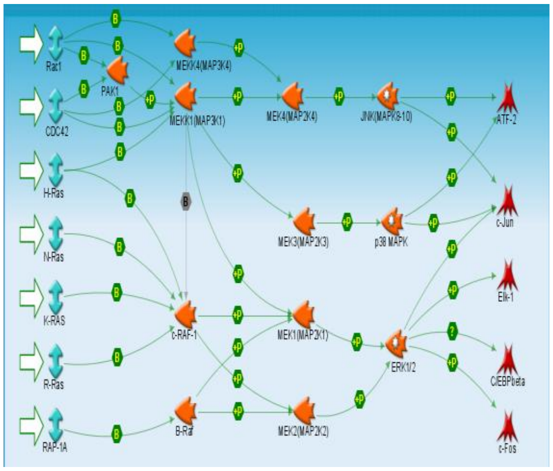


Figure 2: G-Protein signaling_Ras family GTPases in kinase cascades. Image generated by Ingenuity Pathway Analysis.

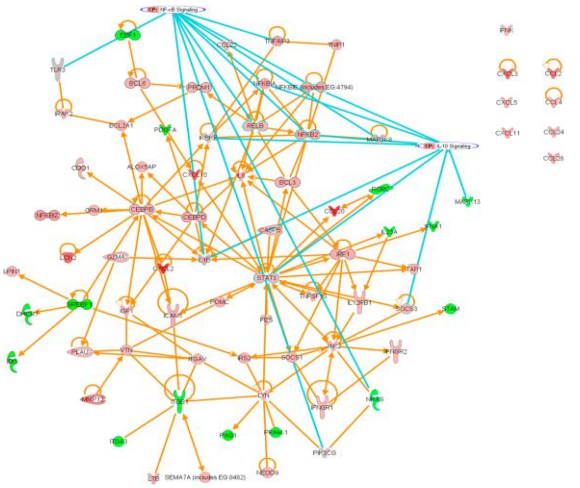


Figure 3: A Network generated by Metacore Software [32]

III. Limitations and Future Developments

After the completion of Human Genome Project, 36,000 genes were discovered, which has potential to synthesize more than 1,00,000 proteins [19], less than 50% of these genes can be assigned a putative biological function on the basis of sequence data [18]. With the advancement of technology many software has been designed to explore biological networks, like protein interactions network, protein-DNA interactions etc., that are based on databases like DIP [31], MINT [30], of mammalian database. The tools represented in this paper are applicable to a wide range of problems and their distinct features make them suitable for a wide range of applications.

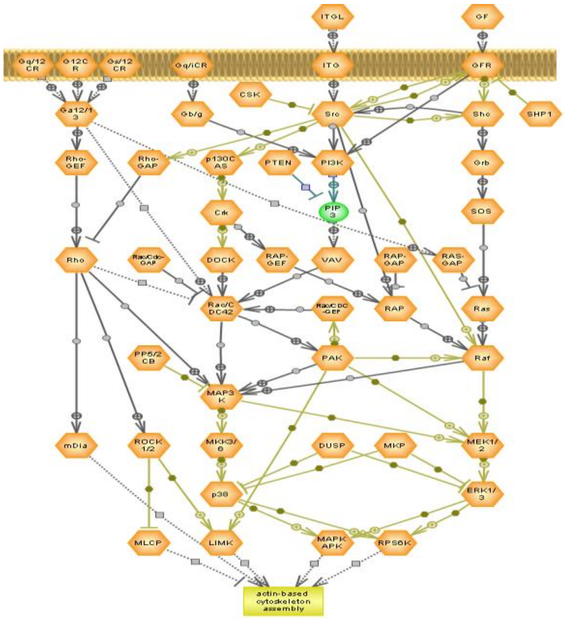


Figure 4: Actin Cytoskeleton Regulation, network generated by Pathway Studio

Clustered graphs are common occurrences in the biological field. Examples of pre-clustered graphs include clustering of proteins or genes based on some biological functionality, structural geometry, expression pattern or other chemical property [11].

In any "post- omic" analysis two essential concepts must be applied to understand biological functions at a systems level. First integrate different levels of information and second, view cells in terms of their underlying network structure. The information about biological entity is scattered in different databases [28]. Hence, the information retrieval from diverse databases is done, which is bit time consuming. Current databases are good for the analysis of a particular protein or small interaction networks. But they are not as useful for integration of complex information on cellular regulation pathways, networks, cellular roles and clinical data and they lack coordination and the ability to exchange information between multiple data sources [20]. There is need of software that can integrate information from different database as well as from diverse sources.

To analyze data, at present many software are there like Ingenuity Pathway analysis, Metacore and Pathway Studio. They work on owner curetted database at the same time there high price make them unaffordable for academic institute to use them. At the same time Cytoscape which has many properties for visualization of high throughput data, can be alternative for users [15].

However network constructed with Cytoscape are sometimes liable to show errors. So there is need to improve quality of available curetted databases and also to develop integrative knowledge bases that are especially designed to construct biological networks.

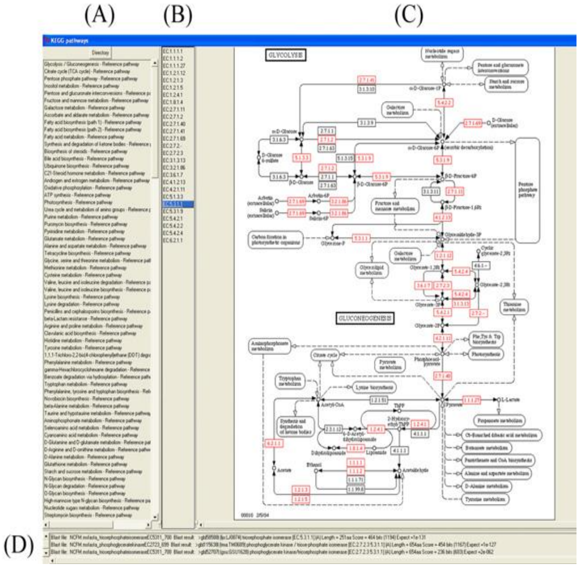


Figure 5: Interactive KEGG Pathway display. The screenshot illustrates KEGG pathway mapping for the glycolysis / gluconeogenesis pathway using the predicted ORFeome of the GAMOLA annotated *L. acidophilus* NCFM genome as query template [19].

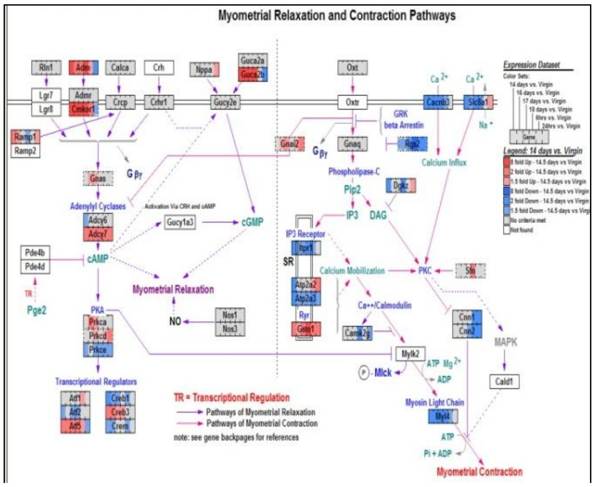


Figure 6: Myometrial Relaxation and contraction Pathway, Image generated by GenMapp.

A metabolic network is a reliable source of information. The reconstruction of GRNs is largely promoted by advances in high-throughput technologies, which enable to measure the global response of a biological system to specific interventions. For instance, large-scale gene expression monitoring using DNA microarrays is a popular technique for measuring the abundance of mRNAs. However, integration of different types of 'omics' data from genomics, proteomics and metabolomic studies can be undertaken. Although the metabolic network has important features for drug discovery, its use in case of human is very limited [18]. Further, Proteomics may yield crucial information on the regulation of biological functions and the mechanism of diseases. In this sense it is a highly promising area for drug discovery. Hence, additional efforts will be required for metabolic network reconstruction and analysis.

IV. CONCLUSION

From the above tools it can be concluded that metabolic pathways stored as directed acyclic graphs can be considered a basic concept for the visualization tool for metabolic pathway. With respect to visualization, single network views provide little more than brief glimpses of the large datasets. Visualization tools need to support many different types of views, each network view at a different level of detail. Dynamic navigation from one view to another will be a key to showing the connection between different views. Navigating from one time series point to another, for instance, could involve a view showing only the differences between the two time points. If the time points are consecutive, the number of differences will tend to be quite small. A similar approach could be applied to sub-cellular localization information as well. To adequately address each of these issues, active cooperation is required between a variety research fields including graph drawing, information visualization, network analysis and of course biology. Though all the mentioned tool differs significantly in its approach for pathway reconstructions. Hence for future a tool is needed which describe all pathways that make up a cell and how they interact as a system in the overall physiology of an organism is required.

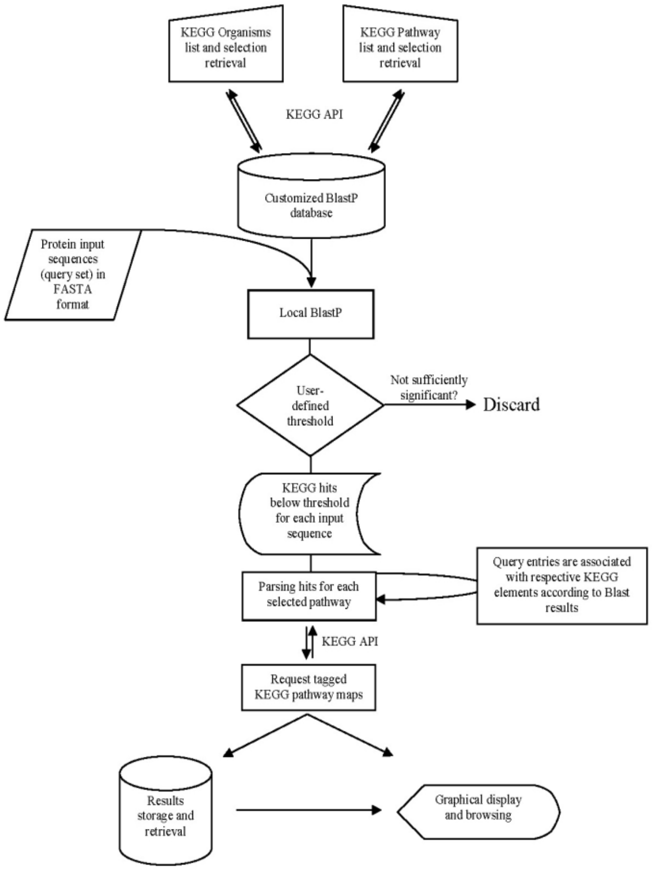

KEGG Organisms list and selection retrieval
KEGG Pathway list and selection retrieval
KEGG API
Customized BlastP database
Protein input sequences (query set) in FASTA format
Local BlastP
User-defined threshold
Not sufficiently significant?
Discard
KEGG hits below threshold for each input sequence
Parsing hits for each selected pathway
Query entries are associated with respective KEGG elements according to Blast results
KEGG API
Request tagged KEGG pathway maps
Results storage and retrieval
Graphical display and browsing

Figure 7: Pathway voyager mapping procedure [17].

REFERENCES

[1] Lenzerini M: Data Integration: A Theoretical Perspective. PODS 2002 , 243-246.

[2] B., Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.Fraenkel, E., Jaakkola, T.S., Young, R.A., Gifford, D.K. (2003). Computational discovery of gene module and regulatory networks. Nat Biotechnol. 21(11):1337-42.

[3] Fields S, Song O: A novel genetic system to detect protein-protein interactions. Nature 1989, 340:245-246.

[4] Spirin, V. and Mirny, L.A. (2003) PNAS, 100,12123-128

[5] Jeong H, Mason SP, Barabasi AL, Oltvai ZN: Lethality and centrality in protein networks. Nature 2001, 411:41-42.

[6] Batagelj V, Mrvar A: Pajek – Program for Large Network Analysis. Connections 1998, 21:47-57.

[7] The Gene Ontology Consortium: Gene Ontology: tool for the unification of biology. Nat Genet 2000, 25:25-29.

[8] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003, 13(11):2498-2504.17

[9] Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R: IntAct: an open source molecular interaction database. Nucleic Acids Res 2004, (32 Database):D452-455

[10] Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update Nucleic Acids Res, . 33, D418–D42.

[11] Rashmi Rameshwari and Dr. T. V. Prasad, "A Survey of Various Protein-Protein Interaction Tools", Proc. of National Conference on

Advances in Knowledge Management NCAKM '10, Lingaya's University, Faridabad, 2010

[12] Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5:101-13

[13] Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, *et al*. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 2004;430:88-93.

[14] Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet 2002;31:19-20.

[15] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, *et al*. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res 2003;13:2498-504.

[16] Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I. Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics 2004;20:604-11.

[17] Altermann E, Klaenhammer TR. PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. BMC Genomics 2005;6:60.

[18] Overington JP, Al Lazikani B, Hopkins AL. How many drug targets are there? Nat Rev Drug Discov 2006;5:993-6.

[19] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, *et al*. Initial sequencing and analysis of the human genome. Nature 2001;409:860-921.

[20] Tucker CL, Gera JF, Uetz P. Towards an understanding of complex protein networks. Trends in Cell Biology 2001;11:102-6.

[21] Funahashi,A. *et al*. (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico*, **1**, 159–162.

[22] Kell,D.B. (2006) Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Discovery Today*, **11**, 1085–1092.

[23] Kitano,H. *et al*. (2005) Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol.*, **23**, 961–966.

[24] Nikitin,A. et al. (2003) Pathway studio –- the analysis and navigation of molecular networks. Bioinformatics, 19, 2155–2157.

[25] Altermann E, Klaenhammer TR: GAMOLA: a new local solution for sequence annotation and analyzing draft and finished prokaryotic genomes. *OMICS* 2003, **7:**161-169.

[26] Goesmann A, Haubrock M, Meyer F, Kalinowski J, Giegerich R: Path-Finder: reconstruction and dynamic visualization of metabolic pathways**.** *Bioinformatics* 2002, **18:**124-129.

[27] Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. andKanehisa,M. (1999) Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res., 27, 29–34.

[28] Alfarano,C. et al. (2005) The biomolecular interaction network database and related tools 2005 update. Nucleic Acids Res., 33, 418–424.

[29] Kohn,K.W. et al. (2006) Molecular interaction maps of bioregulatory networks: a general rubric for systems biology. Mol. Biol. Cell, 17, 1–13.

[30] Zanzoni,A. et al. (2002) MINT: a molecular interaction database. FEBS Lett., 513, 135–140.

[31] Xenarios,I. et al. (2000) DIP: the database of interacting proteins. Nucleic Acids Res., 28, 289–291.

[32] MetaCore Pathway Analysis Software, available at www.genego.com

AUTHORS PROFILE

Ms. Rashmi Rameshwari has received two master's degree from one from T.M.B.U., Bhagalpur, Bihar and other from Jamia Hamdard, New Delhi in the area of biotechnology and bioinformatics respectively. She is currently associated with Manav Rachna International University as Assistant Professor in Dept. of Biotechnology. Her research interests include systems biology, proteomics, Microarray Technology, Chemoinformatics, etc.

Dr. T. V. Prasad received his master's degree in Computer Science from Nagarjuna University, AP India and doctoral degree from Jamia Milia Islamia University, New Delhi, India. With over 16 years of academic and professional experience, he has a deep interest in planning and executing major IT projects, pursuing research interest in CS/IT and bioinformatics. He has authored over 65 publications in reputed journals and conferences. He has also authored 4 books. He has also held respectable positions such as Deputy Director with Bureau of Indian Standards, New Delhi. His areas of interest include bioinformatics, artificial intelligence, consciousness studies, computer organization and architecture. He is a member of reputed bodies like Indian Society of Remote Sensing, Computer Society of India, APBioNet, International Association of Engineers, etc.

Table 1
COMPARATIVE STATEMENT OF VARIOUS NETWORK VISUALIZATION AND ANALYSIS SOFTWARE

S. No.	Parameters/ Features	Ingenuity Pathway Analysis	Metacore	Pathway Studio	GenMAPP	Cytoscape	Pathway Voyager	PathFinder
1	Developed by	Ingenuity Systems Inc.	GeneGo Inc.	Ariadne Genomics Inc.	Gladstone Institutes	Paul Shanon, Andrew Markiel et al.	Eric Altermann and Todd R Klaenhammer	Alexander Goesmann et.al
2	Description	Database of Biological networks created from millions of relationships, between proteins, genes ,complexes, cells, tissues, drugs and diseases	A manually curetted database of human Protein-protein interaction & Protein-DNA interactions, transcriptional factors, Signaling, metabolism and bioactive molecules	These databases are a collection of eukaryotic molecular interactions generated by MedScan Text to knowledge suit using the entire PubMed database and 43 full text journals. Also works with public database of signaling and biochemical pathways	Archived Maps were drawn based on textbooks, articles and public pathway databases or generated from the public database maintained by the Gene Ontology Project	Software for integrating biomolecular interaction Networks with high-throughput expression data and other molecular states into a unified conceptual framework	Utilizes the KEGG online database for pathway mapping of partial and whole prokaryotic genomes	A tool for the dynamic visualization of metabolic pathways based on annotation data.
3	Availability	Commercial	Commercial	Commercial	Public	Public	Public	Public
4	Based on database	Ingenuity pathways knowledge base	Human Database	Mammalian Database: ResNet 5, ResNet plant database, KEGG, BIND, HPRD	KEGG	KEGG	KEGG	RDBMS is usec along with KEGG
5	Web access	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
6	Platform	Java 1.5 or higher	Java	Python	Java	Java	Perl/TK	Perl/TK
7	Special features	Solution is given to Pharmaceutical, Bio-technology, and Academics	unique ability to concurrently visualize multiple types of experimental data such as gene expression, proteomic, metabolomics, SAGE, MPSS, SNP, HCS, HTS, microRNA and clinical and pre-clinical phenotypic data	Analyze proteomic, Metabolomics and other high throughput data.	has graphics tools for constructing and modifying pathways. Used for analyzing microarray data include statistical filters and pattern finding algorithms such as hierarchical clustering	mRNA Expression profiles, Gene annotations from Gene Ontology (GO) & KEGG. Incorporates statistical analysis	No dedicated hardware or software are necessary to analyze given datasets	Aim at comparing pathways at microscopic level and therefore it can be used for dynamic visualizations of metabolisms from a whole genome perspective
8	Drawbacks	Limited to Human, Mouse, Rat and Canine	Specific server is required	It offers a wizard interface for creating very simple network and data queries and only Biological Networks provides a language interface for expressing such queries.	Generally used to explore Microarray data	Focuses only on high-level networks, low-level models of components and interactions addressed by ongoing projects such as Ecell (Tomita et al. 1999), VirtualCell; mechanisms for bridging high-level interactions with lower level, physico-chemical	For certain selectable pathways (e.g. Ribosomal reference pathway) KEGG does not yet support organism independent marking. For practical reasons, no hits will be displayed for these pathways	Uses RDBMS based internet application. Being integrated into the locally developed genome annotation system GENDB with extended functionality

S. N o.	Parameters/ Features	Ingenuity Pathway Analysis	Metacore	Pathway Studio	GenMAPP	Cytoscape	Pathway Voyager	PathFinder
						models of specific biological processes are required		
9	Integrated with	Only based on Ingenuity Product	Gene Go	MedScan	MAPP	Numerous	GAMOLA	GENDB
10	Storing result	Present	Present	Present	Present	Present	Present	Present
11	Data format	Not Specific.	Not Specific	ResNet Exchange XML formats.	CSV, GPML, WikiPathways , MAPP	PSI-MI, GML/XGMML	Fasta Files	EMBL OR Genbank
12	Graph comparison for species	Present	Present	Present	Present	Present	Present	Present
13	Graphical user Interface	Present	Present	Present	Present	Present	Present	Present
14	Visualization Technique	Present	Present	Present	Present	Present	Present	Present
15	Ease of use and report generation	Excellent	Excellent	Excellent	Excellent	Very Good	Good	Good
16	Graphical Representation	Present	Present	Present	Present	Present	Present	Present
17	Classification Technique	Compare affected pathways and phenotypes across time, dose, or patient population.	Disease, Tissue, Species, Sub cellular localization, Interactions, Metabolites	Interrogate Different species, Multiple genomes	Hierarchical clustering	None	Hierarchical clustering	Chunks and Subway
18	Web browser	Internet Explorer 6.0 or higher	Internet Explorer 6.0 or higher	Internet Explorer 6.0 or higher	Internet Explorer 6.0 or higher	Internet Explorer 6.0 or higher	Internet Explorer 6.0 or higher	Internet Explorer 6.0 or higher
19	Memory	512 Mb (minimum), 1GB (Recommended)	4 GB	1 GB	512 Mb	512 Mb	512 Mb	512 Mb
20	Operating System supported	Vista, Window XP and Macintosh 10.3,10.4, 10.5	Linux 2.1, 3.0 or Red Hat 9.0	Windows	Windows	Window	Windows	UNIX, Windows
21	Reference URL	www.ingenuity.com/	www.genego.com	www.ariadnegenamics.com	www.genmapp.org/	www.cytoscape.org	www.bioinformatics.ai.sri.com/ptools	http://bibiserv.TechFak.UniBielefeld.DE/pathfinder/
22	References	[12]	[32]	[24]	[14]	[15]	[17]	[25], [26]

A Conceptual Framework for an Ontology-Based Examination System

Adekoya Adebayo Felix
Department of Computer Science,
University of Agriculture, Abeokuta,
Nigeria

Akinwale Adio Taofiki
Department of Computer Science,
University of Agriculture, Abeokuta,
Nigeria

Sofoluwe Adetokunbo
Department of Computer Science,
University of Lagos,
Lagos, Nigeria

Abstract— **There is an increasing reliance on the web for many software application deployments. Millions of services ranging from commerce, education, tourism and entertainment are now available on the web, making the web to be the largest database in the world as of today. However, the information available on the web is syntactically structured whereas the trend is to provide semantic linkage to them. The semantic web serves as a medium to enhance the current web in which computers can process information, interpret, and connect it to enhance knowledge retrieval. The semantic web has encouraged the creation of ontologies in a great variety of domains. In this paper, the conceptual framework for an ontology-based examination system and the ontology required for such examination systems were described. The domain ontology was constructed based on the Methontology method proposed by Fernández (1997). The ontology can be used to design and create metadata elements required developing web-based examination applications and can be interoperate-able with other applications. Taxonomic evaluation and the Guarino-Welty Ontoclean techniques were used to assess and refined the domain ontology in other to ensure it is error-free.**

Keywords- semantic web; examination systems; ontology; knowledge bases.

I. INTRODUCTION

The world wide web is a huge library of interlinked and non-machine interpretable documents that are transferred by computers and presented to users. Thus, an information user semi-automatically connects and interprets the information.

The semantic web serves as a medium to enhance the current web in which computers can process information, interpret, and connect it to enhance knowledge retrieval. The semantic web is an XML-based ontological application that provides intelligent access to heterogeneous and distributed information. According to Berners-Lee [5] and [2], the Semantic web enable machine-readable metadata to be added to agents in order to efficiently facilitates useful tasks such as improved search, resource discovery, information brokering and filtering.

The semantic web has encouraged the creation of ontologies in a great variety of domains. Ontologies are being used as formal knowledge representation scheme and are used in designing and creating meta-data elements. It provides taxonomy for domain of discourse, and a set of constraints, relationships and rules between concepts in the taxonomy. Its use has enabled automated acquisition, retrieval and reuse of knowledge and improved software engineering activities through automated code generation, knowledge access etc. Ontology provides a sharable structure and semantics in knowledge management, e-commerce, decision-support and agent communication [6].

In this paper, we described the conceptual framework for an ontology-driven semantic web examination system. Succinctly, the paper described an ontology required for developing examination systems for the semantic web.

II. LITERATURE REVIEW

An examination according to [20] “is an educational activity well organized to test, measure and consequently evaluate the cumulative knowledge of students in their academic endeavours”. It involved an agreed syllabus, same set of conditions and predetermined response time to same questions administered to student(s). [3] stated that “electronic examination system involves the conduct of examinations through the web or the intranet and it reduces the large proportion of workload on examination, training, grading and reviewing”. Web-based examination is on the increase, most educational and professional institutions are progressively using the internet to deliver their services especially examinations to their customers.

The semantic grid uses open standards to enable distributed computers shares computing resources as well as information over the internet [22]. The application of semantic grid in education has provided huge opportunities for academic institutions especially universities to aggregate disparate information technology components to create a unified single system ([11], [9]. The semantic education grid involves development and deployment of large-scale distributed, open and comprehensive end-to-end infrastructure educational applications across academic institutions using web service and grid technology in other to enhance improved teaching and learning quality, and also to expand the global scope of educational services [15].

III. CONCEPTUAL FRAMEWORK

The semantic examination grid is the combination of the electronic examination and the semantic web technologies and is intended to be part of the large semantic education grid presented in Figure 1. It provides a convenient means of designing semantic based question banks that can handle large set of questions while avoiding repetitions in the questions. It also offers an effective way of scheduling examination personnel to examination venues, periods and specific examinations.

An exam ontology which can be used or reused to develop web-based examination applications that would meet the outlined semantic requirements stated above was designed as a core component in the proposed semantic examination grid.

Web-based application developers who are interested in web-based examination delivery services would benefit from the exam ontology. The exam ontology provides a common vocabulary of examination administration with different levels of formality, the meaning of the terms and the relationships between them. The exam ontology is expected to provide common and shared knowledge about examination administration to web developers. It would also allow web-based examination applications to share information. It serves as knowledge bases which can be accessed in a language and platform independent manner to support e-examination. Figure 2 illustrates the conceptual framework for an ontology-driven semantic web examination system.

IV. OWL-DL EXAM ONTOLOGY

The exam ontology was constructed based on the Methontology method proposed by Fernández (1997) and covers five processes namely, specification, conceptualisation, formalisation and implementation and maintenance.

A glossary of terms to be included on the ontology, their natural language definition and their synonyms and acronyms was first developed. The terms were classified into one or more taxonomies of concepts, where a concept is an abstraction for one or more terms.

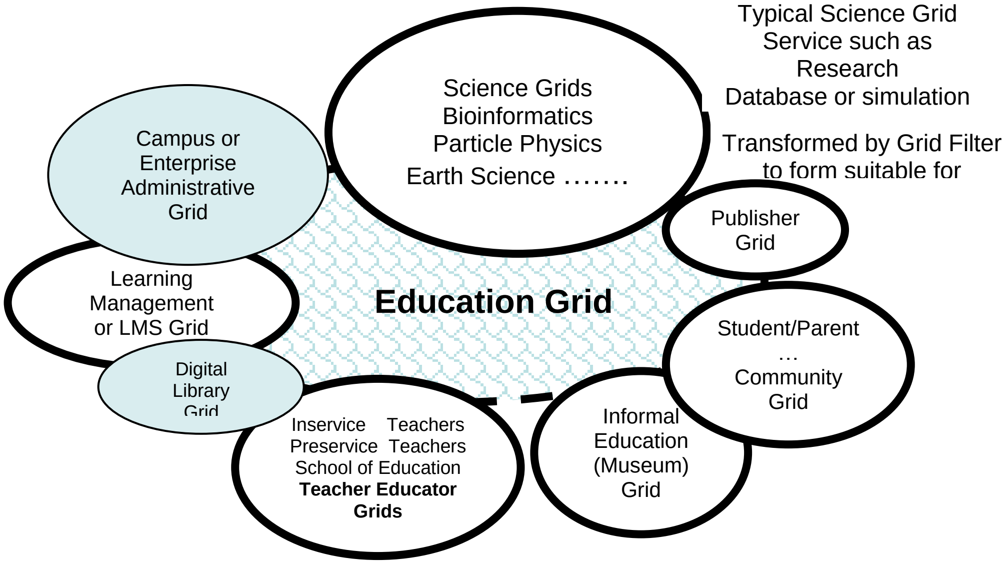


Figure 1: The Semantic Educational Grid as Grid of Grids

Figure 3 illustrates the concepts inherent in the exam ontology using hierarchy of classes by classifying taxonomy with UML. As described in figure 3, ExamAdmin is the parent class in the exam ontology. This parent class is further break down into eight other child classes namely ExamType, ExamFormat, ExamMedia, ExamMaterials, ExamPolicy, ExamTerm, ExamPersonel and ExamMalpractices. Each of these child classes is further sub-divided into child classes and so on as shown in figure 3. This class hierarchy of concepts and relations makes the contents of the exam ontology more manageable and facilitates integration of additional concepts when required in the future. Also, it enables precise and powerful search results when the exam ontology is implemented in a hierarchical application.

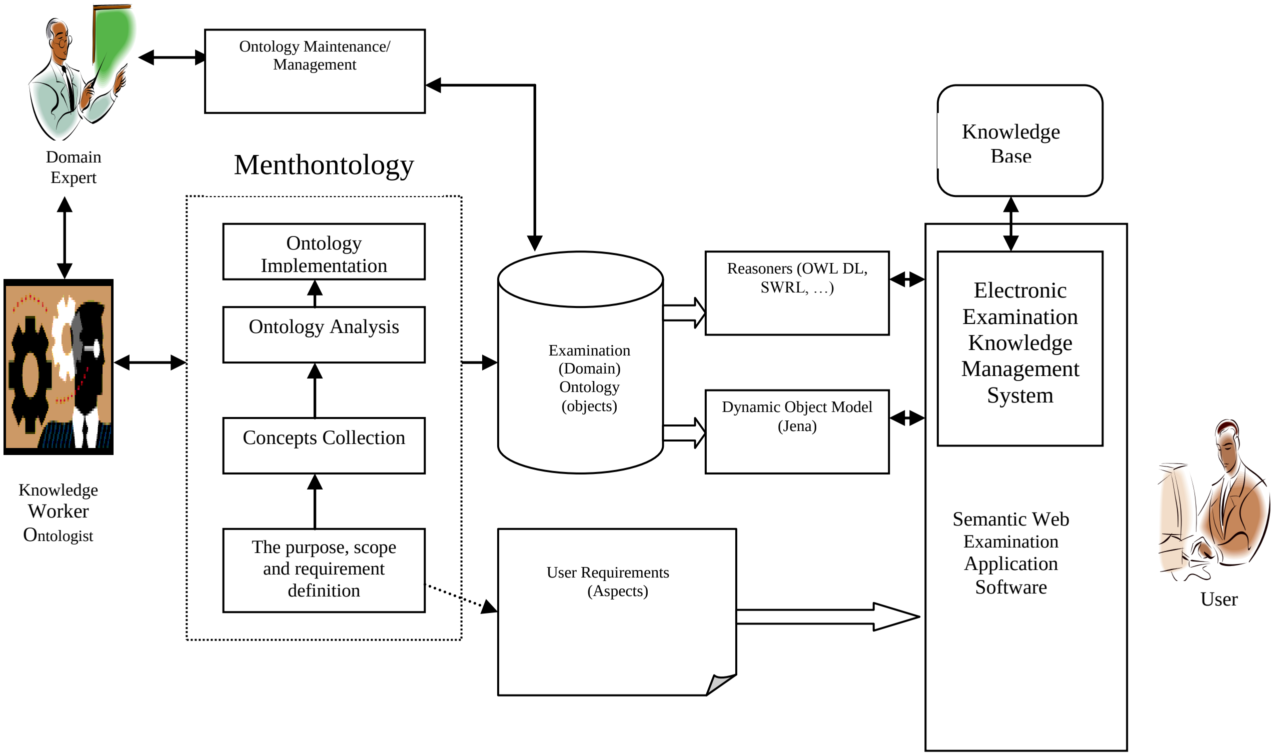

Ontology Maintenance/ Management
Domain Expert
Menthontology
Knowledge Base
Ontology Implementation
Ontology Analysis
Concepts Collection
The purpose, scope and requirement definition
Examination (Domain) Ontology (objects)
Reasoners (OWL DL, SWRL, …)
Dynamic Object Model (Jena)
Electronic Examination Knowledge Management System
Semantic Web Examination Application Software
User Requirements (Aspects)
Knowledge Worker Ontologist
User

Figure 2: An ontology-driven semantic web examination application

The ontology relation diagram designed for the exam ontology is illustrated in Figure 4. The ontology relations defined the ad hoc relations between concepts in the domain ontology and also with concepts in other ontologies. Relations are determined by their names and the source and target concepts.

V. IMPLEMENTATION

Figure 5 shows how the various classes in the exam ontology relate to each other within the Protégé_4.0_alpha application. Unlike the class diagram, the protégé class hierarchy shows the various sub-classes and relations that were added to the parent class in the exam ontology. For instance, ExamPersonnel is a sub-class to ExamAdmin as relation of

ExamAdmin(ExamPersonel). The figure 6 below shows the OWL visualization of a section of the examination ontology which covers the sub-class "ExamPersonnel". The classification of people who are concerns with the administration of examinations would facilitate effective scheduling of personnel to examination venues, subjects and even to supervise other personnel. An instance of any sub-class could be created such as Supervisor, Venue, Invigilator, Question, etc .

The examination ontology was developed with formal semantic flavour in order to improve access to information stored on the web. The examination ontology was designed as a web application which can be integrated with other applications to provide access to information. The sample output of the examination ontology is displayed in the web browser of figure 7 to describe an instance of sub-class supervisor.

The browser output is divided into three parts namely content, all resources and ontology. The content part shows the semantic link to the exam ontology, the resources, classes, and objects and data type properties.

The various classes and attributes are displayed in the all resources part. Each resource is semantically related to the other relevant resources. The ontology part displays how the various classes are represented in the ontology.

For instance, the sub-class "Supervisor" displayed in figure 7 is a sub-class of ExamPersonnel" which in turn is a sub-class of ExamAdmin. The output reveals that the sub-class "Invigilator" is a disjoint class to the sub-class displayed which implies that an individual cannot belong to these two sub-classes at the same time.

The relationship between the sub-classes is also displayed along with other information. For instance, the relationship between "Invigilator" and "Supervisor" is shown in figure 8, which means that there is an individual named "Tony Chris" who belongs to the set of invigilators and who is being supervised by another individual named "Okechukwu Adams".

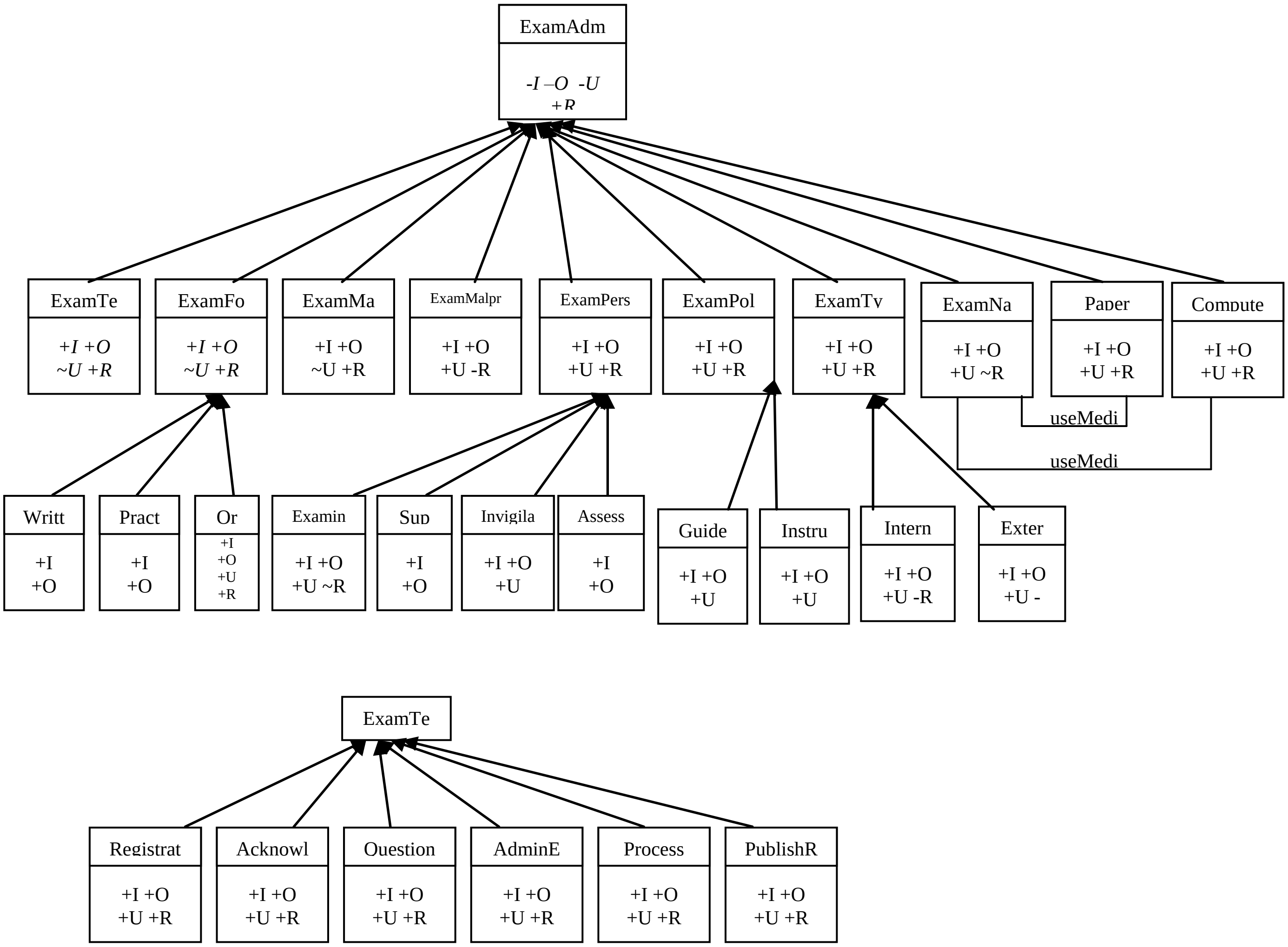


Figure 3: The modified exam ontology taxonomy after applying OntoClean Method

VI. Conclusion

In this paper, we have described an examination ontology based which was developed based on the methontology ontology development technique proposed by Fernández. The relevant concepts which characterize the domain of discourse were identified, appropriately defined along with their binding relationships and slots, and were classified based on the inherent concepts they described. The concepts were presented using a tree-like class hierarchy which shows the relationship between the super-class concept and the sub-class concepts. The ontology was developed with Protégé_4.0_alpha which is based on the OWL-DL. The consistency check and computation of the inferred ontology was done with FaCT++ reasoner and the validity of the ontology was confirmed. The exam ontology was developed to provide a knowledge base for the semantic examination grid. Information regarding a specific examination - persons, questions, date – can be obtained with ease referencing. The ontology was designed with a view to permit integration of additional concepts in the future and at different levels of content granularity. The exam ontology is interoperable which can be reused or integrated into electronic examination applications to facilitate efficient information access and retrieval.

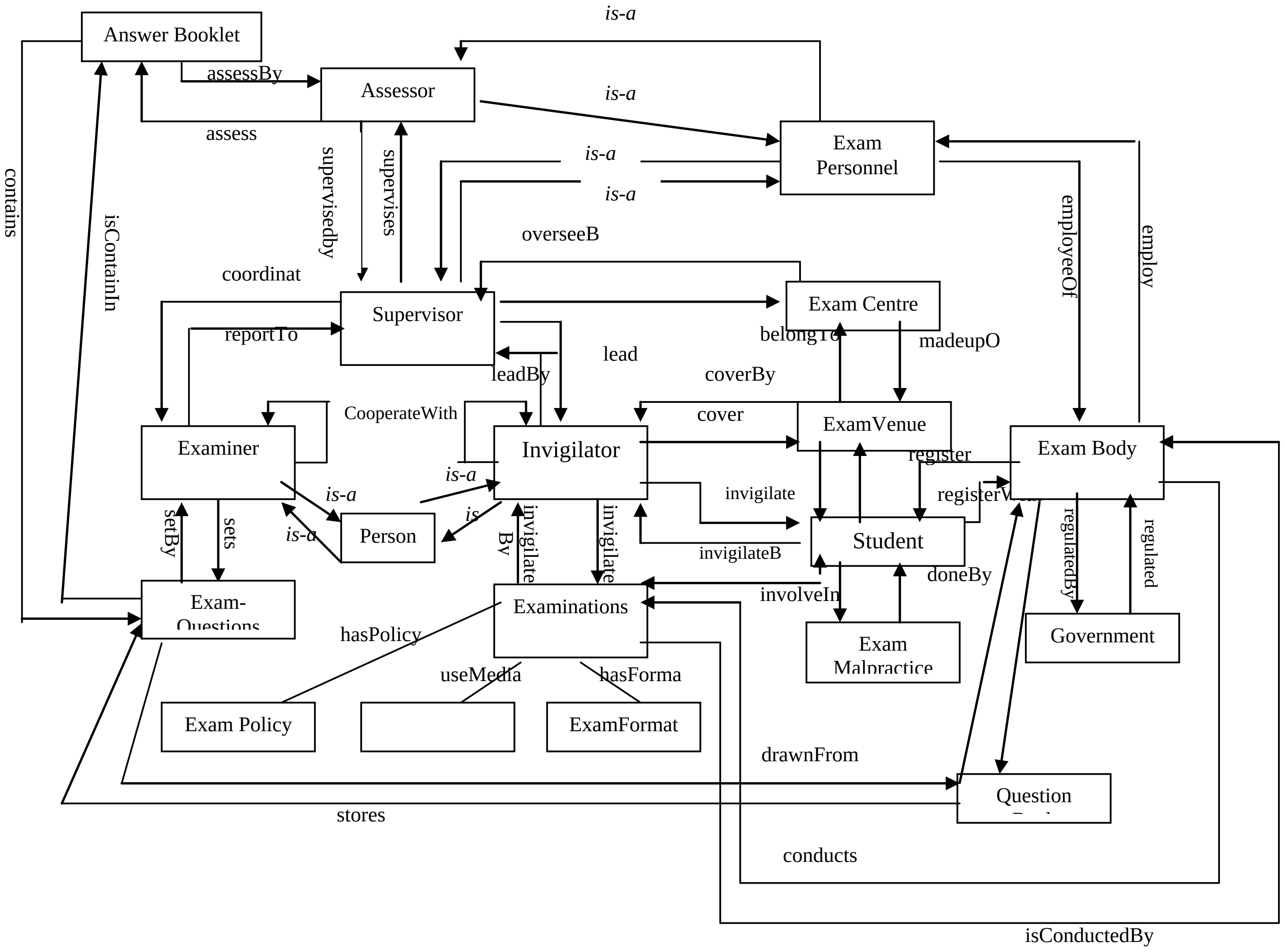
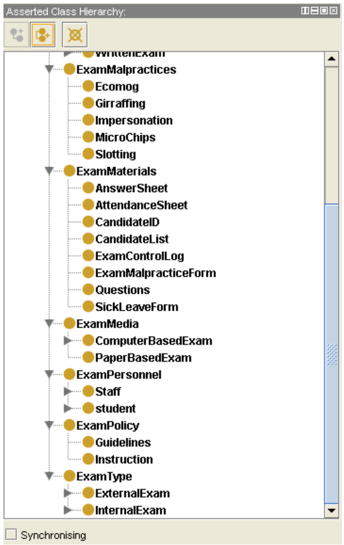


Figure 4. Diagram of Binary Relations *in (KA)*



- WrittenExam
- ExamMalpractices
  - Ecomog
  - Girraffing
  - Impersonation
  - MicroChips
  - Slotting
- ExamMaterials
  - AnswerSheet
  - AttendanceSheet
  - CandidateID
  - CandidateList
  - ExamControlLog
  - ExamMalpracticeForm
  - Questions
  - SickLeaveForm
- ExamMedia
  - ComputerBasedExam
  - PaperBasedExam
- ExamPersonnel
  - Staff
  - student
- ExamPolicy
  - Guidelines
  - Instruction
- ExamType
  - ExternalExam
  - InternalExam

Figure 5: Protégé_4.0_alpha class hierarchy of exam ontology

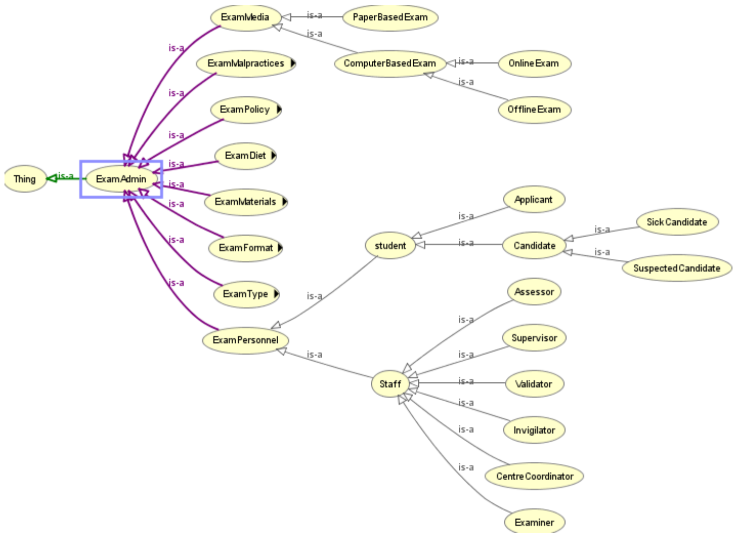
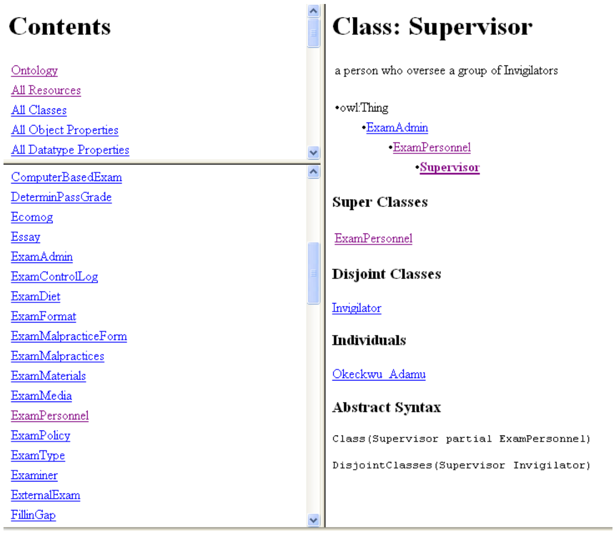


Figure 6: OWL visualization of sub-class “ExamPersonnel” in the exam ontology



Contents

Ontology
All Resources
All Classes
All Object Properties
All Datatype Properties

ComputerBasedExam
DeterminPassGrade
Ecomog
Essay
ExamAdmin
ExamControlLog
ExamDiet
ExamFormat
ExamMalpracticeForm
ExamMalpractices
ExamMaterials
ExamMedia
ExamPersonnel
ExamPolicy
ExamType
Examiner
ExternalExam
FillinGap

Class: Supervisor

a person who oversee a group of Invigilators

•owl:Thing
•ExamAdmin
•ExamPersonnel
•**Supervisor**

Super Classes

ExamPersonnel

Disjoint Classes

Invigilator

Individuals

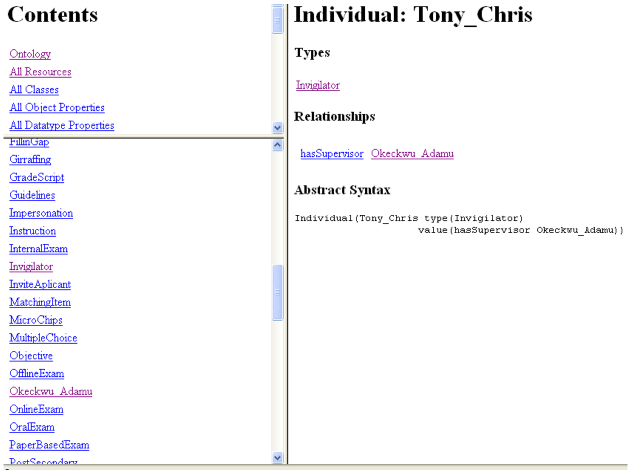
Okeckwu_Adamu

Abstract Syntax

```
Class(Supervisor partial ExamPersonnel)

DisjointClasses(Supervisor Invigilator)
```

Figure 7: web browser of sample output of the exam ontology



Contents

Ontology
All Resources
All Classes
All Object Properties
All Datatype Properties

FillinGap
Girraffing
GradeScript
Guidelines
Impersonation
Instruction
InternalExam
Invigilator
InviteAplicant
MatchingItem
MicroChips
MultipleChoice
Objective
OfflineExam
Okeckwu_Adamu
OnlineExam
OralExam
PaperBasedExam
PostSecondary

Individual: Tony_Chris

Types

Invigilator

Relationships

hasSupervisor Okeckwu_Adamu

Abstract Syntax

```
Individual(Tony_Chris type(Invigilator)
                     value(hasSupervisor Okeckwu_Adamu))
```

Figure 8: an instance of sub-class “Invigilator” and its relationships

REFERENCES

[1] J. C. Arpirez Vega, A. Gomez and H. S Pinto, Reference Ontology and ONTO Agent, Knowledge and Information System, vol. 2(4), pp 387-412, 2000

[2] N. Aussenac-Gilles and D. Sorgel, Supervised Text Analysis for Ontology and Terminology Engineering, Applied Ontology: An Interdisciplinary Journal of Ontological Analysis and Conceptual Modelling, Vol 1(1), pp 35-46, 2005, Netherlands

[3] C.K. Ayo, I.O. Akinyemi, A.A. Adebiyi and U.O. Ekong, The Prospect of E-Examination Implementation in Nigeria Turkish Online, Journal of Distance Education, ISSN 1302-6488, vol., 8(4), 2007

[4] S. Bechhofer, I. Horrocks, C. Goble and R. Stevens, A Reasonable Ontology Editor for the Semantic Web, Lecture notes in Computer Science, 2001

[5] T. Berners-Lee, Weaving the Web, Orion Beusiness Books, 1999, UK.

[6] L. Ceccaroni, Ontowedss – An Ontology-Based Environmental Decision Support System for the Management of Wastewater Treatment Plants. PhD Thesis, Universitat Politechnica, De Catalunya,

[7] D. Fensel, The Semantic Web and its Language, IEEE Computer Society, vol. 15(6), pp 67-73, 2000

[8] A. Gomez-Perez, Knowledge Sharing and Reuse, Ontologies and Applications, A Tutorial on Ontologies Engineering, IJCAI, 1999

[9] Grid Research in Europe: An Overview prepared

Grid Coord. Compiled and edited by the Grid Coord Consortium Office for Official Publications of the European Communities, 2006, http://www.gridcoord.org

[10] T. R. Gruber, A Translation Approach to Portable Ontology Specificaions, Knowledge Acquisition, Vol. 5(2), pp 199-220, 1993

[11] C. Grumgzuo, C. Fei, C. Hu and L. Shufang, OntoEdu : A case study of Ontology-Based Education Grid System for E-Learning, GCCCE2004 International Conference, 2004, Hong Kong

[12] J. Harvey and N. Mogey, Pragmatic issues when integrating technology into the assessment of students, Computer Assisted Assessment, 2006

[13] M. Hogeboom, F. Lin, L. Esmahi and C. Yang, Constructing Knowledge Bases for E-Learning using Protégé 2000 and Web Services, Proceedings of the 19th International Conference on Advanced Information Networking and Application, AINA, 2005

[14] M. Horridge, H. Rector, A. Stevens and C. Wroe A Practical Guide to Building OWL Ontologies using the Protégé OWL Pluging and CO-ODE Tools, Edition 1.0, 2004, University of Manchester, UK

[15] A. Kumar-Das, B. Kanti-Sen and J. Josiah, Open Access to Knowledge and Information: Scholarly Literature and Digital Library Initiatives. The United Nations Educational Scientific and Cultural Organization (UNESCO), B-5/29 Safdarjung Enclave, New Delhi 110029, India, 2008

[16] D. L. McGuiness, R. Fikes, J. Rice and S. Wilder, The Chimavera Ontology Environment, In Proceedings of the 17th National Conference on Artificial Intelligence, 2000, Texas, USA.

[17] M.A. Musen, Dimensions of Knowledge Sharing and Reuse, Computer and Biomedical Research, vol 25, pp 435-467, 1992

[18] F. L. Nov and D. L. McGuiness, Ontology Development 101: A Guide to creating your first Ontology. Standard Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001

[19] R. I. O. Sanni, Educational Measurement and Statistics, Zik Lag Publisher, Lagos, 1998

[20] J. Sowa, Knowledge Represention, Logical Philosophical and Computation Foundations. Brooks Cole, 2000

[21] C. Stergiopoulos, P. Tsiakas, D. Triantis and M. Kaitsa. Evaluating Electronic Examination Methods Applied to Students of Electornics: Effectiveness and Comparison, IEEE International Conference on Sensor Networks , Ubiquitous and Trustworthy Computing, Vol. 2, pp 143-151, 2006

[22] A. H. Tawil, M. Montobello, R. Bahsoon, W.A. Gray and N.J. Fiddian, Interschema correspondence establishement in a cooperative OWL-based Multi-information server grid environment. Information Sciences an International Journal, Elsevier Science Inc., vol. 178 issue 3, 2008

[23] G. VanHeijst, A.T. Schreiber and B.J. Wielinga. Using Explicit Ontologies in Knowledge-Based Systems Development, International Journal of Human and Computer Studies , 1997

[24] M. Uschold and M. Gruinger, Ontologoes: Principles, Methods and Applications. Knowledge Engineering Review, vol. 11(2), pp 93-113, 1996

[25] Doush, I. A. (2011). Annotations , Collaborative Tagging , and Searching Mathematics in E-Learning. International Journal of Advanced Computer Science and Applications - IJACSA, 2(4), 30-39.

[26] Nkenlifack, M., Nangue, R., Demsong, B., & Fotso, V. K. (2011). ICT for Education. International Journal of Advanced Computer Science and Applications - IJACSA, 2(4), 124-133.

Design and Analysis of a Novel Low-Power SRAM Bit-Cell Structure at Deep-Sub-Micron CMOS Technology for Mobile Multimedia Applications

Neeraj Kr. Shukla
E&CE Department
ITM University, Gurgaon
Haryana (India)

R.K.Singh
E&CE Department
BCT-KEC, Dwarahat
Uttarakhand (India)

Manisha Pattanaik
VLSI Group
IIITM Gwalior
MP, (India)

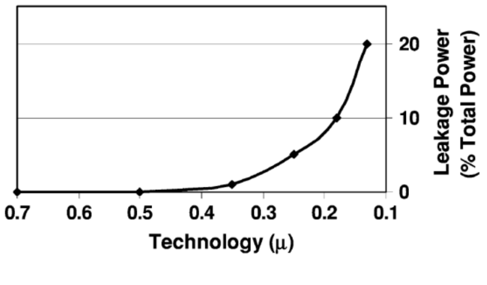
***Abstract*— The growing demand for high density VLSI circuits and the exponential dependency of the leakage current on the oxide thickness is becoming a major challenge in deep-sub-micron CMOS technology. In this work, a novel Static Random Access Memory (SRAM) Cell is proposed targeting to reduce the overall power requirements, i.e., dynamic and standby power in the existing dual-bit-line architecture. The active power is reduced by reducing the supply voltage when the memory is functional and the standby power is reduced by reducing the gate and sub-threshold leakage currents when the memory is idle. This paper explored an integrated approach at the architecture and circuit level to reduce the leakage power dissipation while maintaining high performance in deep-submicron cache memories. The proposed memory bit-cell makes use of the pMOS pass transistors to lower the gate leakage currents while full-supply body-biasing scheme is used to reduce the sub-threshold leakage currents. To further reduce the leakage current, the stacking effect is used by switching off the stack transistors when the memory is ideal. In comparison to the conventional 6T SRAM bit-cell, the total leakage power is reduced by 50% while the cell is storing data '1' and 46% when data '0' at a very small area penalty. The total active power reduction is achieved by 89% when cell is storing data 0 or 1. The design simulation work was performed on the deep-sub-micron CMOS technology, the 45nm, at 25^{0}C with V$_{DD}$ of 0.7V.**

Keywords- SRAM Bit-Cell; Gate Leakage; Sub-threshold Leakage; NC-SRAM; Asymmetric SRAM; PP-SRAM; Stacking Effect.

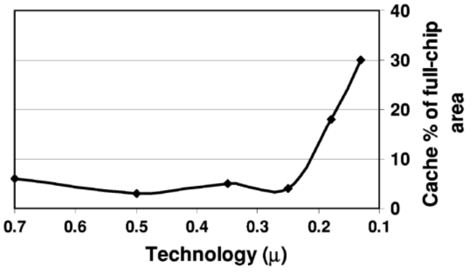
I. INTRODUCTION

Today's mobile/multimedia applications, e.g., a combination of text, audio, still images, graphics (discrete media) and audio, animation, video, video on demand, video recording, interactivity content forms (continuous media), etc. need to be incorporated in one digital system. So, there is a strong need to reduce the standby current leakage while keeping the memory cell data unchanged [1]. In other words, it demands the processor with high processing power, high performance, and low-power on-chip memory. According to the ITRS-2003 (International Technology Roadmap), 90% of the chip-area will be occupied by the memory core by 2014 [2]. This shows the more demand for chips with high functionality and low-power consumption. It is important to focus on minimizing the leakage power of the SRAM structures [3]. The main source for dynamic power consumption is through the switching. But there are several sources for the leakage current, i.e., the sub-threshold current due to low threshold voltage, the gate leakage current due to very thin gate oxides, etc., [4]. The MOS transistor miniaturization also introduces many new challenges in Very Large Scale Integrated (VLSI) circuit designs, such as sensitivity to process variations and increasing transistor leakage. In Fig.1, the leakage power from a high-performance microprocessor has been shown. It increases steadily, as the technology is scaled down [5].

A high-performance VLSI chip also demands ever increasing on-die SRAM to meet the performance needs. This pushes the SRAM scaling towards a more concern domain in today's VLSI design applications. The SRAM cell stability is further degraded by supply voltage scaling. The SRAM leakage power has also become a more significant component of total chip power as a large portion of the total chip transistors directly comes from on-die SRAM. Since the activity factor of a large on-die SRAM is relatively low. So, it is recommended by the researchers in the field to be more effective to put it in a power reduction mechanism dynamically, which modulates the power supply along the memory addressable unit or bank and the need for active/standby mode of operation. In this work, a novel technique of full-supply body-biasing scheme is devised to meet it.



(a)



(b)

Fig.1. (a) Leakage Power Percentage of Total Power. (b) Cache Area Percentage of Total Chip Area [5].

In this work we have presented a novel P4-SRAM Bit-Cell in CMOS deep-sub-micron technology that reduces the active and leakage power in the Conventional SRAM Bit-Cells. In the proposed design, contrary to "...static power is much more important than dynamic power in large memories, static power saving will very well compensate for the increase in dynamic power dissipation [3]" by giving the equal importance to the active mode of power requirements, we focus on both the dynamic and static power dissipations in the active and inactive mode of operations where the SRAM Bit-Cell is fully functional (performing write/read operations) and fully powered ON, i.e., hold state (no read/write operation). The proposed bit-cell utilizes the Gated-V_{DD} technique for transistor stacking in the PP-SRAM along with the full-supply body biasing to reduce the active, standby, and dynamic power in the memory. To the best of my knowledge, the full-supply body biasing along with pMOS stacking is being used for the first time to reduce the overall leakage in the memory cell.

The rest of the paper is organized as follows, in Section II, a basic overview of the SRAM Bit-Cell is presented. In Section III, a review of the related work is presented. In Section IV, the proposed work on a Low-leakage SRAM Bit-Cell is presented which is followed by the Simulation work and Conclusions in Sections V and VI, respectively.

II. The Conventional 6T-SRAM Bit-Cell

The conventional SRAM (CV-SRAM) cell has Six transistors, Fig.2. Unlike DRAM it doesn't need to be refreshed as the bit is latched in. It can operate at lower supply voltages and has large noise immunity. However, the six transistors of an SRAM cell take more space than a DRAM cell made of one transistor and one capacitor thereby increasing the complexity of the cell [6].

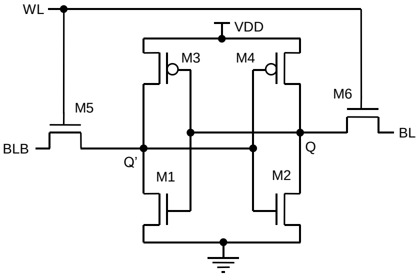


Fig. 2. A 6T-CMOS SRAM Cell [6]

The SRAM Bit Cell

The memory bit-cell has two inverters connected back to back. Two more pass transistors (M5 and M6 in Fig.1) are the access transistors controlled by the Word Line (WL). The cell preserves its one of two possible states to denote '0' and '1', as long as power is available to the bit-cell. Here, Static power dissipation is very small.

Thus the cell draws current from the power supply only during switching. But ideal mode of the memory is becoming of more concern in the deep-sub-micron technology due to is leakage power (gate, sub-threshold leakage, etc.) and data retentation at lower voltages.

The Operation of SRAM Bit-Cell

Although the two nMOS and pMOS transistors of SRAM memory bit-cell form a bi-stable latch, there are mainly the following three states of SRAM memory cell [7], the Write, Read, and Hold states.

A. Standby Operation (Hold)

When WL = '0', M5 and M6 disconnect the cell from Bit-Lines (BL and BLB). The two cross-coupled inverters formed by M1-M4 will continue to reinforce each other as long as they are disconnected from the outside world. The current drawn in this state from the power supply is termed as standby current.

B. Data Read Operation

Read cycle starts with pre-charging BL and BLB to '1', i.e., V_{DD}. Within the memory cell M1 and M4 are ON. Asserting the word line, turns ON the M5 and M6 and the values of Q and Q' are transferred to Bit-Lines (BL and BLB). No current flows through M6, thus M4 and M6 pull BL upto V_{DD}, i.e., BL = '1' and BLB discharges through M1 and M5. This voltage difference is sensed and amplified to logic levels by sense amplifiers.

C. Data Write Operation

The value to be written is applied to the Bit lines. Thus to write data '0', we assert BL=0, BLB = '1' and to write data '1', the BL = '1', BLB = '0', asserted when WL = '1'.

III. A Review of Related Work

As the SRAM bears the low-activity factor, several circuit level techniques have been reported by the researchers to address the low-leakage SRAM design as a major concern.

A. An Asymmetric SRAM Cell to Lower Gate Leakage

In [8], Azizi has proposed an asymmetric SRAM cell design in 70nm technology, Fig.3. In it an nMOS transistor is added to the SRAM bit-cell to reduce the magnitude of the gate voltage when the cell stores data '0', i.e., the cell is in the '0' state. In comparison with the conventional SRAM bit-cell, the gate leakage of the proposed structure decreases in the '0' state while it increases in the '1' state.

Also, the area and long access time are observed as penalty with no change in the dc noise margin (data storage integrity) in the design.

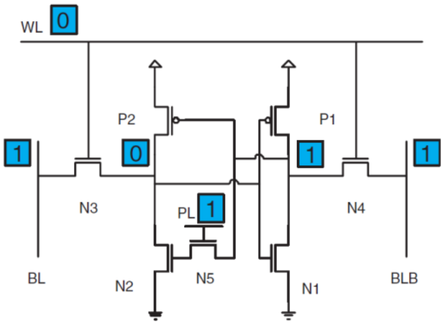


Fig.3: The Structure of Asymmetric Pass-Cell [8]

B. NC-SRAM Bit-Cell Design

Elakkumanan, in [9], has suggested the NC-SRAM design at 65nm technology and 0.8V, Fig. 4. It employs Dynamic Voltage Scaling (DVS) Scheme to reduce the leakage power of the SRAM bit-cells while retaining the stored data during the idle mode.

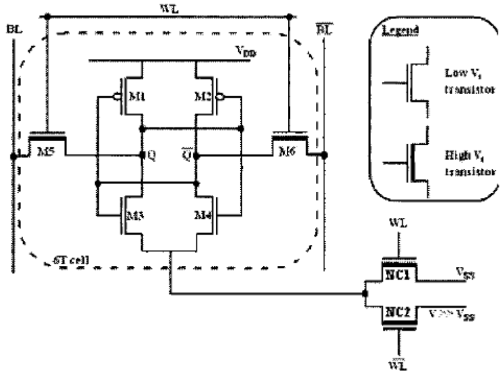


Fig.4. NC-SRAM Circuit: Pass Transistor Control Threshold Voltages of nMOS Transistors in the Cross-Coupled Inverter to Reduce the Leakage [9]

The key idea behind NC-SRAM is the use of two pass transistors NC1 and NC2 which provide different ground levels to the memory cell in the active and idle modes. The positive voltage (virtual ground) reduces the gate leakage and sub-threshold currents of the cell while degrading the read and write performances of the cell [9].

C. Dual-V_t SRAM Bit-Cell

Amelifard, in [10] has suggested a SRAM Bit-Cell topology using dual-gate oxide thicknesses in 65nm technology and at 1.0V supply voltage, Fig.5. It is another approach to reduce the gate leakage current in the SRAM cell [10]. In it, the gate oxide thicknesses of the nMOS pass transistors and the nMOS pull-down transistors are increased. Because the much lower gate leakage of pMOS transistor, no change is made to the gate oxide thickness of the pMOS pull-up transistors. To achieve a lower sub-threshold current, the dual threshold voltage technique has been used. The cell evaluation is performed by using the high threshold voltage for different transistors. In the best case, the power consumption is decreased and the stability is improved but the read access time is degraded [10].

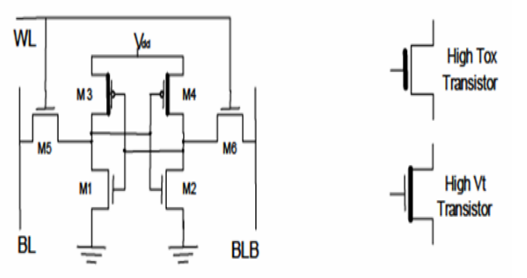


Fig. 5. Dual-V_t SRAM Bit-Cell Structure [10]

D. Gated-V_{DD} SRAM Bit-Cell

Ye, et.al., have proposed a circuit level technique as the name 'Gated-V_{DD}' [11] to gate the supply voltage and reduce the leakage power when the SRAM cells are in ideal mode with a minimal impact on performance, Fig.6. The sub-threshold leakage current and leakage energy dissipation increase exponentially with decreasing threshold voltage. Gated-V_{DD} "turn off" the supply voltage and eliminate virtually all the leakage energy dissipation in the cache's unused sections. The key idea is to introduce an extra transistor in the supply voltage (V_{DD}) or the ground path (GND) of the cache's SRAM cells. The extra transistor is turned ON in the used sections and turned OFF in the unused sections. Thus, the cell's supply voltage is "gated." Gated-V_{DD} maintains the performance advantages of lower supply and threshold voltages with reducing leakage power. The fundamental reason for the reduction in leakage is the stacking effect of self reverse-biasing of series-connected transistors [11]. Gated-V_{DD}'s extra transistor produces the stacking effect in conjunction with the SRAM cell transistors when the gated-V_{DD} transistor is turned off.

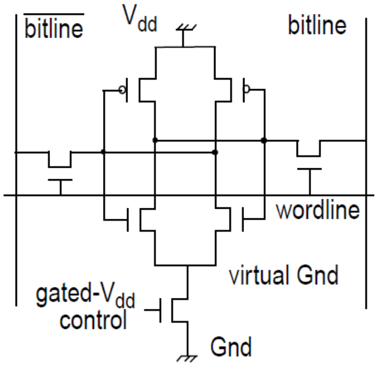


Fig. 6. SRAM with an NMOS Gated-V_{DD}

The role of the area overhead of the Gated Transistor is not much in case of a large cache as it is being used one transistor per row which is clear from Fig.7.

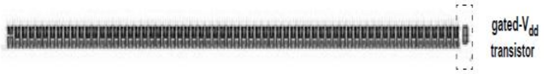


Fig. 7: Layout of 64 SRAM Cells connected to a Single Gated-V_{DD} nMOS Transistor

E. PP-SRAM Cell

G. Razavipour, et.al., in [3] have presented a gate leakage current reduction technique based on the pMOS pass-transistor SRAM bit-cell structure, Fig.8., at 45nm technology and 0.8V supply voltage by giving a name to it as PP-SRAM Bit-Cell. It has lower gate leakage compared to that of the conventional SRAM cell. In order to decrease the gate leakage currents of the SRAM cell, nMOS pass transistors, are replaced by pMOS pass transistors P3 and P4. In the active mode, WL is held at ‘0’ to turn ON the two pass transistors. In the idle mode, WL is charged to V_{DD}, so that the two pMOS pass transistors are OFF. Here, it is being pointed-out that due to the use of the pMOS transistors, there is an increase in the dynamic power of the cell which is consumed during the read and write operation. Also, since static power is much more important than dynamic power in large memories, static power saving will very well compensate for the increase in dynamic power dissipation [3].

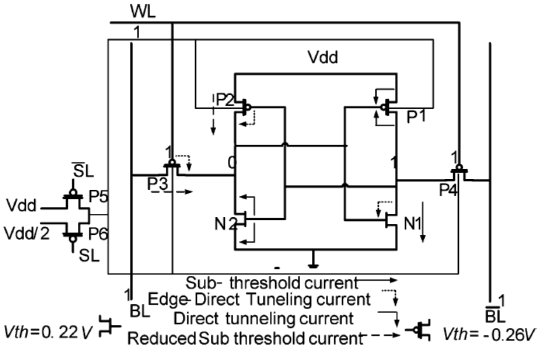


Fig.8. PP-SRAM Cell (Holding ‘0’) with Gate Leakage Current [3]

In addition, the static power consumption induced by the two new inserted pMOS transistors (P5 and P6) is found to be small. The use of pMOS pass transistor, however, may lead to performance degradation due to different mobility coefficients for the nMOS and pMOS transistors. To overcome this problem, the width of pMOS pass transistor is selected as 1.8 times of that of the nMOS for the technology used in this work [3].

IV. P4-SRAM BIT CELL - THE PROPOSED WORK

The proposed SRAM Bit-Cell called as P4-SRAM Bit-Cell is shown in Fig. 9. It mainly considers the fact that the barrier height of the holes is more than that of the electrons, i.e, 4.5eV vs 3.1eV, respectively. So, the gate leakage current will be lower in the off state of the pMOS transistors whereas the full-supply substrate body-bias will involve in the reduction of the sub-threshold leakage during the standby mode of the proposed design. This novel structure of SRAM Bit-Cell uses the concept of the transistor stacking for the leakage power reduction. The fundamental reason for the stacking in the leakage power reduction is the stacking effect of the self-reverse biasing of series connected transistors. To the best of my knowledge pMOS transistor stacking at the Gated-GND has been proposed for the first time at the ‘full-supply voltage body-biasing’ in the SRAM bit-cell. These, extra pMOS transistors of the Gated-GND, produces the stacking effect in conjunction with the SRAM Bit-Cell transistors when the Gated-GND transistors are turned-off by opposing flow of leakage current through them.

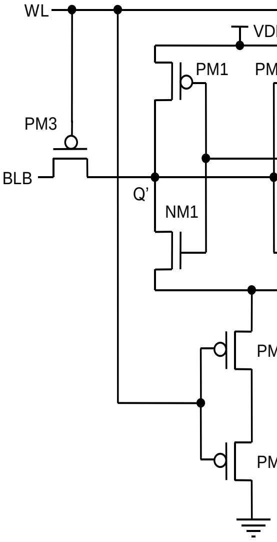


Fig.9. Proposed Novel P4-SRAM Bit-Cell Structure

When WL = ‘0’, the bit-cell behaves as a normal conventional 6T bit-cell by turning the PM2 and PM3 transistors ON, i.e., cell is in the active state and is ready to write/read data. But when the WL = ‘1’, the two access transistor (pMOS) are in OFF state. It also put both the stack pMOS transistors in the OFF state. To sink the current when the bit-cell is active (WL=’0’), stacking transistors need to be ON. Hence, they are need to be made larger in size. But the too large gated-transistors may reduce the stacking effect. Also, bigger the size of the stacking transistors will improve the cell read time but will cost the area. It is a design trade-off and may be considered as per the application demanded. In the proposed design, in order to reduce the negative impact of threshold voltage on the speed of the bit-cell and to reduce the active power consumption, a forward full supply body-biasing is used in the pMOS transistors. In it, the body bias voltage of pMOS transistor in the idle mode and active mode is set to V_{DD}. This work used a single bias voltage (V_{DD}), it directly reduces the circuit complexity demands by the dynamic body bias technique [3]. It also reduces its area overhead. In this work, the external signal can be generated through the row decoder circuit is activated before the WL is activated for the read and write operations or can be applied through WL by assuring the proper write/read access time which further assures the reduction in the area and circuit complexity.

V. ANALYSIS AND SIMULATION WORK

A. Conventional 6T-CMOS SRAM Bit-Cell

Table 1: Power in Standby Mode of Operation – When Bit-Cell Hold Data ‘1’
Active Power in writing ‘1’ is 4.011µW

Cell / Transistor	Power (W)
Conventional 6T Bit-Cell	9.128p
PD NM0	2.716p

PD NM1	882.2z
Pass Gate NM2	2.575p
Pass Gate NM3	2.085p
PU PM0	2.181a
PU PM1	1.751p

Table 2: Power in Standby Mode of Operation – When Bit-Cell Hold Data '0'
Active Power in writing '0' is 3.815µW

Cell / Transistor	Power (W)
Conventional 6T Bit-Cell	4.5p
PD NM0	25.1651f
PD NM1	2.717p
Pass Gate NM2	142.14z
Pass Gate NM3	3.277a
PU PM0	1.751p
PU PM1	6.57005f

B. Proposed SRAM Bit-Cell – P4 SRAM Bit-Cell

Table 3: Standby Mode of Operation – When Bit-Cell Hold Data '1'
Active Power in writing '1' is 418.65nW

Cell / Transistor	Power (W)
Proposed SRAM Bit-Cell	4.5235p
PD NM0	575.342f
PD NM1	1.3476p
Pass Gate PM2	2.5964p
Pass Gate PM3	458.49f
PU PM0	43.67a
PU PM1	337.6121f
Stack PM4	14.091f
Stack PM5	541.5f

Table 4: Standby Mode of Operation – When Bit-Cell Hold Data '0'
Active Power in writing '0' is 418.52nW

Cell / Transistor	Power (W)
Proposed SRAM Bit-Cell	2.435p
PD NM0	6.28892a
PD NM1	577.739f
Pass Gate PM2	962.3f
Pass Gate PM3	305.42z
PU PM0	338.94f
PU PM1	4.95228a
Stack PM4	14.343f
Stack PM5	541.9f

C. Analysis of Power Consumption – Active and Leakage

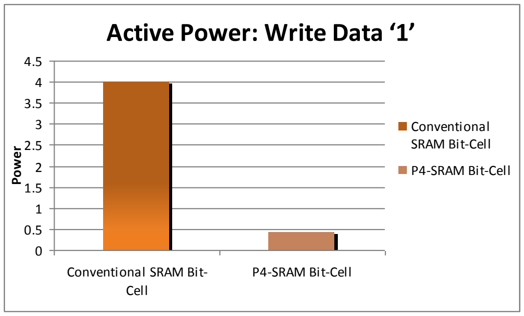


Fig.10: Active Power for Write Data '1'

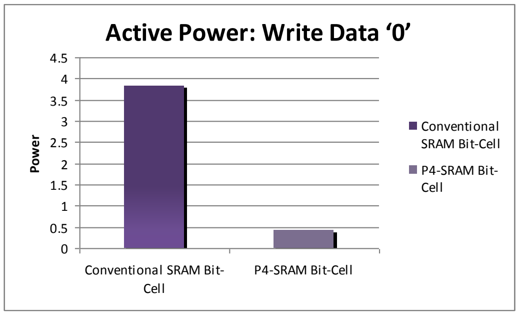


Fig.11: Active Power for Write Data '0'

The use of the pMOS transistors increases the dynamic power of the cell which is consumed during the read/write operations. In [3], considering the static power dissipation more important in large memory, the dynamic power has been compensated by the static power dissipation of the memory. Here, in the proposed work, we have reduced the dynamic power of the bit-cell by 89% for Write Data '0/1', individually, Table 1-4 and Fig. 10, 11.

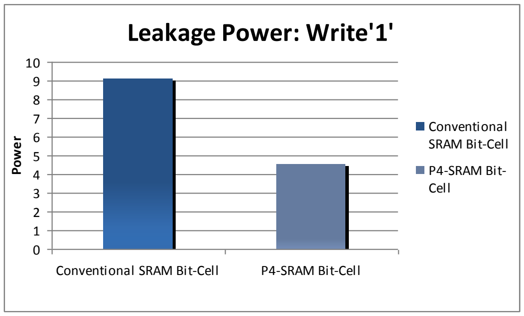


Fig.12: Leakage Power for Hold Data '1'

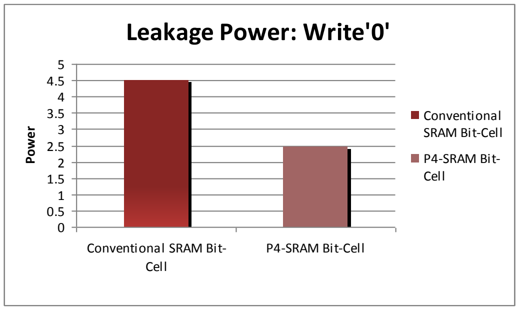


Fig.13: Leakage Power for Hold Data '0'

The leakage power consumption of the proposed P4-SRAM Bit-Cell is reduced by 50% while the cell is storing '1' and 46% when '0' at a small area penalty of two pMOS stack transistors, Table 1-4 and Fig. 12, 13.

The experimental results shows that the best power optimization is achieved at newly reported full-supply body bias of the pMOS transistors in the SRAM bit-cell, Table 5.

Table 5: Substrate Bias-Voltage vs Power Dissipation (Active & Leakage)

Bias Voltage (V)	Active Power (W)		Leakage Power (W)	
	Write Data '0'	Write Data '1'	Store Data '0'	Store Data '1'
0.7	418.52n	418.65n	2.4357p	2.4357p

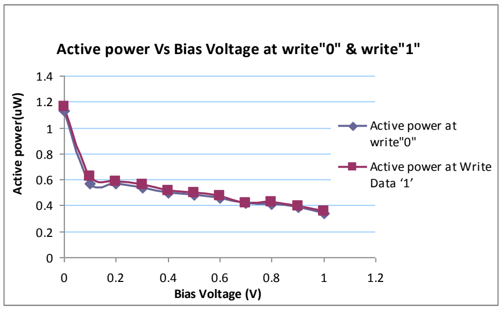


Fig.14: Active Power vs Substrate Bias Voltage for Data Write '0' and '1'

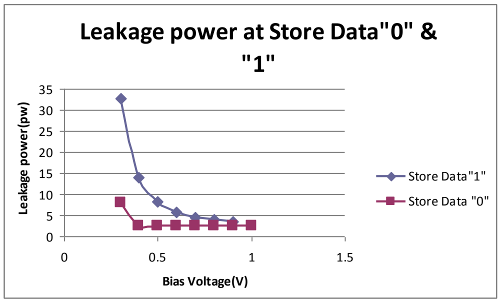


Fig.15: Leakage Power vs Substrate Bias Voltage for Data Hold '0' and '1'

This work utilizes the same body-bias voltage as the supply voltage V_{DD} is called as full-supply body biasing. It simplifies the need for complex dynamic body-bias voltage mechanism at different voltage levels and its control circuit. So that it may further supports area efficiency of the memory, Table 5, and Fig. 14, 15.

VI. CONCLUSIONS

In this paper, a novel structure of the SRAM Bit-Cell, called as P4-SRAM Bit-Cell structure, is presented. The proposed bit-cell utilizes the Gated-V_{DD} technique for transistor stacking in the PP-SRAM along with the full-supply body biasing to reduce the active, standby, and dynamic power in the memory. In comparison to the conventional 6T SRAM bit-cell, the total leakage power is reduced by 50% while the cell is storing data '1' and 46% when data '0' at a very small area penalty. The total active power reduction is achieved by 89% when cell is storing data 0 or 1. The design simulation work was performed on the deep-sub-micron CMOS technology, the 45nm, at 25^0C with V_{DD} of 0.7V for t_{OX} of 2.4nm.

ACKNOWLEDGEMENT

The authors are grateful to their respective organizations for their encouragement and support.

REFERENCES

[1] Neeraj Kr. Shukla, Shilpi Birla, R.K. Singh and Manisha Pattanaik "Analysis of the Effects of the Operating Temperature at the Performance and Leakage Power Consumption in a Conventional CMOS 6T-SRAM Bit-Cell at 65nm, 45nm, and 32nm Technologies," *IACSIT International Journal of Engineering and Technology, Singapore*, Vol.3, No.1, February 2011, Pg. 1-9, ISSN: 1793-8236.

[2] International Technology Roadmap for Semiconductors. Online-Available at http://www.publicitrs.net

[3] G. Razavipour, A. Afzali-Kusha, and M. Pedram, "Design and Analysis of Two Low-Power SRAM Cell Structures", *IEEE Transaction on Very Large Scale Integration (VLSI) Systems*, Vol. 17, No. 10, Oct. 2009, pp. 1551-1555.

[4] K. M. Kao et al., "BSIM4 Gate Leakage Model Including Source-Drain Partition," in *Proc. Int. Electron Devices Meeting*, Dec. 2000, pp. 815–818.

[5] Kevin Zhang, Uddalak Bhattacharya, Zhanping Chen, Fatih Hamzaoglu, Daniel Murray, Narendra Vallepalli, Yih Wang,B. Zheng, and Mark Bohr, "SRAM Design on 65-nm CMOS Technology with Dynamic Sleep Transistor for Leakage Reduction" *IEEE Journal of Solid-State Circuits*, Vol. 40, No. 4, APRIL 2005, pp. 895-901.

[6] Neeraj Kr. Shukla, Shilpi Birla, R.K. Singh, and Manisha Pattanaik, "Speed and Leakage Power Trade-off in Various SRAM Circuits", *International Journal of Computer and Electrical Engineering (IJCEE), Singapore*, VOl.3, No.2, Apr. 2011, pp. 244-249.

[7] Sung-Mo (Steve) Kang, Yusuf Leblebici, "CMOS Digital Integrated Circuits-Analysis and Design", Third Edition Tata McGraw-Hill Edition, New Delhi, India.

[8] Navid Azizi and Farid N. Najm, "An asymmetric SRAM Cell to Lower Gate Leakage," in *Proc. 5th IEEE International Symposium Quality Electronic Design (ISQED'04)*, 2004, pp. 534–539.

[9] P. Elakkumanan, C. Thondapu, and R. Sridhar, "A Gate Leakage Reduction Strategy for Sub-70 nm Memory Circuit," in *Proc. IEEE Dallas/CAS Workshop*, 2004, pp. 145–148.

[10] B. Amelifard, F. Fallah, and M. Pedram, "Reducing the Sub-threshold and Gate-Tunneling Leakage of SRAM Cells using Dual-Vt and Dual-Tox Assignment," in *Proc. DATE*, Mar. 2006, pp. 1–6.

[11] Y. Ye, S. Borkar, and V. De., "A New Technique for Standby Leakage Reduction in High Performance Circuits", In *IEEE Symposium on VLSI Circuits*, pages 40–41, 1998.

[12] Monitoring, R. (2010). Development of a Low-Cost GSM SMS-Based Humidity Remote Monitoring and Control system for Industrial Applications. International Journal of Advanced Computer Science and Applications - IJACSA, 1(4).

AUTHORS PROFILE

[1]**Neeraj Kr. Shukla** (IEEE, IACSIT,IAENG, IETE, IE, CSI, ISTE,VSI-India), a Ph.D. Scholar at the UK Technical University, Dehradun (Uttarakhand) India is an Asst. Professor in the Department of Electrical, Electronics & Communication Engineering, ITM University, Gurgaon, (Haryana) India. He has received his M.Tech. (Electronics Engineering) and B.Tech. (Electronics & Telecommunication Engineering) Degrees from the J.K. Institute of Applied Physics & Technology, University of Allahabad, Allahabad (Uttar Pradesh) India in the year of 1998 and 2000, respectively. His main research interests are in Low-Power Digital VLSI Design and its Multimedia Applications, Open Source EDA, and RTL Design.

[2]**R.K. Singh** (IAENG, IE, ISTE), Professor in the Department of Electronics & Communication Engineering, VCT-Kumaon Engineering College, Dwarahat, Almora (UK) India. He is being honored with the Ph.D. in Electronics Engineering in the Year 2003 from the University of Allahabad, Allahabad (Uttar Pradesh), India. He has received his M.E. (Electronics & Control Engineering) in 1992 from BITS, Pilani, (Rajasthan) India. He has

around thirty five research publications in the conferences and journals at national and international. He has authored seven text-books in the field of VLSI Design, Basic Electronics, and Opto-Electronics. He has worked at various capacities in the Academic domain such as, the Principle, Kumaon Engineering College, Dwarahat in the year 2003-04, Director (O), Directorate of Technical Education, Uttaranchal in the year 2005, and Joint Director, State Project Facilitation Unit, Dehradun for the World Bank TEQIP Project. He is also the recipient of couple of prestigious awards, e.g., Rastriya Samman Puruskar, Jewel of India Award, Rastriya Ekta Award, Life Time Achievement Award, and Arch of Excellence Award. His current areas of interest are VLSI Design, Opto-Electronics and its applications.

[3]**Manisha Pattanaik** (IE, ISTE) has been honored the Ph.D. from Indian Institute of Technology (IIT) Kharagpur, India in the field of VLSI Design from the Department of Electronics and Electrical Communication Engineering in the year 2004. Currently she is an Associate Professor (VLSI Group) at ABV-India Institute of Information Technology & Management (ABV-IIITM), Gwalior, (Madhya Pradesh), India. She shared the responsibility in the capacity of referee for IEEE International Conferences on VLSI Design for two consecutive years, 2003-04. Her areas of interest are Leakage Power Reduction of Nano-Scale CMOS Circuits, Characterization of Logic Circuit Techniques for Low-Power/Low-Voltage and High performance analog & digital VLSI applications and CAD of VLSI Design.

Performance Analysis of GPU compared to Single-core and Multi-core CPU for Natural Language Applications

Shubham Gupta
Master of Technology,
School of Computing Sciences and Engineering,
VIT University, India

Prof. M.Rajasekhara Babu
School of Computing Sciences and Engineering,
VIT University, India

***Abstract*— In Natural Language Processing (NLP) applications, the main time-consuming process is string matching due to the large size of lexicon. In string matching processes, data dependence is minimal and hence it is ideal for parallelization. A dedicated system with memory interleaving and parallel processing techniques for string matching can reduce this burden of host CPU, thereby making the system more suitable for real-time applications. Now it is possible to apply parallelism using multi-cores on CPU, though they need to be used explicitly to achieve high performance. Recent GPUs hold a large number of cores, and have a potential for high performance in many general purpose applications. Programming tools for multi-cores on CPU and a large number of cores on GPU have been formulated, but it is still difficult to achieve high performance on these platforms. In this paper, we compare the performance of single-core, multi-core CPU and GPU using such a Natural Language Processing application.**

***Keywords-* NLP; Lexical Analysis; Lexicon; Shallow Parsing; GPU; GPGPU; CUDA; OpenMP.**

I. INTRODUCTION

In recent times, CPU supports multi-cores each supports improved SIMD instruction sets. And recent GPU supports a large number of cores which run in parallel, and its peak performance outperforms CPU. In [1], comparison of performance on CPU, FPGA and GPU is done using some image processing applications. And in [4], performance analysis of CPU and GPU is performed on some medical image volume rendering application. In this paper, we compare the performance of GPU and CPU (single-core and multi-core) using a NLP application.

In many real-life applications in the areas such as syntactic pattern recognition, syntactic analysis of programming languages etc., the parser speed is an important factor. Since large sum of data is to be processed, efficient low complexity parsers are required. All the techniques of parsing in Natural Language Processing involve string matching as the single most important operation. Traditionally for many years, GPU is just used to accelerate some stages of the graphics rendering pipeline. However, after the programmability available on this chip, GPU opens a door to developers to take advantage of its ALUs besides graphics processing. Compared with CPU in their architectures, GPU is more suitable for stream computations; it can process data elements in parallel with SIMD & MIMD capability. And in many cases, people gain great performance improvement on GPU over CPU. So a totally new technique called **GPGPU** (General Purpose computation on GPU) emerges. And in recent years it became a hot research topic, not only in computer graphics domain, but also in other discipline areas.

II. GPU AND CPU

The graphics processing units (GPU) are highly parallel rapidly gaining maturity as a powerful engine for computationally demanding applications. The GPU's performance and potential will be the future of computing systems. A GPU is basically designed for some particular type of applications with the following characteristics.

- Where Computational requirements are large: GPU must deliver an enormous amount of compute power to cover the requirements of complex real-time applications.
- Parallelism is significant: The graphics

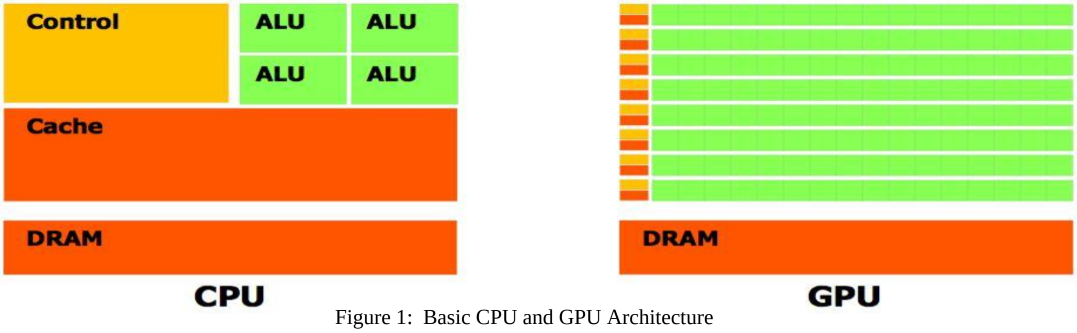


Figure 1: Basic CPU and GPU Architecture

pipeline system architecture is suitable for parallelism.

Few years ago, GPU's were some fixed function processors, built over the three dimensional (3D) graphics pipeline and with very little else to offer. But now, the GPU has evolved into a powerful programmable processor, with both application programming interface (APIs) and the hardware increasingly focusing on the programmability aspects of the GPU. The result is a processor with enormous arithmetic capability and streaming memory bandwidth, both substantially greater than a high-end CPU. [5]

As shown in fig.1 [3], on comparing the GPU with CPU the basic difference is; CPU has few processing units with some cache and control unit, but in GPU there are many more processing units with their own cache and control units with dedicated and specific works defined for them. GPUs are mostly with hundreds of cores which work in parallel to process the data, but in general CPUs processing is done on few cores with very little parallelism.

On architectural comparison with CPU, GPU are more suitable for stream computations, they can process data elements in parallel with SIMD & MIMD capability. So a new technique called **GPGPU** (**G**eneral **P**urpose computation on GPU) emerged and in recent years has become a hot research topic in not only graphics domain but in general computations.[2][5]

III. GPGPU (General Purpose GPU)

GPGPU is a combination between hardware components and software that allows the use of a traditional GPU to perform computing tasks that are extremely demanding in terms of processing power. Traditional CPU architectures available on the market cannot satisfy the processing demands for these specific tasks, and thus the market has moved on to GPGPU in order to achieve greater efficiency.

Few benefits of using a GPU for general purpose processes (GPGPU): [2] [5]

- Large performance benefits in many parallel coded applications. In some situations the GPU clearly performs better compared to a traditional CPU-based high performance computer.
- Purchase price: the prices of GPUs are somewhat similar to the market price of CPUs. This is a large advantage GPGPU has. In some cases, it would take multiple CPUs to match the performance of a GPGPU system. This means that in terms of cost, the GPGPU is a smarter choice.
- Technology refresh rate: GPU manufacturers develop new GPUs with a refresh rate that is much faster compared to that of the CPU market. The advantage the GPU has in this case is rather obvious, as its core technology is updated more frequently than that of the CPUs'.

From the Fig. 2 [2], it is clear that the floating-point operations per second on GPU very much exceed that of CPU. In other words, the computation power of GPU is stronger than that of CPU. The computation power is more reachable than other hardware, and in terms of the computation cost, GPU for per GFLOPS is much lower than CPU.

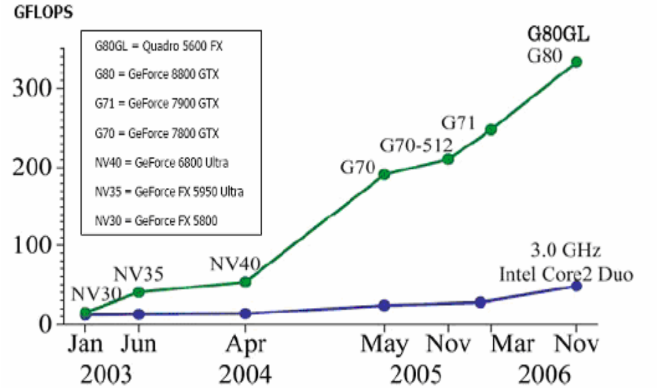


Figure 2: Floating-point operations on CPU and GPU (from NVIDIA)

IV. Architecture of GPU

The architecture of the GPU has progressed in a different direction than that of the CPU. Consider a pipeline of tasks that processes a large number of input elements, the output of each successive task is fed into the input of the next task. Data in multiple pipeline stages can be computed at the same time; that is pipeline shows the task parallelisms. As data in multiple pipeline stages can be computed at the same time; computing more than one element at the same time is data parallelism. To execute such a pipeline, a CPU would take a single element (or group of elements) and process the first stage in the pipeline, then the next stage, and so on. The CPU divides the pipeline in time, applying all resources in the processor to each stage in turn. GPU divides the resources of the processor among the different stages, such that the pipeline is divided in space, not time. The part of the processor working on one stage feeds its output directly into a different part that works on the next stage. In brief, there are many hundred cores on a GPU system; cache memory but with no cache coherency. (Fig. 3)

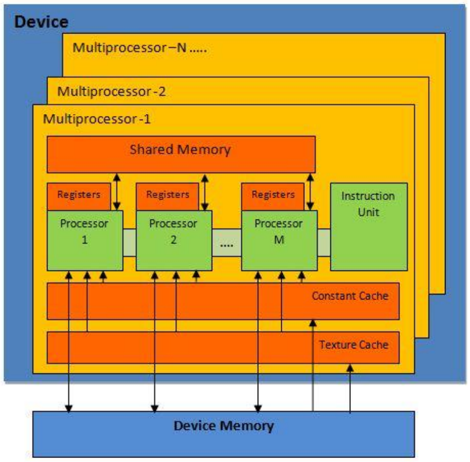


Figure 3: GPU Architecture

V. Programming Model for GPU

The programming over GPU follows a single instruction multiple-data (SIMD) programming model. For efficiency, the GPU processes many elements in parallel using the same program. Each element is independent from the other elements, and in the base programming model, elements cannot communicate with each other. All GPU programs must be structured in this way: many parallel elements each processed in parallel by a single program.

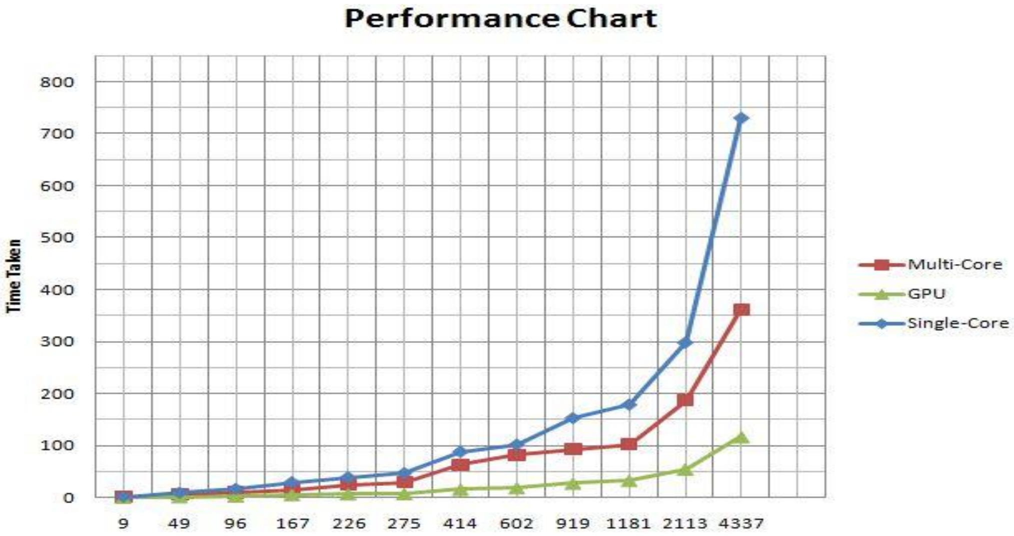


Figure 4: Performance Comparison Graph

NVIDIA has developed **C**ompute **U**nified **D**evice **A**rchitecture (CUDA) which allows the use of the C programming language to code algorithms to execute on the GPU. CUDA enabled GPUs include data parallel cache. Besides the flexible interface for programming, it also supports memory scatter bringing more flexibilities to GPU.

CUDA (Compute Unified Device Architecture)

CUDA provides a C-like syntax for executing on the GPU and compiles offline. CUDA exposes two levels of parallelism, data parallel and multithreading. CUDA also exposes multiple levels of memory hierarchy: per-thread registers, fast shared memory between threads in a block, board memory, and host memory. Kernels in CUDA allow the use of pointers, general load/store to memory allowing the user to scatter data from within a kernel, and synchronization between threads in a thread block. However, all of this flexibility and potential performance gain comes with the cost of requiring the user to understand more of the low-level details of the hardware, notably register usage, thread and thread block scheduling, and behavior of access patterns through memory. All of these systems allow developers to more easily build large applications. [6][7]

VI. NLP APPLICATION

All the techniques of parsing in Natural Language Processing involve string matching as the single most important operation. We are using Lexical Analysis and Shallow Parsing as the NLP application to be implemented on single-core and multi-core CPU system (using OpenMP as programming model) and also on GPU system (using CUDA as their programming model).

As per the algorithm used, system performs lexical analysis and shallow parsing on a text input file in English language. To perform so, algorithm matches it with the knowledge base which is having complete list of possible words in English with their part of speech. The processing time for such input text file is high because of the size of the knowledge base which contains a huge amount of data to be processed for each and every word that is encountered by the algorithm from the input text file.

VII. PERFORMANCE ANALYSIS

For the evaluation, we use the following platforms.

- NVIDIA GeForce G210M 1024MB (800MHz GDDR3, 16cores)
- CUDA version 2.1
- OpenMP
- Intel Core 2 Duo CPU P8600 (2.40GHz, 2CPUs) and Intel C++ Compiler 10.0

As per the algorithm used, for some particular size of file (in terms of number of words in a file) our system processes input file (performs lexical analysis and shallow parsing) and finally provide us with the number of matches and number of part of speech in the provided input file.

TABLE I. TABLE OF EXPERIMENTAL RESULTS (IN SECONDS)

no. of Words	**Time Taken**		
	Single-Core	Multi-Core	GPU
9	1.4538	0.9125	0.2662
49	8.945	5.684	1.5735

no. of Words	**Time Taken**		
96	16.376	9.9218	2.8652
167	29.327	15.6598	5.627
226	38.631	24.6598	6.936
275	46.561	29.3627	7.5648
414	87.871	63.4374	16.0373
602	101.6373	81.9269	18.6092
919	152.655	93.2981	27.514
1181	178.5342	102.2638	32.7635
2113	297.8587	186.8143	54.8736
4337	730.539	361.6552	117.272

Some of the results generated by the NLP algorithm used on processing input file of certain size (in terms of number of words) are shown in the above table (table 1).

The graph generated on some of the data generated (using data from the table given) on the implementation of algorithm used is displayed above (fig.4).

VIII. Conclution

We have compared the performance of GPU with single-core and multi-core CPU (2cores) for a basic NLP application (lexical analysis and shallow parsing). The number of cores in Nvidia GeForce G210M is 16. As the results from the table (table 1) and the graph (fig.4) generated, shows that multi-core CPU has better performance than the single-core CPU but a GPU system has clearly overtaken them with much better performance over CPU for Natural Language Processing (NLP) applications. For the future enhancements, this algorithm can be improved and implemented on programmable GPGPUs more efficiently to give even improved performance.

References

[1] Shuichi Asano, Tsutomu Maruyama and Yoshiki Yamaguchi, University of Tsukuba, Japan; “Performance Comparison of FPGA, GPU and CPU in Image Processing”, IEEE-FPL 2009, pp. 127-131.

[2] Enhua Wu, University of Macau; Youquan Liu, Chinese Academy of Sciences, China; “Emerging Technology about GPGPU”, Circuit and Systems, 2008.APCCAS 2008. IEEE.

[3] Vadali Srinivasa Murty, P.C.Reghu Raj and S.Raman, Department of Computer Science and Engineering, Indian Institute of Technology Madras (IITM); “Design of Language-Independent Parallel String Matching unit for NLP”, 2003 IEEE International Workshop on Computer Architectures for Machine Perception (CAMP).

[4] Mikhail Smelyanskiy and et al. ; “Mapping High-Fidelity Volume Rendering for Medical Imaging to CPU, GPU and Many-Core Architectures”, IEEE Transactions on Visualization and Computer Graphics, Vol. 15, Dec.2009.

[5] John D. Owens, Mike Houston, David Luebke, Simon Green, John E. Stone, and James C. Phillips; “GPU Computing”, Proceedings of the IEEE Vol. 96, No. 5, May 2008.

[6] Nvidia CUDA; “nVIDIA CUDA C Programming Guide”, Version 4.0

[7] John Stratton et al., University of Illinois; “Compute Unified Device Architecture Application Suitability”, Computer Science and Engineering, University of Illinois; IEEE Computer Society & American Institute of Physics.

Authors Profile

Shubham Gupta is Pursuing Master of Technology in Computer Science and Engineering (2009-11) from VIT University, Vellore, India. He did his B.E. (Computer Science and Engineering) from Swami Keshwanand Institute of Technology (Rajasthan University), Jaipur, India (2005-2009). His research interest includes Natural Language Processing, Multi-core Architectures and GPU programming.

Prof. M.Rajasekhara Babu a senior faculty member at the School of Computer Science and Engineering ,VIT University ,Vellore ,India . At, present he is pursing Ph.D from the same University in the area of “Multi-Core Architectures and Natural Language Processing”. He had authored a number of national and international papers and articles in reputed journals and conferences

Main Stream Temperature Control System Based on Smith-PID Scheduling Network Control

Jianqiu Deng
Department of Computer Science and Technology,
Tsinghua University
Beijing, China

Haijun Li,Zhengxia Zhang
Naval Aeronautical Engineering Institute.
Yantai, Shandong ,China.

Abstract—**This paper is concerned with the controller design problem for a class of networked main stream temperature control system with long random time delay and packet losses. To compensate the effect of time delay and packet losses, a gain-scheduling based Smith-PID controller is proposed for the considered networked control systems (NCSs). Moreover, to further improve the control performance of NCSs, genetic algorithm is employed to obtain the optimal control parameters for gain-scheduling Smith-PID controller. Simulation results are given to demonstrate the effectiveness of the proposed methods.**

Keywords- Network control systems (NCSs); Gain-scheduling based Smith-PID; main stream temperature system; time delay; packet loss.

I. Introduction

In recent years, Networked control systems (NCSs) have received increasing attentions. NCSs are a class of cotnrol systems where sensors, actuators, estimator units and control units are connected through communication networks[1]. NCSs have many advantages, such as low cost, simplified maintenance and diagnosis, increased system flexibility, and so on. However, because of the insertion of network, NCSs inevitably have the problems of random time delay, packet out-of-order and packet losses, etc.

The stability and performance of NCSs have been studied by many researchers recently, such as paper [2-4]. In [5] and [6], the problem of designing H∞ controllers for networked control systems with both network-induced time delay and packet out-of-order was studied. Many kinds of time delay in the network system were introduced in [7]. It applied the algorithms of discrete-time sliding mode forecast control to make a simulation on servo system, which is controlled by network control system. In [8], the problem of designing guaranteed cost controller is studied for a class of uncertain time delay networked control system. More Networked controllers were designed in [9-12]. For more details on this subject, we refer the reader to [13-16] and the references therein.

Aiming at the problem of random time delay of the networked control systems, the gain-scheduling based Smith-PID controller is designed in this paper. The Smith controller is used to compensate the long time delay of the system, and the gain-scheduling based PID scheduling controller is used to solve the problem of random time delay of the networked control systems. The two controllers were combined together to control the long time delay networked control systems better.

The structure of this paper is organized as follows. The introduction of the Networked control system is given in Section I. The problem formulation is given in Section II. The controllers design is introduced in Section III. In this section, the Smith control design and the gain-scheduling networked control design are introduced respectively. As an example, the gain-scheduling based Smith-PID controller is used in the main stream temperature control system. The simulation is done in Section IV and the numerical and experimental results are provided. The conclusion is given in Section V.

II. Problem formulation

The main stream temperature control system is composed of six modules. They are temperature detecting module, temperature setting module, alarming module, controlling unit module, temperature displaying module, actuator module. The general block diagram is shown in figure 1.

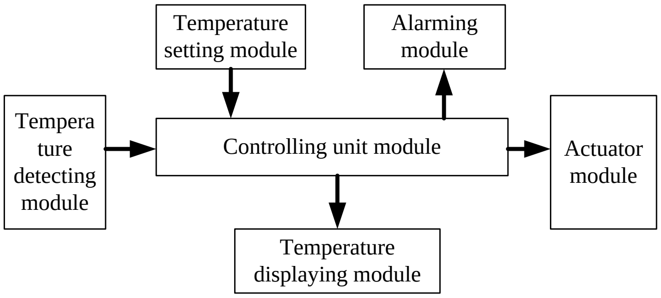

Temperature setting module
Alarming module
Tempera ture detecting module
Controlling unit module
Actuator module
Temperature displaying module

Figure1. The general block diagram of the main stream temperature control system

The main stream temperature system is controlled through network in this paper. Therefore, there are many problems to solve, such as random time delay, packet out-of-order and packet losses. The gain-scheduling based Smith-PID control method is proposed in this paper to solve the aforementioned problems. There is long time delay in the main stream temperature system, and the time delay is different when the system is using different number of sub-loops. Moreover, there is also random time delay in the network. In this paper, we combine the aforementioned two kinds of time delay together. It is worth noting that, during the controller design stage, we also take the packet out-of-order and packet losses of network into consideration.

III. Controller Design

A. Smith-PID Control Design

The single loop control system with pure delay is shown in figure 2, whose closed-loop transfer function is shown as follows:

$$\phi(s)=\frac{Y(s)}{R(s)}=\frac{G_c(s)G_0(s)e^{-\tau s}}{1+G_c(s)G_0(s)e^{-\tau s}} \qquad (1)$$

Its characteristic equation is:

$$1+G_c(s)G_0(s)e^{-\tau s}=0 \qquad (2)$$

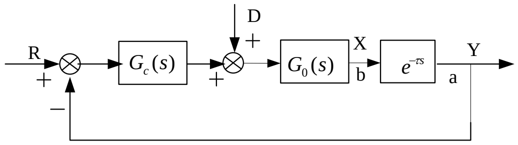


Figure2. The single loop control system with pure delay

Obviously, there is pure delay in the characteristic equation. If τ is big enough, the system will be unstable. That is the essence that the long time delay process is hard to control[20]. $e^{-\tau s}$ appears in the characteristic equation, because the feedback signal is cited from the point a of the system. If the feedback signal is cited from the point b, the pure delay part is removed outside of the control circuit. As shown in figure 3. After time delay τ, the controlled variable Y will repeat changes of X.

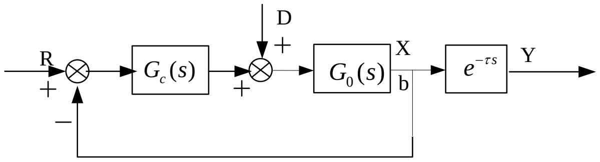


Figure 3. Improved single loop control system with pure delay

Because the feedback signal X has no delay, the response of the system is greatly improved. Point b is not exist or limited by the physical conditions in the practical system. So, the feedback signal can't be cited from the point b. According to this problem, Smith proposed artificial model method. The structure is shown in figure 4.

If the model is accurate, such as $G_0(s)=G_m(s)$, $\tau=\tau_m$, and there is no load disturbance(D=0). $Y=Y_m$, $E_m=Y-Y_m=0$, $X=X_m$. So, X_m can change X as the first feedback loop and the pure delay part is moved outside of the control loop. If the model is inaccurate or there is load disturbance in the control process, X is not equal to X_m, $E_m=Y-Y_m\neq 0$, and the control precision is not a great satisfaction. So, E_m is used to realize the second feedback loop. This is the control strategy of Smith predictor controller.

B. Scheduling Network Control Design

The time-delay of the networked control systems is varying all the time.

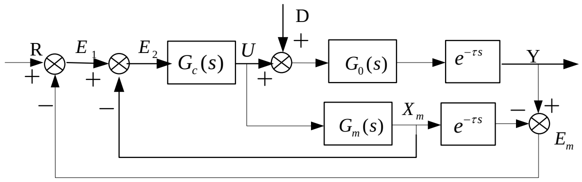


Figure 4. Smith predict control system

We can use scheduling network control to solve this problem. The block diagram of the networked control system is shown in figure 5.

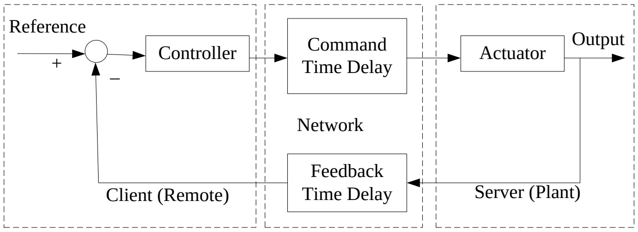


Figure 5. The block diagram of the networked control system

Suppose that: the time delay τ is varying in the three rang [a, a+h], [a+h, a+2h], [a+2h, a+3h](Thereinto, h and a are arbitrary real numbers). The parameters of the PID controller are k=f{Kp, Ki, Kd}. When τ is varying in the rang [a, a+h], the k is big. When τ is varying in the rang [a+h, a+2h], the k is middle. When τ is varying in the rang [a+2h, a+3h], the k is little. With the parameter τ is varying in the three rang, different parameters of the PID controller can be selected to make the networked control systems running quickly and stable. That is so called Scheduling PID Network Controller.

The different parameters suit different τ can be optimized by the adaptive online genetic algorithm.

C. gain-scheduling based Smith-PID Network Control Design

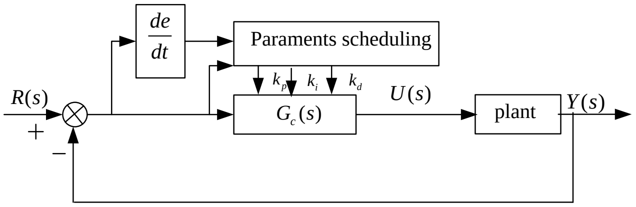


Figure 6. The structure of the Smith-PID scheduling network control

The Smith-PID controller is used to compensate the long time delay of the system, and the gain-scheduling based PID controller is used to solve the problem of packet out-of-order and packet losses of the system. The two controllers were combined together to control the long time delay networked control systems better. The structure of the gain-scheduling based Smith-PID network control is shown in figure 6.

D. Genetic Algorithm

The parameters of PID are optimized by the genetic algorithm. The genetic algorithm is introduced as follows.

The setting principles of the parameters of PID based on the genetic algorithm are proposed as below:

1) The determination and the expression of parameters

First of all, the ranges of parameters should be determined. The ranges are usually given by users, and then been coded according to the demanding of precision[17-18]. Every parameter is expressed by selecting binary word serial, and relationships between them are established[19]. Then, the binary word serials are combined together to a long binary word serial. The long word serial is the operation object of genetic algorithm.

2) The selection of the initial population

The process of genetic algorithm is realized by programming. So, the initial population will be generated randomly by computer. For the binary programming, the random numbers between 0~1 are generated. Then, the random number between 0~0.5 will express 0, and the random number between 0.5~1 will express 1. What's more, the complex degree of the calculation will be considered when determining the size of the population.

3) The determination of the fitness function

For the normal optimizing method, one group parameters can be obtained under the restrained conditions. The best one can be selected in the group of parameters. There are three aspects of index to measure a control system: stability, accuracy, and rapidity. The rise time reflects the rapidity of the system. The shorter of the rise time, the quicker of the control process, the better of the quality of the system.

If the dynamic performance of the system is pursued only, the obtained parameters will probably making the control signal too big. This will lead to the unstable of the system because of the saturation characteristic of the system. We will add control variable into the objective function to prevent the oversize of the control energy. To control the system better, the control variable, error and rise time are given as constraints. Because the fitness function is related with the objective function, the objective function will be taken as the fitness function to optimizing parameters after determination.

4) The operation of the genetic algorithm

First of all, the fitness proportion method is used to copy, and the copy probability is obtained. The product of copy probability and the number of every generation of word serial is taken as the copy number of the next generation. The one who has bigger copy probability will has more descendants in the next generation. The one who has smaller copy probability will has smaller descendants in the next generation.

Through copying, crossing and variation, the new population is obtained from the initial population. The population is induced into the fitness function after decoding to observe if the termination conditions are satisfied or not. If the termination conditions are not satisfied, the operations above are repeated until they are satisfied.

The termination conditions are decided by specific problems. Only if all the target parameters are in the specified ranges, the calculation will be terminated.

The operating process above can be expressed in figure 7.

The steps of optimizing the parameters k_p, k_i, k_d using the genetic algorithm is introduced as follows:

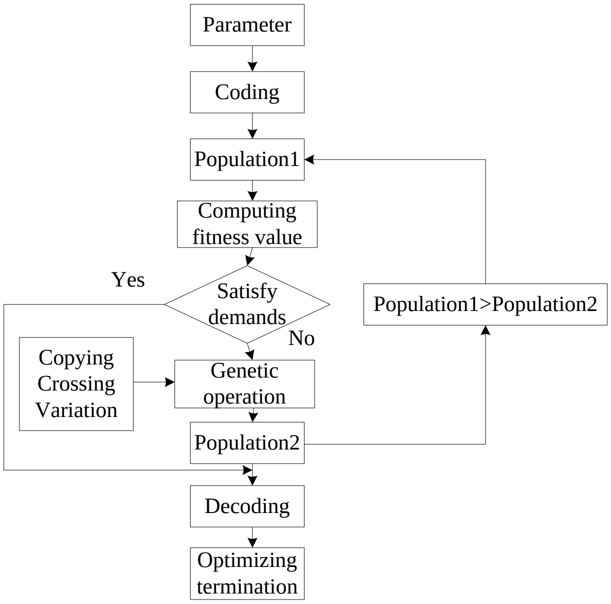


Figure 7. The flow chart of the genetic algorithm

1) Determining the range and the length of every parameter and coding.

2) N individuls are generated randomly and they composed of the initial population P(0).

3) Decoding the individuls of the population to the responding parameters. These parameters are used to solve the cost function value J and the fitness function value f, f=1/J.

4) The operation is done to the population P(t) using copying operator, crossing operator and variation operator. So, the next generation of population P(t+1) is generated.

5) The steps 3) and 4) are repeated until the parameters are converging or reaches the predetermined index.

The cost function is designed as follows:

$$fitness = \min(\int_0^\infty t^2 \times |e(t)|\, dt) \qquad (3)$$

where e(t) is the error between the output signal of NCS and the expected signal. The binary word serials are decoded into the control parameters in reality. These parameters are used in the discrete NCS controller. Then, simulation about NCS is done. Last, the fitness degree of individual is computed based on the time domain response curve of the controlled object.

There are four advantages for the genetic algorithm when setting the parameters of PID:

1) Compared with the simplex method, the genetic algorithm has good optimizing parameters. It overcomes the sensitivity of the simplex method to the initial values of the parameters. In the situation of improper selection of the initial condition, genetic algorithm can also find the proper parameters to make the control target meets the requirement. What's more, the simplex method will cost too long time or cause failure optimization. The genetic algorithm can solve the questions very well.

2) Compared with the expert setting method, it has the advantages of simple and quick. It needs no complex rules, and can receive optimal value through simple copy, cross, and variation. It avoids plenty of collection work of knowledge database and simulation in the expert setting method.

3) The genetic algorithm is doing parallel operations from many points. It overcomes the blindness of begging from one point. So, its speed is faster and it avoids falling into the local optimal solution too early.

4) The genetic algorithm is not only used in single target optimization but also multiple target optimizations. The genetic algorithm will find the proper parameters in the prescribed scopes based on different system.

IV. SIMULATION

Take the main stream temperature control system as an example. The transfer function of the main stream temperature control system is $W(s) = \frac{e^{-300s}}{1500s^2 + 1501s + 1}$. Suppose that: the time delay τ is 300s at the begging, turns to 280s from 200s, and turns to 260s from 400s.

The normal Smith-PID controller is used to control the main stream temperature control system first. Through repeated debugging, there are three result curves shown as below.

When the parameters of the PID controller are Kp=6.2, Ki=0.01, Kd=1.1, the input-output curve of the system is shown in figure 8.

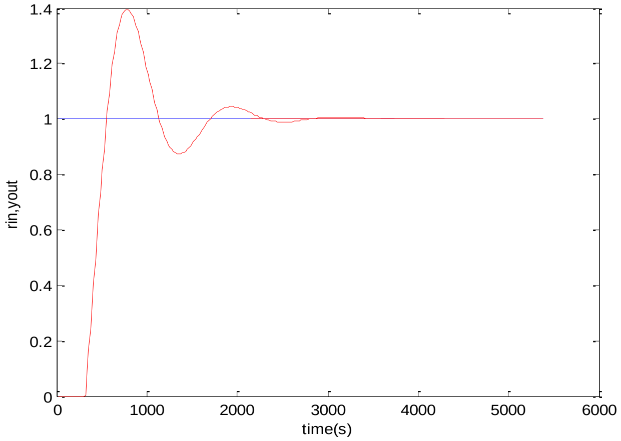


Figure 8. The simulation result with the normal Smith-PID controller

The gain-scheduling based Smith-PID network controller is used to control the main stream temperature control system. Through optimizing by the genetic algorithm, the parameters of the PID controller are Kp=16.81, Ki=0.0114, Kd=13.31 at the begging. When the time delay of the system turns to 280, the parameters of the PID controller turn to Kp=21.81, Ki=0.01529, Kd=17.31. When the time delay of the system turns to 260, the parameters of the PID controller turn to Kp=33.01, Ki=0.01852, Kd=19.31. The simulation result of the system is shown in figure 9.

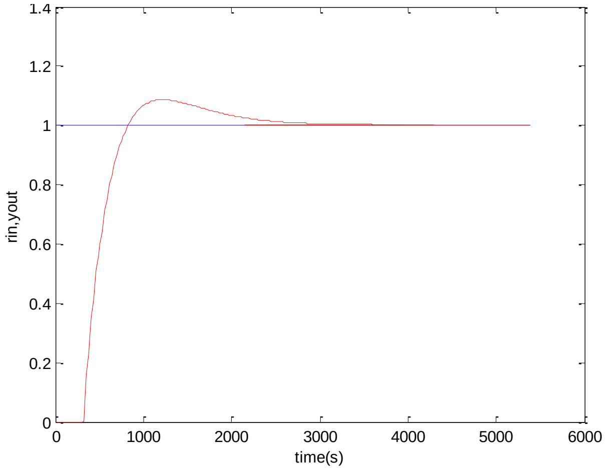


Figure 9. The simulation result with the proposed controller

From the simulation results we can see that the proposed gain-scheduling based Smith-PID network controller shows better control performance than normal Smith-PID controller, which demonstrates the effectiveness of the proposed method.

V. CONCLUSIONS

The result curves of the simulation show that the normal Smith-PID controller can not adapt the varying of the time delay very well. Its control effect to the networked control system is not very satisfied. The Smith-PID scheduling network controller proposed in this paper can adjust the parameters of the PID controller on time with the varying of the time delay. So, it can fit the varying of the time delay better. Because the parameters of the PID controller are the best in every period, the anti-interference ability of the system is very strong too. The input-output curve of the system with Smith-PID scheduling network controller shows that the Smith-PID scheduling network controller can obtain satisfied control effect.

In future work, we will consider the stability analysis of the considered NCSs during the controller design stage. Moreover, we will also extend the results to the case where full state measurement is not available.

ACKNOWLEDGMENT

The authors would like to thank the editor and the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. This work was supported by the National Science Foundation of China under Grants 61004021.

REFERENCES

[1] Zhang lihua, Wu yuqiang, Gong lei, chen haojie. A Novel Research Approach on Network Control Systems. 2008 International Conference on Internet Computing in Science and Engineering. pp.262-265.

[2] S. Hu and W.-Y. Yan, “Stability of Networked Control Systems Under a Multiple-Packet Transmission Policy,” *IEEE Transactions on Automatic Control*, vol. 53, no.7, pp. 1706-1711, 2008.

[3] X. Jiang, Q.-L. Han, S. Liu, and A. Xie, “A New H*1* Stabilization Criterion for Networked Control Systems,” *IEEE Transactions on Automatic Control*, vol. 53, no.4, pp. 1025-1032, 2008.

[4] L. Dritsas and A. Tzes, “Robust stability bounds for networked controlled systems with unknown, bounded and varying delays,” *IET Control Theory & Applications* vol. 513, no.3, pp. 270-280, 2009.

[5] Y.-L. Wang and G.-H. Yang. $H\infty$ control of networked control systems with time delay and packet disordering. IET Control Theory Appl., Vol. 1, No. 5, September 2007. pp.1344-1354.

[6] Guo Xijin, Li Haigang, Zhang Qian, Zhang Qinying. Robust $H\infty$ □Control of Network Control System with Random Time-delay. Proceedings of the 7th World Congress on Intelligent Control and Automation June 25 - 27, 2008, Chongqing, China. pp.2536-2541.

[7] Guang-yi Chen, Zhao-yong He. The Design of Network Control System Controller Based on Variable Structure Silding Mode Control. Proceedings of Sixth International Conference on Machine Learning Cybernetics, Hong Kong, 19-22 August 2007. pp. 465-468.

[8] Junfeng Liu, Tao Ju, Baoan Hu, Aischeng Xia, Jie Wei. Guaranteed Cost Controller Design for Time-delay Network Control System. 2008 Chinese Control and Decision Conference(CCDC 2008). pp. 65-67.

[9] L. Zhang, Y. Shi, T. Chen, and B. Huang, “A New Method for Stabilization of Networked Control Systems With Random Delays,” *IEEE Transactions on Automatic Control*, vol. 50, no.8, pp. 1706-1711, 2005.

[10] H. Gao and T. Chen, “Network-Based H1 Output Tracking Control,” IEEE Transactions on Automatic Control, vol. 53, no.3, pp. 655-667, 2008.

[11] H. Li, M.-Y. Chow, and Z. Sun, “EDA-based speed control of a networked DC motor system,” IEEE Transactions on Industrial Electronics, vol. 56, no.5, pp. 1727-1735, 2009.

[12] H. Li, Z. Sun, B. Chen, and H. Liu, “Intelligent Scheduling Control of Networked Control Systems with Networked-induced Delay and Packet Dropout.” International Journal of Control, Automation and Systems. vol. 6, no.6, pp. 915-927, 2008.

[13] T. C. Yang, “Networked control system: a brief survey,” IET Control Theory & Applications, vol. 153, pp. 403-412, 2006.

[14] J. P. Hespanha, P. Naghshtabrizi, and Yonggang Xu. “A survey of recent results in networked control systems,” Proceedings J. P. Hespanha, P. Naghshtabrizi, and Yonggang Xu. “A survey of recent results in networked control systems,” Proceedings.

[15] P. Wen, J. Cao, and Y. Li, “Design of high-performance networked real-time control systems,” IET Control Theory & Applications, vol. 1, no.5, pp. 1329-1335, 2007.

[16] D. Huang, and S. K. Nguang, “Robust disturbance attenuation for uncertain networked control systems with random time delays,” IET Control Theory & Applications, vol. 2, no.11, pp. 1008-1023, 2008.

[17] E. Tian, D. Yue, and X. Zhao, “Quantised control design for networked control systems,” IET Control Theory & Applications, vol. 1, pp. 1693-1699, 2007.

[18] HOLLAND J H. Adap tation in natural and artificial systems[M].Ann Arbor: University of Michigan Press, 1975: 502-507.

[19] Michalewiz Z, Genetic algorithm+data=evolution programs[M].3rd edition. New York: Springer Verlag, 1996

[20] Waters G A, Smith D K. Evolutionary design algorithm for optimal layout of tree networks[J]. Engineering Optimization,1995, 24: 261-281.

[21] Salim Labiod, Thierry Marie Guerra, “Adaptive fuzzy control of a class of SISO nonaffine nonlinear systems”, Fuzzy Sets and Systems, vol.158.2007, pp. 1126-1137.

AUTHORS PROFILE

Jianqiu Deng (Changde city, Hunan province, China, August 4 1974) Master’s degree, Navigation, Guidance and Control, Naval Aeronautical Engineering Institute, Shandong, China, 2000. Ph.D candidate, Tsinghua University, Beijing, China. The major field of study: cybernation. He teaches at Naval Aeronautical Engineering Institute now. The institute is located in Shandong province of China.

FPGA-Based Design of High-Speed CIC Decimator for Wireless Applications

Rajesh Mehra
ECE Department,
National Institute of Technical Teachers‘ Training & Research
Chandigarh, India

Rashmi Arora
M.E. Student, ECE Department,
National Institute of Technical Teachers‘ Training & Research
Chandigarh, India

Abstract— **In this paper an efficient multiplier-less technique is presented to design and implement a high speed CIC decimator for wireless applications like SDR and GSM. The Cascaded Integrator Comb is a commonly used decimation filter which performs sample rate conversion (SRC) using only additions/subtractions. The implementation is based on efficient utilization of embedded LUTs of the target device to enhance the speed of proposed design. It is an efficient method used to design and implement CIC decimator because the use of embedded LUTs not only increases the speed but also saves the resources on the target device. The fully pipelined CIC decimator is designed with Matlab, simulated with Xilinx AccelDSP, synthesized with Xilinx Synthesis Tool (XST), and implemented on Virtex-II based XC2VP50-6 target FPGA device. The proposed design can operate at an estimated frequency of 276.6 MHz by consuming considerably less resources on target device to provide cost effective solution for SDR based wireless applications.**

Keywords- CIC; FPGA; FPGA; GSM; LUT; SDR.

I. INTRODUCTION

The widespread use of digital representation of signals for transmission and storage has created challenges in the area of digital signal processing [1]. The applications of digital FIR filter and up/down sampling techniques are found everywhere in modem electronic products. For every electronic product, lower circuit complexity is always an important design target since it reduces the cost [2]. There are many applications where the sampling rate must be changed. Interpolators and decimators are utilized to increase or decrease the sampling rate. This rate conversion requirement leads to production of undesired signals associated with aliasing and imaging errors. So some kind of filter should be placed to attenuate these errors [3].

Recently, there is increasingly strong interest on implementing multi-mode terminals, which are able to process different types of signals, e.g. WCDMA, GPRS, WLAN and Bluetooth. These versatile mobile terminals favor simple receiver architectures because otherwise they’d be too costly and bulky for practical applications [4]. The answer to the diverse range of requirements is the software defined radio. Software defined radios (SDR) are highly configurable hardware platforms that provide the technology for realizing the rapidly expanding digital wireless communication infrastructure. Many sophisticated signal processing tasks are performed in SDR, including advanced compression algorithms, power control, channel estimation, equalization, forward error control, adaptive antennas, rake processing in a WCDMA (wideband code division multiple access) system and protocol management.

Today’s consumer electronics such as cellular phones and other multi-media and wireless devices often require digital signal processing (DSP) algorithms for several crucial operations[5] in order to increase speed, reduce area and power consumption. Due to a growing demand for such complex DSP applications, high performance, low-cost Soc implementations of DSP algorithms are receiving increased attention among researchers and design engineers. Although ASICs and DSP chips have been the traditional solution for high performance applications, now the technology and the market demands are looking for changes.

On one hand, high development costs and time-to-market factors associated with ASICs can be prohibitive for certain applications while, on the other hand, programmable DSP processors can be unable to meet desired performance due to their sequential-execution architecture [6]. In this context, embedded FPGAs offer a very attractive solution that balance high flexibility, time-to-market, cost and performance.

The digital signal processing application by using variable sampling rates can improve the flexibility of a software defined radio. It reduces the need for expensive anti-aliasing analog filters and enables processing of different types of signals with different sampling rates. It allows partitioning of the high-speed processing into parallel multiple lower speed processing tasks which can lead to a significant saving in computational power and cost. Wideband receivers take advantage of multi-rate signal processing for efficient channelization and offers flexibility for symbol synchronization.

II. CIC DECIMATORS

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. If you are using A4-sized paper, please close this file and download the file for “MSW A4 format”.

The Cascaded Integrator Comb (CIC), first introduced by Hogenauer, presents a simple but effective platform for implementation of decimations. It is a commonly used decimation filter which performs sample rate conversion (SRC) using only additions/subtractions. It then has experienced some modifications toward improvements in power consumption and frequency response [7]-[8].

It consists of two main sections: an integrator and a comb, separated by a down-sampler [9]-[10]. An integrator is simply a single-pole IIR filter with a unity feedback coefficient:

$$y[n] = y[n-1] + x[n] \qquad (1)$$

This system is also known as an accumulator. The transfer function for an integrator on the z-plane is

$$H_I(z) = \frac{1}{1 - z^{-1}} \qquad (2)$$

The power response of integrator is basically a low-pass filter with a -20 dB per decade (-6 dB per octave) rolloff, but with infinite gain at DC [11]. This is due to the single pole at z = 1; the output can grow without bound for a bounded input. In other words, a single integrator by itself is unstable and shown in Figure 1

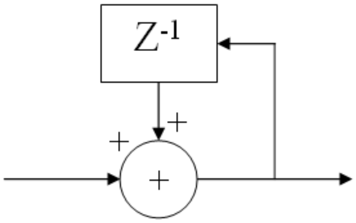


Figure 1. Basic Integrator

A comb filter running at the high sampling rate, fs, for a rate change of R is an odd symmetric FIR filter described by

$$y[n] = x[n] - x[n - RM] \qquad (3)$$

Where M is a design parameter and is called the differential delay. M can be any positive integer, but it is usually limited to 1 or 2. The corresponding transfer at fs

$$H_c(z) = 1 - z^{-RM} \qquad (4)$$

When R = 1 and M = 1, the power response is a high-pass function with 20 dB per decade (6 dB per octave) gain (after all, it is the inverse of an integrator). When RM ≠ 1; the power response takes on the familiar raised cosine form, with RM cycles from 0 to 2π. The basic comb is shown in Figure 2.

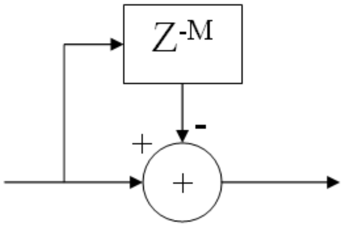


Figure 2. Basic Comb

When we build a CIC filter, we cascade, or chain output to input, N integrator sections together with N comb sections. This filter would be fine, but we can simplify it by combining it with the rate changer. Using a technique for multi-rate analysis of LTI systems from [13], we can “push" the comb sections through the rate changer, and have them become at the slower sampling rate fs/R.

$$y[n] = x[n] - x[n - M] \qquad (5)$$

The transfer function for a CIC filter at fs is

$$H(z) = H_I^N(z) H_C^N(z)$$
$$H(z) = \frac{(1 - z^{-RM})^N}{(1 - z^{-1})^N} = (\sum_{k=0}^{RM-1} z^{-k})^N \qquad (6)$$

This equation shows that even though a CIC has integrators in it, which by themselves have an infinite impulse response, a CIC filter is equivalent to N FIR filters, each having a rectangular impulse response. The CIC filter has a high pass-band droop and a low stop-band attenuation, which can be improved by increasing the number of the cascaded CIC filters [14]. Sharpening based methods generally improve both the pass-band and the stop-band characteristic of the CIC filter at an expense of the increased complexity [15].

Since all of the coefficients of these FIR filters are unity, and therefore symmetric, a CIC filter also has a linear phase response and constant group delay [16]. The magnitude response at the output of the filter can be shown to be:

$$|H(f)| = \left| \frac{Sin\,\pi MF}{Sin\,\frac{\pi f}{R}} \right|^N \qquad (7)$$

By using the relation sin x ≈ x for small x and some algebra, we can approximate this function for large R as

$$|H(f)| \approx \left| RM\,\frac{Sin\,\pi MF}{\pi MF} \right|^N \; for\, 0 \le f \le \frac{1}{M} \qquad (8)$$

We can notice a few things about the response. One is that the output spectrum has nulls at multiples of f = 1/M. In addition, the region around the null is where aliasing/imaging occurs. If we define fc to be the cutoff of the usable passband, then the aliasing/imaging regions are at

$$(i - f_c) \le f \le (i + f_c) \qquad (9)$$

for $f \le \frac{1}{2}$ and i = 1, 2,……[R/2]. If $f_c \le \frac{M}{2}$,

then the maximum of these will occur at the lower edge of the first band, 1-fc. The system designer must take this into consideration, and adjust R, M, and N as needed. Another thing we can notice is that the passband attenuation is a function of the number of stages. As a result, while increasing the number of stages improves the imaging/alias rejection, it also increases the passband "droop."

III. PROPOSED CIC DECIMATOR DESIGN & SIMULATION

In this proposed work fully pipelined 3-stage CIC decimator is designed using Matlab and Xilinx AccelDSP by taking filter R as 8 and M as 2.

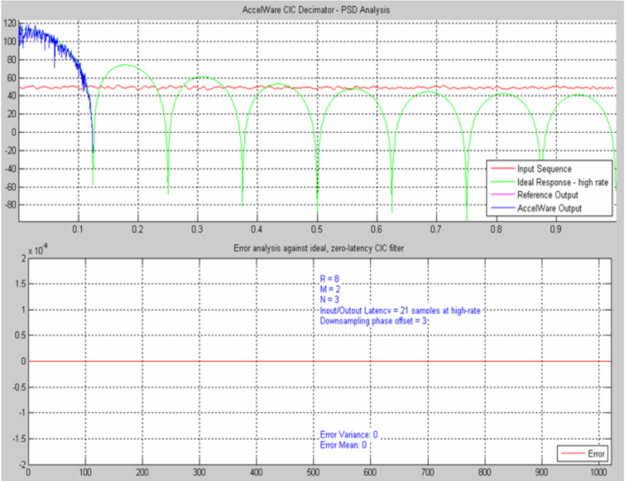


Figure 3. Floating Point Output of CIC Decimator

The Matlab based floating point output of proposed design is shown in Fig 3. Then the equivalent fixed point file is generated and verified by AccelDSP whose output is shown in Fig4. The red wave shows the input sequence, green wave shows the ideal response and blue plot is the output from CIC decimator.

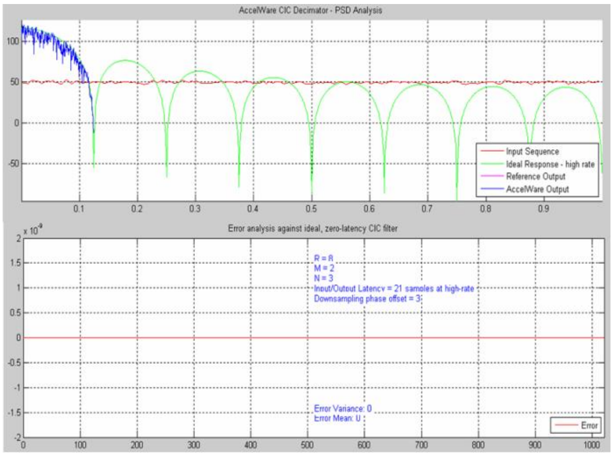


Figure 4. Fixed Point Output of CIC Decimator

The 3 stage CIC decimator is designed to accomplish three things here. First, we have slowed down half of the filter and therefore increased efficiency. Second, we have reduced the number of delay elements needed in the comb sections. Third, and most important, the integrator and comb structure are now independent of the rate change. This means we can design a CIC filter with a programmable rate change and keep the same filtering structure. A CIC decimator would have N cascaded integrator stages clocked at fs, followed by a rate change by a factor R, followed by N cascaded comb stages running at fs/R as shown in Figure 5 The complete Matlab to AccelDSP design flow is shown in Fig6.

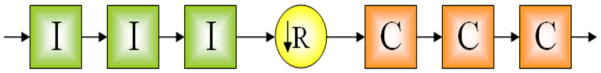


Figure 5. CIC Decimator

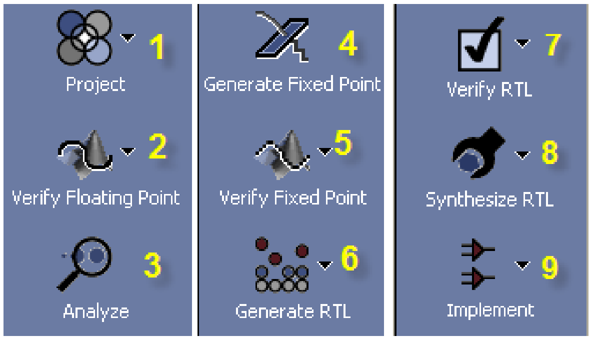


Figure 6. AccelDSP Design Flow

The proposed design shows an efficient realization of CIC decimator by using embedded LUTs of target FPGA to provide high speed operation. The multiplier less LUT based technique consist of input registers, 4-input LUT unit and shifter/accumulator unit as shown in Fig7.

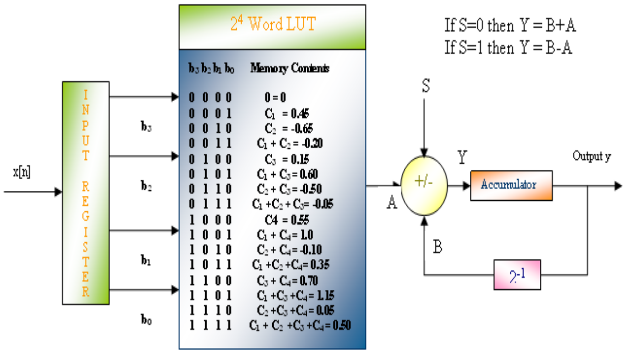
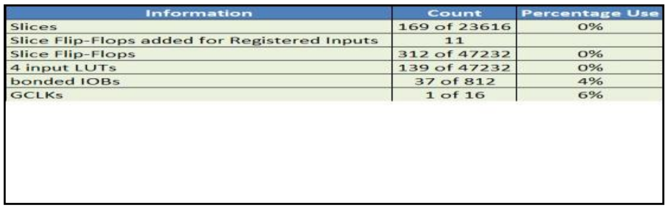


Figure 7. LUT based Multiplier Less Implementation

IV. FPGA IMPLEMENTATION RESULTS

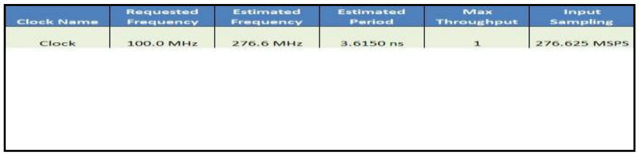
To observe the speed and resource utilization, RTL is generated, verified and synthesized. The proposed CIC decimator filter is implemented on Virtex-II Pro based XC2VP50-6 target device using fully pipelined LUT based multiplier less technique. The resource utilization of proposed implementation is shown in table I.

TABLE I. VIRTEX-II PRO BASED RESOURCE UTILIZATION



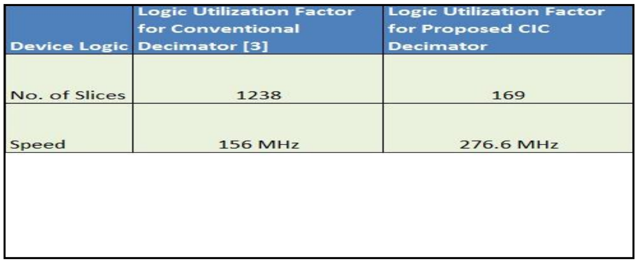
Information	Count	Percentage Use
Slices	169 of 23616	0%
Slice Flip-Flops added for Registered Inputs	11	
Slice Flip-Flops	312 of 47232	0%
4 input LUTs	139 of 47232	0%
bonded IOBs	37 of 812	4%
GCLKs	1 of 16	6%

TABLE II. TRANSPOSED FORM PERFORMANCE EVALUATION



Clock Name	Requested Frequency	Estimated Frequency	Estimated Period	Max Throughput	Input Sampling
Clock	100.0 MHz	276.6 MHz	3.6150 ns	1	276.625 MSPS

TABLE III. LOGIC UTILIZATION COMPARISON ON VIRTEX-II PRO BASED XC2VP50-6 FPGA



Device Logic	Logic Utilization Factor for Conventional Decimator [3]	Logic Utilization Factor for Proposed CIC Decimator
No. of Slices	1238	169
Speed	156 MHz	276.6 MHz

As shown in table III, the proposed LUT based design can work at an estimated frequency of 276.6 MHz as compared to 156 MHz in case of [3] by using considerable less resources of target FPGA. The speed performance of proposed design is shown in table II.

V. CONCLUSION

In this paper, a Xilinx AccelDSP based approach is presented for a CIC Decimator to minimize the time to market factor. The proposed fully pipelined CIC decimator filter is designed by using embedded LUTs of target device. The results show enhanced performance in terms of speed and area utilization. The proposed transposed design can operate at an estimated frequency of 276.6 MHz by consuming considerably less resources available on target device to provide cost effective solution for SDR based wireless communication applications.

REFERENCES

[1] Vijay Sundararajan, Keshab K. Parhi, "Synthesis of Minimum-Area Folded Architectures for Rectangular Multidimensional",IEEE Transactions on Signal Processing, pp. 1954-1965, Vol. 51, No.7, July 2003.

[2] ShyhJye Jou, Kai-Yuan Jheng*, Hsiao-Yun Chen and An-Yeu Wu, "Multiplierless Multirate Decimator I Interpolator Module Generator", IEEE Asia-Pacific Conference on Advanced System Integrated Circuits, pp. 58-61, Aug-2004.

[3] Amir Beygi, Ali Mohammadi, Adib Abrishamifar. "AN FPGA-BASED IRRATIONAL DECIMATOR FOR DIGITAL RECEIVERS", in 9th IEEE International Symposium on Signal Processing and its Applications, pp. 1-4, ISSPA-2007.

[4] K. B. Huang, Y. H. Chew, and P. S. Chin "A Novel DS-CDMA Rake Receiver: Architecture and Performance",IEEE International Conference on Communications, pp-2904-2908, ICC-2004.

[5] D.J. Allred, H. Yoo, V. Krishnan, W. Huang, and D. Anderson, "A Novel High Performance Distributed Arithmetic Adaptive Filter Implementation on an FPGA", in Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP'04), Vol. 5, pp. 161-164, 2004.

[6] Patrick Longa and Ali Miri "Area-Efficient FIR Filter Design on FPGAs using Distributed Arithmetic", IEEE International Symposium on Signal Processing and Information Technology, Vol. 4, pp.248-252, 2006.

[7] F.J.A. de Aquino, C.A.F. da Rocha, and L.S. Resende, "Design of CIC filters for software radio system", IEEE International Conference on Acoustics, Speech, Signal Processing, May 2006.

[8] E. B. Hogenauer, "An economical class of digital filters for decimation and interpolation", IEEE Transactions on Acoustics,Speech, Signal Processing, Vol. ASSP-29, pp. 155-162, April 1981.

[9] R. Uusikartano, and J. Takala, "Power-efficient CIC decimator architecture for fs/4-downconverting digital receivers", IEEE International Midwest Symposium on Circuits and Systems, Dec. 2004.

[10] G. Jovanovic Dolecek, S.K. Mitra, " Efficient Comb-Rotated Sinc (RS) Decimator With Sharpened Magnitude Response", IEEE International Midwest Symposium on Circuits and Systems,2004.

[11] Matthew P. Donadio, "CIC Filter Introduction", IEEE International Symposium on Communications , July 2000.

[12] W.A. Abu-Al-Saud, and G.L. Stuber, "Modified CIC filter for sample rate conversion in software radio systems", IEEE Signal Processing Letters, Vol. 10, Issue: 5, pp. 152-154, May 2003.

[13] G.J. Dolecek, and J.D. Carmona, "A new cascaded modified CIC-cosine decimation filter", IEEE International Symposium on Circuits and Systems, May 2005.

[14] M. Becker; N. Lotze et al; "Implementation of a Power Optimized Decimator for Cascaded Sigma-Delta A/D Converters", in ICSES, pp. 83-86, 2004.

[15] Quan Liu; Jun Gao; "On Design of Efficient Comb Decimator with Improved Response for Sigma-Delta Analog-to-Digital Converters", International Conference on image and signal processing, pp. 1-5, 2009.

[16] G. Jovanovic Dolecek; S.K. Mitra; "Two-stage CIC-based decimator with improved characteristics", in IET Signal Process., Vol. 4, pp. 22-29, Oct 2010.

[17] Alfonso Fernandez-Vazquez, Gordana Jovanovic Dolecek, "Passband and Stopband CIC Improvement Based on Efficient IIR Filter Structure", IEEE Transactions on Circuits and Systems, Vol. 52, No. 7, July 2005.

[18] H. Aboushady; Y. Dumonteix; M. Louerat; and H. Mehrez; "Efficient polyphase decomposition comb decimation filters in sigma–delta analog–digital converter", IEEE Trans. Circuits Syst. II, Analog Digit.Signal Process., Vol. 48, No. 10, pp. 898–903, Oct. 2001.

[19] Gordana Jovanovic-Dolecek, Sanjit K. Mitra, "A New Two-Stage Sharpened Comb Decimator", IEEE Transactions on Circuits and Systems, Vol. 52, No. 7, July 2005.

[20] Y. Gao; L. Jia, J. Isoaha; and H. Tenhunen; "A comparison design of comb decimators for sigma–delta analog-to-digital converters", Analog Integr. Circuits Signal Process., Vol. 22, No. 1, pp. 51–60, Jan. 2000.

[21] Chakraborty, R. (2011). FPGA Based Cipher Design & Implementation of Recursive Oriented Block Arithmetic and Substitution Technique (ROBAST). International Journal of Advanced Computer Science and Applications - IJACSA, 2(4), 54-59.

AUTHORS PROFILE

Mr. Rajesh Mehra is currently Assistant Professor at National Institute of Technical Teachers' Training & Research, Chandigarh, India. He is pursuing his PhD from Panjab University, Chandigarh, India. He has completed his M.E. from NITTTR, Chandigarh, India and B.Tech. from NIT, Jalandhar, India. Mr. Mehra has 15 years of academic experience. He has authored more than 15 research papers in reputed International Journals and 35 research papers in National and International conferences. Mr. Mehra's interest areas include VLSI Design, Embedded System Design, Advanced Digital Signal Processing, Wireless & Mobile Communication and Digital System Design.

Ms Rashmi Arora received her B.E. Degree with Honors in Electronics & Communication engineering from M.J.P. Rohilkhand University, Bareilly, U.P., India, in 2002. She is currently pursuing her M.E. degree from NITTTR, Chandigarh, India. She has 7 years of academic experience. Her interest areas are Signal Processing, Embedded Systems and VLSI Design.

Implementation and performance analysis of Video Edge Detection system on Multiprocessor Platform

Mandeep Kaur
Department of Electronics and Communication
Thapar University
Patiala, Punjab, India

Kulbir Singh
Department of Electronics and Communication
Thapar University
Patiala, Punjab, India

***Abstract*— This paper presents an agile development, implementation of Edge Detection on SMT8039 based Video And Imaging module. With the development of video processing techniques its algorithm becomes more complicated. High resolution and real time application cannot be implemented with single CPU or DSP. The system offers significant performance increase over current programmable DSP-based implementations. This paper shows that the considerable performance improvement using the FPGA solution results from the availability of high I/O resources and pipelined architecture. FPGA technology provides an alternative way to obtain high performance. Prototyping a design with FPGA offer some advantages such as relatively low cost, reduce time to market, flexibility. Another capability of FPGA is the amount of support of logic to implement complete systems/subsystems and provide reconfigurable logic for purpose of application specific based programming. DSP's to provide more and more power and design nearly any function in a large enough FPGA, this is not usually the easiest, cheapest approach. This paper designed and implemented an Edge detection method based on coordinated DSP-FPGA techniques. The whole processing task divided between DSP and FPGA. DSP is dedicated for data I/O functions. FPGA's task is to take input video from DSP to implement logic and after processing it gives back to DSP. The PSNR values of the all the edge detection techniques are compared. When the system is validated, it is observed that Laplacian of Gaussian method appears to be the most sensitive even in low levels of noise, while the Robert, Canny and Prewitt methods appear to be barely perturbed. However, Sobel performs best with median filter in the presence of Gaussian, Salt and Pepper, Speckle noise in video signal.**

Keywords-Multiprocessor platform; Edge detection; Performance evaluation; noise.

I. INTRODUCTION

Video processing has been used in many fields such as industry, military, medical image processing, surveillances recording etc. Video and imaging applications demand a range of processes to be performed in single applications. Edge detection is one of the basic characteristics of the image [1]. It is an important basis for the field of image analysis such as: the image segmentation, target area identification, extraction and other regional forms .It is widely used in image segmentation, image recognition, and texture analysis of them. Edge detection[2] technology must not only detect the image gray value of the non-continuity, but also to determine their exact location .Although you can use multiple DSP's to provide more and more power and design nearly any function in a large enough FPGA, this is not usually the easiest, cheapest approach[3]-[6]. The obvious result is to mix the two technologies benefits of co-processing. But DSP and FPGA designs are quite disparate disciplines, involving very different techniques, skills and tools [7][8]. But the differences in DSP and FPGA create obstacles to a fluid co-design process rather unpalatable to a specialist in one of the two fields and even more so to an expert of neither. Integrating the hardware [8]-[14] it also presents a significant amount of work that you could avoid if you stuck with just one technology.

II. EDGE DETECTION

Edge has two properties--the direction and the magnitude [1], [2]. Usually the change of the gray level along the edge is flat, but the pixels perpendicular to the edge change dramatically. According to the characteristics of intensity change, it can be divided into step-type and roof- type. In step type, both sides of the pixel in value have changed significantly, and roof type, it is located in the gray scale to reduce the rate of change from the turning point. This paper introduces edge detection for video [10]-[16] on DSP-FPGA system i.e. SMT8039. These algorithms are based on the detection of discontinuities in the values of the grey levels of the image. The most widely used techniques are the generation of **a** differential image by means of Sobel, Prewitt, Robert, Canny and LOG operator[17]-[21]. The characteristics of these operators, regularity and efficiency, make them adequate for its implementation in an application specific architecture. These operators [5] are based on the differential approach to edge detection. With this approach, a differential image G is generated from the original image F**,** where the changes in grey levels are accentuated. After this, the edges are detected [10]-[13] by means of the comparison of the amplitude values to a predefined threshold level. These are based on the gradient operator. The first derivative of the digital image is based on various approximations to the 2-D gradient. The gradient of the image f(x, y) at location (x, y) is defined as the vector.

$$\Delta f = \begin{bmatrix} \partial f/\partial x \\ \partial f/\partial y \end{bmatrix} \qquad (1)$$

We know that gradient vector points in the direction of maximum rate of change of at coordinates(x,y). An important quantity in edge detection is magnitude of this vector. Δf= mag(Δf)=$[G_x^2+G_y^2]^{1/2}$ .The direction of the gradient vector also is an important quantity. Let α(x,y) represents the direction [1],[2] angle of the vector Δf at (x, y) then from vector analysis:

$$\alpha(x,y) = \tan^{-1}\left(\frac{Gy}{Gx}\right) \qquad (\mathbf{2})$$

Computational of the gradient of an image is based on obtaining the partial derivatives ∂f/∂x and ∂f/∂y at every pixel location. The 3X3 area mask in Fig. 1 for Sobel in Fig. 2 and Prewitt in Fig. 3 operations mask of 3X3, and for Robert operation 2X2 mask is shown in Fig. 4 are used to convolve with each pixel values of the image

Z_1	Z_2	Z_3
Z_4	Z_5	Z_6
Z_7	Z_8	Z_9

Figure 1. 3x3 neighboring of pixels in an image

-1	-2	-1
0	0	0
1	2	1

-1	0	1
-2	0	2
-1	0	1

Figure 2. Sobel matrix x and y directional

-1	-1	-1
0	0	0
1	1	1

-1	-1	-1
0	0	0
1	1	1

Figure 3.Prewitt matrix x and y directional

-1	0
0	1

0	-1
1	0

Figure 4.Robert matrix x and y directional

For Sobel equation is

$$\mathrm{Gx} = (\mathrm{Z7} + 2\mathrm{Z8} + \mathrm{Z9}) - (\mathrm{Z1} + 2\mathrm{Z2} + \mathrm{Z3})$$

$$Gy = (Z3 + 2\,Z6 + Z9) - Z1 + 2Z4 + Z7) \qquad (\mathbf{3})$$

For prewitt equation is

$$Gx = (Z7 + Z8 + Z9) - (Z1 + Z2 + Z3)\ \ Gy = (Z3 +\ Z6 + Z9) - (Z1 + Z4 + Z7) \qquad (\mathbf{4})$$

And for Robert equation is

$$\mathrm{Gx} = \ (\mathrm{Z1} - \mathrm{Z4})$$

$$Gy = (Z2 - Z3\,) \qquad (\mathbf{5})$$

In this formulation, the difference between the first and third rows of the 3X3 image region approximates the derivative in x-direction and the difference between the third and first columns approximates the derivatives in the y-direction. However this implementation is not always desirable because of the computational burden required by Squares and Square root. The equation is much more attractive computationally, and it still preserves relative changes in gray levels. The laplacian is not used in original form because its magnitude produces double edges. The purpose of this operator is to provide image with zero crossing used to establish the location of edges.

III. System Architecture

System architecture includes:

1. CCD camera for PAL or NTSC standard video input.

2. TMS320DM642 DSP board is used as executing image processing algorithms [4].

3. Video processing board is shown as dashed frame in Fig. 6. FPGA is used as logic unit. Virtex 4 FPGA is connected to the DSP's EMIF[4][5]. This allows high speed transfers to be initiated at request. The Module features a single Philips Semiconductors SAA7109AE/108AE video decoder/encoder that accept most PAL and NTSC standards, and can output processed images in PAL/NTSC or VGA (1280x1024, or HDTV Y/Cb/Cr) The DM642 has 128 Mbytes of high speed SDRAM (Micron MT48LC64M32F2S5) available onboard for image processing and an 8Mbytes FLASH device is fitted to store programs and FPGA configuration information. The function SAA7109AE/108AE is to change analog video signals from CCD into digital signal and the image data with the format of YUV 4:2:2 are stored in SDRAM.

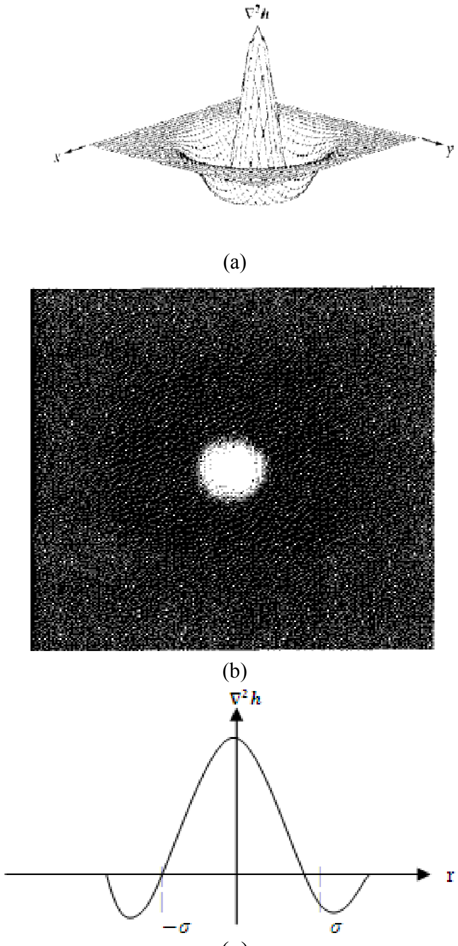
4. VGA display is used as displaying output Image.

IV. Design and Implementation of Edge Detection System on Software

The software 3L Diamond for SMT339 provides a model describing multiprocessor system as a no. of independent tasks that communicate together over a channel[4][5]. Weather these tasks are executing on DSP or FPGA Diamond manages the interconnection and programming so that you can concentrate on the application In this system, different module (tasks) are created. These connections are logically defined for communication between different tasks for DSP and FPGA [14][15][21]. In DSP, a Task Dsp_pal which is written in c language, In DSP, frames information like no. of input frames, no. of output frame, video memory [1] for channel A, B video capture registers, FIFO registers are defined in the library are imported, there are 3 video ports: Vp1 is used for input the video from the camera, Vp2 is undefined and video port Vp0 is used for displaying video on VGA display. For RGB656 format this involves a single EDMA channel, so DMA transfer 64 bit data and for YCbCr, it contains 3 separate channels for initialization.

V. VIDEO EDGE DETECTION SYSTEM TESTS

In this paper we take a frame of video and we perform different edge detection techniques on this frame. In this we added noise like Gaussian, salt and pepper, speckle[22]-[25]. In this paper, we estimate the effect of noise on different edge detection algorithms that which one is more sensitive to the noise, Original video frame is shown in Fig. 10. Fig. 11 shows effects of noise on different edge detection techniques at different PSNR values.



(a)

(b)

(c)

0	0	-1	0	0
0	-1	-2	-1	0
-1	-2	16	-2	-1
0	-1	-2	-1	0
0	0	-1	0	0

(d)

Figure 5.Laplacian of Gaussian) 3D- plot (a) Image black is negative, gray is zero and white is positive. (c) Cross section showing zero crossing (d) 5x5 mask of LOG.

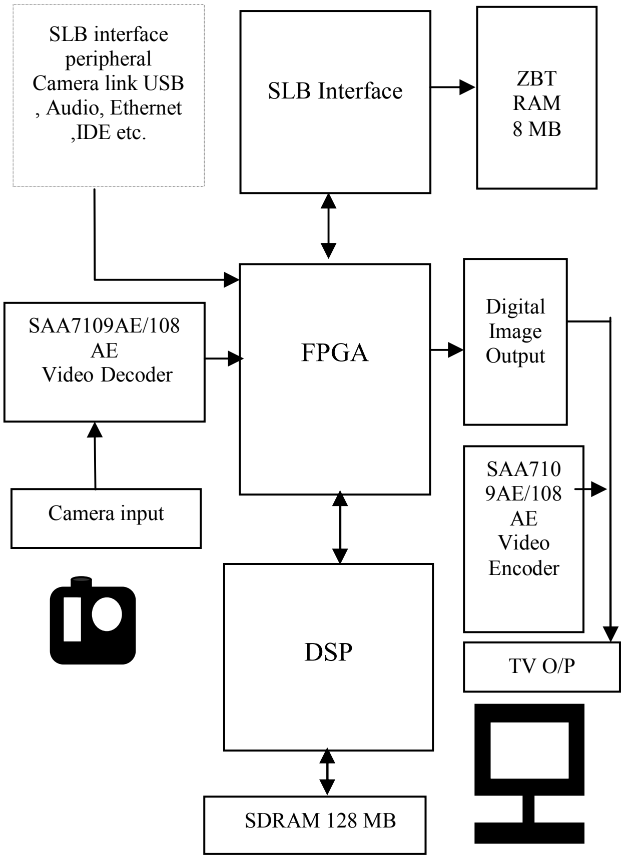

SLB interface peripheral Camera link USB , Audio, Ethernet ,IDE etc.
SLB Interface
ZBT RAM 8 MB
SAA7109AE/108 AE Video Decoder
FPGA
Digital Image Output
SAA710 9AE/108 AE Video Encoder
Camera input
DSP
TV O/P
SDRAM 128 MB

Figure 6. System Architecture

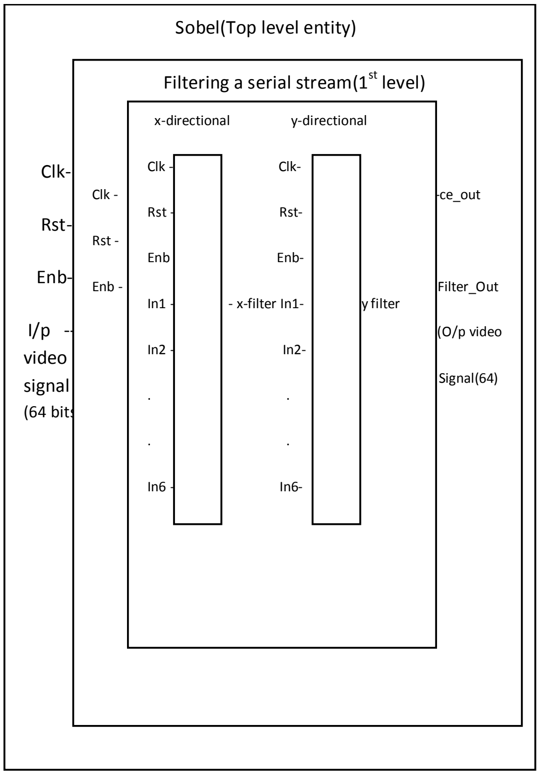


Figure 7.Entity for edge detection (Sobel, Prewitt)

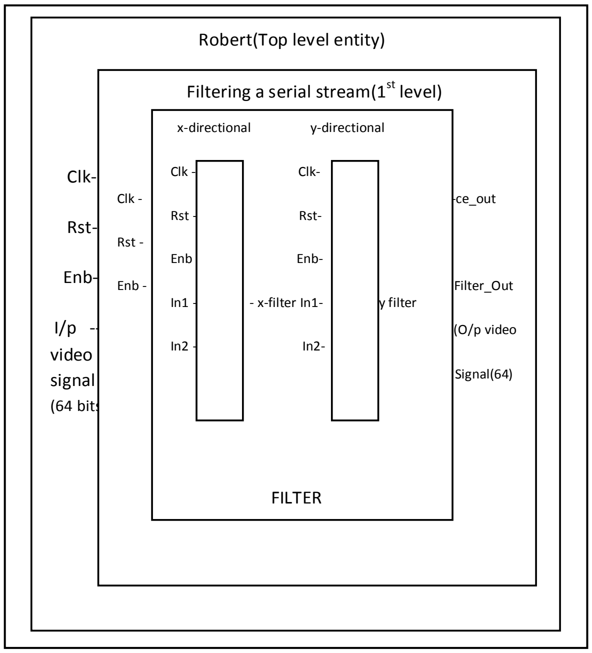


Figure 8.Entity for edge detection (Roberts)

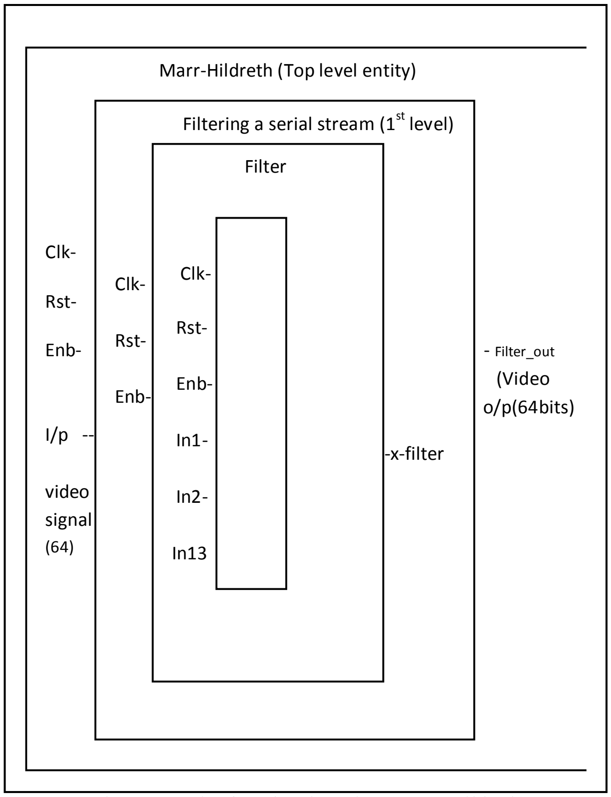


Figure 9.Entity for edge detection (Marr -Hildreth)s

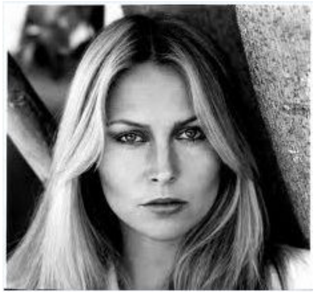
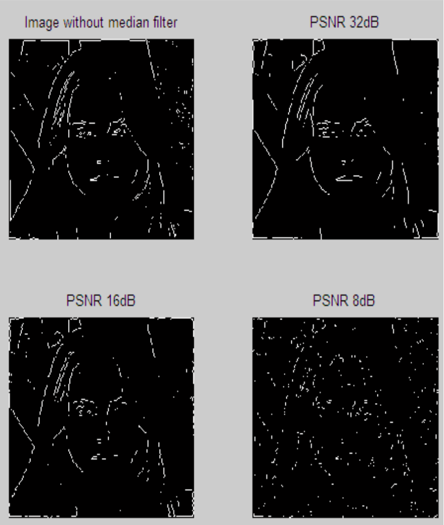
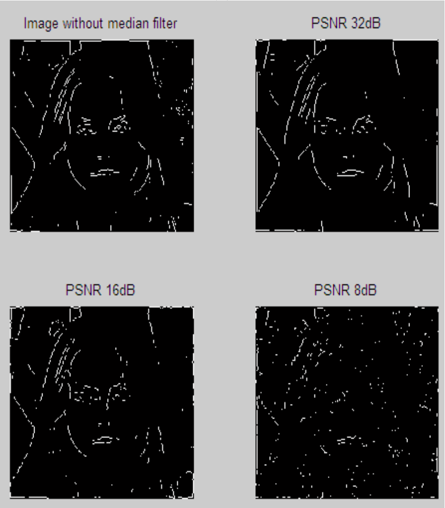


Figure 10.Original Frame



(a)



(b)

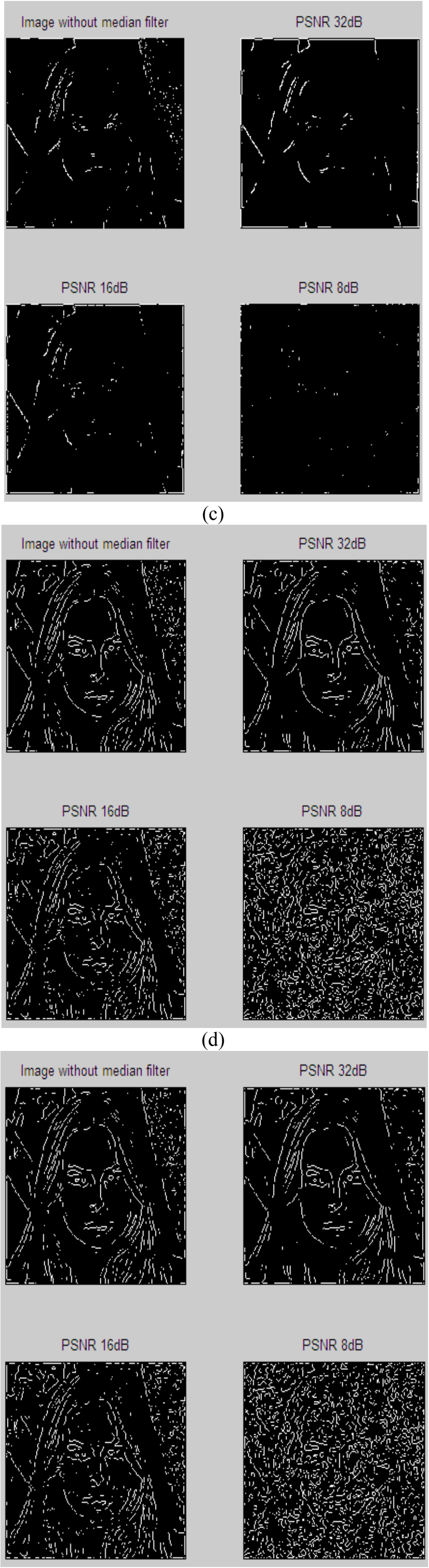


Image without median filter
PSNR 32dB
PSNR 16dB
PSNR 8dB
Image without median filter
PSNR 32dB
PSNR 16dB
PSNR 8dB
Image without median filter
PSNR 32dB
PSNR 16dB
PSNR 8dB

(c)

(d)

(e)

Figure 11. (a) Sobel operation at different PSNR (b) Prewitt operation at different PSNR (c) Roberts operation at different PSNR (d) Marr-Hildreth operation at different PSNR (e) Canny operation at different PSNR

TABLE I.

EFFECT OF GAUSSIAN NOISE ON DIFFERENT EDGE DETECTION TECHNIQUES				
PSNR of noisy image(dB)	***PSNR of noisy image after median filter(dB)***	***Edge Detection Technique (dB)***	***PSNR of edge detection before Median filter(dB)***	***PSNR Of edge detection after Median filter(dB)***
38.3437	24.4731	Sobel	3.1254	4.1513
		Prewitt	11.0582	12.058
		Robert	12.6772	11.0199
		Marr-Hildreth	10.7655	10.2409
		Canny	10.3832	11.2856
16.0414	21.4583	Sobel	7.1792	13.0138
		Prewitt	8.6966	13.1051
		Robert	15.1037	12.8259
		Marr-Hildreth	10.9128	9.88
		Canny	4.9631	11.2809
10.487	17.297	Sobel	0.0508	12.7839
		Prewitt	9.6073	13.2758
		Robert	15.5064	13.772
		Marr-Hildreth	10.7196	10.0412
		Canny	9.9918	11.1472
8.1645	14.7276	Sobel	-3.857	11.6824
		Prewitt	-3.5126	11.5007
		Robert	13.2864	11.2624
		Marr-Hildreth	10.404	10.2732
		Canny	10.2075	10.8719

PSNR in dB for Gaussian Noise

We also compare the PSNR values of the all the edge detection techniques which are listed above with different kinds of noise levels and noise type [21] [26] [27]. Out of five operators, Sobel edge detection method is found as the best in detecting the edges in noisy images. By applying median filter to the noisy image, noise is removed from the images and then all techniques are applied to filtered frame [28]. So the paper concludes that Sobel edge detector with the Median filter performs well in detecting the edges, when compared to other edge detector with median filter [28][23].

In Fig. 11a, shows Sobel operation, Fig. 11b shows Prewitt operation, Fig 11c shows Robert's operation and Fig. 11d shows LOG operation and Fig. 11e shows Canny. Fig 11a shows image with median filter original and second image at 32dB PSNR, 3rd at 16dB and last one at 8dB. Here PSNR is calculated by comparing the mean of the pixel values with the mean of the additive Gaussian noise [29][23]. The noise is multiplied by the proper scale so that it has a mean value of 0.016 for the 32 dB case. At this PSNR level, all methods return acceptable results. As the values of PSNR decreased, performance decreased [30]-[35].

TABLE II.

EFFECT OF SALT & PEPPER NOISE ON DIFFERENT EDGE DETECTION TECHNIQUES

PSNR of noisy image(dB)	***PSNR Of Noisy Image After median filter(dB)***	***Edge Detection Techniques (dB)***	***PSNR Of edge detection before Median filter (dB)***	***PSNR edge detection after Median filter (dB)***
16.139	21.5725	Sobel	1.921	4.2946
		Prewitt	6.6144	12.5597
		Robert	12.8492	11.0199
		Marr-Hildreth	10.0447	10.5141
		Canny	10.1012	11.963
12.7221	19.8018	Sobel	2.4532	12.5432
		Prewitt	2.9922	12.4228
		Robert	12.7041	12.9375
		Marr-Hildreth	10.0029	10.3087
		Canny	9.9916	10.8565
19.5703	23.1778	Sobel	7.9475	12.5191
		Prewitt	10.23	12.5832
		Robert	12.7041	12.285
		Marr-Hildreth	10.1879	10.6876
		Canny	10.2506	11.1916
13.4067	20.0276	Sobel	3.155	12.5415
		Prewitt	3.7613	12.464
		Robert	12.8773	12.8604
		Marr-Hildreth	9.9325	10.5383
		Canny	10.0141	10.9018

PSNR in dB for Salt and Pepper noise

TABLE III.

EFFECT OF SPECKLE NOISE ON DIFFERENT EDGE DETECTION TECHNIQUES

PSNR Of noisy image(dB)	***PSNR Of Noisy Image After Median Filter(dB)***	***Edge Detection Techniques (dB)***	***PSNR of edge detection before Median filter (dB)***	***PSNR edge detection after Median filter (dB)***
23.9577	23.417	Sobel	3.3957	4.332
		Prewitt	12.4686	12.6079
		Robert	14.3418	12.563
		Marr-Hildreth	10.3874	11.3048
		Canny	10.2656	10.78
21.0908	23.3807	Sobel	10.6997	13.1004
		Prewitt	12.7342	13.143
		Robert	15.2956	12.563
		Marr-Hildreth	10.3882	11.3595
		Canny	10.2664	10.834
15.02323	20.8912	Sobel	7.9304	13.7925
		Prewitt	12.7663	13.0254
		Robert	18.2607	13.2232
		Marr-Hildreth	10.3815	11.704
		Canny	10.2504	11.3389
18.2588	22.5847	Sobel	10.1194	13.438
		Prewitt	12.1019	13.8302
		Robert	16.5047	12.7849
		Marr-Hildreth	10.3638	11.4982
		Canny	10.2387	11.0434

PSNR in dB for Speckle Noise

TABLE IV
RESOURCES USED

BUFG	**DCM**	**RAM16**	**SLICES**
4 out of 32	***2 out of 12***	***1 out of 232***	***953 out of 25280***
12%	16%	1%	13%

TOTAL TIME FOR BUILDING THE APPLICATION:

Synthesis time: 70 second (1 min and 10 sec)
Translate: 19 second (0 min and 19 sec)
Map: 60 second (1 min and 10 sec)
Place and route: 61 second (1 min and 1 sec)
Generate bit stream: 32 second (0 min and 32 sec)

After the verification of this design in Xilinx FPGA development board, video edge detection system achieves the desired test results. Both filtering and edge detection perform well.

VI. CONCLUSION AND FUTURE WORK

This paper realizes a DSP-FPGA based video edge detection system and combines the respective strengths of FPGA's and DSP's can be starting with a scalable system. Supported by a comprehensive software environment, such complex hardware can become both adaptable and accessible Verification on SMT8039 development board and on VGA display indicates that the system can accurately detect the images edge and satisfy requirements of the real-time video image edge detection. Finally it achieves the desired experimental results. Out of five operators, Sobel edge detection method is found to be the best in detecting the edges in noisy images. The Laplacian of Gaussian method appears to be the most sensitive to even low levels of noise, while the other methods appear to be barely perturbed. In fact, even though the other methods appear to be returning nearly perfect results.

ACKNOWLEDGMENT

The authors would like to thank Mr. Ankush Kansal for willingly answering numerous questions regarding the Video algorithms. The authors would also like to thank Mr. Parminder Singh Reel for many valuable discussions; and the authors of the edge detectors for making the implementations available. Early results of this work appear in [8], [9],[17].

REFERENCES

[1] Rafael C. Gonzalez, Richard E. Woods, Digital Image Processing, Beijing: Publishing House of Electronics Industry, 2nd Edition 2003.

[2] Kenneth R. Castleman, Digital Image Processing, Pearson Education Asia Limited and Tsinghua University Press, 2003.

[3] J. Batlle, J. Marti, P. Ridao. "A New FPGA/DSP-Based Parallel Architecture for Real-time Image Processing", Real-Time Imaging, vol. 8, pp 345-356, 2002

[4] TMS320DM642 Video/ Imaging Fixed-Point Digital Signal Processor Data Manual[EB/OL].Texas Instruments Incorporated,2003.

[5] SMT339 User Manual (QCF42); Version 1.2, 8/11/00 Sundance Multiprocessor Technology Ltd. 1999.

[6] C.Qi, Y.H. Chen, T.S. Huang. The Real-time Image Processing Technique Based on DSP[J]. Wuhan University Journal of Natural Sciences, 2005,10(6):1025-1029 Daniel Baumgarthner, Peter Rossler and Wilfried Kubingger "performance Benchmark of DSP and FPGA Implementations of Low-Level Vision Algorithms," Computer Vision and Pattern Recogniton, 2007. CVPR '07. IEEE Conference, Digital Object Identifier, pp. 17-22, June 2007.

[7] Miguel A. Vega-Rodriguez, Juan M. Sanchez-Perez, Juan A. Gomez-Pulido, "Real Time Image Processing with Reconfigurable Hardware", Electronics Circuits and Systems. The 8th IEEE International Conference on Digital Object Identifier, pp 213-216 Vol.1, Sept 2001.

[8] Jincheng Wu, Jingrui Sun, Wenying Liu, "Design and Implementation of Video Image edge Detection System Based on FPGA" 2010 3rd International Conference on Image and Signal Processing (CISP2010), 2010.

[9] Duan Jinghong, Deng Yanling, Liang Kun "Development of Image Processing System Based on DSP and FPGA," Electronic Measurement and Instruments, 2007. ICEMI '07. 8th International Conference on Digital Object Identifier, vol. 2 pp. 791 -794, 18 July, 2007.

[10] C. Vivas, J. Gomez Ortega and M. Vargas "parallel DSP based implementation of an automated visual quality inspection system", Emerging Technologies and Factory Automation, 1999. proceedings. ETFA'99.199 7th IEEE International Conference, Digital Object Identifier: 10.1109/ ETFA. 1999.815429, pp.743-749 vol.1, Barcelona, Spain, 10/18/1999- 10/21/1999.

[11] Patrick Murphy, Vinay Bharadwaj, Erik Welsh, J.Patrick Frantz, "A DSP-Based Platform for Wireless Video Compression", Global Telecommunications Conference, 2002. GLOBE COM' 02.IEEE, Digital Object Identifier: 10.1109/ GLOCOM. 2002. 1188499, pp.1754-1758 vol.2, 17-21 Nov. 2002.

[12] Manish Kumar Birla "FPGA Based Reconfigurable Platform for Complex Image Processing," Electro/information Technology, 2006 IEEE International Conference, Digital Object Identifier: 10.1109/EIT.2006.252111, pp. 204-209, East Lansing, MI, 7-10 May 2006.

[13] Miguel A. Vega-Rodriguez, Juan M. Sanchez-Perez, Juan A. Gomez-Pulido "Real Time Image Processing with Reconfigurable Hardware", Electronics, Circuits and Systems, 2001. ICECS 2001. The 8th IEEE International Conference, Digital Object Identifier 10.1109/ ICECS. 2001. 957718, pp.213-216 vol.1, 2-5 Sept. 2001.

[14] Daniel Baumgarthner, Peter Rossler and Wilfried Kubingger "performance Benchmark of DSP and FPGA Implementations of Low-Level VisionAlgorithms," Computer Vision and Pattern Recogniton, 2007. CVPR '07. IEEE Conference, Digital Object Identifier: 10.1109/ CVPR. 2007. 38342, pp. 1-8,June 2007.

[15] Duan Jinghong, Deng Yanling, Liang Kun "Development of Image Processing System Based on DSP and FPGA," Electronic Measurement and Instruments, 2007. ICEMI '07. 8th International Conference, Digital Object Identifier ICEMI..4350799, pp. 2-791 - 2-794, July 18 2007.

[16] Luxin Yan, Tianxu Zhang and Sheng Zhong "A DSP/FPGA-Based Parallel Architecture for Real-time Image Processing," Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress, Digital Object Identifier 10.1109/ WCICA.2006.1713959, pp.10022-10025, 21-23, June 2006.

[17] T. Peli and D. Malah, "A study of edge detection algorithms", IEEE transaction on Computer Graphics and Image Processing. vol. 20, no. 1, pp. 1–21, Sept. 1982.

[18] W. Lunscher and M. Beddoes, "Optimal edge detector evaluation", IEEE Transaction on System Manufacturing and Cybernetics, vol. SMC-16, pp. 304–312, Apr. 1986.

[19] Renyan Zhang, Guoliang Zhao and Li Su, "A New Edge Detection Method in Image Processing", Proceedings of ISCIT, pp 430-433, 2005.

[20] A. Rosenfeld and M. Thurston, "Edge and curve detection for visual scene analysis," IEEE Transaction On Computing, vol. C-20, no. 5, pp. 562-569,1971.

[21] M. Boo, E. Antelo and J.D. Bruguera, "VLSI Implementation of an Edge Detector Based on Sobel Operator", Department of Electronics University of Santiago de Compostela Santiago de Compostela Spain, 1994.

[21] Yan Lei, Zho Gang, "The performance of image acquisition and processing system based on DSP-FPGA", International Conference on Smart Manufacturing And Application,2008 Korea.

[22] John Canny, "A Computational Approach to Edge detection", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol-6, pp 679-698, 1986.

[23] G.Padamvathi, P Subashini, "Performance evaluation of the various edge detectors and filters for the noisy IR images", IEEE proceedings on Sensors, Signals, Visualization, Imaging, Simulation And Materials, pp 199-203, 2009.

[24] I.yasri, N.H hamid, "Performance analysis of FPGA based Sobel Edge Detection operator", International Conference on Electronic Design Malaysia,2008.

[25] V. Ramesh , R. M. Haralick, "Performance characterization of edge detectors," SPIE Application and Artificial Intelligence and Machine Vision on Robotics , vol.1708, pp. 252–266, April. 1992.

[26] D. Marr and E. Hildreth, "Theory of Edge Detection", Proc. of the Royal Society of London B, vol. 207, pp. 187-217, 1980.

[27] W. Lunscher and M. Beddoes, "Optimal edge detector evaluation," IEEE Trans. Syst., Man, Cybern., vol. SMC-16, pp. 304–312, Apr. 1986.

[28] M.Shin,D. Goldgof, and K. W. Bowyer, "An objective comparison methodology of edge detection algorithms for structure from motion task," in Proc. IEEE Conf. Comput. Vision Pattern Recognit., Santa Barbara,CA, pp. 190–195, 1998.

[29] I. E. Abdou and W. K. Pratt, "Quantitative design and evaluation of enhancement/thresholding edge detectors," Proc. IEEE, vol. 67, pp. 753–763, May 1979.

[30] G.B.Shaw, "Local and regional edge detectors: Some comparison", Computer Graphics an Image Processing, vol. 9, pp. 135–149, Feb. 1979.

[31] L. Rosenthaler, F.Heitger, O.Kubler, and R.von der Heydt, "Detection of general edges and key points," in Proceeding of ECCV, G.Sandini, pp. 78–86, 1992.

[32] F. Heijden, "Edge and line feature extraction based on covariance models IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 17, pp. 16–33, Jan. 1995.

[33] Huang Yu-cheng, Hu Guo-qing, Wu Xion-gying, "A Study of De-noising Methods in Face Recognition System[J].Control & Automation, pp. 187-188, 2005.

[34] Xu Xian-ling, Lin Yi-shui, "Comparison and Analysis for Image Edge Detection Algorithms" [J]. Automation & Information Engineering. pp. 44-46, 2007.

[35] S.P.Ierodiaconou, N, Dahnoun, L.-Q. Xu "Implementation and Optimisation of a Video Object Segmentation Algorithm on an Embedded DSP Platform," Crime and Security. 2006. The Institution of

Engineering and Technology Conference, INSPEC Accession Number: 9518513, pp. 432-437, 13-14 June 2006.

AUTHORS PROFILE

Mandeep Kaur received the BTech. degree, in Electronics and Communication Engineering from Shaheed Bhagat Singh College of Engineering and Technology, Firozepur in 2008. She is currently pursuing the M.E degree in Electronics and Communication at Thapar University, Patiala.Her research interests include Video Compression and Evaluation of algorithms.

Dr. Kulbir Singh was born in Batala (Pb) India. He received his BTech degree in 1997 from PTU, Jalandhar. He obtained his ME and PhD degree from Thapar Institute of Engineering and Technology, Patiala in 2000 and 2006 respectively. He worked as lecturer from 2000 to 2007 in Electronics and Communication Engineering Department, Thapar University, Patiala. Presently he is working as Assistant Professor in Thapar University, Patiala since July 2006. He has published about 60 research articles in refereed international journals, international conference and national conference. He is life time member of IETE and ISTE. His research interest include Digital Signal Processing, Image Processing, Fractional Fourier Transform and Communication systems.

A robust multi color lane marking detection approach for Indian scenario

L N P Boggavarapu, R S Vaddi,
K R Anne
Department of Information Technology,
V R Siddhartha Engineering College,
Vijayawada.

H D Vankayalapati
Department of Computer Science Engineering
V R Siddhartha Engineering College, Vijayawada.

J K Munagala
Department of Electronics & Computers
PVP Siddhartha Institute of Technology, Vijayawada

***Abstract*— Lane detection is an essential component of Advanced Driver Assistance System. The cognition on the roads is increasing day by day due to increase in the four wheelers on the road. The cognition coupled with ignorance towards road rules is contributing to road accidents. The lane marking violence is one of the major causes for accidents on highways in India. In this work we have designed and implemented an automatic lane marking violence detection algorithm in real time. The HSV color-segmentation based approach is verified for both white lanes and yellow lanes in Indian context. Various comparative experimental results show that the proposed approach is very effective in the lane detection and can be implemented in real-time.**

Keywords- Color segmentation; HSV; Edge orientation; connected components.

I. INTRODUCTION

Traffic accidents have become one of the most serious problems in today's world. Roads are the choicest and most opted modes of transport in providing the finest connections among all other modes [1]. Due to increase in vehicles from 3,00,000 in 1951 to about 7,30,00,000 in 2004 [2] as shown in Fig 1, traffic accidents especially road accidents have become predominant. According to official statistics 105,725 people were killed in road traffic crashes in India in 2006 (NCRB, 2007) [3]. During recent years, traffic fatalities increased by about 5 percent per year from 1980 to 2000 [2] and since then have increased by about 8 percent per year for the four years for which statistics are available as shown in Fig 2.

The major factors that contribute to road accidents are due to negligence of the driver. Reducing the accidents on road is possible by improving the road safety. Increasing the safety and saving lives of human beings is one of the basic features in Advanced Driver Assistance System (ADAS), a component in Intelligent Transportation System (ITS). A real time computer vision based system plays an important role in providing a useful and effective information like lane marking [20], departure and front and side images etc. The present paper deals with the detection of lanes on roads especially Indian typical roads.

Many researchers have shown lane detectors based on a wide variety of techniques. Techniques used varied from using monocular to stereo vision using low level morphological operations to using probabilistic grouping and B-snakes [22]. All the techniques are classified into two main categories namely feature based techniques and model based techniques. The feature based technique combines low level features like color; shape etc. in order to detect the lane and the model-based scheme is more robust in lane detection when different lane types with occlusions or shadows are handled. Road and lane markings can vary greatly, making the generation of a single feature-extraction technique is difficult. So, we combined the features of both color based and edge based techniques [23], [24].

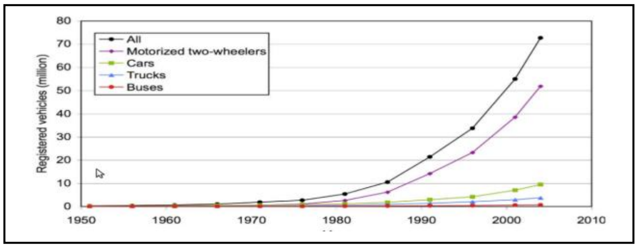


Figure 1. Registered vehicles, 1951 through 2004

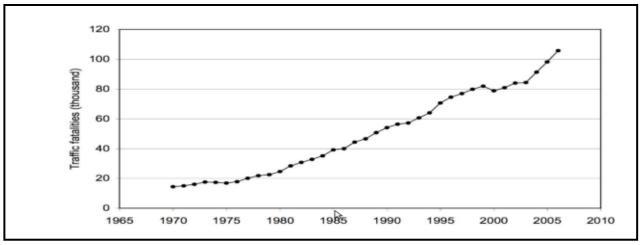


Figure 2. Traffic fatalities, 1970 through 2006

II. STATE OF ART

During the Recent years, many techniques for processing and analyzing images for lane detection have been proposed which involves, Template Matching and Ellipse Modeling Approach to Detecting Lane Markers [4]. Lakshmanan and

Kluge [5] applied deformable template model of lane structure to locate lane boundaries without thresholding the intensity gradient information. Yim and Oh [6] developed a three-feature-based automatic lane-detection algorithm using the starting position, direction, and gray-level value of a lane boundary as features to recognize the lane. ZuWhan Kim [7] developed new method for lane detection which involves lane boundary hypotheses generation and probabilistic lane grouping. M. Bertozzi and A. Broggi [8] were proposed a method for lane-detection which is based on morphological filter. It is observed from literature study that only very few attempts have been made to work on color images [9]. Yu-Chi Leng and Chieh-Li Chen [10] proposed a method for lane-detection which is based on urban traffic scenes [20]. Yue Wang, Eam Khwang Teoh and Dinggang Shen developed Lane detection and tracking using B-Snake [22]. H. Zhang, et al proposed a Novel Lane Detection Algorithm Based on Support Vector Machine.[21]. In the present study, two lane features, lane width and lane boundary continuity, are proposed to obtain reliable and quality lane detected results.

III. Lane marking detection using HSV-Edge based Approach

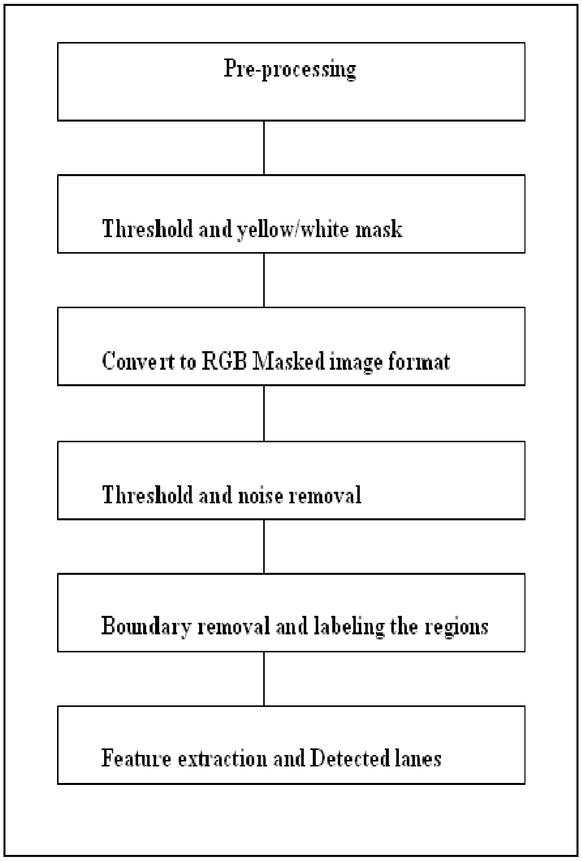

Pre-processing
Threshold and yellow/white mask
Convert to RGB Masked image format
Threshold and noise removal
Boundary removal and labeling the regions
Feature extraction and Detected lanes

Figure 3. Proposed approach

Many lane detection approaches [17], [18], [19] use color model in order to segment the lane line from background images. However, the color feature [2] is not sufficient to decide an exact lane line in images depicting the variety of road markings and conditions. If there are many lanes or obstacle which is similar to lane color, it will be difficult to decide an exact lane. Similarly, some lane detection method uses only edge information. The proposed method involves the combination of both color segmentation and edge orientation to detect lanes of roads of any color (especially yellow and white which are the common colors for the lane).

IV. Color Segmentation

In color segmentation [11-12], we will find out the objects or part of the image that were of the lane color. That is the image at the end possesses only those parts of the image which has the lane color (yellow or white). The color segmentation deals with the color feature of lane markings. The method works out with the hue saturation value color model rather than the red green blue color model.

A. Pre-processing

The pre-processing includes reading the input image, conversion to HSV format and split into individual H, S and V bands. The first step in the process is reading the input image. If the image taken is a gray scale image then the image is read as a two dimensional array. If the input image file contains a true color (RGB) image, it is a three-dimensional array mxnx3. The RGB image is converted into HSV format [13] as hue, saturation and value are properties of a particular color in an image whereas Red, Green and Blue are the primary colors which when combined gives rise to a particular color. Then we segment the areas of a particular color of HSV image by applying thresholds to Hue, Saturation and Value from the HSV image.

B. Threshold and yellow/ white mask

The values of the Hue, saturation and value of the Lane color are set as the thresholding values. Table 1 describes the set of threshold values are used for yellow and while colored lanes. The values are taken after through experimentation on different types of road images under various conditions.

C. Boundary removal and labeling the regions by connected components

We employ the logical and operation of the masks in the above step to obtain the Single Lane colored mask and then apply the colored masks to RGB bands by typecast the yellow object mask/white object mask, as the case may be, into the class of RGB image by using cast function. Now this RGB image obtained is split into individual Red, Green and Blue bands .To each Individual band we multiply the “yellow Objects Mask/ White Objects Mask” so as to obtain masked red green blue images separately. Now we concatenate these masked images into a single masked RGB image. The final masked RGB image consists of the desired lane colored portions only.

TABLE I. Threshold values for detecting yellow and white color lanes

Table Head	**Yellow color lane**		**White color lane**	
	Low	***High***	***Low***	***High***
Hue	0.10	0.14	0.0	0.14

Table Head	**Yellow color lane**		**White color lane**	
	Low	***High***	***Low***	***High***
Saturation	0.4	1.0	0.0	0.2
Value	0.8	1.0	0.8	1.0

V. Edge Orientation

The lane colored objects that are identified in the color segmentation step are subjected to the edge orientation by using the eccentricity [14] property of shape. After the image is subjected to these steps we finally will be able to detect the lane of any color on the road.

A. *Threshold and noise removal*

The masked image obtained at the end of Color Segmentation process is considered as the input image in this step. In the thresholding step [15-16], the input image is converted into a gray scale image. The intensity of a gray scale image varies from 0 to 255(0 for black and 255 for white).Since the lanes to be detected are in white color, the pixel intensity at the region of the lane is closer to 255 (>200). Therefore we set a threshold value above 200. The pixels whose intensity is above threshold are made white and the pixels below the threshold are eliminated. Thus a gray scale image is converted into a binary image that has intensity values 0 or 1 of which our lanes are an integral part.

The next step in our process is removal of unnecessary pixels. These unnecessary pixels include the noise and the boundary objects existing in the image. We performed these operations by performing certain morphological operations on the image. We first morphologically open the binary image and eliminate all the connected components of the binary image that have the number of pixels less than the amount specified by us.

B. *Remove boundary objects & labeling the regions*

The next step in the process is identifying the boundaries of the lighter regions inside the binary image. Through this the number of regions existing in the image is identified and is also labeled accordingly. The regions in the image are visualized by assigning the colors to all the regions existing uniquely.

C. *Feature extraction and detecting the lanes*

After identifying different regions inside the binary image, we need to measure the properties like region like Area, Eccentricity, Major Axis Length, Minor Axis Length, Orientation and Perimeter etc., of each and every region of the image in order to find the lanes present in the road image. In this paper we are mainly concerned with the property Eccentricity. The value of eccentricity varies from 0 to 1. The value 0 indicates that the region is in the form of a circle and if it is 1, then the region is a straight line. Since our lanes are straight, they may have the eccentricity value closer to 1. We store the eccentricity values of each region in an array and compare them with a value 0.98 and the regions having eccentricity more than 0.98 are identified and stored separately in another array.

The final step of the system is detecting the lanes. This can be easily done by mapping the original image with the binary image on which the lanes have been identified. We iterate the array of straight lines produced in the above step and mark each pixel of the region being iterated is marked with the required color thus identifying the lanes in the image.

VI. Performance Analysis for yellow colored images and white colored images

The images shown in Figure 4 and Figure 5 represent the way to detect the yellow colored lanes and white colored lanes respectively by using the HSV values shown in Table 1. In Figure 4, we consider yellow colored lane image as input and after applying the said series of steps we have an output image with identified yellow colored lines and in Figure 5, a typical Indian road with white colored lane is considered for the test and after applying the said series of steps, the exact location of the white lines are detected and identified.

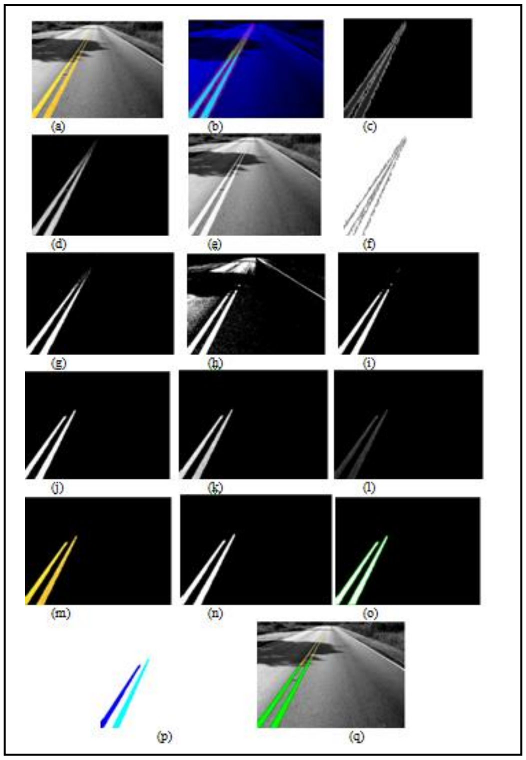


Figure 4. (a) Original Image (b) HSV image(c) Hue Image(d) Saturation image(e) Value image (f) Masked Hue(g) Masked Saturation(h) Masked Value(i) Yellow Masked image (j) Masked Red Image(k) Masked green image(l) Masked blue image(m) Masked yellow image (n) Threshold Image (o) Boundaries identified (p) Regions labeled (q) Identified lanes.

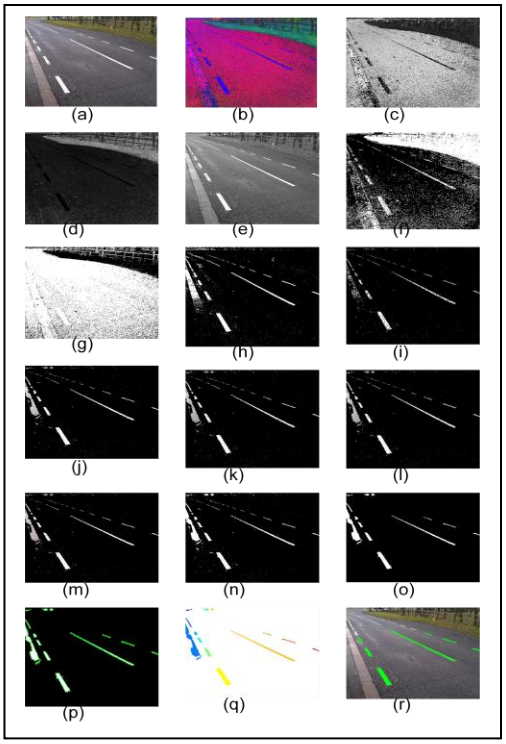


Figure 5. (a) Original RGB Image (b) HSV image (c) Hue Image (d) Saturation Image (e)Value Image (f) Hue Mask (g) Saturation Mask (h) Value Mask (i) White portions left (j) Masked Red Image (k) Masked Green Image (l) Masked Blue Image (m) Masked RGB Image (n) Thresholded Image (o) Noise Removed image (p) Boundaries Identified (q) Regions Labeled (r) Lanes Identified

VII. CONCLUSION

In this paper, we have developed and implemented a novel algorithm to detect white and yellow colored lanes on the road. The lane detection method is robust and effective in finding the exact lanes by using both color and edge orientations. The main contributions in this paper are the color segmentation procedure identifies the yellow of white colored lanes followed by edge orientation in which the boundaries are eliminated, regions are labeled and finally the lanes are detected. As the height of the camera is relatively constant with respect to the road surface, the road portion of the image can be exclusively cropped by providing the coordinates, so that identifying the lanes becomes much more efficient. The experimental results show the effectiveness of the proposed method in cases of yellow and white colored lanes. The entire work is done in a static way that is on an image. We can extend this to detect lanes in a video.

ACKNOWLEDGMENT

This work was supported by Major Research Project on Robust non-intrusive multi-modal real-time “vehicle black box” to monitor driver and vehicle state funded by UGC, New Delhi.

REFERENCES

[1] S Sathees,Man Made Disasters-Road Accidents, Highways Research Station, Chennai

[2] Dinesh Mohan, Omer Tsimhoni, Micheal Sivak, Michael J Flannagan: The University of Michigan Strategic Worldwide Transportation 2020-ROAD SAFETY IN INDIA: CHALLENGES AND OPPORTUNITIES (2009)

[3] NCRB [National Crime Records Bureau], (2007), Accidental deaths and suicides in India - 2006, New Delhi: Ministry of Home Affairs, National Crime Records Bureau.

[4] Amol Borkar, Monson Hayes and Mark T. Smith: A Template Matching and Ellipse Modeling Approach to Detecting Lane Markers, Lecture Notes in Computer Science (2010)

[5] K. Kluge, S. Lakshmanan: A deformable-template approach to lane detection, in Proc. IEEE Intell. Vehicle Symp (1995)

[6] Y. U. Yim , S. Y. Oh: Three-feature based automatic lane detection algorithm (TFALDA) for autonomous driving, IEEE Trans. Intell. Transp. Syst., vol. 4, no. 4 (2003)

[7] ZuWhan Kim: Lane Detection and Tracking in Challenging Scenarios IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 9, NO. 1 (2008)

[8] M. Bertozzi, A. Broggi: GOLD: a parallel real-time stereo vision system for generic obstacle and lane detection. IEEE Transactions on Image Processing (1997)

[9] F. Diebolt: Reconnaissance des marquages routiers. PhD thesis, L'Univ. Louis Pasteur de Strasbourg, Strasbourg, France(1996)

[10] Yu-Chi Leng , Chieh-Li Chen: Vision-Based Lane Departure Detection System in Urban Traffic Scenes 2010 11th Int. Conf. Control, Automation, Robotics and Vision (2010)

[11] Y. He, H. Wang, B. Zhang, Color-Based Road Detection in Urban Traffic Scenes IEEE Transactions on ITS, vol. 5, pp. 309-318 (2004)

[12] H.Y. Cheng et al : Lane detection with moving vehicle in the traffic scenes IEEE Transactions on ITS, Vol. 7, pp. 571-582(2006)

[13] Jin-Wook Lee, Jae-Soo Cho, Effective lane detection and tracking method using statistical modeling of color and lane edge-orientation in Advanced in Information Sciences and Service Sciences Volume 2, Number 3 (2010)

[14] Amol Borkar, Monson Hayes and Mark T. Smith, A Template Matching and Ellipse Modeling Approach to Detecting Lane Markers Lecture Notes in Computer Science, 2010, Volume 6475/2010, 179-190, DOI: 10.1007/978-3-642-17691-3_17

[15] S. Nedevschi, R. Schmidt, T. Graf, R. Danescu, D. Frentiu, T. Marita, F. Oniga, and C. Pocol : 3D lane detection system based on stereovision in Proc. IEEE Intelligent Transportation Systems Conf., Washington, DC, pp. 161–166 (2004)

[16] D.J. Kang , M.-H. Jung: Road lane segmentation using dynamic programming for active safety vehicles - Pattern Recognition Letters,Vol . 2 4 , pp. 3177-3185 (2003)

[17] Hong Wang Qiang Chen “Real-time lane detection in various conditions and night cases”, IEEE Conference on ITS, ITSC '06. pp.1226-1231

[18] Lue-Wei Tsai, Jun-Wei Hsieh, Chi-Hung Chuang, and Kuo-Chin Fan. “Lane Detection Using Directional Random Walks.” IEEE Intelligent Vehicles Symposium, 2008.

[19] J. W. Lee, “A machine vision system for lane-departure detection,” Com-put. Vis. Image Underst., vol. 86, no. 1, pp. 52–78, Apr. 2002.

[20] M. Aly, "Real time detection of lane markings in urban streets," presented at the IEEE Intelligent Vehicles Symposium, Eindhoven, The Netherlands, 2008.

[21] H. Zhang, et al., "A Novel Lane Detection Algorithm Based on Support Vector Machine," presented at the Progress in Electromagnetics Research Symposium, Hangzhou, China, 2005

[22] Yue Wang ,Eam Khwang Teoh, and Dinggang Shen Lane detection and tracking using B-Snake ELSEVIER Image and Vision Computing 22 (2004) 269–280

[23] Road Lane Detection with Elimination of High-Curvature Edges Krzysztof Ślot, Michał Strzelecki, Agnieszka Krawczyńska, Maciej Polańczyk ICCVG 2008 Proceedings of the International Conference on Computer Vision and Graphics

[24] An Effective and Fast Lane Detection Algorithm Chung-Yen Su, Gen-Hau Fan ISVC '08: Proceedings of the 4th International Symposium on Advances in Visual Computing, Part II.

[25] Chiu, K.-yu, & Lin, S.-fuu. (2010). A Face Replacement System Based on Face Pose Estimation. International Journal of Advanced Computer Science and Applications - IJACSA, 1(6), 147-156.

[26] Mahjoub, M. A. (2011). Image segmentation by adaptive distance based on EM algorithm. International Journal of Advanced Computer Science and Applications - IJACSA, (Special Issue), 19-25.

[27] Kekre, H. B. (2010). Texture Based Segmentation using Statistical Properties for Mammographic Images. International Journal of Advanced Computer Science and Applications - IJACSA, 1(5), 102-107.

AUTHORS PROFILE

L N P Boggavarapu received Master degree in Computer Science & Engineering from Acharya Nagarjuna University. Presently he is working as a Lecturer in the Department of Information Technology VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India

R.S Vaddi received both Bachelor and Master degrees from Acharya Nagarjuna University. Presently he is working as a Lecturer in the Department of Information Technology VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India.

J K Munagala received both Bachelors and Masters from Acharya Nagarjuna University. Presently he is working as Assistant Professor in the Department of Electronics & Computers, P V P Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India.

H.D Vankayalapati received Masters Degree from University of Klagenfurt, AUSTRIA. Presently she is working as a Lecture in the Department of Computer Science and Engineering, VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India.

K.R Anne received his Bachelors degree from Bharatiyar University, Masters from University of Hannover, Germany and PhD from University of Klagenfurt, AUSTRIA. Presently he is working as a Professor in the Department of Information Technology, V R Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India.

A Comprehensive Analysis of Materialized Views in a Data Warehouse Environment

Garima Thakur
M.Tech (IT), Department of IT
USIT, Guru Gobind Singh Indraprastha University
Delhi, India
thakur_garima_27@yahoo.co.in

Anjana Gosain
Associate Professor, Department of IT
USIT, Guru Gobind Singh Indraprastha University
Delhi, India
anjana_gosain@hotmail.com

Abstract— **Data in a warehouse can be perceived as a collection of materialized views that are generated as per the user requirements specified in the queries being generated against the information contained in the warehouse. User requirements and constraints frequently change over time, which may evolve data and view definitions stored in a data warehouse dynamically. The current requirements are modified and some novel and innovative requirements are added in order to deal with the latest business scenarios. In fact, data preserved in a warehouse along with these materialized views must also be updated and maintained so that they can deal with the changes in data sources as well as the requirements stated by the users. Selection and maintenance of these views is one of the vital tasks in a data warehousing environment in order to provide optimal efficiency by reducing the query response time, query processing and maintenance costs as well. Another major issue related to materialized views is that whether these views should be recomputed for every change in the definition or base relations, or they should be adapted incrementally from existing views. In this paper, we have examined several ways o performing changes in materialized views their selection and maintenance in data warehousing environments. We have also provided a comprehensive study on research works of different authors on various parameters and presented the same in a tabular manner.**

Keywords- Materialized views; view maintenance; view selection; view adaptation; view synchronization.

I. INTRODUCTION

Data warehouse is referred as a subject-oriented, non-volatile & time variant centralized repository that preserves quality data [1]. A data warehouse extracts and integrates information from diverse operational systems prevailing in an organization under a unified schema and structure in order to facilitate reporting and trend analysis. Information sources which are integrated in the data warehouse are dynamic in nature i.e. they may transform or evolve in terms of their instances and schemas. Moreover, requirements specified by the various stakeholders and developers frequently change owing to numerous reasons as mentioned below [17] [18] 19]:

1. Ambiguous or insufficient requirements during the developmental phase [17].
2. Change in the requirements during the operational phase of the Data Warehouse which results in the structural evolution of the data warehouse [18].
3. Reorganization of the data warehouse schema during the operational phase of the data warehouse as a result of different design solutions that are decided upon [18].
4. New user or business requirements arise or new versions need to be created [18] [19].
5. Periodical revisions are made in order to eliminate the errors & redundancies [17][18].
6. The data warehouse must be adapted to any changes which occur in the underlying data sources [18] [19].

Hence, data warehouse and views present in warehouse must evolve whenever there is any modification or update in the requirements or base relations, in order to fulfill the needs and constraints allocated by the various users who need the assistance of data warehouse system. In fact, data warehouse evolution process never ceases. Appropriate techniques should be devised to handle the above mentioned changes in the data sources as well as view definitions to keep the warehouse in its most consistent state.

Whenever any user poses a query, the query is processed directly at this repository thereby, eliminating the need to access the actual source of information. The resulting datasets that are generated in the response to the queries raised by the users are called as *views*, which represent functions derived from the base relations to support viewing of snapshots of stored data by the users according to their requirements. These derived functions are recomputed every time the view is called upon. Re-computing and selection of views becomes impossible for each and every query especially; when the data warehouse is very large or the view is quite complex or query execution rate is high. Thus, we accumulate some pre-calculated results (or views) in our central repository (i.e. data warehouse) in order to provide faster access to data and enhance the query performance. This technique is referred as *materialization of views.*

Materialized views act as a data cache that gather information from distributed databases and support faster and reliable availability of already computed intermediate result sets (i.e. responses to queries). Data sources in current scenario are becoming quite vast and dynamic in nature i.e. they change rapidly. Consequently, frequency of deletion, addition and update operations on the base relations rises unexpectedly. Whenever the underlying base relation is modified the corresponding materialized view also evolves in reaction to those changes so that it can present quality data at the view level. Hence, we need certain techniques to deal with the problem of keeping a materialized view up-to date in order to propagate the changes from remote data source to the destined materialized view in the warehouse. These techniques can be broadly classified as- *view selection, view maintenance, view synchronization* and lastly, *view adaptation.* Each one of them is explained in more detail in the next section.

The layout of the paper is as follows. In section 2, we address the above mentioned techniques and also give a brief on the literatures being reviewed for the same. Section 3, presents a comparative study of the various research works explored in the previous section. Lastly, we conclude in section 4.

II. State of the Art

In this section, we describe the various techniques designed to handle the evolution of a materialized view in response to the modifications in data sources it originated from. In addition, we also discuss the literatures being reviewed in context of each and every technique.

The tasks involved in evolution of materialized views in a data warehouse can be categorized as follows:

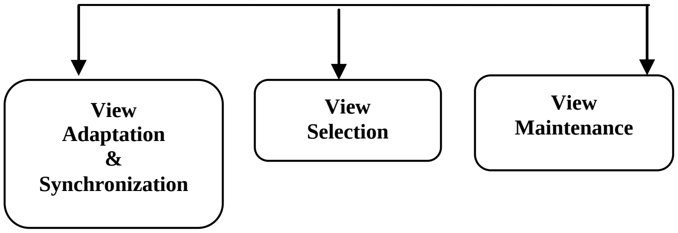


Figure 1. Tasks in materialized view evolution

A. *View Adaptation & Synchronization*

One of the factors that contribute to the changes in a materialized view is rewriting of views that leads to changes in the original view definition itself. This problem is addressed as *view adaptation*. Re-writing of view definitions generates the need to adapt the view schema to match it up with the most current view definition being referenced [2, 3]. View adaptation can be done either in incremental fashion or by performing full re-computation of the views [3]. If re-computation results in equivalent views then, there is no need to implement adaptation techniques because data is preserved. Non-equivalent definitions create new schema for the same view resulting in evolution of the original view. Some of the examples are listed below:

TABLE I. EXAMPLES OF SCHEMA CHANGES IN VIEW ADAPTATION

Schema changes	Description
Rename	Data preserving, no adaptation required.
Drop/Delete	Data deleted, hence non-equivalent views might be generated
Normalization	Schema structure and data preserved, hence no adaptation done.

In [2] the authors have provided a comprehensive study on various adaptation techniques. They have also provided re-definitions of all SQL clauses and views when local changes are made to view definitions. But they have only handled single materialized view changes.

In [13] author has discussed various view adaptation techniques where only the changes in view definitions cause adaptation in the views. Relation algebra binary operators can be added to SQL clauses to adapt the views. Expression trees are used to evaluate view definitions.

Another technique employed to handle materialized views is *view synchronization.* This technique changes the view definition when the structure of its base relations changes. It addresses both equivalent & non-equivalent view re-definitions [3]. Some of the changes that result in creation of new schema definitions are as follow [3]:

TABLE II. EXAMPLES OF CHANGES THAT RESULT IN SCHEMA CHANGES

Schema changes	Description
Rename	Renames the attributes and tables in the original view
Drop/Delete	Deleted attributes or tuples or tables in original views

EVE (Evolvable View Environment), a general framework has been developed in [4] to handle view synchronization in large distributed dynamic environments like- WWW. A view definition language, *E-SQL*, has also been designed along with some replacement strategies to propagate the changes in affected view components.

B. *View Selection*

The most important issue while designing a data warehouse is to identify and store the most appropriate set of materialized views in the warehouse so that they optimise two costs included in materialization of views: the query processing cost and materialized view maintenance cost.

Materialization of all possible views is not recommended due to memory space and time constraints [6]. The prime aim of view selection problem is to minimize either one of the constraints or a cost function as shown below:

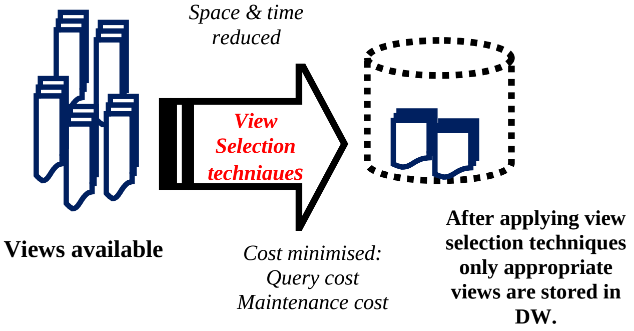

Space & time reduced
View Selection techniques
Views available
Cost minimised: Query cost Maintenance cost
After applying view selection techniques only appropriate views are stored in DW.

Figure 2. View Selection Process

Hence, view selection problem is formally defined as a process of identifying and selecting a group of materialized views that are most closely-associated to user-defined requirements in the form of queries in order to minimize the query response time, maintenance cost and query processing time under certain resource constraints [6].

In [5] authors have developed a AND/OR graph based approach to handle view selection problem in data cubes present in the data warehouse by taking an example of TPC-D benchmark database. They have also proposed an optimization algorithm to select certain views, but, this algorithm does not perform well in some of the cases.

Another graph based approach has been discussed in [6] in order to select a set of views for special cases under disk-space and maintenance cost constraints. AND view graphs have been discussed to evaluate the global plan for queries and OR view graphs focus on data cubes. They have proposed greedy heuristics based algorithms to handle the same. But still the approach has certain limitations like very little insight into the approximation of view-selection problem in AND/OR view graphs. Problem in AND view graphs is still not known to be NP-hard.

In [7] the authors have proposed two algorithms, one for view selection and maintenance and the second one for node selection for fast view selection in distributed environments. They have considered various parameters: query cost, maintenance cost, net benefit & storage space.

In [8] presented a framework for automatically selecting materialized views and indexes for SQL databases that has been implemented as a part of performance tuning in SQL Server 2000.

In [9] authors have presented a framework for selection of views to improve query performance under storage space constraints. It considers all the cost metrics in order to provide the optimal set of views to be stored in the warehouse. They have also proposed certain algorithms for selecting views based on their assigned weightage in the storage space and query.

In [14] a clustering based algorithm ASVMRT, based on clustering. Reduced tables are computed using clustering techniques and then materialized views are computed based on these reduced tables rather than original relations.

C. View Maintenance

Re-computation of materialized views is quite a wasteful task in data warehousing environments. Instead, we can only update a part of the views which are affected by the changes in the base relations. Hence, *View maintenance* incrementally updates a view by evaluating the changes to be incorporated in the view so that it can evolve over time. If views are maintained efficiently then, the overhead incurred while performing expensive joins and aggregations is eliminated to a larger extent.

In [10] authors have proposed a framework for dynamic environments called *DyDa*, for view maintenance in order to handle both concurrent schema and data changes. They have identified three types of anomalies and also proposed some dependency detection and correction algorithms to resolve any violation of inter-dependencies occurring between the maintenance processes.

An algorithmic approach has been implemented in [11] for incremental materialized view maintenance. The authors have employed the concept of version store so that the older versions of relations can be preserved and retrieval of correct data in the desired state is available round the clock. They have further proposed architecture to support of DW augmented with a View Manager.

In [12] authors have designed algebra based algorithm for incremental maintenance of views by schema restructuring. They have proposed a *SchemaSQL* language to handle data

updates and schema changes. Moreover, transformation operators have also been proposed to propagate data and schema changes easily.

View maintenance problem has been dealt in [3] by means of a compensation algorithm that eliminates interfering update anomalies encountered during incremental computations. Version numbers have been assigned to the updates occurring on the base relations to arrange them in a proper order. These numbers also help in detecting update notification messages that might be lost in the whole process of propagating the changes from source relation to views.

In [15] authors have presented PNUTS to handle asynchronous view maintenance in VLSD databases. The main approach is to defer expensive views by identifying RVTs & LVTs. PNUTS is also supported by a consistency model to hide details for replication of views. They have also listed the supported as well as unsupported views. Evaluation also reveals the performance of PNUTS on fault tolerance, throughput, complexity, query cost, maintenance, view staleness, latency, etc.

In [16] authors have discussed issues related to materialized views and their maintenance in Peer Data Management systems by using schema mappings (SPDMS). They have designed a hybrid peer architecture that consists of peers and super peers. Also, concepts of local, peer and global views have been developed to handle global view maintenance by handling peer vies in local PDMS, where, relations are numbered. Mapping rules guide the changes to map one version number to a new version. A push-based algorithm for view maintenance has been developed to handle view maintenance in a distributed manner.

III. COMPARATIVE STUDY

We have analyzed the various research works on several parameters and presented their comparison in the table below.

Table III. COMPARISON OF VARIOUS RESEARCH WORKS

Features Authors	Technique	Issues addressed	Changes handled	Proposed work	Query Language Supported	Meta Data supported	Advantages	Disadvantages	Tool support / implementation
Gupta, Mumick, Rao & Ross (2001) [2]	View adaptation	Re-materialization + In-place adaptations + Non-in place adaptations	Handling of local changes in view definition	Redefinitions of all SQL clauses + Guidelines for users & DBA	SQL based view definition language	Additional information kept with materialization	Query & cost optimization	Only single materialized view changes addressed	Not addressed
Mohania (1997) [13]	View Adaptation	Materialized view adaptation in distributed DW	Changes in view definition	Expression trees + Relational binary operators + Join & derive count values	SQL clauses: Select, From, Where	Additional results also materialized	Re-computation not needed + Cost of computing decreased	Overheads in maintain additional materialized results	Not addressed
Lee, Nica & Elke (2002) [4]	View synchronization	Synchronization in distributed dynamic environments	Schema changes Of data sources	EVE framework + replacement strategies + Algorithms	E-SQL view definition language	Meta Knowledge base	Handling changes in large dynamic environment (WWW) + A general framework	Only addressed schema changes in sources + No cost and quality issues addressed	JAVA + JDBC + MS-Access
Dhote & Ali (2007) [5]	View Selection	Selection of views to minimize query response time	Data cube changes	AND/OR DAG to minimize the query response time + Optimization algorithm	SQL based	✘	Heuristic based algorithm + Simple approach	Algorithm does not works well on certain cases + Cant be used on whole data cube (works only on lattice)	Not addressed
Gupta & Mumick (2005) [6]	View Selection	View selection under disk space & maintenance cost constraints.	Global evaluation plan for queries + Data cubes	AND/OR view graphs + Greedy heuristics based algorithms	SQL based	✘	Optimal solution for special cases (AND/OR views) + Polynomial time heuristics	Approximation in view-selection problem not addressed + Problem in AND view graphs not NP-hard + Solution fairly close to optimum	Not mentioned
Karde & Thakare (2010)	View Selection	Query cost, maintenance cost, storage space &	In distributed environments	Algorithm for creation and maintenance of views	Not mentioned	✘	Query performance improved	Only distributed environments highlighted	Not addressed

[7]				+ Algorithm for node selection					
Agrawal, Chaudhari & Narasayya (2000) [8]	View Selection	Automated view and index selection	✘	Framework for index & view selection + Candidate selection & enumeration techniques	SQL based	✘	Robust tool support + Both indexes & view selected	Only a part of physical design space addressed	SQL Server 2000
Ashadevi & Balasubramanian [9]	View selection	Cost-effective view selection under storage space constraints	✘	Framework for selecting views + Algorithm for the same + Cost metrics	Not addressed	✘	All cost metrics considered	Query response time not considered + Threshold value not indicated clearly	Algorithms implemented in JAVA
Yang & Chung (2006) [14]	View selection	Attribute-value density + Clustered tables + Selection of views based on clustered /reduced tables	Related dimensions or relations	ASVMRT algorithm for view selction	SQL based	✘	Faster computation time + Reduced storage space + 1.8 times performance better than conventional algorithms	Maintenance of reduced tables not addressed + Updating Reduced tables needs attention	In pubs database + ETRI
Chen, Zhang & Elke (2006) [3]	View maintenance	Source Data updates + Preserving & non-preserving schema changes + 3 types of anomalies	Source schema & data updates	DyDa Framework + Dependency & Correction algorithms	SQL based maintenance & compensation queries	✘	Can handle concurrent & interleaved data and schema changes	Extra cost on data updates + Cannot maintain mixed updates in single process	JAVA & Oracle 8i
Almazyad & Siddiqui (2010) [10]	View maintenance	Incremental view maintenance + synchronization between DW and source + lost update notifications	Source relation changes	Framework with version store	Not mentioned clearly	Version store provides needed metadata	Synchronization between source and DW + Detection of update notification messages	Process becomes a bit lengthy + more space needed + Version numbers should be handled properly	Not addressed
Koeller & Rundensteiner (2004) [11]	View maintenance	Schema restructuring of views	Data + Schema changes	Algebra based maintenance + transformation operators	Schema SQL queries	Not clearly mentioned	Algebra-based can be adapted to other query languages easily	Time consuming process	JAVA + Oracle 8 (JDBC)
Ling & Sze (2001) [12]	View maintenance	Update anomalies + Notification messages	Modifications in base relations	Compensating algorithms + Version numbers	Not addressed	Present in form of log files + version numbers	Algorithms does not require quiescent state before views can be refreshed	Time consuming + Version numbers should be handled	Not addressed

							+ Update notifications handled efficiently + version numbers reflect state of relations	properly	
Agrawal, Silberstein, Cooper, Srivastava & Ramakrishnan (2009) [15]	View maintenance	Asynchronous view (Remote view Tables RVT, Local View Tables LVT) + Replication of views	Deferred Indexes & Views in Very Large Scale Distributed databases, horizontally partitioned	PNUTS + consistency model + RVTs &LVTs + Group-by views, select views, indexes & equi-join views	SQL based	Metadata for view definitions & partitions	Improved client latency + Improved scalability + Balanced view staleness, system complexity + Improved query cost	Views maintenance adds load to system + VLSD are complex + Decreased throughput + Complex failure recovery	C++ + FreeBSD 6.3 (Linux can also be used)
Qin, Wang & Du (2005) [16]	View Maintenance	Global view maintenance by maintaining local views in PDMS	Schema changes + Schema mappings amongst peers	Hybrid peer architecture (P2P & Peer-super peer) + Local, global & peer views + Rules to use updategrams & boosters + Push-based Algorithm	Not mentioned	One kind of Super peer maintains metadata for mapping schemas in intre-peers or inter-peer changes in local PDMS	Decentralised maintenance strategy + Higher efficiency + Parallelism + Efficient in 80-20 distribution + Central bottlenecks avoided	Information sharing is complex & difficult in PDMS + Querying not addressed	Simulation system developed in JAVA

IV. Conclusion

In this paper we have presented an analysis of different approaches being proposed by various researchers to deal with the materialized views in data warehouse namely- view adaptation & synchronization, view selection and view maintenance. We have examined these techniques on various parameters and provided a comparative study in a tabular manner.

V. Future Work

As future work, we will direct our research towards batch-oriented view maintenance and selection strategies. A thorough investigation of the methodologies to handle materialized views in highly distributed environments for query processing and analysis seems worth attention.

References

[1] W. Inmon, “Building the data warehouse”, Wiley publications, pp 23, 1991.

[2] A. Gupta, I. Mumick, J. Run, and K. Ross, “Adapting materialized views after redefinitions: techniques and a performance study”, In Elsevier Science Ltd., pp 323-362, 2001.

[3] S. Chen, X. Zhang, and E. Rundensteiner, “A compensation based approach for view maintenance in distributed environments”, In IEEE transactions and data engineering, 18, 2006.

[4] A. Lee, A. Nica, and E. Rundensteiner, “The EVE approach view synchronization in dynamic distributed environments”, In IEEE Transactions and Data Engineering, 14, 2002.

[5] C. Dhote, and M. Ali, “Materialized view selection in data warehouse”, In International Conference on Information Technology, 2007.

[6] A. Gupta, and I. Mumick, “Selection of views to materialize in a data warehouse”, In IEEE Transactions on Knowledge and Data Engineering, vol. 17, 2005.

[7] P. Karde, and V. Thakare,” Selection of materialized views using query optimization in database management: An efficient methodology”, In International Journal of Management Systems, vol. 2.

[8] S. Agrawal, S. Chaudhari, and V. Narasayya, “Automated selection of materialized views and indexes for SQL databases”, In Proceedings of 26th International Conference on Very Large Databases, 2000.

[9] B. Ashadevi, and R. Balasubramanian, “Cost effective approach for materialized views selection in data warehouse environment”, In International Journal of Computer Science and Network Security, vol. 8, 2008.

[10] A. Almazyad, and M. Siddiqui, “Incremental view maintenance: an algorithmic approach”, In International Journal of Electrical & Computer Sciences, vol. 10, 2010.

[11] A. Koeller, and A. Rundensteiner, “Incremental maintenance of schema–restructuring view in SchemaSQL”, In IEEE Transactions and Data Engineering, 16, 2004.

[12] T. Ling, and E. Sze,” Materialized view maintenance using version numbers”, In Proceeding of Springer Berlin, 2001.

[13] M. Mohania, “Avoiding re-computation: View adaptation in data warehouses”, in australian research council, 1997.

[14] J. Yang, and I. Chung, “ASVMRT: Materialized view selection algorithm in data warehouse”, In International Journal of Information Processing System, 2006.

[15] P. Agrawal, A. Silberstein, B. Cooper, U. Srivastava, and R. Ramakrishnan, “Asynchronous view maintenance in vlsd databases”, In SIGMOD International Conference on Management of Data, 2009.

[16] B. Qin, S. Wang, and X. Du, “Effective maintenance of materialized views in peer data management systems”, In Proceedings of First International Conference on Semantics, Knowledge and Grid, 2005.

[17] D. Sahpaski, G. VelInov, B. Jakimovski, and M. Kon-Popovska, “Dynamic evolution and improvement of data warehouse design”, In Balkan Conference in Informatics, 2009.

[18] B. Bebel, J. Eder, C. Koncilia, T. Morzy, and R. Wrembel, “Creation and management of versions in multiversion data warehouse”, In Proc. ACM SAC, 717–723, 2004.

[19] B. Bebel, J. Eder, C. Koncilia, T. Morzy, and R. Wrembel, “Formal approach to modeling data warehouse”, In bulletin of the Polish Academy of Sciences, 54, 1, 2006.

[20] Vashishta, S. (2011). Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm. International Journal of Advanced Computer Science and Applications - IJACSA, 2(4), 77-80.

[21] Hanandi, M., & Grimaldi, M. (2010). Organizational and collaborative knowledge management : a Virtual HRD model based on Web2 . 0. International Journal of Advanced Computer Science and Applications - IJACSA, 1(4), 11-19.

AUTHORS PROFILE

Garima Thakur is pursuing her M.Tech in Information Technology from Guru Gobind Singh Indraprastha University, Delhi, India. She has done her B.Tech in Computer Science branch from the same university in the year 2009. She is doing her research work in the field of Data warehouse & Data Mining and Knowledge Discovery.

Dr. (Mrs.) Anjana Gosain is working as reader in University school of information technology. She obtained her Ph.D. from GGS Indraprastha University & M.Tech in Information Systems from Netaji Subhas Institute of Technology (NSIT) Delhi. Prior to joining the school, she has worked with computer science department of Y.M.C.A institute of Engineering, Faridabad (1994-2002). She has also worked with REC kurukshetra. Her technical and research interests include data warehouse, requirements engineering, databases, software engineering, object orientation and conceptual modeling. She has published 18 research papers in International / National journals and conferences.

A Routing Scheme for a New Irregular Baseline Multistage Interconnection Network

Mamta Ghai
Department of Computer Science & Engineering
Punjabi University
Patiala(India)

Abstract— **Parallel processing is an efficient form of information processing system, which emphasizes the exploitation of concurrent events in the computing process. To achieve parallel processing it's required to develop more capable and cost-effective systems. In order to operate more efficiently a network is required to handle large amount of traffic. Multi-stage Interconnection Network plays a vital role on the performance of these multiprocessor systems. In this paper an attempt has been made to analyze the characteristics of new class of Irregular Fault-Tolerant Multistage Interconnection network named as Irregular Modified Baseline Multistage Interconnection network IMABN and an efficient routing procedure has been defined to study the fault-tolerance of the network. Fault-Tolerance in an interconnection network is very important for its continuous operation over a relatively long period of time. Fault-Tolerance is an ability of the network to operate in presence of multiple faults. The behavior of the Proposed IMABN has been analyzed and compared with regular network MABN under fault free conditions and in presence of faults. In IMABN there are six possible paths between source and destinations whereas MABN has only four. Thus the proposed IMABN is more Fault-tolerant than existing regular Modified Augmented Baseline multistage interconnection network (MABN).**

***Keywords-* Multistage Interconnection network; Fault-Tolerance; Augmented Baseline Network.**

I. INTRODUCTION

With the present state of technology building multiprocessor system with hundreds of processors is feasible. A vital component of these systems is the interconnection network(IN) that enables the processors to communicate among themselves or with the memory units. Multipath nature of multistage interconnection networks become more popular. Many ways of providing fault-tolerance to multistage interconnection networks(MINs) have been proposed. The basic idea for fault-tolerance is to provide multiple paths between source-destination pair so that alternate paths can been used in case of faults. Sufficient work has been done on the regular type of MINs,but little attention has been paid to the irregular type of MIN.

In this paper, a new class of irregular Baseline multistage interconnection network named as irregular modified augmented baseline network(IMABN) is proposed. In this paper we present methods of increasing fault-tolerance of an network by introducing the extra stage. Hence with the additional stage more paths available between each source and destination, as compared to existing network MABN. The proposed Irregular Modified Augmented Baseline Network(IMABN) is an Modified augmented baseline network(MABN) with additional stage. In an IMABN, there are Six possible paths between any source-destination pair, whereas MABN has only Four. Fault-Tolerance in an Interconnection network is very important for its continuous operation over a relatively long period of time. It is the ability of the network to operate in presence of multiple faults, although at a degraded performance. There are many ways to increase the fault-tolerance of the network.

(1) Increasing the no of stages.
(2) Providing multiple links between stages.
(3) Increasing size of switch.
(4) Incorporating multiple copies of a basic network.

This paper has been organized into five sections whose details are as follows:

Section I introduces the subject under study. Section II describe the structure and design of the networks. Section III focus on the routing tags. Section IV describe the routing procedure for the proposed IMABN.Finally, some concluding remarks are given in section V.

II. STRUCTURE OF NETWORKS

A. MABN (Modified Augmented Baseline Network)

To construct an MABN of size N, two identical groups of N/2 sources and N/2 destinations need to be formed first. Each source is linked to both the groups via multiplexers. There is one 4 x 1 MUX for each input link of a switch in stage 1 and one 1 x 4 DEMUX for each output link of a switch in stage n-2. MABN consists of two identical sub-networks which are denoted by G^i. Switches A, B, C, D belonging to stage 1 of a sub-network (G^i) form a conjugate subset,switches A and B form a conjugate pair, and switches A and C form a conjugate loop. An MABN of size 16X16 is shown in Figure 1.

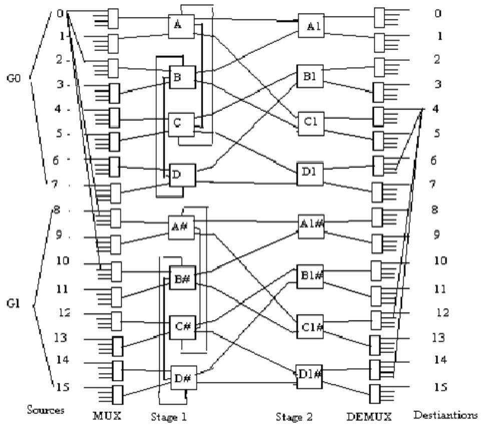


Figure 1: An MABN of size 16 X 16.

B. IMABN(Irregular MABN)

IMABN (Irregular modified augmented baseline network) is a Modied Augmented Baseline Network with one additional stage, increase size of switch and more auxiliary links. To construct an IMABN of size N i.e. N sources and N destinations, two identical groups of N/2 sources and N/2 destinations need to be formed first. Each source is linked to both the groups via multiplexers. There is one 4 x 1 MUX for each input link of a switch in stage 1 and one 1 x 4 DEMUX for each output link of a switch in stage n-1. IMABN consists of two identical sub-networks which are denoted by G^i. For example, in Figure 2, switches A, B, C, D belonging to stage 1 of a subnetwork (G^i) form a conjugate subset, switches A and B form a conjugate pair, and switches A and C form a conjugate loop.Thus an IMABN of size N consists of N number of 4 x 1 MUXs, N number of 1 x 4 DEMUXs, and switches in the last stage of size 2 x 2, switches in the middle stage of size 5 x 5 and switches in the first stage of size 3 x 3. IMABN as its name suggest is an irregular network in which middle (additional) stage doesn't have equal number of switches as of other stages. The irregular topology of IMABN varies the number of switching elements encountered in the way of data transfer through an input-output pair depending on the path chosen, which makes the average rate of failure of the network less as compared to that of regular ABN and MABN.IMABN is a dynamically re-routable irregular MIN and provides multiple paths of varying lengths between a source-destination pair.

Observe that this construction procedure has two benefits. First, the network can tolerate the failure of any switch in the network. And, secondly it provides a topology which lends itself to on-line repair and maintainability, as a loop can be removed from any stage of the IMABN without disrupting the operation of the network. Since the sub-networks are identical, so the VLSI implementation of the network becomes simple. IMABN of size 16x16 is shown in Fig 2.

III. ROUTING TAGS

A source selects a particular subnetwork (G^i) based upon the most significant bit of the destination. each source is connected to two switches (primary and secondary) in a subnetwork.

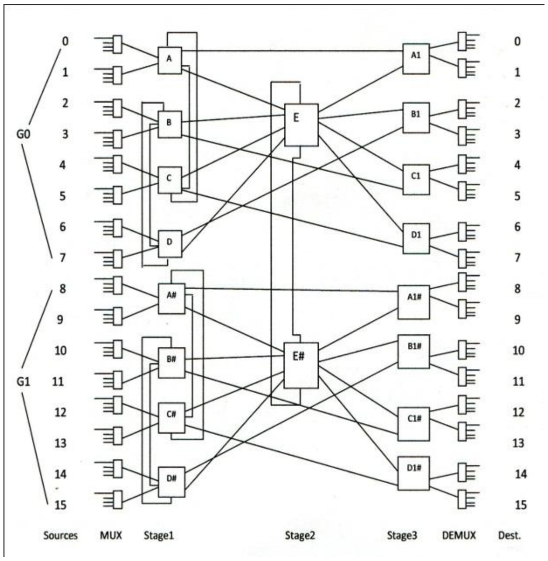


Figure 2: An IMABN of size 16 X 16

Let the source S and destination D be represented in binary code as:

$$S=s_0,\ s_1,\ldots,s_{n-2},s_{n-1}$$
$$D=d_0,\ d_1,\ldots,d_{n-2},d_{n-1}$$

(i) Source S is connected to the ($s_1,\ldots,s_{n-2}$) primary switch in both the sub-networks through the multiplexers.

(ii) Source S is also connected to the $[\{(s_1,\ldots,s_{n-2})+1\}\text{mod } N/4]$ secondary switch in both the sub-networks through the multiplexers.

IV. ROUTING PROCEDURE

Routing procedure for a network tells that how a request from any source S travels through the various intermediate stages and reaches to a given destination D. Following assumptions have been made for the algorithm.

- Each source destination pair tries to utilize only one path at a time.
- Source & switches have the ability to detect faults in the switches to which they are connected.

A. Routing Scheme for IMABN(Irregular MABN)

A request from any source S to a given destination D is routed through the IMABN as:

1) *For each source:* The source S selects one of the sub-network G^i based on the most significant bit of the destination D ($i=d_0$). Each source attempts entry into the IMABN via its primary path. If the primary path is faulty, then the request is routed to secondary path. If the secondary path is also faulty then the request is routed to the other subnetwork of the IMABN, via auxiliary links of stage 2. If still request doesn't get matured, then the request is rerouted to the secondary sub-network, in which same routing is followed as in the case of primary sub-network.

2) *For each switch in stage n - 3:* After the MUX, the routing of the request in the first (n-3) stage of the sub-

network depends upon one tag bit, which depends on d_1d_2 destination address bits. Routing tag bit for stage 1 is calculated as follows :

If $d_1d_2 = 00$,

then both conjugate pairs in the sub-network will have tag bit = 0.

If $d_1d_2 = 01$,

then first conjugate pair(A/A#, B/B#) will have tag bit = 1,and Second conjugate pair(C/C#, D/D#) will have tag bit = 0.

If $d_1d_2 = 10$,

then both conjugate pairs in the sub-network will have tag bit = 1.

If $d_1d_2 = 11$,

then first conjugate pair(A/A#, B/B#) will have tag bit = 0,and Second conjugate pair(C/C#, D/D#) will have tag bit = 1.

3) For each switch in stage n - 2: For a request at a switch in stage n-2, value of tag bit is given as below:

If $d_1d_2 = 00$,

then at E ,if request comes from(A/A#,B/B#) then it goes to Node A1(use tag bit 00) and if request comes from (C/C#,D/D#) then it goes to node B1(use tag bit 01).

If $d_1d_2 = 01$,

then at E ,if request comes from(A/A#,B/B#) then it goes to Node C1(use tag bit 10) and if request comes from (C/C#,D/D#) then it goes to node B1(use tag bit 01)

If $d_1d_2 = 10$,

then at E ,if request comes from(A/A#,B/B#) then it goes to Node C1(use tag bit 10) and if request comes from (C/C#,D/D#) then it goes to node D1(use tag bit 11)

If $d_1d_2 = 11$,

then at E ,if request comes from(A/A#,B/B#) then it goes to Node A1(use tag bit 00) and if request comes from (C/C#,D/D#) then it goes to node D1(use tag bit 11)

4) For each switch in stage n - 1: For a request at a switch in stage n-1, use bit d_{n-1} of the routing tag and route the request accordingly to one of the output links. If the required output link is busy, then repeat step two and three in the secondary sub-network. If again the required output link is busy in stage n-2, then drop the request.

5) For each demultiplexer at the output of stage n - 1: For routing a request through a DEMUX, following concept is used.

If destination and MUX are in same sub-network, then 1st MUX uses output line 00 and 2nd MUX uses output line 10.

If destination and MUX are in different sub-networks, then 1st MUX uses output line 01 and 2nd MUX uses output line 11.

Multiple paths between S=0000 and D = 0100 of an IMABN are shown below.

Case 1:Routing in IMABN

Primary Path:

0-> MUX(0) – A – E – C1 – DEMUX(4) – 4
0-> MUX(0) – A – C – D1 – DEMUX(6)-4

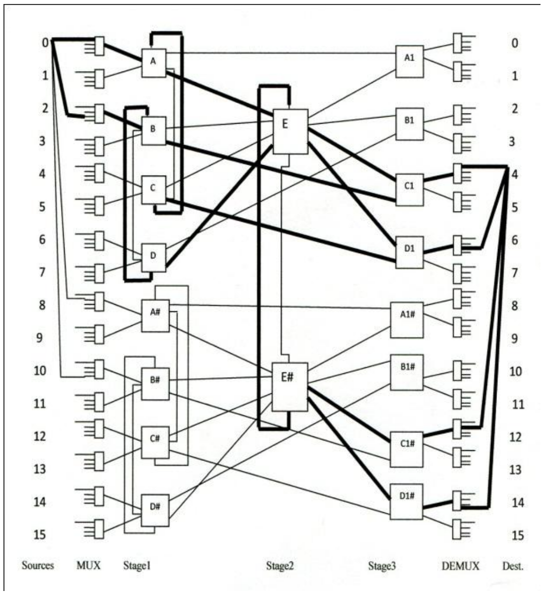


Figure 3: Routing in IMABN

Primary path using auxiliary links of stage 2:-

0-> MUX(0) – A – E – E# - C1# - DEMUX(12) – 4

Secondary path:

0-> MUX(2) – B – C1 – DEMUX(4) – 4

0-> MUX(2) – B – D – E – D1 – DEMUX(6) – 4

Secondary path using auxiliary links of stage 2:-

0-> MUX(2) – B – D – E – E# - D1# - DEMUX(14) – 4

Case2: Routing in IMABN

Primary# Path:

0->MUX(8) – A# - E# - C1# - DEMUX(12) – 4

0->MUX(8) – A# - C# - D1# - DEMUX(14) – 4

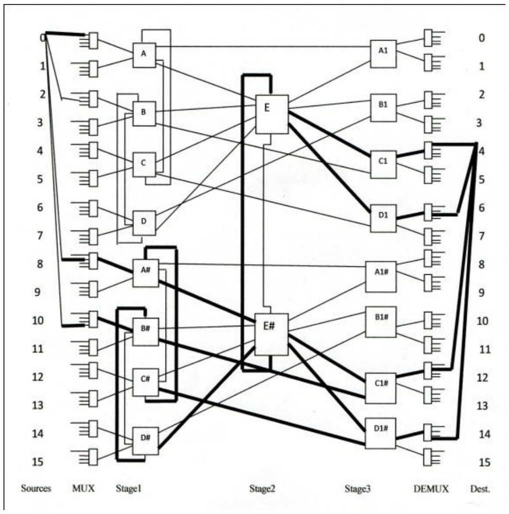


Figure 4: Routing in IMABN

Primary # path using auxiliary links of stage2:-

0->MUX(8) – A# - E# - E –C1 - DEMUX(4) – 4

Secondary # path:

0->MUX(10) – B# - C1# - DEMUX(12)
0->MUX(10) –B# - D# - E# - D1# - DEMUX(14) – 4

Secondary # path using auxiliary links of stage2:-

0->MUX(10) – B# -D# - E# - E – D1 - DEMUX(6) – 4

V. Conclusions

An Irregular Modified Augmented Baseline Network (IMABN) is designed from regular Modified Augmented Baseline network (MABN).IMABN is dynamically re-routable and providing multiple paths of varying lengths between source and destination pairs. It has been found that IMABN has six possible paths whereas in MABN has only four. Thus IMABN is more fault-tolerant than MABN.

References

[1] Bhuyan Laxmi N., Yang Qing and Aggarwal P. Dharma, "*Performance of Multiprocessor Interconnection Networks*", Proceeding of IEEE, February 1989, pp. 25-37.

[2] Bhogavilli Suresh K. and Abu-Amara Hosame, "*Design and Analysis of High Performance Multistage Interconnection Networks*", IEEE Transactions on Computers, vol. 46, no. 1, January 1997, pp. 110 -117.

[3] Bansal P.K, Singh Kuldeep and Joshi R.C, " *On Fault tolerant Multistage Interconnection Network*", Conference on Computer Electrical Engineering, vol. 20, no.4, 1994, pp. 335-345.

[4] Blaket James T. and Trivedi Kishor S., "*Reliabilities of Two Fault-Tolerant Interconnection Networks*", Proceeding of IEEE, 1988, pp. 300-305.

[5] Charles Chenggong Charles and Bruck Jehoshua, "*Tolerating Multiple Faults in Multistage Interconnection Networks with Minimal Extra Stages*", IEEE Transactions on Computers, vol. 49, no. 9, September 2000, pp. 998-1004.

[6] Cheema Karamjit, Aggarwal Rinkle, " *Design Scheme and Performance Evaluation of new Fault-Tolerant Multistage Interconnection Network",* IJCSNS International Jounal of Computer Science and Network Security , VOL.9 NO.9,September 2009.

[7] Mamta Ghai et al., " Performance Analysis of Fault-Tolerant Irregular baseline multistage interconnection network" , IJCSE International journal on computer science & engineering Vol. 02, No. 09, 2010, 3079-3084

[8] Mamta Ghai et al., " Design and Reliability Analysis of New Fault-Tolerant Irregular Multistage Interconnection network", IJCA International Journal of Computer Applications(0975-8887) Volume 10, No 1, November 2010.

[9] Nitin, "*On Analytic Bounds of Regular and Irregular Fault-tolerant Multi-stage Interconnection Networks",* Proceedings of International Conference, 2006.

[10] Sadawarti Harsh and Bansal P.K., " *Fault Tolerant Irregular Augmented Shuffle Network*", Proceeding of the 2007 WSEAS International Conference on Computer Engineering and Applications, Australia, January 17-19,2007. pp. 7-12.

[11] Sengupta J. and Bansal P.K, "*Performance of Regular and Irregular Dynamic MINs*", Proceeding of International Conference IEEE TENCON, 1999, pp. 427-430.

[12] Sharma Sandeep and Bansal P.K., "*A New Fault Tolerant Multistage Interconnection Network*", Proceeding of International Conference IEEE TENCON, 2002, pp. 347-350.

[13] Rao, D. S. (2010). Performance Evaluation of Node Failure Prediction QoS Routing Protocol (NFPQR) in Ad Hoc Networks. International Journal of Advanced Computer Science and Applications - IJACSA, 1(6), 54-59.

[14] A Survey on Attacks and Defense Metrics of Routing Mechanism in Mobile Ad hoc Networks. International Journal of Advanced Computer Science and Applications - IJACSA, 2(3), 7-12.

[15] Beebi, P. K. A. (2011). A Study on Cross Layer MAC design for performance optimization of routing protocols in MANETs. International Journal of Advanced Computer Science and Applications - IJACSA, 2(2), 11-16.

[16] Suri, P. K. (2011). Simulation of Packet Telephony in Mobile Adhoc Networks Using Network Simulator. International Journal of Advanced Computer Science and Applications - IJACSA, 2(1), 87-92.

Application of Fuzzy Logic Approach to Software Effort Estimation

Prasad Reddy P.V.G.D
Dept. of CSSE
Andhra University
Visakhapatnam, INDIA

Sudha K. R
Dept. of EE
Andhra University
Visakhapatnam, INDIA

Rama Sree P
Dept. of CSSE
Aditya Engineering College, JNTUK
Kakinada, INDIA

Abstract— **The most significant activity in software project management is Software development effort prediction. The literature shows several algorithmic cost estimation models such as Boehm's COCOMO, Albrecht's' Function Point Analysis, Putnam's SLIM, ESTIMACS etc., but each model do have their own pros and cons in estimating development cost and effort. This is because project data, available in the initial stages of project is often incomplete, inconsistent, uncertain and unclear. The need for accurate effort prediction in software project management is an ongoing challenge. A fuzzy model is more apt when the systems are not suitable for analysis by conventional approach or when the available data is uncertain, inaccurate or vague. Fuzzy logic is a convenient way to map an input space to an output space. Fuzzy Logic is based on fuzzy set theory. A fuzzy set is a set without a crisp, clearly defined boundary. It is characterized by a membership function, which associates with each point in the fuzzy set a real number in the interval [0, 1], called degree or grade of membership. The membership functions may be Triangular, GBell, Gauss and Trapezoidal etc. In the present paper, software development effort prediction using Fuzzy Triangular Membership Function and GBell Membership Function is implemented and compared with COCOMO. A case study based on the NASA93 dataset compares the proposed fuzzy model with the Intermediate COCOMO. The results were analyzed using different criterions like VAF, MARE, VARE, MMRE, Prediction and Mean BRE. It is observed that the Fuzzy Logic Model using Triangular Membership Function provided better results than the other models.**

Keywords- ***Development Effort; EAF; Cost Drivers; Fuzzy Identification; Membership Functions; Fuzzy Rules; NASA93 dataset.***

I. INTRODUCTION

In algorithmic cost estimation [1], costs and efforts are predicted using mathematical formulae. The formulae are derived based on some historical data [2,19]. The best known algorithmic cost model called COCOMO (COnstructive COst MOdel) was published by Barry Boehm in 1981[3]. It was developed from the analysis of sixty three (63) software projects. Boehm projected three levels of the model called Basic COCOMO, Intermediate COCOMO and Detailed COCOMO [3, 5]. In the present paper we mainly focus on the Intermediate COCOMO.

A. Intermediate COCOMO

The Basic COCOMO model [3] is based on the relationship: Development Effort, DE = a*(SIZE)^b; where, SIZE is measured in thousand delivered source instructions. The constants a, b are dependent upon the 'mode' of development of projects. DE is measured in man-months. Boehm proposed 3 modes of projects [3]:

a) Organic mode – simple projects that engage small teams working in known and stable environments.

b) Semi-detached mode – projects that engage teams with a mixture of experience. It is in between organic and embedded modes.

c) Embedded mode – complex projects that are developed under tight constraints with changing requirements.

The accuracy of Basic COCOMO is limited because it does not consider the factors like hardware, personnel, use of modern tools and other attributes that affect the project cost. Further, Boehm proposed the Intermediate COCOMO[3,4] that adds accuracy to the Basic COCOMO by multiplying 'Cost Drivers' into the equation with a new variable: EAF (Effort Adjustment Factor) shown in Table I.

TABLE I. DE FOR THE INTERMEDIATE COCOMO

Development mode	Intermediate Effort Equation
Organic	DE = EAF * 3.2 * (SIZE)$^{1.05}$
Semi-detached	DE = EAF * 3.0 * (SIZE)$^{1.12}$
Embedded	DE = EAF * 2.8 * (SIZE)$^{1.2}$

The EAF term is the product of 15 Cost Drivers [5, 11] that are listed in Table II. The multipliers of the cost drivers are Very Low, Low, Nominal, High, Very High and Extra High. For example, for a project, if RELY is Low, DATA is High , CPLX is extra high, TIME is Very High, STOR is High and rest parameters are nominal then EAF = 0.75 * 1.08 * 1.65 *1.30*1.06 *1.0. If the category values of all the 15 cost drivers are "Nominal", then EAF is equal to 1.

TABLE II. INTERMEDIATE COCOMO COST DRIVERS WITH MULTIPLIERS

S. No	Cost Driver Symbol	Very low	Low	Nominal	High	Very high	Extra high
1	RELY	0.75	0.88	1.00	1.15	1.40	—

2	DATA	—	0.94	1.00	1.08	1.16	—
3	CPLX	0.70	0.85	1.00	1.15	1.30	1.65
4	TIME	—	—	1.00	1.11	1.30	1.66
5	STOR	—	—	1.00	1.06	1.21	1.56
6	VIRT	—	0.87	1.00	1.15	1.30	—
7	TURN	—	0.87	1.00	1.07	1.15	—
8	ACAP	—	0.87	1.00	1.07	1.15	—
9	AEXP	1.29	1.13	1.00	0.91	0.82	—
10	PCAP	1.42	1.17	1.00	0.86	0.70	—
11	VEXP	1.21	1.10	1.00	0.90	—	—
12	LEXP	1.14	1.07	1.00	0.95	—	—
13	MODP	1.24	1.10	1.00	0.91	0.82	—
14	TOOL	1.24	1.10	1.00	0.91	0.83	—
15	SCED	1.23	1.08	1.00	1.04	1.10	—

The 15 cost drivers are broadly classified into 4 categories [3, 5].

a) *Product*: RELY - Required software reliability
DATA - Data base size
CPLX - Product complexity

b) *Platform*: TIME - Execution time
STOR- Main storage constraint
VIRT - Virtual machine volatility
TURN - Computer turnaround time

c) *Personnel*: ACAP - Analyst capability
AEXP - Applications experience
PCAP - Programmer capability
VEXP - Virtual machine experience
LEXP - Language experience

d) *Project*: MODP - Modern programming
TOOL - Use of software tools
SCED - Required development schedule

Depending on the projects, multipliers of the cost drivers will vary and thereby the EAF may be greater than or less than 1, thus affecting the Effort [5].

II. Fuzzy Identification

A fuzzy model [13,15] is used when the systems are not suitable for analysis by conventional approach or when the available data is uncertain, inaccurate or vague [7]. The point of Fuzzy logic is to map an input space to an output space using a list of if-then statements called rules. All rules are evaluated in parallel, and the order of the rules is unimportant. For writing the rules, the inputs and outputs of the system are to be identified. To obtain [18] a fuzzy model from the data available, the steps to be followed are,

- Select a Sugeno type Fuzzy Inference System.
- Define the input variables mode, size and output variable effort.
- Set the type of the membership functions (TMF or GBellMF) for input variables.
- Set the type of the membership function as linear for output variable.
- The data is now translated into a set of if–then rules written in Rule editor.
- A certain model structure is created, and parameters of input and output variables can be tuned to get the desired output.

A. *Fuzzy Approach for Prediction of Effort*

The Intermediate COCOMO model data is used for developing the Fuzzy Inference System (FIS)[10]. The inputs to this system are MODE and SIZE. The output is Fuzzy Nominal Effort. The framework [8] is shown in "Fig. 1".

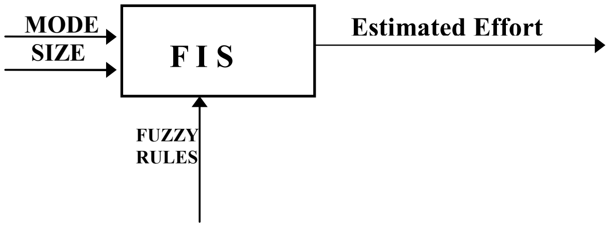


Figure 1: Fuzzy Framework

Fuzzy approach [9] specifies the SIZE of a project as a range of possible values rather than a specific number. The MODE of development is specified as a fuzzy range .The advantage of using the fuzzy ranges[14] is that we will be able to predict the effort for projects that do not come under a precise mode i.e. comes in between 2 modes. This situation cannot be handled using the COCOMO. The output of this FIS is the Fuzzy Nominal Effort. The Fuzzy Nominal Effort multiplied by the EAF gives the Estimated Effort. The FIS[16] needs appropriate membership functions and rules.

B. *Fuzzy Membership Functions*

A membership function (MF) [9, 12] is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. The input space is also called as the universe of discourse. For our problem, we have used 2 types of membership functions:

1) *Triangular membership function*
2) *Generalized Bell membership function*

Triangular membership function (TMF):

It is a three-point function [17], defined by minimum (α),Maximum (β) and modal (m) values, that is, TMF (α, m, β), where (α ≤ m ≤β). The parameters α and β locate the "feet" of the triangle and the parameter m locates the peak. The triangular curve is a function of a vector, x, and depends on three scalar parameters α, m, and β as given by

$$f(x;\alpha,m,\beta) = \max\left(\min\left(\frac{x-\alpha}{m-\alpha},\frac{\beta-x}{\beta-m}\right),0\right) \quad (1)$$

Please refer to "Fig. 2" for a sample triangular membership function.

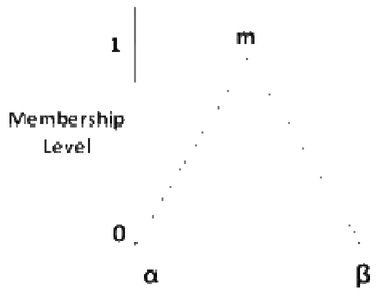


Figure 2: A Sample Triangular Membership Function

The fuzzy set definitions for the MODE of development appear in "Fig. 3" and the fuzzy set [8] for SIZE appear in "Fig. 4".

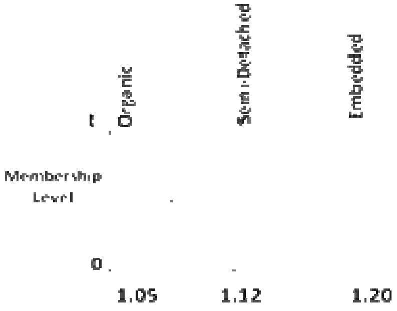


Figure 3: Fuzzy Set for Mode

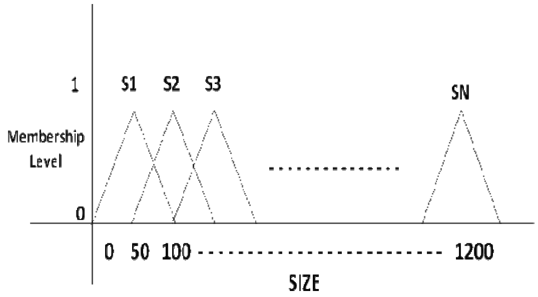


Figure 4: Fuzzy set for SIZE

Generalized Bell Membership Function (GBellMF):

It is a three-point function, defined by minimum (α), maximum (β) and modal (m) values, that is, GBellMF(α, m, β), where (α ≤ m ≤β). Please refer to "Fig. 5" for a sample Generalized Bell membership function. The generalized bell function depends on three parameters α, m, and β as given by

$$f(x;\alpha,m,\beta)=\frac{1}{1+\left|\frac{x-\beta}{\alpha}\right|^{2m}} \quad (2)$$

where the parameter m is usually positive. The parameter β locates the center of the curve.

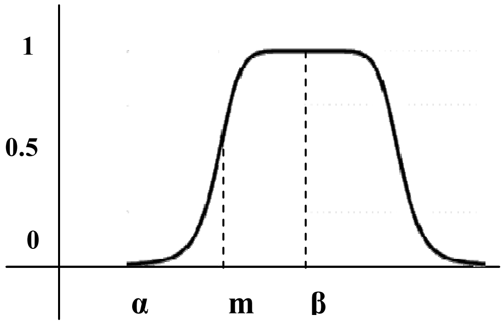


Figure 5: A Sample Generalized Bell Membership Function

We can get the Fuzzy sets for MODE, SIZE and Effort for GBellMF in the same way as in triangular method, but the difference is only in the shape of the curves.

C. Fuzzy Rules

Our rules based on the fuzzy sets [9, 20] of MODE, SIZE and EFFORT appears in the following form:

If MODE is organic and SIZE is s1 then EFFORT is EF1
If MODE is semidetached and SIZE is s1 then EFFORT is EF2
If MODE is embedded and SIZE is s1 then EFFORT is EF3
If MODE is organic and SIZE is s2 then EFFORT is EF4
If MODE is semidetached and SIZE is s2 then EFFORT is EF5
If MODE is embedded and SIZE is s3 then EFFORT is EF5
If MODE is embedded and SIZE is s4 then EFFORT is EF3
If MODE is organic and SIZE is s3 then EFFORT is EF4
If MODE is embedded and SIZE is s5 then EFFORT is EF6
If MODE is organic and SIZE is s4 then EFFORT is EF4
......

III. Various Criterions for Assessment of Software Effort

a) Variance Accounted For (VAF)

$$\text{VAF (\%)} = \left(1-\frac{\operatorname{var}\left(E-\hat{E}\right)}{\operatorname{var}E}\right)\times 100 \quad (3)$$

b) Mean Absolute Relative Error (MARE)

$$\text{MARE (\%)} = \frac{\sum f\left(\mathrm{R}_{\mathrm{E}}\right)}{\sum f}\times 100 \quad (4)$$

c) Variance Absolute Relative Error (VARE)

$$\text{VARE (\%)} = \frac{\sum f\left(R_E - meanR_E\right)^2}{\sum f}\times 100 \quad (5)$$

d) Prediction (n)

Prediction at level n is defined as the % of projects that have absolute relative error less than n.

e) Balance Relative Error (BRE)

$$\text{BRE} = \frac{\left|E-\hat{E}\right|}{\min(E,\hat{E})} \quad (6)$$

Where, E = Estimated effort $\hat{E}$ = Actual effort

Absolute Relative Error (RE) = $$\frac{\left|E-\hat{E}\right|}{\left|E\right|} \qquad (7)$$

f) Mean Magnitude of Relative Error (MMRE)

$$MMRE(\%) = \frac{1}{N}\sum_{i=1}^{N} MRE_i \times 100 \qquad (8)$$

Where $$MRE = \left|\frac{\hat{E}-E}{\hat{E}}\right| \qquad (9)$$

N = No. of Projects, E = Estimated effort, $\hat{E}$ = Actual effort

A model which gives higher VAF is better than that which gives lower VAF. A model which gives higher Pred(n) is better than that which gives lower Pred(n). A model which gives lower MARE is better than that which gives higher MARE [11, 12]. A model which gives lower VARE is better than that which gives higher VARE [6]. A model which is having less MMRE is better than the model which is having higher MMRE. A model which gives lower BRE is better than that which gives higher BRE. A model which gives lower MMRE is better than that which gives higher MMRE.

IV. Experimental Study

In carrying out our experiments, we have chosen 93 projects of the NASA 93 dataset. The estimated efforts using Intermediate COCOMO, Fuzzy using TMF and GBellMF are shown in Table III. Table IV and "Fig. 6" to "Fig. 12" shows the comparisons of various models basing on different criterions.

TABLE III. Estimated Effort in Man Months of Various Models

Project ID	Actual Effort	Estimated Effort using		
		COCOMO	Fuzzy-TriMF	Fuzzy-GbellMF
1	8.40	2.30	12.61	14.05
9	36.00	27.80	42.98	45.31
12	42.00	32.30	35.86	32.37
13	42.00	35.40	42.14	43.83
17	50.00	36.30	52.23	60.74
19	60.00	50.60	67.46	80.99
26	72.00	33.00	93.12	94.04
36	120.00	116.70	134.45	113.77
41	192.00	574.20	192.00	172.59
45	239.00	224.70	245.30	268.04
49	300.00	290.50	355.46	354.73
53	352.80	290.50	355.46	354.73
59	420.00	436.90	448.12	483.07
61	432.00	615.50	438.00	269.83
68	576.00	821.10	533.08	652.35
77	882.00	1332.30	882.03	881.34
81	1248.00	1113.60	1183.38	1004.62
83	1368.00	1139.60	1358.98	1358.40
92	4560.00	24726.50	4559.60	4581.44

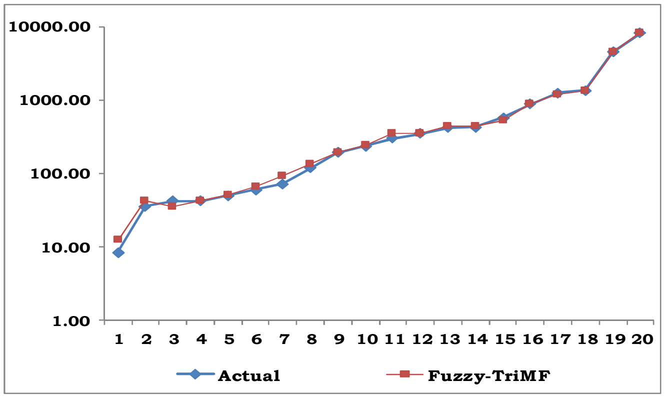


Figure 6: Estimated Effort using Fuzzy-TriMF versus Actual Effort

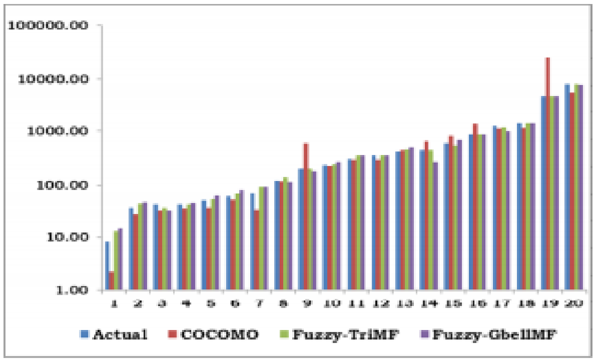


Figure 7: Estimated Effort using Various Models versus Actual Effort

TABLE IV. Comparison of various models

Model	VAF	MARE	VARE	Mean BRE	MMRE	Pred(30)%
Fuzzy-TriMF	96.53	28.53	10.51	0.61	54.81	62
Fuzzy-GBellMF	95.90	23.78	32.59	0.59	63.16	65
COCOMO	33.65	47.22	46.89	0.78	59.50	53

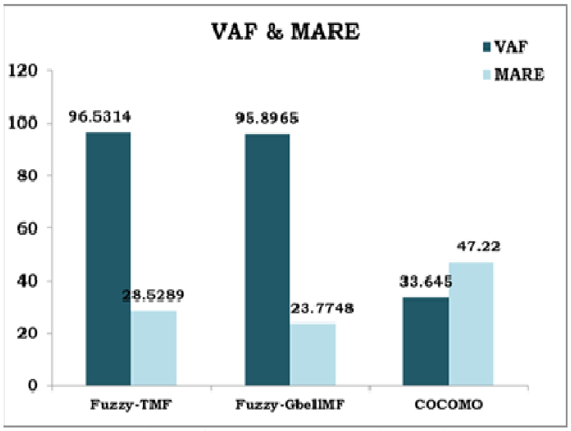


Figure 8: Comparison of VAF & MARE against various models

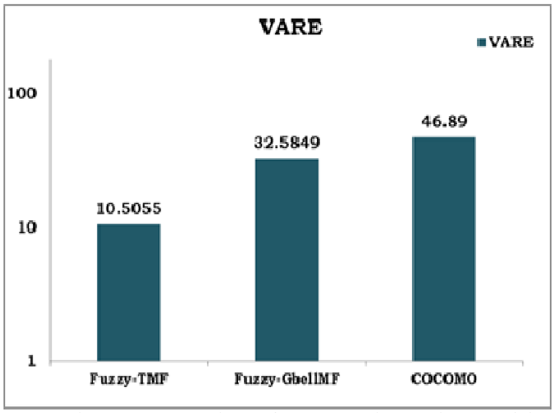


Figure 9: Comparison of VARE against various models

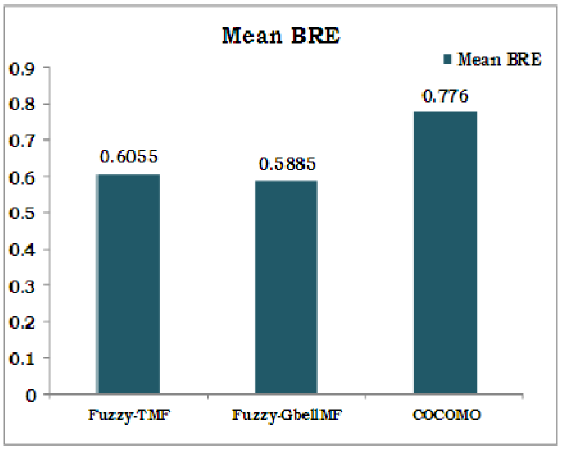


Figure 10: Comparison of Mean BRE against various models

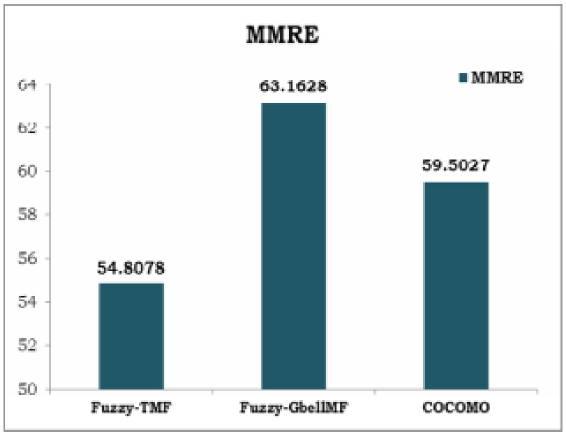


Figure 11: Comparison of MMRE against various models

V. CONCLUSION

Referring to Table 4, we see that Fuzzy using TMF yields better results for maximum criterions when compared with the other methods. Thus, basing on VAF, MMRE, VARE, MARE, Pred(30) & Mean BRE, we come to a conclusion that the Fuzzy method using TMF (triangular membership function) is better than Fuzzy method using GBellMF or Intermediate COCOMO. It is not possible to evolve a method, which can give 100 % VAF. By suitably adjusting the values of the parameters in FIS we can optimize the estimated effort.

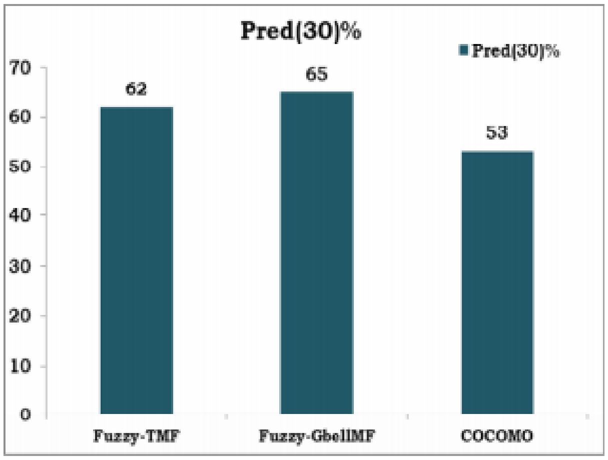


Figure 12: Comparison of Pred (30) % against various models

REFERENCES

[1] RamilJ.F., Algorithmic cost estimation for software evolution, Software Engg. (2000) 701-703.

[2] Angelis L, Stamelos I, Morisio M, Building software cost estimation model based on categorical data, Software Metrics Symposium, 2001- Seventh International Volume

(2001) 4-15.

[3] B.W. Boehm, Software Engineering Economics, Prentice-Hall, Englewood Cli4s, NJ, 1981

[4] Kirti Seth, Arun Sharma & Ashish Seth, Component Selection Efforts Estimation– a Fuzzy Logic Based Approach, IJCSS-83, Vol (3), Issue (3).

[5] Zhiwei Xu, Taghi M. Khoshgoftaar, Identification of fuzzy models of software cost estimation, Fuzzy Sets and Systems 145 (2004) 141–163

[6] Harish Mittal, Harish Mittal, Optimization Criteria for Effort Estimation using Fuzzy Technique, CLEI Electronic Journal, Vol 10, No 1, Paper 2, 2007

[7] R. Babuska, Fuzzy Modeling for Control, Kluwer Academic Publishers, Dordrecht, 1999

[8] Moshood Omolade Saliu, Adaptive Fuzzy Logic Based Framework for Software Development Effort Prediction, King Fahd University of Petroleum & Minerals, April 2003

[9] Iman Attarzadeh and Siew Hock Ow, Software Development Effort Estimation Based on a New Fuzzy Logic Model, IJCTE, Vol. 1, No. 4, October2009

[10] Xishi Huang, Danny Ho,Jing Ren, Luiz F. Capretz, A soft computing framework for software effort estimation, Springer link, Vol 10, No 2 Jan-2006

[11] Prasad Reddy P.V.G.D, Sudha K.R , Rama Sree P & Ramesh S.N.S.V.S.C, Software Effort Estimation using Radial Basis and Generalized Regression Neural Networks, Journal of Computing, Vol 2, Issue 5 May 2010

[12] Prasad Reddy P.V.G.D, Sudha K.R , Rama Sree P & Ramesh S.N.S.V.S.C, Fuzzy Based Approach for Predicting Software Development Effort, International Journal of Software Engineering, Vol 1, Issue 1, June2010

[13] Zonglian F. and Xihui L., "f-COCOMO: Fuzzy Constructive Cost Model in Software Engineering", Proc. of IEEE Int. Conf. On Fuzzy Systems, IEEE, 1992, 331-337.

[14] Ryder J., "Fuzzy Modeling of Software Effort Prediction", Proc. of IEEE Information Technology Conference, Syracuse, NY, 1998.

[15] Idri A. and Abran A., "COCOMO Cost Model Using Fuzzy Logic", 7th International Conference on Fuzzy Theory & Technology, Atlantic City, New Jersey, March 2000.

[16] M. W. Nisar, Yong-Ji Wang, M. Elahi and I.A Khan, "Software Development Effort Estimation Using Fuzzy Logic", Information

Technology Journal, 2009 Asian Network for Scientific Information, 2009.

[17] http://www.mathworks.com/help/toolbox/fuzzy/fp351dup8.html

[18] M. Braz and S. Vergilio, "Using Fuzzy Theory for Effort Estimation of Object-Oriented Software", Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2004.

[19] B. A. Kitchenham and N. R. Taylor "Software Cost Models", ICL Technical Journal, May 1984

[20] W. Pedrycz, J.F. Peters and S. Ramanna, "A Fuzzy Set Approach to Cost Estimation of Software Projects", Proceedings of the 1999 IEEE Canadian Conference on Electrical and Computer Engineering, Shaw Conference Center, Edmonton, Alberta, Canada May 9-12, 1999.

[21] Ahmad, Y., & Husain, S. (2010). Applying Intuitionistic Fuzzy Approach to Reduce Search Domain in an Accidental Case. International Journal of Advanced Computer Science and Applications - IJACSA, 1(4).

[22] Barik, S., Mishra, D., Mishra, S., Satapathy, S. K., Rath, A. K., & Acharya, M. (2010). Pattern Discovery using Fuzzy FP-growth Algorithm from Gene Expression Data. International Journal of Advanced Computer Science and Applications - IJACSA, 1(5), 50-55.

[23] Khan, A. R., Rehman, Z. U., & Amin, H. U. (2011). Application of Expert System with Fuzzy Logic in Teachers ' Performance Evaluation. International Journal of Advanced Computer Science and Applications - IJACSA, 2(2), 51-57.

A Load Balancing Policy for Heterogeneous Computational Grids

Said Fathy El-Zoghdy
Mathematics and Computer Science Department
Faculty of Science, Menoufia University
Shebin El-Koom, Egypt.
Email: Elzoghdy@yahoo.com

***Abstract*—Computational grids have the potential computing power for solving large-scale scientific computing applications. To improve the global throughput of these applications, workload has to be effectively balanced among the available computational resources in the grid environment. This paper addresses the problem of scheduling and load balancing in heterogeneous computational grids. We proposed a two-level load balancing policy for the multi-cluster grid environment where computational resources are dispersed in different administrative domains or clusters that existed physically in various LANs. The proposed load balancing policy reflects the heterogeneity of the computational resources in deciding load distributions decisions. It balances the system's load according to the computing nodes capacity. Therefore, system's overall job response time and utilization are minimized and maximized respectively. An analytical model is developed to gauge the performance of the proposed load balancing policy. The results obtained analytically are validated by simulating the model using Arena simulation package. The results show that the overall mean job response time obtained by simulation is very close to that obtained analytically. Also, the results revealed that the performance of the suggested load balancing strategy outperforms that of the random and uniform distribution load balancing strategies in terms of mean job response time. The improvement ratio increases as the system workload increases and the maximum improvement ratio obtained is about 72% within the studied system parameters values.**

Keywords- Computational grids; resource management; load distribution; queuing theory; simulation model.

I. INTRODUCTION

The rapid development in computing resources has enhanced the performance of computers and reduced their costs. This availability of low cost powerful computers coupled with the advances and popularity of the Internet and high speed networks has led the computing environment to be mapped from the traditionally distributed systems and clusters to the Grid computing environments. The Grid computing has emerged as an attractive computing paradigm [1,2]. The Computing Grid, a kind of grid environments, aims to solve the massive computation problems. It can be defined as hardware and software infrastructure which provides dependable, consistent, pervasive and inexpensive access to geographically widely distributed computational resources. These resources may belong to various individuals and institutions to solve large-scale scientific applications. Such applications may contain Nano-materials, massive data, DNA research and simulated meteorology systems.

Basically, grid resources are physically distributed workstations or servers, which are gathered to works as an integrated processing system. The primary motivation of grid computing system is to support clients and programs with universal and continuous access to enormous set of high performance computational resources [1-4]. Computational grids offer many types of services. These services are provided by the servers in the grid computing system. The servers are generally heterogeneous as they may have different CPUs computing power, storage size, etc. [4].

As a consequence of the unequal task arrival rates and difference of computing capacities and capabilities, the computers in one grid site may be heavily loaded while others in a different grid site may be lightly loaded or even idle. It is therefore needed to shift some jobs from the heavily loaded computers to others from the lightly loaded set aiming to efficiently employ the resources and consequently minimize the average job response time. The load shifting process is recognized as load balancing (LB)[4,5,6].

In general, LB policies can be categorized into centralized or distributed. In centralized policies, the system has only one LB decision maker which has a global view of the system load information. In such polices, the system's incoming jobs are automatically forwarded to the decision maker, which balances the load among different processing nodes aiming to improve system average response time. These strategies are favorable if the communication cost is unneglectable or not important as in shared memory multiprocessor systems. Various scholars claim that, the centralized policies are not scalable as if the number of processing nodes in the system increases, the decision maker may fail [6-9,16].

In the distributed (decentralized) LB policies on contrary, all computers (nodes) in the system participate in taking the load distribution decisions. As a result, the decisions of load redistribution are not centralized in one node. Therefore, various scholars think that, the distributed LB strategies are better from the scalability and fault tolerance points of view than the centralized ones. But at the same time, it is very costly to enable every computer in a distributed system from collecting the state information of the entire system. As a consequence, in the distributed load distribution strategies, every computing node receives its incoming tasks and after that, it decides to shift a part of its load based on the partial or complete information it has about the overall system's load distribution [17-19]. It appears that this policy is closely related to the individually optimal policy, in that each job (or its user) optimizes its own expected mean response time independently of the others [4-10].

Although the problem of balancing loads in conventional distributed environments has been studied massively [6-14], new challenges in Grid computing still make it an interesting topic and many research projects are interested in this problem.

In this paper, we present a distributed LB policy for the grid computing environment. The proposed policy tends to improve grid resources utilization and hence maximizes throughput. It concentrates on studying the proposed model in its steady state. In this state, the total number of admitted jobs to the computational grid is adequately large and the incoming jobs rate cannot surpass the entire processing capacity of system [15]. As in [15], steady-state mode will help us to derive optimality for the proposed LB policy. The suggested LB strategy addresses the problem's class of massive computation and entirely independent jobs that has no in between communications. An analytical model is presented. This model is based on queuing theory. We are interested in computing the overall mean job response time. The results obtained analytically are validated by simulating the model using Arena simulation package.

The structure of this paper's remaining sections is as follows: Section II gives major and recent related works. Section III presents the architecture of suggested computational grid model. Section IV introduces the proposed grid LB policy. Section V discusses the analytical queuing model. In Section VI, we assess the performance evaluation of the proposed LB policy. Lastly, Section VII concludes this paper.

II. Related Work and Motivations

LB has been studied massively in the conventional distributed systems literature for more than two decades. Various policies and algorithms have been suggested, analyzed and implemented in a number of studies [6-14]. It is more challenging to achieve LB in Grid systems than in conventional distributed computing ones because of the heterogeneity and the complicated dynamic nature of the Grid systems. The problem of LB in grid architecture is addressed by assigning loads in a grid without neglecting the communication overhead in collecting the load information. It considers load index as a decision factor for scheduling of jobs in a cluster and among clusters.

Many papers have been published recently to address the problem of LB in Grid computing environments. Some of the proposed computational grids LB policies are modifications or extensions to the conventional distributed systems LB policies. In [23], a decentralized model for heterogeneous grid has been proposed as a collection of clusters. In [1], the authors employed the tree structure in representing a computational grid model. Their suggested model considers the heterogeneity of system's computational nodes but it is entirely autonomous of any real grid structure. Though, they did not offer any job assigning algorithm. Their resource controlling strategy relies on the periodic gathering of node's information via manager node. Such strategy suffers from having massive communication overhead. Indeed, the manager node may represent a single point of failure to the system. The authors in [24] suggested utilizing ring topology in guiding managers of computational grids. These managers are in charge of controlling a dynamic set of computing nodes (computers or processors). The process of taking workload balancing decisions in their model relies on real load of computing nodes in the system. In [21], the authors proposed a hierarchical structure for grid managers rather than ring topology to improve scalability of the grid computing system. They also proposed a job allocation policy which automatically regulates the job flow rate directed to a given grid manager.

In this paper we propose a decentralized LB policy that can cater for the next exclusive features of applied computational grids systems:

- **Large-scale.** As a grid can involve a huge set of advanced computational nodes that really existed in various distributed sites; where it is impossible for the centralized systems to deal with the problems of having enormous communication overhead and remotely administrating distant stations.
- **Heterogeneous grid sites.** There might be various hardware specifications, OS and processing speeds in different sites.
- **Effects from considerable transfer delay.** The communication overhead involved in capturing load information of sites before making a dispatching decision can be a major issue negating the advantages of job migration. We should not ignore the considerable dynamic transfer delay in disseminating load updates on the Internet.

III. Grid COMPUTING Service Structure

The studied computational grid model is a large-scale service one and it relies on a geographical hierarchy decomposition arrangement. Every user submits his computing jobs and their hardware requirements to the Grid Computing Service (GCS). The GCS will reply to the user by sending the results when it finishes the execution of the jobs. In the GCS, jobs pass through four phases which can be summarized as follows:

A. Task submission phase

Grid clients can admit the jobs via any web explorer. This facilitates the job admission procedure and makes the system reachable to all users.

B. Task allocation phase

Once the GCS receives a job, it looks for the available resources (computers or processors) and allocates the suitable resources to the task.

C. Task execution phase

Once the needed resources are allocated to the task, it is scheduled for execution on that computing site.

D. Results collection phase

The GCS informs the user by his task's results immediately upon the execution is completed.

Three-level Top-Down view of the considered grid computing model is shown in Fig. 1 and can be explained as follows:

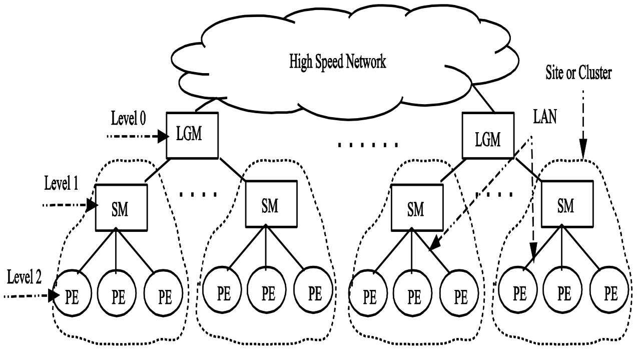


Figure 1. Grid Computing Model Structure

- **Level 0:** Local Grid Manager (LGM)

Any LGM manages a pool of Site Managers (SMs) in its geographical area. The role of LGM is to collect information about the active resources managed by its corresponding SMs. LGMs are also involved in the task allocation and LB process in the grid. New SMs can join the GCS by sending a *join* request to register themselves at the nearest parent LGM.

- **Level 1:** Site Manager (SM)

Every SM is in charge of controlling a set of computing nodes that are configured dynamically (i.e., any computing node can enter or disuse the system as desired). A new joining computing node to the site should register itself within the SM. The role of the SM is to collect information about active processing elements in its pool. The collected information mainly includes CPU speed and other hardware specifications. Also, any SM has the responsibility of allocating the incoming jobs to any processing element in its pool according to a specified LB algorithm.

- **Level 2:** Processing Elements (PE)

Any private or public PC or workstation can join the grid system by registering within any SM and offer its computing resources to be used by the grid users. When a computing element joins the grid, it starts the GCS system which will report to the SM some information about its resources such as CPU speed.

Within this hierarchy, the addition or removal of a SMs or PEs is an easy process and ensures scalability of suggested model of computational grids.

The LGMs represent the entry points of computing jobs in the proposed grid computing model. Any LGM works like a server in the web for the grid model. Any client can admit his jobs to the associated LGM using the web explorer. According to the available LB information, the LGM will pass the arrived jobs to the appropriate SM. The SM in turn distributes these computing jobs according to the available site LB information to a chosen processing element for execution. LGMs allover the world may be interconnected using a high-speed network as shown in Fig. 1.

As explained earlier, the information of any processing element joining or leaving the grid system is collected at the associated SM which in turn transmits it to its parent LGM. This means that a communication is needed only if a processing element joins or leaves its site. All of the collected information is used in balancing the system workload between the processing elements to efficiently utilize the entire system resources aiming to minimalize user's jobs response time. This policy minimizes the communication overhead involved in capturing system information before making a LB decision which improves the system performance. .

IV. GRID LOAD BALANCING POLICY

We proposed a two-level LB policy for the multi-cluster grid environment where clusters are located in different local area networks. The proposed LB policy takes into account the heterogeneity of the computational resources. It balances the system's load according to capacity of computing nodes. We assume that the jobs admitted to the grid system are entirely independent ones with no inter-process communication in between and that they are massive computation jobs.

To formalize the LB policy, we define the following parameters for grid computing service model:

1. **Job:** Every job is represented by a job Id, number of job instructions NJI, and a job size in bytes JS.
2. **Processing Element Capacity (PEC_{ij}):** Number of jobs that can be executed by the jth PE at full load in the ith site per second. The PEC can be calculated using the PEs CPU speed and assuming an Average Number of job Instructions ANJI.
3. **Site Processing Capacity (SPC_i):** Number of jobs that can be executed by the ith site per second. Hence, the SPC_i can be calculated by summing all the PECs for all the PEs managed the ith site**.**
4. **Local grid manager Processing Capacity (LPC):** Number of jobs that can be executed under the responsibility of the LGM per second. The LPC can be calculated by summing all the SPCs for all the sites managed by that LGM.

The proposed LB policy is a multi-level one as it could be seen form Fig 2. This policy is explained at each level of the grid architecture as follows:

A. Local Grid Manager Load Balancing Level

Consider a Local Grid Manager (LGM) which is responsible of a group of site managers (SMs). As mentioned earlier, the LGM maintains information about all of its SMs in terms of processing capacity SPCs. The total processing capacity of a LGM is LPC which is the sum of all the SPCs for all the sites managed by that LGM. Based on the total processing capacity of every site SPC, the LGM scheduler distributes the workload among his sites group members (SMs). Let *N* denotes the number of jobs arrived at a LGM in the steady state. Hence, the ith site workload (S_iWL) which is the number of jobs to be allocated to ith site manager is obtained as follows:

$$S_i WL = N \times \frac{SPC_i}{LPC} \qquad (1)$$

B. Site Manager Load Balancing Level

As it is explained earlier every SM manages a dynamic pool of processing elements (workstations or processors). Hence, it has information about the PECs of all the processing elements in its pool. The total site processing capacity SPC is obtained by summing all the PECs of all the processing elements in that site. Let M be the number of jobs arrived at a SM in the steady state. The SM scheduler will use a LB policy similar to that used by the LGM scheduler. This means that the site workload will be distributed among his group of processing elements based on their processing capacity. Using this policy, the throughput of every processing element will be maximized and also its resource utilization will be improved. Hence, the ith PE workload (PE_iWL) which is the number of jobs to be allocated to ith PE is obtained as follows:

$$PE_i WL = M \times \frac{PEC_i}{SPC} \qquad (2)$$

Example: Let N =1500 *j/s* (job/second*)* arrive at a LGM with five SMs having the following processing capacities:

SPC_1=440 j/s, SPC_2=260 j/s, SPC_3=320 j/s, SPC_4=580 j/s, and SPC_5=400 j/s.

Hence, LPC= 440+260+320+580+400=2000 *j/s.* So, the workload for every site will be computed according to equation 1 as follows:

$$S_1 WL = 1500 \times \frac{440}{2000} = 330 \quad j/s$$

$$S_2 WL = 1500 \times \frac{260}{2000} = 195 \quad j/s$$

$$S_3 WL = 1500 \times \frac{320}{2000} = 240 \quad j/s$$

$$S_4 WL = 1500 \times \frac{580}{2000} = 435 \quad j/s$$

$$S_5 WL = 1500 \times \frac{400}{2000} = 300 \quad j/s$$

Then workload of every site will be allocated to the processing elements managed by that site based on equation 2. As an example, suppose that the fifth site contains three PEs having the processing capacities of 90*j/s*, 200*j/s*, and *150j/s* respectively. Hence the SPC= 90+200+150= 440 *t/s.* Remember that this site workload equals to 300 *t/s* as computed previously. So, the workload for every PE will be computed according to equation 2 as follows:

$$PE_1 WL = 300 \times \frac{180}{400} = 135 \quad j/s$$

$$PE_2 WL = 300 \times \frac{120}{400} = 90 \quad j/s$$

$$PE_3 WL = 300 \times \frac{100}{400} = 75 \quad j/s$$

From this simple numerical example, one can see that the proposed LB policy allocates more workload to the faster PEs which improves the system utilization and maximizes system throughput.

V. Analytical Model

To compute the mean job response time analytically, we consider one LGM section as a simplified grid model. In this model, we will concentrate on the time spent by a job in the processing elements. Consider the following system parameters:

- λ is the external job arrival rate from grid clients to the LGM.
- λ_i is the job flow rate from the LGM to the ith SM which is managed by that LGM.
- λ_{ij} is the job flow rate from the ith SM to the jth PE managed by that SM.
- μ is the LGM processing capacity.
- μ_i is processing capacity of the ith SM.
- μ_{ij} is the processing capacity of the jth PE which is managed by the ith SM.
- $\rho=\lambda/\mu$ is the system traffic intensity. For the system to be stable ρ must be less than 1.
- $\rho_i = \frac{\lambda_i}{\mu_i}$ is traffic intensity of the ith SM.
- $\rho_{ij} = \frac{\lambda_{ij}}{\mu_{ij}}$ is traffic intensity of the jth PE which is managed by ith SM.

We assume that the jobs arrive from clients to the LGM according to a time-invariant Poisson process. Jobs arrive at the LGM sequentially, with inter-arrival times which are independent, identically, and exponentially distributed with the arrival rate λ *j/s*. Simultaneous arrivals are excluded. Every PE in the dynamic site pool will be modeled by an M/M/1 queue.

Since jobs that arrive to the LGM will be automatically distributed on the sites managed by that LGM with a routing probability $\mathrm{PrS}_i = \frac{\mathrm{SPC}_i}{\mathrm{LPC}}$ according to the LB policy, where *i* is the site number, hence $\lambda_i = \lambda \times \Pr S_i = \lambda \times \frac{\mathrm{SPC}_i}{\mathrm{LPC}}$. Again the site *i* arrivals will also automatically be distributed on the PEs managed by that site with a routing probability $\mathrm{PrE}_{ij} = \frac{\mathrm{PEC}_{ij}}{\mathrm{SPC}_i}$ based on the LBP, where *j* is the PE number and *i* is the site number. Hence, $\lambda_{ij} = \lambda_i \times \mathrm{PrE}_j = \lambda_i \times \frac{\mathrm{PEC}_{ij}}{\mathrm{SPC}_i}$.

Since the arrivals to LGM are assumed to follow a Poisson process, then the arrivals to the PEs will also follow a Poisson process. We also assume that the service times at the jth PE in the ith SM is exponentially distributed with fixed service rate μ_{ij} *j/s*. Note that μ_{ij} represents the PE's processing capacity (PEC) in our LB policy. The service discipline is First Come First Serviced. This grid queueing model is illustrated in Fig 2.

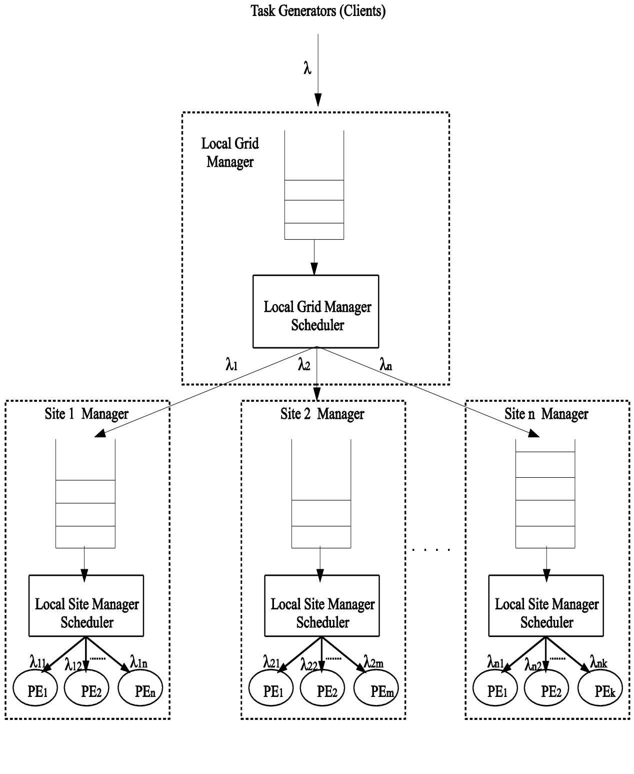

Task Generators (Clients)
λ
Local Grid Manager
Local Grid Manager Scheduler
λ1
λ2
λn
Site 1 Manager
Site 2 Manager
Site n Manager
. . . .
Local Site Manager Scheduler
Local Site Manager Scheduler
Local Site Manager Scheduler
λ11
λ12
λ1n
PE1
PE2
PEn
λ21
λ22
λ2m
PE1
PE2
PEm
λn1
λn2
λnk
PE1
PE2
PEk

Figure 2. Grid Computing Queueing Model

The state transition diagram of the jth PE in ith site manager is shown in Fig. 3.

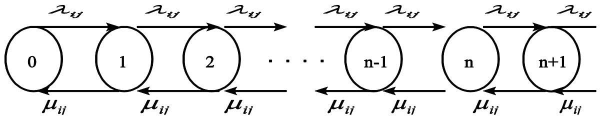


Figure 3. A state transition diagram of jth PE in ith site manager.

As mentioned earlier, we are interested in studying the system at the steady state that is the traffic intensity is less than one i.e., $\rho \prec 1$. To compute the expected mean job response time, the Little's formula will be used. Let E[T$_g$] denotes the mean time spent by a job at the grid to the arrival rate λ and E[N$_g$] denotes the number of jobs in the system. Hence by Little formula, the mean time spent by a job at the grid will be given by equation 3 as follows:

$$E[N_g] = \lambda \times E[T_g] \qquad (3).$$

$E[N_g]$ can be computed by summing the mean number of jobs in every PE at all the grid sites. So, $E[N_g] = \sum_{i=1}^{m} \sum_{j=1}^{n} E[N_{PE}^{ij}]$, where i=1,2,..m, is the number of site managers managed by a LGM, *j=1,2,…,n* is the number of processing elements managed by a SM and $E[N_{PE}^{ij}]$ is the mean number of jobs in a processing element number *j* at site number *i*. Since every PE is modeled as an M/M/1 queue, then $E[N_{PE}^{ij}] = \frac{\rho_{ij}}{1-\rho_{ij}}$, where $\rho_{ij} = \frac{\lambda_{ij}}{\mu_{ij}}$, μ_{ij} =PEC$_{ij}$ for PE number *j* at site number *i*. From equation 3, the expected mean job response time is given by:

$$E[T_g] = \frac{1}{\lambda} \times E[N_g] = \frac{1}{\lambda} \times \sum_{i=1}^{m} \sum_{j=1}^{n} E[N_{PE}^{ij}]$$

Note that the stability condition for PE$_{ij}$ is $\rho_{ij} \prec 1$.

VI. Results and Discussion

A. Experimental Environment

The simulation was carried out using the great discrete event system simulator Arena [25]. This simulator allows modeling and simulation of entities in grid computing systems users, applications, resources and resource load balancers for design and evaluation of LB algorithms.

To gauge the performance of grid computing system under the proposed LB policy, a simulation model is built using Arena simulator. This simulation model consists of one LGM which manages a number of SMs which in turn manages a number of PEs (Workstations or Processors). All simulations are performed on a PC (Core 2 Processor, 2.73GHz, 1GB RAM) using Windows xp OS.

B. Simulation Results and Analysis

We assume that the external jobs come to the LGM in a sequential fashion and their inter-arrival times are independent and they follow the exponential distribution with mean 1/λ *j/s*. no Instantaneous arrivals is allowed. We also assume that the service times of LGMs follow the exponential distribution with mean 1/μ *j/s*.

The performance of the grid computing system under the proposed LB policy is compared with two other policies namely; Random distribution LB policy and Uniform distribution LB policy.

In the Uniform distribution LB policy the job flow rate (routing probability) from LGM to its SMs is fixed to the value $\frac{1}{n_s}$, where n_s is the number of SMs in the grid computing service model. Also the job flow rate (routing probability) from any SM to its PEs is fixed to the value $\frac{1}{n_{PE}}$, where n_{PE} is the number of PEs which are managed by that site.

In the Random distribution LB policy a resource for job execution is selected randomly without considering any performance metrics to that resource or to the system. This policy is explained in [26]. However, in the proposed LB policy all the arriving jobs from clients to the LGMs are distributed on the SMs based on their processing capacity to improve utilization aiming to minimize mean job response time.

The grid system built in our simulation experiment has 1 LGM, 3 SMs having 4, 3, and 5 PEs respectively. We fixed the total grid system processing capacity μ=LPC=1700 *j/s*. First, the mean job response time under the proposed LB policy is computed analytically and by simulation as shown in Table 1. From that table, we can see that the response times obtained by the simulation approximate that obtained analytically. The obtained simulation results satisfy 95% confidence level.

Also, from table 1, we can notice that the proposed LB policy is asymptotically optimal because its saturation point (λ/μ)≈1 is very close to the saturation level of the grid computing model.

Using the same grid model parameters setting of our simulation experiment, the performance of the proposed LB policy is compared with that of the Uniform distribution, and Random distribution as shown in Fig. 4. From that figure we can see that proposed LBP outperforms the Random distribution and Uniform distribution LBPs in terms of system mean job response time. It is also noticed that the system mean response time obtained by the uniform LBP lies between that of the proposed and random distribution LBPs.

To evaluate how much improvement obtained in the system mean job response time as a result of applying the proposed LBP, we computed the improvement ratio $(T_U - T_p)/T_U$, where T_U is the system mean job response time under uniform distribution LBP and T_P is the system mean job response time under proposed LBP, see Fig. 5.

TABLE 1: COMPARISON BETWEEN ANALYTIC AND SIMULATION MEAN TASK RESPONSE TIMES USING THE PROPOSED LBP

Arrival rate λ	Traffic Intensity ρ=λ/μ	Analytic Response Times	Simulation Response Times
400	0.235294	0.009231	0.009431
500	0.294118	0.010000	0.010210
600	0.352941	0.010909	0.010709
700	0.411765	0.012000	0.012032
800	0.470588	0.013333	0.012833
900	0.529412	0.015000	0.015401
1000	0.588235	0.017143	0.017023
1100	0.647059	0.020000	0.019821
1200	0.705882	0.024000	0.024025
1300	0.764706	0.030000	0.029903
1400	0.823529	0.040000	0.040240
1500	0.882353	0.060000	0.058024
1600	0.941176	0.120000	0.119012
1650	0.970588	0.240000	0.238671
1660	0.976471	0.300000	0.297401
1670	0.982353	0.400000	0.401202
1680	0.988235	0.600000	0.610231
1685	0.991176	0.800000	0.798502
1690	0.994118	1.200000	1.201692

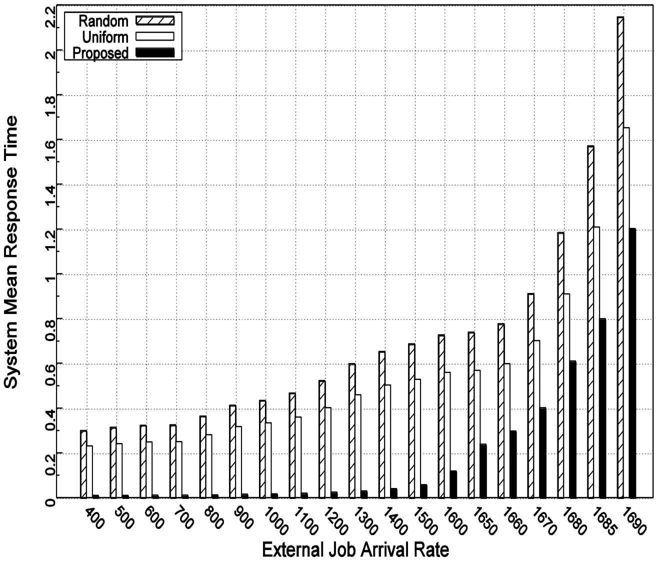


Figure 4. System mean job response time versus job arrival rate

From that figure, we can see that the improvement ratio increases as the system workload increases and it is about 72% in the range of parameter values examined. This result was anticipated since the proposed LBP balances the system's load according to the capacity of computing nodes which leads to maximizing system resources utilization ratio and as a result system mean job response time is minimized. In contrast, the Random distribution policy distributes the system workload randomly on the system PE without putting any performance metric in mind which may lead to unbalanced system workload distribution which leads to poor resources utilization and hence, the system performance is affected. This situation appears clearly as the system workload increases. Also, the Uniform distribution policy distributes the system workload equally on the PEs without putting their processing capacity or any workload information in mind which repeats the same situation as the random distribution LBP. To be fair, we must say that according to the obtained simulation results, the performance of the Uniform distribution LBP is much better that that of the Random distribution LBP.

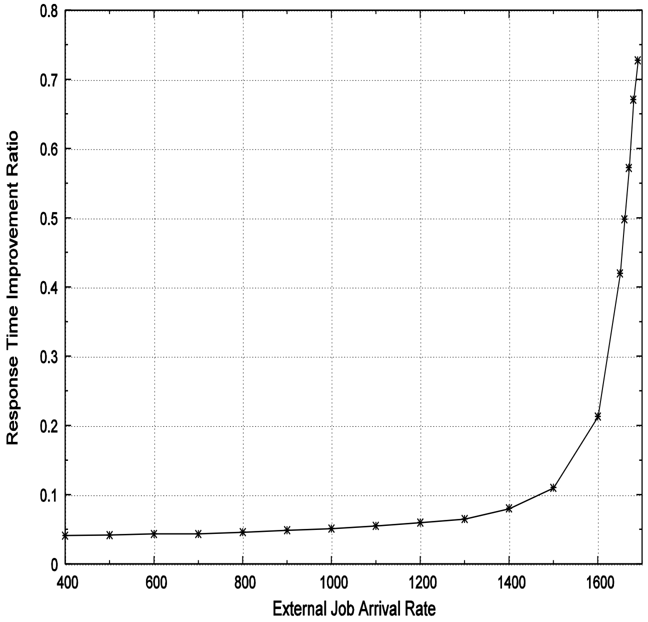


Figure 5. System mean job response time improvement ratio

VII. CONCLUSION

This paper addresses the load balancing problem for computational grid environment. We proposed a two-level load balancing policy for the multi-cluster grid environment where clusters are located in different local area networks. The proposed load balancing strategy reflects the heterogeneity of the computing nodes. It balances system's load according to capacity of computing nodes. Consequently, the system's overall job response time, utilization are minimized and maximized respectively.

An analytical model is developed to compute the expected mean job response time in the grid system. To evaluate the performance of the proposed load balancing policy and validate the analytic results a simulation model is built using Arena simulator. The results show that the overall mean job response time obtained analytically is very close to that obtained by the simulation.

Also, the results showed that the performance of the proposed load balancing outperforms that of the Random and Uniform distribution load balancing policies in terms of mean job response time. It improves the overall job mean response time. The improvement ratio increases as the system workload increases and the maximum improvement ratio obtained is about 72% in the range of system parameter values examined.

REFERENCES

[1] B. Yagoubi and Y. Slimani, "Task Load Balancing Strategy for Grid Computing," *Journal of Computer Science*, vol. 3, no. 3: pp. 186-194, 2007.

[2] K. Lu, R. Subrata, and A. Y. Zomaya,"On The Performance-Driven Load Distribution For Heterogeneous Computational Grids," *Journal of Computer and System Science,* vol. 73, no. 8, pp. 1191-1206, 2007.

[3] S. Parsa and R. Entezari-Maleki," RASA: A New Task Scheduling Algorithm in Grid Environment," *World Applied Sciences Journal 7* (Special Issue of Computer & IT), pp. 152-160, 2009

[4] K. Li, "Optimal load distribution in nondedicated heterogeneous cluster and grid computing environments," *Journal of Systems Architecture*, vol. 54, pp. 111–123, 2008.

[5] Y. Li, Y. Yang, M. Ma, and L. Zhou, "A hybrid load balancing strategy of sequential jobs for grid computing Environments," *Future Generation Computer Systems*, vol. 25, pp.) 819-828, 2009.

[6] H. Kameda, J. Li, C. Kim, and Y. Zhang, "Optimal Load Balancing in Distributed Computer Systems," Springer, London, 1997.

[7] S. F. El-Zoghdy, H. Kameda, and J. Li, "Numerical Studies on Paradoxes in Non-Cooperative Distributed Computer Systems," *Game Theory and Applications*, vol. 9, pp. 1-16, 2003.

[8] S. F. El-Zoghdy, H. Kameda, and J. Li, "Numerical Studies on a Paradox for Non-Cooperative Static Load Balancing in Distributed Computer Systems," *Computers and Operation Research*, vol. 33, pp. 345-355, 2006..

[9] S. F. El-Zoghdy, "Studies on Braess-Like Paradoxes for Non-Cooperative Dynamic Load Balancing in Distributed Computer Systems," *Proc. of the IASTED Inter. Conf. on Parallel and Distributed Computing and Networks*, pp. 238-243, 2006.

[10] S. F. El-Zoghdy, H. Kameda, and J. Li, "A comparative study of static and dynamic individually optimal load balancing policies," *Proc. of the IASTED Inter. Conf. on Networks, Parallel and Distributed Processing and Applications*, pp. 200-205. 2002.

[11] A. N. Tantawi and D. Towsley, "Optimal static load balancing in distributed computer systems," *J. ACM*, vol.32, no.2, pp.455-465, Apr 1985.

[12] J. Li and H. Kameda, "A Decomposition Algorithm for Optimal Static Load Balancing in Tree Hierarchy Network Configurations," *IEEE Trans. Parallel and Distributed Systems*, vol. 5, no. 5, pp.540-548, 1994.

[13] J. Li and H. Kameda, "Load Balancing Problems for Multiclass Jobs in Distributed/Parallel Computer Systems," *IEEE Trans. Comput.*, vol. 47, no. 3, pp322-332, 1998.

[14] R. Mirchandaney, D. Towsley, and J. A. Stankovic, "Adaptive Load Sharing in Heterogeneous Distributed Systems", *J. Parallel and Distributed Computing*," vol. 9, pp.331-346, 1990.

[15] O. Beaumont, A. Legrand, L. Marchal and Y. *Robert. "Steady-State Scheduling on Heterogeneous Clusters," Int. J. of Foundations of Computer Science*, vol. 16, no.2,pp. 163-194, 2005.

[16] M. J. Zaki, W. Li, and S. Parthasarathy *"Customized dynamic load balancing for network of Workstations," In Proc. of the 5th IEEE Int. Symp. HDPC*: p. 282-291, 1996.

[17] A. Barak, O. La'adan, "The MOSIX multicomputer operating system for high performance cluster computing," *J. Future Gener. Comput. Systems,* vol. 13, no. (4-5), pp. 361–372, 1998.

[18] H.-U. Heiss, M. Schmitz, "Decentralized dynamic load balancing: The particles approach," *Inform. Sci.*, vol. 84, no. (1–2), pp. 115-128, 1995.

[19] M.H. Willebeek-LeMair, A.P. Reeves, "Strategies for dynamic load balancing on highly parallel computers," *IEEE Trans. Parallel Distrib. Systems*, vol. 4, no. 9, pp. 979–993, 1993.

[20] E. Saravanakumar and P. Gomathy," A novel load balancing algorithm for computational grid," *Int. J. of Computational Intelligence Techniques*, vol. 1, no. 1, 2010

[21] A. Touzene, H. Al Maqbali, "Analytical Model for Performance Evaluation of Load Balancing Algorithm for Grid Computing," Proc. of the 25th *IASTED Inter. Multi-Conference: Parallel and Distributed Computing and Networks*, pp. 98-102, 2007.

[22] Y. Wu, L. Liu, J. Mao, G. Yang, and W. Zheng, "Analytical Model for Performance Evaluation in a Computational Grid," *Proc. of the 3rd Asian Tech. Info. Program's (ATIP'S) on High performance computing*: solution approaches to impediment performance computing, pp. 145-151, 2007.

[23] J. Balasangameshwara, N. Raju, "A Decentralized Recent Neighbour Load Balancing Algorithm for Computational Grid," *Int. J. of ACM Jordan*, vol. 1, no. 3, pp. 128-133, 2010.

[24] A. Touzene, S. Al Yahia, K.Day, B. Arafeh, "Load Balancing Grid Computing Middleware," IASTED Inter. Conf. on Web Technologies, Applications, and Services, 2005.

[25] Arena simulator <http://www.ArenaSimulation.com>).

[26] Zikos, S., Karatza, H.D., “Resource allocation strategies in a 2-level hierarchical grid system,” *Proc*. of the 41st Annual Simulation Symposium (ANSS), April 13–16, 2008. IEEE Computer Society Press, SCS, pp. 157–164.

AUTHORS PROFILE

Dr. Said Fathy El-Zoghdy Was born in El-Menoufia, Egypt, in 1970. He received the BSc degree in pure Mathematics and Computer Sciences in 1993, and MSc degree for his work in computer science in 1997, all from the Faculty of Science, Menoufia, Shebin El-Koom, Egypt. In 2004, he received his Ph. D. in Computer Science from the Institute of Information Sciences and Electronics, University of Tsukuba, Japan. From 1994 to 1997, he was a demonstrator of computer science at the Faculty of Science, Menoufia University, Egypt. From December 1997 to March 2000, he was an assistant lecturer of computer science at the same place. From April 2000 to March 2004, he was a Ph. D. candidate at the Institute of Information Sciences and Electronics, University of Tsukuba, Japan., where he was conducting research on aspects of load balancing in distributed and parallel computer systems. From April 2004 to 2007, he worked as a lecturer of computer science, Faculty of Science, Menoufia University, Egypt. From 2007 until now, he is working as an assistant professor of computer science at the Faculty of Computers and Information Systems, Taif University, Kingdom of Saudi Arabia. His research interests are in load balancing in distributed/parallel systems, Grid computing, performance evaluation, network security and cryptography.

Generating PNS for Secret Key Cryptography Using Cellular Automaton

Bijayalaxmi Kar
Dept. of Computer Science & Engg.
College of Engineering Bhubaneswar
Bhubaneswar, Odisha, INDIA

D.Chandrasekhra Rao
Dept. of Computer Sc. & Engg.
College of Engineering Bhubaneswar
Bhubaneswar, Odisha, INDIA

Dr. Amiya Kumar Rath
Dept. of Computer Sc. & Engg.
College of Engineering Bhubaneswar
Bhubaneswar, Odisha, INDIA

Abstract **- The paper presents new results concerning application of cellular automata (CAs) to the secret key using vernam cipher cryptography.CA are applied to generate pseudo-random numbers sequence (PNS) which is used during the encryption process. One dimensional, non-uniform CAs is considered as a generator of pseudorandom number sequences (PNSs) used in cryptography with the secret key. The quality of PNSs highly depends on a set of applied CA rules. Rules of radius *r* = 1 and 2 for non-uniform one dimensional CAs have been considered. The search of rules is performed with use of evolutionary technique called cellular programming. As the result of collective behavior of discovered set of CA rules very high quality PNSs are generated. The quality of PNSs outperforms the quality of known one dimensional CA-based PNS generators used in the secret key cryptography. The extended set of CA rules which was found makes the cryptography system much more resistant on breaking a cryptography key.**

Keywords - Cellular automata; Cellular programming; Random number generators; Symmetric key; cryptography; Vernam cipher.

I. Introduction

Confidentiality is mandatory for a majority of network applications for example commercial uses of the internet. Two classes of algorithms exist on the market for Data encryption: secret key systems and public key systems. An emerging cryptography techniques used in both types of system. One of such a promising cryptography techniques are cellular automata. Cellular automata are highly parallel and distributed systems which are able to perform complex computations. New perspectives in this area have been opened when evolutionary techniques appeared and have been used to design automatically CA based system.

CAs were proposed for public key cryptosystems by Guan [15] and Kari [9] .In such systems two keys are required: one key is used for encryption and other is used for decryption, and one of them is held in private, other is published. However the main concern of this paper is secret key cryptosystems. In such system the same key is used for encryption and decryption. The encryption process is based on the generation of pseudorandom bit sequences, and CA is used for this purpose. In the context of secret key systems, CA were first studied by wolfram [17], and later by Nandi et al. [20] and Gutowitz [8]. Recently they were a subject of study by Tomassini and his colleagues [12]. This paper extends these recent studies and describes the application of one-dimensional (1D) CAs for the secret key cryptography.

The paper is organized as follows. The following section presents the idea of an encryption process based on Vernam cipher and used in CA-based secret key cryptosystems. Section 3 outlines the main concepts of CAs, overviews current state of applications of CAs in secret key cryptography and states the problem considered

In this paper Section IV outlines evolutionary technique called cellular programming and shows how this technique is used to discover new CA rules suitable for encryption process. Section V contains the analysis of results and Section VI concludes the paper.

II. Vernam Cipher and Secret Key Cryptography

Let P be a plain-text message consisting of m bits P_1 P_2 . . . $P_{m,}$ and k_1 k_2 . . . k_m be a bit stream of a key K. Let C_i be the *ith* bit of a cipher-text obtained by applying a $\oplus$ (exclusive-or) enciphering operation: $C_i = P_i \oplus K_i$

The original bit P_i of a message can be recovered by applying the same operation $\oplus$ on c_i with use of the same bit stream key k: $P_i = C_i \oplus K_i$

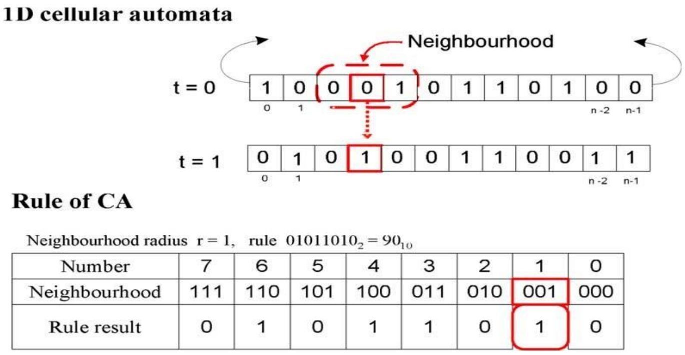
The enciphering algorithm called Vernam cipher is known to be [5, 9] perfectly safe if the key stream is truly unpredictable and is used only one time.

III. Cellular Automata and Cryptography

One-dimensional CA is in a simplest case a collection of two-state elementary automata arranged in a lattice of the length N, and locally interacted in a discrete time t. For each cell i called a central cell, a neighborhood of a radius r is defined, consisting of $n_i = 2r + 1$ cells, including the cell i. When considering a finite size of CAs a cyclic boundary condition is applied, resulting in a circle grid as shown in Figure 1.

It is assumed that a state q_i^{t+1} of a cell i at the time $t + 1$ depends only on states of its neighborhood at the time t, i.e. $q_i^{t+1} = f\,(q_i^{\,t}, q_{i1}^{\,t}, q_{i2}^{\,t}, . . . , q_{in}^{\,t})$, and a transition function f, called a *rule*, which defines a rule of updating a cell i. A length L of a rule and a number of neighborhood states for a binary uniform CAs is $L = 2^n$, where $n = n_i$ is a number of cells of a given neighborhood, and a number of such rules can be expressed as 2^L. For CAs with e.g. $r = 2$ the length of a rule is equal to $L = 32$, and a number of such rules is 2^{32} and grows very fast with *L*. When the same rule is applied to update cells of CAs, such CAs are called uniform CAs, in contrast with non uniform CAs when different rules are assigned to cells and used to update them.

Wolfram was the first to apply CAs to generate PNSs. He used uniform, 1D CAs with r = 1, and rule 30. Hortensius and Nandi et al. [20] used nonuniform CAs with two rules 90 and 150, and it was found that the quality of generated PNSs was better than the quality of the Wolfram system. Recently Tomassini and Perrenoud [12] proposed to use nonuniform, 1D CAs with r = 1 and four rules.



Number	7	6	5	4	3	2	1	0
Neighbourhood	111	110	101	100	011	010	001	000
Rule result	0	1	0	1	1	0	1	0

Figure 1. 1D cellular automata with neighbourhood = 1

90, 105, 150 and 165, which provide high quality PNSs and a huge space of possible secret keys which is difficult for cryptanalysis. Instead to design rules for CAs they used evolutionary technique called cellular programming (CP) to search for them.

In this study we continue this line of research. We will use finite, 1D, non uniform CAs. However, we extend the potential space of rules by consideration of two sizes of rule neighborhood, namely neighborhood of radius r = 1 and 2. To discover appropriate rules in this huge space of rules we will use CP.

IV. Cellular Programming Environment

A. *Cellular programming*

CP is an evolutionary computation technique similar to the diffusion model of parallel genetic algorithms and introduced [13] to discover rules for non uniform CAs. Fig.2 shows a CP system implemented [2] to discover such rules. In contrast with the CP used in [12] the system has a possibility to evaluate non uniform rules of two types. The system consists of a population of *N* rules (left) and each rule is assigned to a single cell of CAs (right). After initiating states of each cell, i.e. setting an initial configuration, the CAs start to evolve according to assigned rules during a predefined number of time steps. Each cell produces a stream of bits, creating this way a PNS.

After stopping evolving CAs all PNSs are evaluated. The entropy E_h is used to evaluate the statistical quality of each PNS. To calculate a value of the entropy each PNS is divided into subsequences of a size *h*. In all experiments the value *h* = 4 was used. Let *l* be the number of values which can take each element of a sequence (in our case of binary values of all elements *l* = 2) and l^h a number of possible states of each sequence (l^h = 16). E_h can be calculated in the following way:

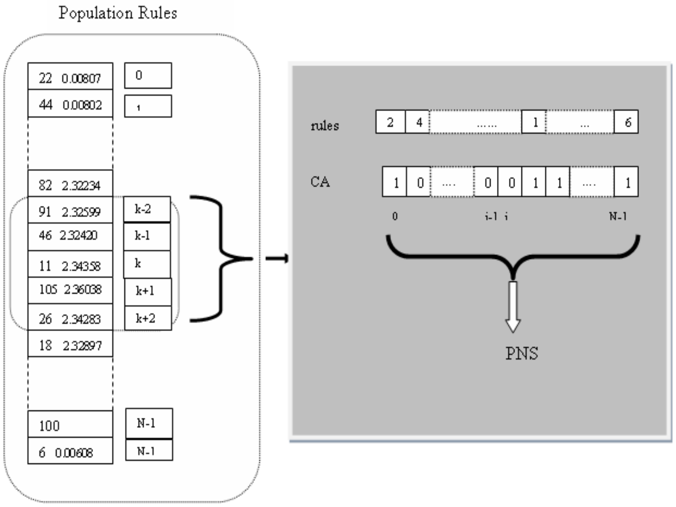


Figure 2. CP environment for evolution of rules of nonuniform CAs.

$$E_h = -\sum_{j=1}^{l^h} P_{hj} \log_2 P_{hj} \qquad (1)$$

where P_{hj} is a measured probability of occurrence of a sequence h_j in a PNS. The entropy achieves its maximal value $E_h = h$ when the probabilities of the k_h possible sequences of the length *h* are equal to $1/l^h$. The entropy will be used as a fitness function of CP.

A single PNS is produced by a CA cell according to assigned rules and depends on a configuration ci of states of CAs. To evaluate statistically reliable value of the entropy, CAs run with the same set of rules C times for different configurations ci, and finally the average value of entropy is calculated and serves as a fitness function of each rule from the population of rules.

After evaluation of a fitness function of all rules of the population genetic operators of selection, crossover and mutation are locally performed on rules. The evolutionary algorithm stops after some predefined number of generations of CP.

The algorithm can be summarized in the following way:

1. Initiate randomly *population* of *N* rules of type 1 (*r* = 1) or type 2 (r = 2), or both types, and create CAs consisting of *N* cells
2. Assign *k*ith rule from the CP population to *k*ith cell of CAs
3. **for** *i* = 1 . . . *C* **do** { create randomly configuration c_i of CAs evolve CAs during M time steps evaluate entropy of each PNS }
4. Evaluate fitness function of each rule
5. Apply locally to rules in a specified sequence genetic operators of selection, cross-over and mutation
6. If STOP condition is not satisfied return to 2.

B. Discovery of rules in 1D, non uniform CAs

In all conducted experiments a population of CP and the size of non uniform CAs were equal to 50 and the population was processing during 50 generations. The CAs with initial random configuration of states and a set of assigned rules evolved during *M* = 4096 time steps. Running CAs with a given set of rules was repeated for *C* = 300 initial configurations. Fig. 3 shows an example of running CP for the evolutionary neighborhood *i* - 3, *i* - 2, *i*, *i* + 2, *i* + 3. One can see that whole CAs is able to produce very good PNSs after about 40 generations (see, the average value avg of the entropy close to 4).

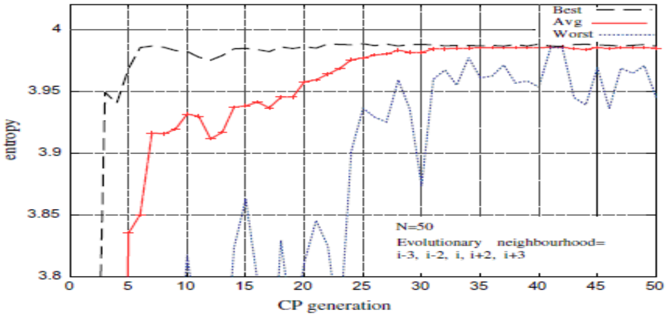


Figure 3. A single run of CP evolutionary process

A typical result of a single run of an evolutionary process starting with a random rules assigned to cells of CAs is discovering by CP a small set of good rules which divide the cellular space of CAs into domains-areas where the same rules, short (*r* = *1*) or long (*r* = *2*), live together (see Table 1). Evolutionary process is continued on borders of domains where different rules live. This process may result in increasing domains of rules which are only slightly better than neighboring rules, which domains will decrease and finally disappear.

This happens in particular when two neighboring domains are occupied respectively by the same short rules and the same long rules. The search space of short rules is much smaller than the search space of the long rules. Therefore better short rules are discovered faster than better long rules, and for this reason long rules are gradually replaced by short rules. To limit this premature convergence of short rules, the short and long rules are initially randomly assigned to cells in the proportion of 1:3 in all subsequent experiments.

The purpose of the experiments which followed was to discover an enlarged set of rules (to enlarge the key space of cryptography system) which working collectively would produce very high quality PNSs. It was noticed that in a single run of CP the evolutionary algorithm produces typically a small set of rules with a very high value of the entropy. In the result of evolutionary searching process a set of 8 short rules (including 5 rules found by [16]) and a set of 39 long rules was found.

TABLE I. DOMAINS IN A FINAL POPULATION OF A EVOLUTIONARY PROCESS

Rule	**Rule name**	**Fitness value**
Generation 50 01011010	90	3.98924
01011010	90	3.98943
01011010	90	3.98920
01011010	90	3.98981
00110011110011000011001111001100	869020620	3.98924
00110011110011000011001111001100	869020620	3.98959
00110011110011000011001111001100	869020620	3.98940
00110011110011000011001111001100	869020620	3.98906
00110011110011000011001011001100	869020364	3.94157
00110011110011000011001111001100	869020620	3.98960
00110011110011000011001111001100	869020620	3.98952
00110011110011000011001111001100	869020620	3.98929
00110011110011000011001111001100	869020620	3.98931
00110011110011000011001111001100	869020620	3.98933
00110011110011000011001111001100	869020620	3.98955
00110011110011000011001111001100	869020620	3.98964
00110011110011000011001111001100	869020620	3.98911
00110011110011000011001111001100	869020620	3.98941
00110011110011000011001111001100	869020620	3.98952
00110011110011000011001111001100	869020620	3.98933
00110011110011000011001111011100	869020636	3.97190
00110011110011000011001111001100	869020620	3.98981
00110011110011000011001111001100	869020620	3.98940
00110011110011000011001111001100	869020620	3.98930
00110011110011000011001111001100	869020620	3.98978
00110011110011000011001111001100	869020620	3.98922
00110011110011000011001111001100	869020620	3.98922
00110011110011000011001111001100	869020620	3.98957
01011010		3.98977
00110011110011000011001111001100	869020620	3.98949
01011010		3.98971
01011010		3.98988
01011010		3.98950
01011010		3.98945
01011010		3.98934
01011010		3.98935
01011010		3.98897
01011010		3.98942
01011010		3.98961
01011010		3.98962
01011010		3.98960
01011010		3.98970
01011010		3.98962
01011010		3.98933
01011010		3.98943
01011010		3.98955
01011010		3.98927
01011010		3.98925
01011010		3.98935
01011010		3.98948
Global fitness of automata: 3.99976		

V. ANALYSIS AND COMPARISON OF RESULTS

The entropy used as the fitness function for evolution CA rules producing high quality PNSs is only one of existing statistical tests of PNSs. None of them is enough strong to claim statistical randomness of a PNS in the case of passing a given test. For this purpose uniform CAs consisting of 50 cells evolved during 65536 time steps with each single discovered rule. Each PNS produced by CAs was divided into 4-bit words and tested on general statistical tests such as the entropy, v2 test, serial correlation test [6] (some weaker rules after this testing were removed).

The best scores were achieved by rules 30, 86, 101, 153 and by 8 long rules. Rules 90,105,150 and 65 working separately in uniform CA obtained good results in test of entropy and long runs test, quite good results in serial correlation test and monobit test but were week in X2 test ,poker test, runs test.sult weak in v2 test, poker test and runs test. However this set of rules work-ing collectively in non uniform CAs achieves good results (see, Table 2). For this reason only 10 rules were removed from discovered set of rules which have passed the FIPS 140-2 standard testing. These rules were worse than Tomassini and Perrenoud rules. However passing all statistical tests does not exclude a possibility that the PNS is not suitable for cryptographic purposes. Before a PNS is accepted it should pass special cryptographic tests. Therefore rules which passed tests were next submitted to a set of Marsaglia tests [7]—a set of 23 very strong tests of randomness implemented in the Diehard program. Only 11rules passed all 23 Marsaglia tests. These are short rules 30, 86, 101, and long rules 869020563, 1047380370, 1436194405, 1436965290, 1705400746, 1815843780, 2084275140 and 2592765285.

The purpose of the last set of experiments was a selection of a small set of short and long rules for non uniform CAs which working collectively would provide a generation of very high quality PNSs suitable for the secret key cryptography. Simple combination of different rules which passed all Marsaglia tests in non uniform CAs have shown that resulting PNSs may have worse statistical characteristic than PNSs obtained using uniform CAs. On the other hand, experiments with Tomassini and Perrenoud rules show that rules that separately are working worse can provide better quality working collectively. For these reasons rules 153 and some long rules which obtained very good results in general tests but not passed all Marsaglia tests were also accepted for the set of rules to search a final set of rules. In the result of combining rules into sets of rules and testing collective behavior of these sets working in no non uniform CAs the following set of rules has been selected: 86, 90, 101, 105, 150, 153, 165 (*r = 1*), and 1436965290 (*r = 2*). Among the rules are 4 rules discovered in [16]. The set of found rules have been tested again on statistical and cryptographic tests using non uniform CAs with random assignment of rules to CA cells. Table II presents the results of testing this new set of rules and compares the results with ones obtained for Tomassini and Perrenoud rules. One can see that results of testing both sets on general tests and FIPS 140-2 tests are similar. However, the main difference between these results can be observed in passing Marsaglia test.

The secret key K which should be exchanged between two users of considered CA- based cryptosystem consists of a pair of randomly created vectors: the vector R_i informing about assigning 8 rules to N cells of CAs and the vector C(0) describing an initial binary state of CA cells. The whole key space has therefore the size 8^N x 2^N. The key space is much larger than the key space (4^N x 2^N) of 1D CA-based system [16]. Therefore the proposed system is much more resistant for cryptographic attacks.

TABLE II. COMPARISON OF RULES FOUND BY TOMASSINI AND PERRENOUD [12] AND NEW SET OF DISCOVERED RULES

Test	**Tomassini and Perrenoud rules (90, 105, 150, 165)**	**Discovered rules (86, 90, 101,105, 150, 153, 165,1436965290)**
Min entropy	3.9988	3.9987
Max entropy	3.9998	3.9997
Min v^2	5.0254	6.998
Max v^2	26.396	30.805
Min correlation	0.00007	–0.00006
Max correlation	0.02553	0.01675
Monobit test	50	50
Poker test	50	50
Run test	50	50
Long run test	50	50
Number of passed Marsaglia tests	11	23

VI. CONCLUSIONS

CA are an attractive approach for cryptographic applications. They are simple, modular logic systems that can generate good quality pseudorandom bit streams as required in robust cryptographic systems. In the paper we have reported results of the study on applying CAs to the secret key cryptography. The purpose of the study was to discover a set of CA rules which produce PNSs of a very high statistical quality for a CA-based cryptosystem which is resistant on breaking a cryptography key. The main assumption of our approach was to consider non uniform 1D CAs operating with two types of rules. Evolutionary approach called CP was used to discover suitable rules. After discovery of a set of rules they were carefully selected using a number of strong statistical and cryptographic tests. Finally, the set consisting of 8 rules has been selected. Results of experiments have shown that discovered rules working collectively are able to produce PNSs of a very high quality outperforming the quality of known 1D CA-based secret key cryptosystems, which also are much more resistant for breaking cryptography keys that know.

REFERENCES

[1] A. Menezes, P. van Oorschot, S. Vanstone, Handbook of Applied Cryptography, CRC Press, 1996

[2] A. Mroczkowski, Application of cellular automata in cryptography, Master Thesis, Warsaw University of Technology, 2002 (in Polish).

[3] A.K Das, A. Sanyal, and P.P. Chaudhuri, “On the Characterization of Cellular Automata, ” Information Science, 1991.

[4] A.K Das and P.P Chaudhuri, “Efficient Characterization of Cellular Automata”, Proc. IEE(Part E), IEE, Stevenge, U.K., Vol. 137, Jan. 1990, pp. 81-87.

[5] B. Schneier, Applied Cryptography, Wiley, New York, 1996.

[6] D.E. Knuth, The Art of Computer Programming, in: Seminumerical Algorithms, vols. 1 and 2, Addison-Wesley, 1981.

[7] G Marsaglia, Diehard. Available from http://stat.fsu.edu/~geo/diehard.html (1998).

[8] H. Gutowitz, Cryptography with dynamical systems, in: E. Goles, N. Boccara (Eds.), Cellular Automata and Cooperative Phenomena, Kluwer, 1993. [16]. H. Nishio. "Real Time Sorting of Binary Numbers by One-dimensional Cellular Automata", Technical Report, Kyoto Univ., Japan, 1981.

[9] J. Kari, Cryptosystems based on reversible cellular automata, Personal communication, 1992.

[10] M. Mitchell, An Introduction to Genetic Algorithms (Complex Adaptive Systems), MIT Press, ISBN: 0262133164.

[11] M. Sipper, M. Tomassini, Generating parallel random number generators by cellular programming, International Journal of Modern Physics C 7 (2) (1996) 181-190.

[12] M. Tomassini, M. Perrenoud, Stream ciphers with one- and two-dimensional cellular automata, in: M. Schoenauer et al. (Eds.), Parallel Problem Solving from Nature--PPSN VI, LNCS 1917, Springer, 2000, pp. 722-731.

[13] P. Sarkar, A brief history of cellular automata, ACM Computing Surveys 32 (1) (2000) 80-107.

[14] P. Guan, Cellular automaton public-key cryptosystem, Complex Systems 1 (1987) 51-56.

[15] P. Pal Chaudhuri, D. Roy Chowdhury, S. Nandi, and S. Chatterjee, "Additive Cellular Automata – Theory and Applications", volume 1. IEEE Computer society Press, CA, USA, ISBN 0-8186-7717-1,1997.

[16] R. Sommerhalder and S.C van Westrhenen, "Parallel Language Recognition in Constant Time by Cellular Automata", Acta Informatica, Vol. 6, 1983, ppl 397-407.

[17] S. Wolfram, Cryptography with cellular automata, in: Advances in Cryptology: Crypto'85 Proceedings, LNCS 218, Springer,

[18] S. Wolfram. "Statistical Mechanics of Cellular Automata". Rev. Mod. Phys., Vol. 55, July 1983, pp. 601–644.

[19] S. Chakraborty, D.R. Chowdhury, and P.P. Chaudhuri, "Theory and Application of Non-Group Cellular Automata for Synthesis of Easily Testable Finite State Machines", IEE Trans. Computers, Vol. 45, No. 7, July 1996, pp. 769-781.

[20] S. Nandi, B.K. Kar, P.P. Chaudhuri, Theory and applications of cellular automata in cryptography, IEEE Transactions on Computers 43 (1994) 1346-1357.

[21] Lalith, T. (2010). Key Management Techniques for Controlling the Distribution and Update of Cryptographic keys. International Journal of Advanced Computer Science and Applications - IJACSA, 1(6), 163-166.

[22] Hajami, A., & Elkoutbi, M. (2010). A Council-based Distributed Key Management Scheme for MANETs. International Journal of Advanced Computer Science and Applications - IJACSA, 1(3).

[23] Meshram, C. (2010). Modified ID-Based Public key Cryptosystem using Double Discrete Logarithm Problem. International Journal of Advanced Computer Science and Applications - IJACSA, 1(6).

[24] Chakraborty, R. (2011). FPGA Based Cipher Design & Implementation of Recursive Oriented Block Arithmetic and Substitution Technique (ROBAST). International Journal of Advanced Computer Science and Applications - IJACSA, 2(4), 54-59.

AUTHORS PROFILE

Bijayalaxmi Kar is a scholar of M.Tech (CSE) at College of Engineering, BijuPattanaik University of Technology, Bhubaneswar, Odisha, INDIA. Her research areas include Algoritm analysis, automata theory, Soft Computing Techniques etc.

D Chandrasekhar Rao is as Assistant Professor in the department of Computer Sc. & Engineering, College of Engineering Bhubaneswar (CEB). He received his Masters degree from Biju Pattanaik University Technology, Bhubaneswar. His research area includes Compiler design automata theory etc.

Dr.Amiya Kumar Rath obtained Ph.D in Computer Science in the year 2005 from Utkal University for the work in the field of Embedded system. Presently working with College of Engineering Bhubaneswar (CEB) as Professor of Computer Science & Engg. Cum Director (A&R) and is actively engaged in conducting Academic, Research and development programs in the field of Computer Science and IT Engg. Contributed more than 30 research level papers to many national and International journals. and conferences Besides this, published 4 books by reputed publishers. Having research interests include Embedded System, Adhoc Network, Sensor Network, Power Minimization, Biclustering, Evolutionary Computation and Data Mining.

Computerised Speech Processing in Hearing Aids using FPGA Architecture

V. Hanuman Kumar
Department of CS & SE
Andhra University,
Visakhapatnam -530003, India

Prof. P. Seetha Ramaiah
Department of CS & SE
Andhra University,
Visakhapatnam -530003, India

Abstract— **The development of computerized speech processing system is to mimic the natural functionality of human hearing, because of advent of technology that used Very Large Scale Integration (VLSI) devices such as Field Programmable Gate Array (FPGA) to meet the challenging requirement of providing 100% functionality of the damaged human hearing parts. Here a computerized laboratory speech processor based on Xilinx Spartan3 FPGA system was developed for hearing aids research and also presented comparison details of the commercially available Hearing Aids. The hardware design and software implementation of the speech processor are described in detail. The FPGA based speech processor is capable of providing high-rate stimulation with 12 electrodes against conventional 8 electrodes in earlier research. Using short biphasic pulses presented either simultaneously or in an interleaved fashion. Different speech processing algorithms including the Continuous Interleaved Sampling (CIS) strategy were implemented in this processor and tested successfully.**

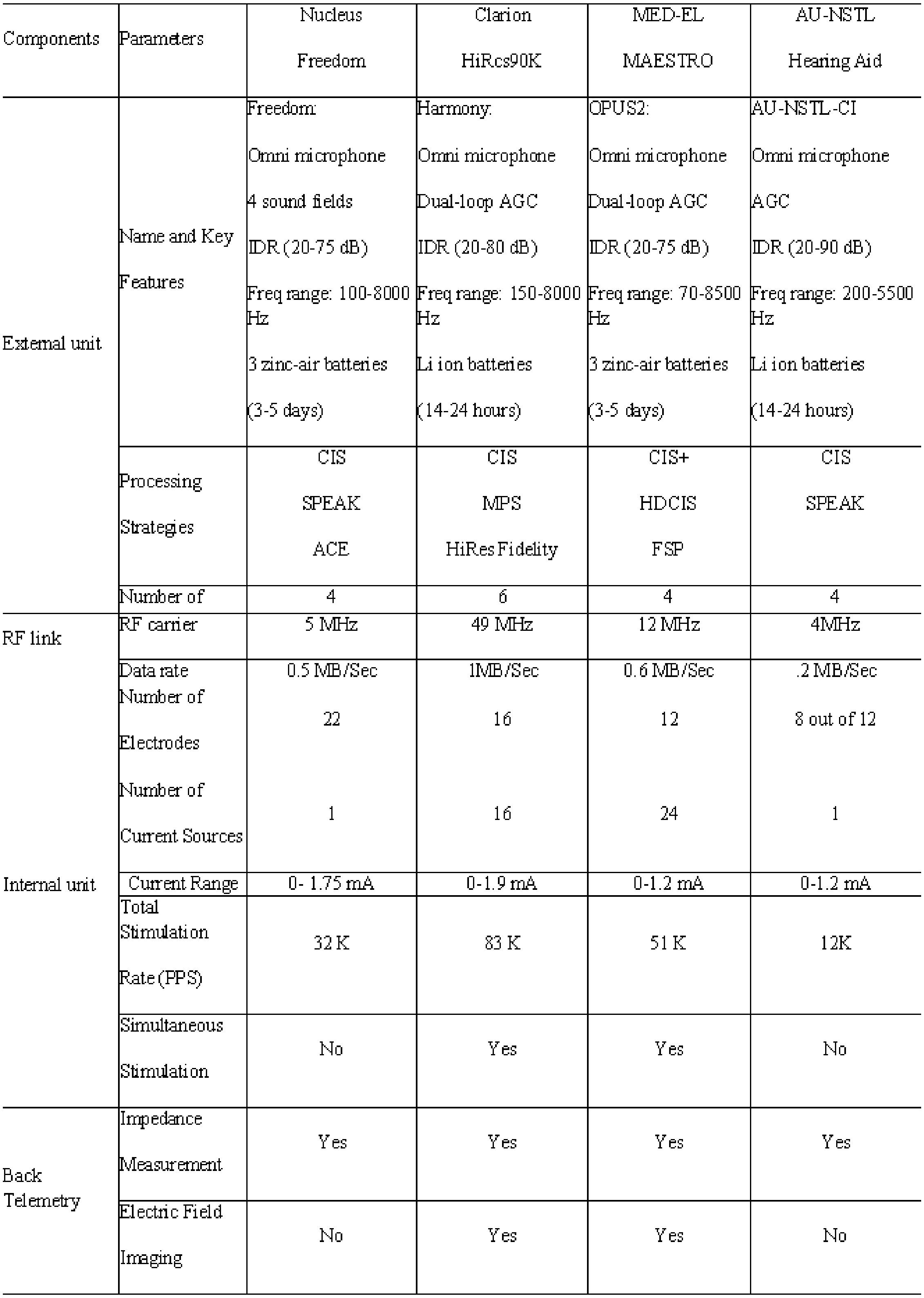
Keywords- Speech processing system; VLSI; FPGA; CIS.

I. Introduction

Today digital signal processing applications (DSP), like e.g. speech processing or image processing, need considerable computing performance in order to work in real time. In some cases the performance of DSPs (Digital Signal Processor) is not sufficient to guarantee real time behavior. In these cases the problem is often solved by using dedicated hardware to do some pre-processing of the data or even implement the whole signal processing system. Since hard-wired ASICs (Application-Specific Integrated Circuits) are too expensive and not flexible enough, FPGAs (Field-Programmable Gate-Arrays) have proven to be a viable alternative for implementing DSP algorithms in hardware in the last decade. FPGAs belong to the class of programmable logic devices (PLD) and are reprogrammable or reconfigurable. The reason why hardware implementations of DSP algorithms show a superior performance in many cases compared to a software implementation on a DSP is a different computing paradigm. Compared to a DSP, which implements basically the sequential “von Neumann” computing principle (computing- in-time), FPGAs, can implement algorithms with much stronger parallelism; since the hardware is inherently parallel (computing-in-space). Unfortunately the design of FPGA hardware needs normally more effort than a software implementation on a DSP. Furthermore the development tools like hardware description languages (HDL) e.g. VerilogHDL or VHDL, simulators and synthesis and the way in which the implementation is described differs much from what the system or software engineers are used to, like e.g. describing the problem in tools like Mat lab or in languages like C/C++.

Hearing Aids are devices that provide partial hearing to deaf people through electrical stimulation of the auditory nerve (Wilson *et al*, 1988). A typical device consists of a speech processor, a behind-the-ear unit and an implantable device (Tierney*et al*, 1994). Sound is picked up by a microphone in the behind the- ear unit and sent to the speech processing device. Depending on the speech processing algorithm being implemented, the speech processor extracts various parameters from the input signal and determines the amplitude of the current pulses to be sent to the implant device (Loizou 1998). The amplitude information is transmitted to a receiver/stimulator circuit implanted under the scalp of the patient. This circuit sends current pulses to the electrodes implanted in the cochlea of the patient. As a result, the auditory nerve is excited and transmits nerve pulses to the brainstem and brain, where they are interpreted as sound (Ay *et al*, 1997). Hearing Aids (Cochlear Implants) are not available to most of the world’s deaf and severely hearing impaired people due to the high cost (around $30,000 USD).

Table 1: Comparison of Various Cochlear Implant Systems



Components	Parameters	Nucleus Freedom	Clarion HiRcs90K	MED-EL MAESTRO	AU-NSTL Hearing Aid
External unit	Name and Key Features	Freedom: Omni microphone 4 sound fields IDR (20-75 dB) Freq range: 100-8000 Hz 3 zinc-air batteries (3-5 days)	Harmony: Omni microphone Dual-loop AGC IDR (20-80 dB) Freq range: 150-8000 Hz Li ion batteries (14-24 hours)	OPUS2: Omni microphone Dual-loop AGC IDR (20-75 dB) Freq range: 70-8500 Hz 3 zinc-air batteries (3-5 days)	AU-NSTL-CI Omni microphone AGC IDR (20-90 dB) Freq range: 200-5500 Hz Li ion batteries (14-24 hours)
	Processing Strategies	CIS SPEAK ACE	CIS MPS HiRes Fidelity	CIS+ HDCIS FSP	CIS SPEAK
	Number of	4	6	4	4
RF link	RF carrier	5 MHz	49 MHz	12 MHz	4MHz
	Data rate	0.5 MB/Sec	1MB/Sec	0.6 MB/Sec	.2 MB/Sec
Internal unit	Number of Electrodes	22	16	12	8 out of 12
	Number of Current Sources	1	16	24	1
	Current Range	0- 1.75 mA	0-1.9 mA	0-1.2 mA	0-1.2 mA
	Total Stimulation Rate (FPS)	32 K	83 K	51 K	12K
	Simultaneous Stimulation	No	Yes	Yes	No
Back Telemetry	Impedance Measurement	Yes	Yes	Yes	Yes
	Electric Field Imaging	No	Yes	Yes	No

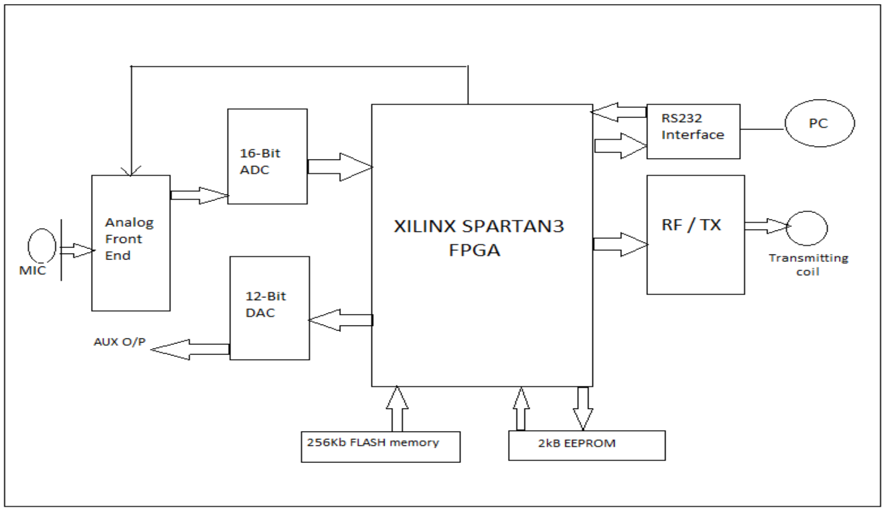


fig:1. Functional Block Diagram for Speech Processing

It is most unfortunate, as approximately 80% of the world's hearing impaired people live in developing countries, with highly limited healthcare budgets and widespread poverty with annual household income around $600 USD, cannot possibly afford the high cost of currently available implant systems. Many researchers are attempting to develop low cost but effective cochlear implant devices [Wilson, 1998], [Kim, 2007]. The three commercially available cochlear implant systems namely, Nucleus Freedom, Clarion HiRes 90K, MEDEL MAESTRO are compared in respect of the parameters: processing strategies, number of maps, RF carrier, data rate, number of electrodes, number of current sources, current range, total stimulation rate, simultaneous stimulation, impedance measurement and electric filed imaging with our developed AU-NSTL Cochlear Implant System [Fan-Gang Zing, 2008]. It is identified that the performance is similar to compared Cochlear implant Systems and parameter comparison information is given in Table 1.

II. Speech Processor Hardware

A laboratory design of computerized speech processor for the Hearing Aids can accept speech or audio signals and transforms them into human understandable processed speech or audio to an implantable cochlear receiver-stimulator of 12 electrodes for making the deaf person to understand the speech or audio tones is designed, developed and evaluated. The working principle of the speech processing system involves capturing sound from the environment, processing sound into digital signals and delivering sound to the hearing nerve via electrode array in cochlea. The core of the system is Speech Processor (Xilinx spartan3 FPGA), 16-bit ADC,12-bit DAC, analog front end circuit includes a fixed gain amplifier, programmable gain amplifier ,an EEPROM and a serial programming interface shown in fig1. The speech processing system can drive a hearing aid receiver stimulator and excite 8-channel electrode array. In a typical application the system works in the following way. Sound waves are converted to electrical signals by the microphone and then fed to analog front-end circuit. An electric condenser microphone can be connected to the front panel auxiliary input socket.

The sound signals are amplified by fixed gain amplifier with a fixed gain of 30dB and based on volume control of speech processing device, programmable gain amplifier amplifies the output of the fixed gain amplifier and then amplified signal is attenuated by the sensitivity potentiometer and then passes to a fixed gain amplifier in the analog front-end circuit with a fixed gain of 30dB. The signal is filtered to eliminate noise before being converted to a 16-bit digital value by the 16-bit ADC, and transmitted to the Xilinx spartan3 FPGA based speech processor in 16-bit zero-padded frames via a serial interface.

A software-controlled programmable gain stage allows independent gain for each channel going into the ADC. The ADC's output can be digitally mixed with the DAC input. The ADC samples the filtered signals at a rate of 64K samples / sec. The 16-bit ADC and 12-bit DAC section is interfaced to Xilinx spartan3 FPGA via Serial Communication Interface (SCI). The FPGA based digital speech processor continually runs a program individually configured for the hearing-impaired user and their hearing prosthesis, is interrupted for every new data sample is received by the processor's SCI. The Xilinx spartan3 FPGA typically stores this sample in memory for future processing and may transfer a modified sample back to the SCI to be transmitted to the DAC, where the sample is converted to an analog signal to drive a hearing aid receiver. For auditory prostheses use, the Programmable Xilinx spartan3 FPGA based processor periodically construct data frames in the special format required for the cochlear implant receiver stimulator which determines the active electrodes and their current levels and then sends the encoded data frames serially with 171 Kbps rate to the RF transmitter.

The RF transmitter is based on ASK modulation and operates at 4MHz carrier frequency. The RF transmitter modulates the incoming encoded serial data and fed to the RF transmitting coil. The RF transmitting coil is seventeen turns, 175 strands Litz wire with High Q value. RF transmitter sends the data frames via the transcutanious inductive coupling to the receiver coil in the laboratory model receiver-stimulator. The receiver stimulator decodes the data and activates the specified electrodes, which stimulate nearby auditory neurons, giving the user the sensation of hearing. The Xilinx spartan3 FPGA is used as the central processor running at a rate of 326 MHz of core clock frequency and Densities as high as 74,880 logic cells. Up to 1872 Kbits of total block RAM ,up to 520Kbits of distributed RAM, Three separate power supplies for the core (1.2V), IOs (1.2V to 3.3V), and Special function(2.5V) eliminating the need for power- consuming external RAM in cochlear implant applications. The on-chip peripherals consist of SCI-a Serial Communications Interface, a parallel Host Interface, and a Timer Module and relevant control signals are used to access external memories, as well as the encoder in this application. Programs are stored in an external EEPROM and the Xilinx spartan3 FPGA is programmed to boot from EEPROM. The mode pin logic levels are automatically altered when the programming cable is connected to the processor. The speech processing Software CIS for DSPMAP is developed in personal computer (PC) using VerilogHDL and configured Xilinx spartan3 FPGA using JTAG cable using Xilinx ISE Environment.

III. SPEECH PROCESSOR SOFTWARE

The Hearing Aids are mainly classified into 2 categories Single channel and multi channel Aids.

A. Single Channel Implants

Single channel Hearing Aids consist of a single pair of electrodes inserted in the cochlea. The main advantage of such a device is its simplicity, low cost and small size. The small size could enable it to be developed as a complete Behind the Ear (BTE) system. The two commercially produced single channel Implantable Hearing Aids were the House/3M single channel implants and Vienna/3M single channel implants.

B. Multi Channel Implants

Single electrode cochlear implants are failed mainly due to stimulate only a particular place in the cochlea due to the single electrode used. Thus single electrode cochlear implants can only provide very limited frequency information, since they use only one electrode and perform simple spectral analysis. To better exploit the place/frequency mechanism found in the peripheral auditory system, multi-channel cochlear implants were developed. Multi channel implants provide electrical stimulation at multiple sites in the cochlea using an array of electrodes. An electrode array is used so that different auditory nerve fibers can be stimulated at different places in the cochlea, thereby exploiting the place mechanism for coding frequencies. Different electrodes are stimulated depending on the frequency of the signal. Electrodes near the base of the cochlea are stimulated with high-frequency signals, while electrodes near the apex are stimulated with low-frequency signals.

When researcher's starts designing of multi-channel implants, several issues are raised such as (i) how many electrodes should be used and how many channels are needed to obtain maximum speech understanding? And (ii) as more than one electrode will be stimulated, what kind of information should be transmitted to each electrodes? Researchers experimented with different numbers of electrodes, some are used with more number of electrodes (22) with selected stimulation, while other used a fewer number of electrodes (4-8) and stimulated all of them. Based on how researchers tried to address the second question, different types of speech processing algorithms are evolved. The various speech processing algorithms developed for multi channel implantable Hearing Aids can be divided into three main categories: waveform based speech processing algorithms, feature-extraction based speech processing algorithms and hybrid approach. These algorithms differ in the way the information is extracted from the incoming speech signal and presented to the electrodes for stimulation. The waveform based speech processing algorithms try to present some type of waveform (in analog or pulsatile form) derived from the speech signal by filtering into different frequency bands, while the feature-extraction based speech processing algorithms are try to present some type of spectral features, such as formants, derived using feature extraction algorithms. Hybrid algorithms presents the utilizing both algorithms. A brief coverage of these speech processing algorithms is given in Figure 2.Here the speech processor design using FPGA architecture system software comprises of two important modules, namely Programmable Speech Processing modules and Speech Data Encoding modules as shown in figure3.

The Programmability of the speech processing system design described herein provides the means to develop and test 8 Channel CIS speech processing algorithms. It provides flexibility and programmability according to patient's active electrodes. By using the impedance telemetry and clinical programming software, the audiologist identifies the active electrodes and sends this information to the Speech Processing module via Speech Data Encoding module of Xilinx spartan3 FPGA device. Based on this information Speech Processing Module implements the CIS algorithm for 4 to 8 channel. For example Encoding module sends the information to process 5 channel CIS algorithm, the speech processor changes the coefficients for 5-channel Band Pass filter and process the 5 channel CIS algorithm. The Band pass filters are configured so that the cutoff frequencies are adjusted depending on the number of active electrodes as shown in Table 2 and observe the block diagram of CIS algorithm in the figure4.

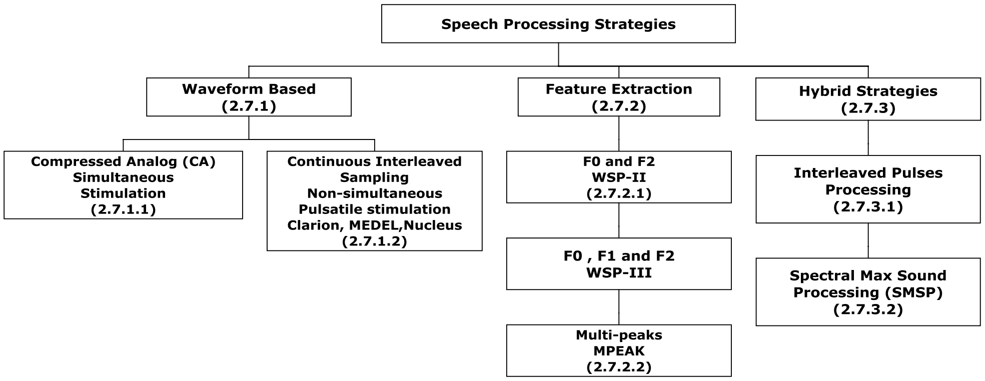

Speech Processing Strategies
Waveform Based (2.7.1)
Feature Extraction (2.7.2)
Hybrid Strategies (2.7.3)
Compressed Analog (CA) Simultaneous Stimulation (2.7.1.1)
Continuous Interleaved Sampling Non-simultaneous Pulsatile stimulation Clarion, MEDEL,Nucleus (2.7.1.2)
F0 and F2 WSP-II (2.7.2.1)
F0 , F1 and F2 WSP-III
Multi-peaks MPEAK (2.7.2.2)
Interleaved Pulses Processing (2.7.3.1)
Spectral Max Sound Processing (SMSP) (2.7.3.2)

Figure2: Various Types of Speech Processing Algorithms

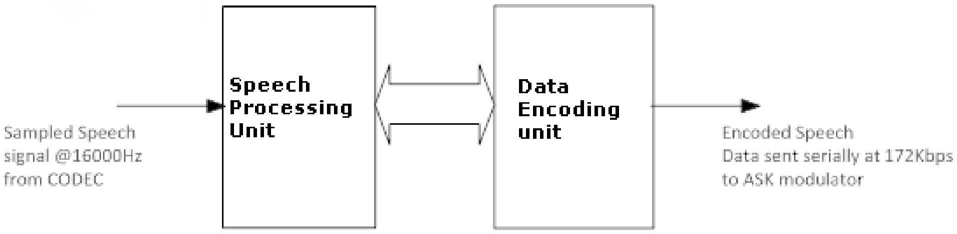

Sampled Speech signal @16000Hz from CODEC
Speech Processing Unit
Data Encoding unit
Encoded Speech Data sent serially at 172Kbps to ASK modulator

Figure3: Functional Modules of the Speech Processing Design

Table 2: Cut-off frequencies of Channels

Channel number	Lower frequency	Upper frequency	Center frequency
1	200	310	255
2	310	479	394
3	479	742	611
4	742	1149	946
5	1149	1779	1464
6	1779	2754	2266
7	2754	4263	3508
8	4263	6600	5432

Acoustic inputs from a microphone are directed to analog front-end (AFE) which amplifies the weak acoustic signals. The output from the AFE is Sampled with the sampling frequency of 64 KHz by the AD7683 and operated at 80MHz clock. The input samples which are collected by AD7683 converted to 16-bit digital data and send to Xilinx spartan3 FPGA (1.2V) via programmable interface between ADC and FPGA. The AD7683 interrupts Xilinx spartan3 FPGA whenever the input sample is available. These samples are windowed to produce a frequency response for the spectral analysis comparable with that of the commercially available auditory prostheses.

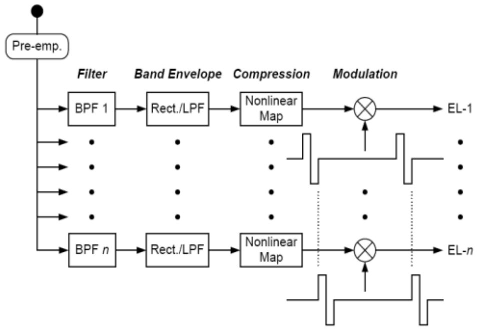


Figure4**:** Block Diagram of CIS Algorithm

The input signals are digitally filtered into 8 band-pass filters using Hamming window finite impulse response (FIR) filtering. Mat lab's Filter Design and Analysis tool (FDA Tool) is used to generate filter coefficients by using Hamming window of length 128bits. The input samples are stored in the circular buffer managed by Xilinx spartan3 FPGA. Each input acoustic sample is fed to a bank of band-pass channels. Each channel includes the stages of band-pass filtering, envelope detection, compression and modulation. The temporal envelope in each channel is extracted with full-wave rectifier followed by 32nd order low pass Hamming window FIR filter. The low-pass filters were designed to smooth the amplitude estimates in time, and were set to a cutoff frequency of 200Hz to allow maximal transmission of temporal envelope cues while avoiding aliasing when a relatively low carrier rates are used.

IV. PERFORMANCE

A test program was run on FPGA based speech processor with core frequency at 326 MHz as the main processing unit, the ADC sampling rate at 64 kHz and the encoding module of FPGA formatting and transmitting two data frames via the RF coil. A 128th order FIR program containing 8 band-pass filters runs on Xilinx Spartan3 FPGA processing unit. The input signal "FPGA" to the speech processor is received from the microphone and observed in digital Scope Coder as shown in fig-5 and corresponding responses from 8 channels is observed as shown in fig-6. Since the sampling rate is fixed at 64 KHz, we get 64000 samples for every second (i.e. 64 samples for every 1ms). These samples are fed to the filter bank, containing 8 band pass filters with frequencies ranging from 200Hz to 6600Hz. Rectified signals are generated from each filter, which are then fed to the low pass filter with cutoff frequency of 200Hz for envelope outputs. These 8 envelope signals are compressed using power law compression. These compressed signals of eight band-pass filters are transmitted to the encoding module in a base to apex order. This process is continuous and repeated for every 1.15ms. The FPGA based speech processor using Continuous Interleaved Sampling strategy has been run and tested with a laboratory model implant module.

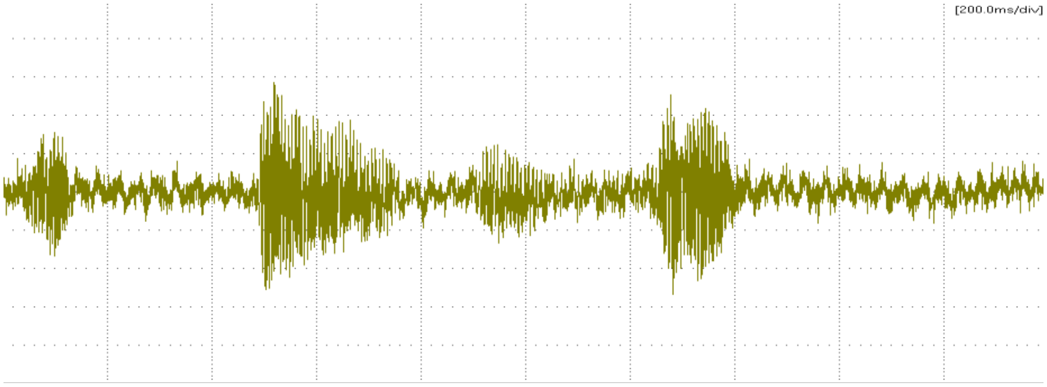


Fig.5: Input signal "FPGA"

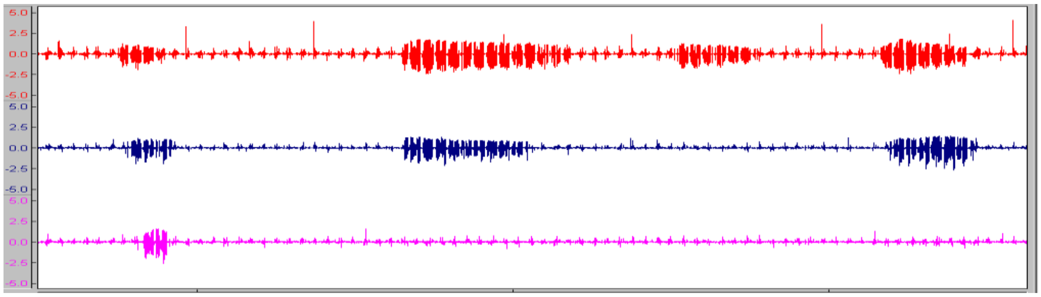


Fig6: Channel responses for signal "FPGA"

V. Conclusions

The Computerized Speech Processing in Hearing Aids using FPGA Architecture has been developed particularly for use with a hearing Prostheses. The flexibility and computational power of the developed system allows speech processing using CIS strategy, which is tested and evaluated. Speech processing schemes may be improved by including environmental noise, increase the no of channels and speech intelligibility optimization.

VI. Acknowledgements

The funding and technical support provided by NSTL, Visakhapatnam is acknowledged.

References

[1] Matlab-Homepage "www.mathworks.com"

[2] Ay, S., Zeng, F. -G. and Sheu, B. (**1997**). "Hearing with bionic ears: Speech processing strategies for cochlear implant devices,"IEEE Circuits & Devices Magazine, May 1997, 18-23.

[3] Boëx, C., Pelizzone, M. and Montandon, P. (**1996**). "Speechrecognition with a CIS strategy for the Ineraid multichannelcochlear implant," Am. J. Otol. **17**, 61-68.

[4] François, J., Tinembart, J., Bessat, C., Leone, P., Rossman, F.and Pelizzone, M. (**1994**). "Implants cochle aires: Un processeurportable pour le développement de l'algorithme CIS," in Actesde le conférence DSP 94, Paris

[5] Loizou, P. (**1998**). "Mimicking the human ear: An overview ofsignal-processing strategies for converting sound into electricalsignals in cochlear implants," IEEE Signal Processing Magazine**5**, 101-130.

[6] Loizou, P. and Poroy, O. (**1999**). "A parametric study of the CISstrategy," presented at the 1999 Conference on ImplantableAuditory Prostheses, Asilomar, Monterey, CA.

[7] [Clark, 2006] Graeme M. Clark, "The multiple-channel cochlear implant: the interface between sound and the central nervous system for hearing, speech, and language in deaf people—a personal perspective", Phil. Trans. R. Soc. B , 2006, Vol. 361,pp 791–810

[8] [Greenwood, 1990] Greenwood, D. D. (1990). "A cochlear frequency-position function for several species: 29 years later", Journal *of the Acoustical Society of America, Vol. 87, 1990, pp* 2592-2605.

[9] [Hirshorn, 1986] Michael S . Hirshorn, Dianne J. Mecklenburg, Judith A. Brimacombe, "Nucleus 22-channel cochlear implant: Preliminary observations", Journal of Rehabilitation Research and Development, April 1986, Vol . 23, No . 2, pp 27-33.

[10] Dua, S. (2010). PC And Speech Recognition Based Electrical Device Control. International Journal of Advanced Computer Science and Applications - IJACSA, 1(July), 46-49.

[11] Anifowose, F. A. (2010). A Comparative Study of Gaussian Mixture Model and Radial Basis Function for Voice Recognition. International Journal of Advanced Computer Science and Applications - IJACSA, 1(3), 1-9.

A Neural Approach for Reliable and Fault Tolerant Wireless Sensor Networks

Vijay Kumar (Member, IEEE)[1]
Department of Computer Engineering,
M. M. University, Mullana,
Ambala, Haryana, India- 133207

R. B. Patel (Member, IEEE)[2]
Department of Computer Engineering,
M. M. University, Mullana,
Ambala, Haryana, India- 133207

Manpreet Singh (Member,IEEE)[3]
Department of Computer Engineering,
M. M. University, Mullana,
Ambala, Haryana, India- 133207

Rohit Vaid[4]
Department of Computer Engineering,
M. M. University, Mullana,
Ambala, Haryana, India- 133207

Abstract— **This paper presents a neural model for reliable and fault tolerant transmission in Wireless Sensor Networks based on Bi-directional Associative Memory. The proposed model is an attempt to enhance the performances of both the cooperative and non cooperative Automatic Repeat Request (ARQ) schemes in terms of reliability and fault tolerance. We have also demonstrated the performances of both the schemes with the help of suitable examples.**

Keywords- Reliability; Fault tolerance; Bi-directional Associative Memory; Wireless Sensor Network.

I. INTRODUCTION

Wireless sensor networks (WSNs) [9][2] are the topic of intense academic and industrial studies. Research is mainly focused on energy saving schemes to increase the lifetime of these networks [4][5]. There is an exciting new wave in sensor applications-wireless sensor networking- which enables sensors and actuators to be deployed independent of costs and physical constraints of wiring. For a wireless sensor network to deliver real world benefits, it must support the following requirements in deployment: scalability, reliability, responsiveness, power efficiency and mobility.

The complex inter-relationships between these characteristics are a balance; if they are not managed properly, the network can suffer from overhead that negates its applicability. In order to ensure that the network supports the application’s requirements, it is important to understand how each of these characteristics affects the reliability.

A. *Scalability and Reliability*

Network reliability and scalability are closely coupled and typically they act against each other. In other words, it is very difficult to build a reliable ad hoc network as the number of nodes increases [7]. This is due to network overhead that comes with increased size of network. In ad hoc network, there is no predefined topology or shape. Therefore, any node wishing to communicate with other nodes should generate more control packets than data packets. Moreover, as network size increases, there is more risk that communication links get broken, which will end up with creating more control packets. In summary, more overhead is unavoidable in a larger scale wireless sensor network to keep the communication path intact.

B. *Reliability and power efficiency*

Power efficiency also plays a very important role in this complex equation. To design a low power wireless sensor network, the duty cycle of each node needs to be reduced. The drawback is that as the node stays longer in sleep mode [3] to save the power, there is less probability that the node can communicate with its neighbors and may also lower the reliability due to lack of exchange of control packets and delays in the packet delivery.

C. *Reliability and responsiveness*

Ability of the network to adapt quickly the changes in the topology is known as responsiveness. For better responsiveness, there should be more issue and exchange of control packets in ad hoc network, which will naturally result in less reliability.

D. *Mobility and reliability*

A wireless sensor network that includes a number of mobile nodes should have high responsiveness to deal with the mobility. The mobility effect on responsiveness will compound the reliability challenge.

Many applications for wireless sensor networks require immediate and guaranteed action; for example medical emergency alarm, fire alarm detection, instruction detection [6]. In these situations packets has to be transported in a reliable way and in time through the sensor network. Thus, besides the energy consumption, delay and data reliability becomes very relevant for the proper functioning of the network.

Direct communication between any node and sink could be subject only to just a small delay, if the distance between the source and the destination is short, but it suffers an important energy wasting when the distance increases. Therefore often mutihop short range communications through other sensor nodes, acting as intermediate relay, are preferred in order to reduce the energy consumption in the network. In such a scenario it is necessary to define efficient technique that can ensure reliable communication with very tight delay constraint. In this work we focus attention on the control of data and reliability in multihop scenario.

A simple implementation of ARQ is represented by the Stop and Wait technique that consists in waiting the acknowledgement of each transmitted packet before transmitting the next one, and retransmit the same packet in case it is lost or wrongly, received by destination [8].

We extend here this analysis by introducing the investigation of the delay required by the reliable data delivery task. To this aim we investigate the delay required by a cooperative ARQ mechanism to correctly deliver a packet through a multihop linear path from a source node to the sink. In particular we analyze the delay and the coverage range of the nodes in the path, therefore the relation between delay and the number of cooperative relays included in the forwarding process.

II. NEURAL NETWORK IN WSN

Recently there is a strong interest to use neural network in energy efficient methods of WSNs. Algorithm developed in artificial neural network can be easily developed to wireless sensor network platforms and can satisfy the requirement for sensor networks kike: simple parallel distributed computation distributed storage, data robustness & auto classification of sensor readings, fault tolerance & low computation. Neural networks can help through dimensionality reduction obtain from the outputs of the neural networks clustering algorithms, leads to lower communication cost & energy saving. The other reason to use neural network based methods in WSNs is the analogy between WSNs & ANNs. As authors [1] strongly believe that ANNs exhibit exactly the same architecture WSNs since neural networks compared to sensor nodes & communications corresponds to radio links classification techniques.

Cluster based routing are most frequently used energy efficient routing protocols in WSNs which avoid single gateway architecture through developing of network into several clusters, while cluster head of each cluster play the role of a local base station.

Sensor node spends maximum energy in data communication by radio unit. Data communication in WSN must be reliable & must consume maximum power. Every sensor node consists of multiple sensors embedded in the node & this is a source of data. Rather communicating these data streams directly to the neighbouring node or base station, these data streams used to be classified.

III. PERFORMANCE OF COOPERATIVE ARQ USING MARKOV MODEL

Fig. 1 shows the network structure with linear multihop path consist of source node (node n =1), destination (node n = N) and (N-2)*t intermediate relay nodes deployed at equal distance where t is the number of parallel path of intermediate relay nodes between source and destination. Each path is composed by Z = N – 1 links. Suppose that all the nodes have circular radio coverage with the same transmission range R_t. When a sensor transmits a packet, it is received by all the sensors in a listen state inside the coverage area of the sender.

When a packet is transmitted, it can be forwarded towards the destination by only those nodes which are closer to the destination than the transmitter.

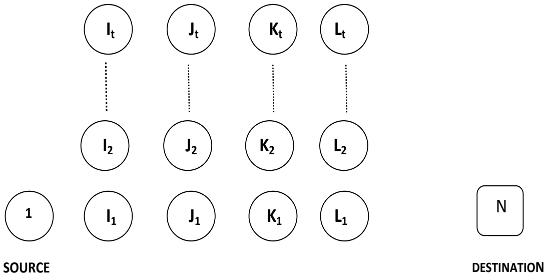


Fig.1 Network structure with Linear multihop path

A. *Discrete Parameter MARKOV CHAIN with Absorbing State*

Packet transfer from source to destination via intermediate forwarders can be treated as a state diagram of discrete parameter Markov chain with absorbing state. An absorbing state is a state from which there is zero probability of exiting. An absorbing Markov system is a Markov system that contains at least one absorbing state, and is such that it is possible to get from each non absorbing state to some absorbing state in one or more time steps. Consider p be the probability of successful transmission of a packet to an intermediate relay node inside the coverage range. Therefore 1-p will be the probability of unsuccessful transmission of packet.

For each node n, the probability to correctly deliver a packet to a node that is R_t links distant is equal to p. So the probability that the packet is not correctly received will be (1 – p), while it is correctly received from the immediately previous node with a probability p. Therefore the packet will be forwarded by the previous node with a probability (1 – p) p. If this node also has not correctly received the packet send by node n then the packet will be forwarded by the node previous to previous with a probability $(1 – p)^2$ p. If none of the node in the coverage area of the transmitter receives a correct packet it is necessary to ask the retransmission of the packet by the source node. It is possible to describe the process concerning one data packet forwarding from the source node n = 1 to the destination n = N with a discrete time Markov chain with absorbing state. Packet transmitted by a node will be further forwarded by a node in the coverage range of the transmitter which is furthest node from the source and has correctly received the packet.

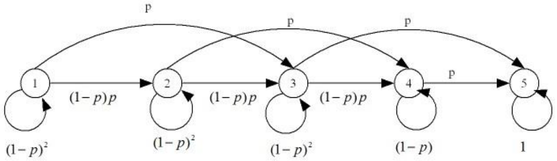


Fig. 2 Packet transmission in Cooperative ARQ

Consider a single multihop linear path consisting five sensors with four links as shown in Fig. 2. Assume transmission range of each sensor node is R_t=2 unit. State transition probability matrix for a successful transmission of a packet under cooperative automatic repeat request will be as under:

$$P_{Success} = \begin{bmatrix} (1-p)^2 & p(1-p) & p & 0 & 0 \\ 0 & (1-p)^2 & p(1-p) & p & 0 \\ 0 & 0 & (1-p)^2 & p(1-p) & p \\ 0 & 0 & 0 & (1-p) & p \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Similarly we can find the probability matrix for link error by replacing (1-p) with q. In fig. 2 states 1 through 4 are transient state while state 5 is an absorbing state.

In general, we consider a Markov chain with n states, s_1, s_2, …, s_n. S_n will be the absorbing state, and the remaining state will be transient. The transition probability matrix of such a chain may be partitioned so that

$$P = \left[--\frac{Q}{0}-- \mid --\frac{C}{1}-- \right]$$

Where Q is an (n-1) by (n-1) substochastic matrix, describing the probabilities of transition only among the transient states. C is a column vector and 0 is a row vector of (n-1) zeros. Now the k-step transition probability matrix P^k has the form

$$P^k = \left[--\frac{Q^k}{0}-- \mid --\frac{C^{'}}{1}-- \right]$$

Where C' is a column vector whose elements will be of no further use and hence need not be computed. The (i, j) entry of matrix Q^k denotes the probability of arriving in (transient) state s^j after exactly k steps starting from (transient) state s^i. It can be shown that $\sum_{k=o}^{t} Q^k$ converges as t approaches infinity. This imply that the inverse matrix $(I\text{-}Q)^{-1}$, called the fundamental matrix, M, exists and is given by

$$M = (I-Q)^{-1} = I + Q + Q^2 + \ldots = \sum\nolimits_{k=0}^{\infty} Q^k \quad .$$

The fundamental matrix is used for calculating the expected no. of steps to absorption. The number of times, starting in state i, and expected to visit state j before absorption is the ij^th entry of M. The total no. of steps expected before absorption equals the total no. of visits expected to make to all the non absorption states. This is the sum of all the entries in the i^th row of M.

Suppose p=0.8, then Q will be as under

$$Q_{Success} = \begin{bmatrix} .04 & .16 & .8 & 0 \\ 0 & .04 & .16 & .8 \\ 0 & 0 & .04 & .16 \\ 0 & 0 & 0 & .2 \end{bmatrix}$$

Therefore fundamental matrix $M = (I-Q)^{-1}$

$$= \begin{bmatrix} 25/24 & 25/144 & 775/864 & 305/864 \\ 0 & 25/24 & 25/144 & 155/144 \\ 0 & 0 & 25/24 & 5/24 \\ 0 & 0 & 0 & 5/4 \end{bmatrix}$$

Thus the states 1, 2, 3 and 4 are respectively executed 25/24, 25/144, 775/864, 305/864 times on the average. If t_1, t_2, t_3 and t_4 respectively is the time for one time execution of the states 1, 2, 3 and 4 then total time required to transmit a packet from source node 1 to destination node 5 is equal to :

T=25/24 t_1 + 25/144 t_2 + 775/864 t_3 + 305/864 t_4 unit times.

If t_1=t_2=t_3=t_4=t then T=2.4645 unit times.

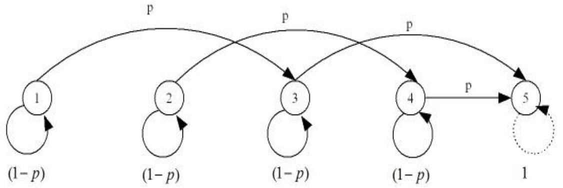


Fig 3. Packet transmission in Non-Cooperative ARQ as a discrete parameter Markov Chain with absorbing state

In non-Cooperative ARQ, a packet transmitted by source node is received by a node at distance R_t towards the destination from source and is forwarded by the node if packet received correctly otherwise transmitter is forced for retransmission. Other intermediate nodes between the transmitter and the node at distance R_t remains in sleep mode as they will never be involved in packet forwarding process. State transition probability matrix for successful transmission of the packet for non-cooperative ARQ will be as under:

$$P_{Success} = \begin{bmatrix} 1-p & 0 & p & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 1-p & 0 & p \\ 0 & 0 & 0 & (1-p) & p \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Suppose p=0.8, then Q will be as under

$$Q_{Success} = \begin{bmatrix} .2 & 0 & .8 & 0 \\ 0 & .2 & 0 & .8 \\ 0 & 0 & .2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Therefore fundamental matrix

$$M = (I-Q)^{-1} = \begin{bmatrix} 5/4 & 0 & 5/4 & 0 \\ 0 & 5/4 & 0 & 1 \\ 0 & 0 & 5/4 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Thus the states 1, 2, 3 and 4 are respectively executed 5/4, 0, 5/4, 0 times on the average if source node is considered as node 1. If t_1, t_2, t_3 and t_4 respectively is the time for one time execution of the states 1, 2, 3 and 4 then total time required to transmit a packet from source node 1 to destination node 5 is equal to:

T=5/4 t_1 + 5/4 t_3 units times.
If t_1=t_2=t_3=t_4=t then T=2.5 unit times.

It has been observed that packet delivery is more reliable and timely in case of cooperative ARQ, where as non cooperative ARQ is better in terms of power efficiency of sensor nodes as most of the sensors do not participate in packet forwarding process.

IV. System Model for Fault Tolerant and Reliable Transmission

In Co-operative ARQ scheme, a packet may not be successfully transmitted to the destination because of the following reasons; 1. Probable forwarders inside the coverage are in good working condition but have not received the packet correctly because of noise etc. and 2. All the forwarders are in failure condition.

Our proposed model is an attempt to still improve the fault tolerance & reliability of packet delivery process in WSN. This model addresses both the reasons of unsuccessful packet delivery in co-operative ARQ scheme. Proposed model is most suitable for those networks where the events that can occur in the network are from a predefined set of events. Assume that sensor node has to sense a packet from a set P={p_1, p_2, p_3, p_4, p_5, …, p_{10}} where the size of each packet is k-bit. Larger is the size of a packet, more will be the chances of faulty delivery of packet to the next probable forwarder/destination because of noise etc. Two possible solutions for reliable packet delivery at the next probable forwarder/destination are as under-

1. Data compression before transmission at the source for reducing the size of the packet to minimize the chances of faulty packet delivery at next probable forwarder/destination.

   If we compress the packet using ART1network, the packet needs to be splitted into small frames. These frames are fed into an ART1 network for self organization, and the output represents the class indices for these frames. After training, these indices and their corresponding prototypes are transmitted to the destination. The packet is then reconstructed according to these indices and prototypes. The benefit from this operation is measured by the compression ratio (Q), which is computed as follows: Q=NF/(C.N+F.$\log_2$C).

   Where N is the dimension of the frames, F is the total number of frames and C is the total number of classes formed during learning.

   On the other hand, the price we pay is measured by the distortion ratio defined as:

   $E = \sum_{j=1}^{F} \sum_{i=1}^{N} (I_{ij} - I'_{ij})^2$ Where I_{ij} is the value in original packet and I'_{ij} is the corresponding value in reconstructed packet. In packet compression, vigilance parameter is used to control the trade-off between Q and E.

2. Transmitting a small size packet (vector) associated to the original packet of very large size after encoding the associations between these hetro-associative vectors. Our proposed model is based on neural network concept. Transmitting a associated vector of small size has better reliability & fault tolerance over a compressed packet.

Discrete Bi-directional Associative Memory (BAM) neural network is used for encoding the associations. Bi-directional associative memory is a hetro associative recurrent neural network consisting of two layers as shown in Fig. 4. Layer-1 has n units and layer-2 has m units. Synaptic strength of various connections from the layer-1 to layer-2 is represented by the weight matrix W_{nXm} and the weight matrix from layer-2 to layer-1 is W_{nxm}^T. This is a fully connected network, wherein the inputs are different from the outputs. After training the network, if vector A is presented over the network then it recalls the associated vector B and similarly vector B recalls vector A if vector B is presented over the transpose weight matrix W^T.

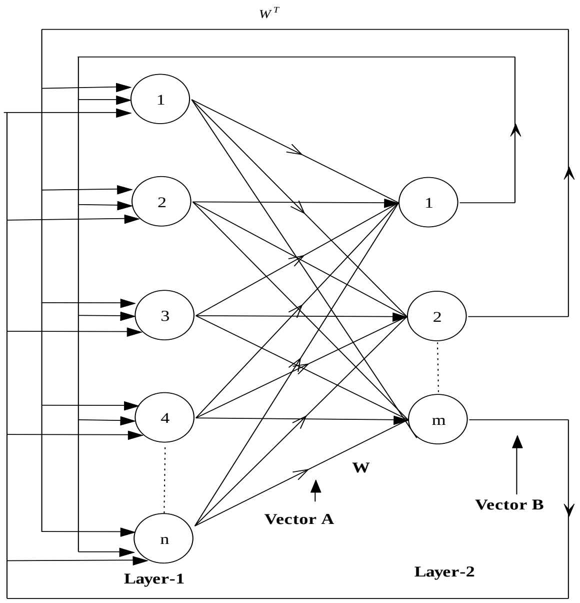
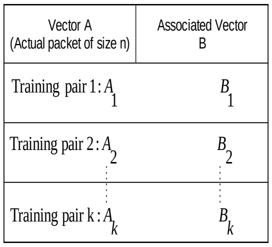


Fig4. Architecture of Bi-directional associative memory

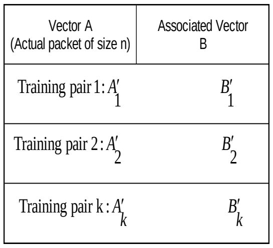
Training procedure (Encoding the associations):

Suppose we want to encode the association between the hetro-associative patterns for the given problem space.



Vector A (Actual packet of size n)	Associated Vector B
Training pair 1: A_1	B_1
Training pair 2: A_2	B_2
Training pair k: A_k	B_k

Binary Version



Vector A (Actual packet of size n)	Associated Vector B
Training pair 1: A'_1	B'_1
Training pair 2: A'_2	B'_2
Training pair k: A'_k	B'_k

Bipolar Version

After the training, weight matrix $W = \sum_{i=1}^{k} (A'_i)^T \; B'_i$.

In this proposed model BAM is embedded in every source node and destination node. At all the source nodes, before transmitting a packet A, it is presented over the BAM network through weight matrix W to get a smaller associated vector B_i. Associated vector B_i is transmitted and more successfully delivered at the probable next forwarder/destination, where it is again presented over the embedded BAM network through W^T to ensure retrieval of the actual packet A_i. If receiver is not the final destination then associated vector B_i is further transmitted towards the destination until destination is reached.

Activation function used in BAM is described as:-

$$\text{Output} = \text{f(NET)} = \begin{cases} 1, \text{if NET} > 0 \\ 0, \text{if NET} < 0 \\ \text{Previous output,if NET} = 0 \end{cases}$$

Where $\text{NET} = \begin{cases} \text{A.W} \quad \text{at Layer - 2} \\ \text{B.W}^{\text{T}} \quad \text{at Layer - 1} \end{cases}$

The memory capacity of BAM network is defined as min (n, m); where n is the number of units in layer-1 (size of vector A i.e. the original packet) and m is the number of units in layer-2 (size of vector B i.e. the vector to be transmitted). Size of the associated vector B_i corresponding to the original packet A_i depends on the total number of different packets (events) to encounter in the network irrespective of the size of original packet. Even if the packet size is very large (say n=10000 bits) and the total number of different packet to be handled by the network is small (say K=10) then associated vector B to be transmitted should be of the size m ≤ 10 bits for the reliable transmission of data. Even if the associated vector B has been received incorrectly at the next forwarder/destination, network may recall the actual packet despite of the corrupted signal received at next forwarded/destination (input signal to the BAM network) as network has the power of generalization. Moreover, BAM net is more fault tolerant with corrupted signal due to “missing” bits than that of “mistaken” bits.

This can be illustrated with the help of example 1:

Example 1:

Original packet	Associated packet for the transmission
A_1=[1 0 1 0 1 0 1 1]	B_1=[1 1 1]
A_2=[1 1 0 1 1 1 0 1]	B_2=[1 0 0]

Bi-Polar versions

A_1' =[1 -1 1 -1 1 -1 1 1]	B_1' =[1 1 1]
A_2' =[1 1 -1 1 1 1 -1 1]	B_2' =[1 -1 -1]

Weight matrix $W = \sum_{i=1}^{k} (A_i')^T \; B_i'$

$$= \begin{bmatrix} 2 & 0 & 0 \\ 0 & -2 & -2 \\ 0 & 2 & 2 \\ 0 & -2 & -2 \\ 2 & 0 & 0 \\ 0 & -2 & -2 \\ 0 & 2 & 2 \\ 2 & 0 & 0 \end{bmatrix}$$

Associated pattern corresponding to packet A_1 to be transmitted from the source node is computed as:

$NET = A_1'.W = [$ 6 10 10]
Output=F[NET]= [1 1 1] = B_1.

This pattern B_1 is transmitted towards the destination and is more reliably received by the next probable forwarder as the size of the pattern B_1 is smaller than the actual packet A_1. Received pattern is presented over the BAM network embedded at the next probable forwarder to ensure the correct delivery of the packet. Retrieval of the associated actual packet at the forwarder is shown as under:

$$NET = B_1'.W^T = [2 \; -4 \; 4 \; -4 \; 2 \; -4 \; 4 \; 2]$$
$$Output = [1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1] = A_1$$

In case forwarder/Destination receives the corrupted signal, BAM network has the power of generalization to recall the actual associated packet within its fault tolerance limit. This is illustrated in the example2.

Example 2:

Suppose signal received is faulty at the second missing bit i.e. B_1= [1 0 1]. The packet retrieval by the destination is shown as under

$$NET = B_1'.W^T = [2 \; -2 \; 2 \; -2 \; 2 \; -2 \; 2 \; 2]$$
$$Output = [1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1] = A_1$$

This shows that the packet has been retrieved correctly at the destination despite of corrupted signal. Similarly for fault at third missing bit i.e. B_1=[1 1 0] .

$$NET = B_1'.W^T = [2 \; -2 \; 2 \; -2 \; 2 \; -2 \; 2 \; 2]$$
$$Output = [1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1] = A_1$$

V. CONCLUSION

In this work we have presented the neural approach to analyze the performance of cooperative and non cooperative ARQ schemes in terms of delay and power efficiency. This model explores the advantages of both the schemes. With the embedded BAM at each sensor, packet delivery is more reliable and timely than that of the normal cases of cooperative and non cooperative ARQ as BAM is capable to retrieve the actual packet despite of corrupted associated pattern delivery at the forwarder within its fault tolerance limit.

VI. FUTURE SCOPE

Fault tolerance power of the proposed model may still be improved by cascading BAM structure with a recurrent network in asynchronous mode. Corrupted input pattern received at the next forwarder may be recovered partially by recurrent network bit by bit before directly applying to the BAM structure. A classification based neural network may also be used to further improve the efficiency of the proposed model.

REFERENCES

[1] Oldewurtel, Frank and Mahonen, Petri, (2006), "Neural Wireless Sensor Networks", International Conference on systems & Networks Communications, ICSNS'06, pp. 28-28 .

[2] Tubaishat M & Madria S, "Sensor networks: an overview" *IEEE Potentials Volume 22*, Issue 2, April- May 2003 pp: 20-23, 2003.

[3] C. F. Chiasserini and M. Garetto, "An analytical model for wireless sensor networks with sleeping nodes", *IEEE Trans. Mobile Computing*, vol. 5, no. 12, pp: 1706-1718, 2006.

[4] Pal, Y., Awasthi, L.K., Singh, A.J., "Maximize the Lifetime of Object Tracking Sensor Network with Node-to-Node Activation Scheme", *in Proceeding of Advance Computing Conference*, pp: 1200 – 1205, 2009.

[5] Yan-liang Jin, Hao-jie Lin, Zhu-ming Zhang, Zhen Zhang, Xu-yuan Zhang, "*Estimating the Reliability and Lifetime of Wireless Sensor Network", in Proceeding of Wireless Communications, Networking and Mobile Computing (WiCOM 2008)*, pp: 2181 – 2186, 2008.

[6] L. Bernardo, R. Oliveria, R. Tiago, P. Pinto, "A Fire Monitoring Application for Scattered Wireless Sensor Network", *in the proceeding of WinSys 2007*, on July 28-31, 2007.

[7] Wenyu Cai, Xinyu Jin, Yu Zhang, Kangsheng Chen, Jun Tang, "Research on Reliability Model of Large-Scale Wireless Sensor Networks", *in Proceeding of Wireless Communications, Networking and Mobile Computing (WiCOM 2006)* pp: 1-4, 2006.

[8] AboElFotoh, H.M.F., Iyengar, S.S., Chakrabarty, K, "Computing reliability and message delay for Cooperative wireless distributed sensor networks subject to random failures", *in IEEE Transactions on Reliability*, pp:145 – 155, 2005.

[9] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A Survey on sensor networks", *IEEE Communication Magazine*, Volume: 40 Issue:8, pp: 102-114, August 2002.

[10] Prasad, D. (2011). A Reliable Security Model Irrespective of Energy Constraints in Wireless Sensor Networks. International Journal of Advanced Computer Science and Applications - IJACSA, 2(4), 20-29.

[11] Jaigirdar, F. T. (2011). Grid Approximation Based Inductive Charger Deployment Technique in Wireless Sensor Networks. International Journal of Advanced Computer Science and Applications - IJACSA, 2(1).

AUTHORS PROFILE

Prof. Vijay Kumar born in Kanpur, India, on 30th June 1972. He received his B.E & M.E. degrees from Kumaon University Nainital (U.P) and Thapar University Patiala (Punjab) respectively. He has supervised 8 M. Tech and 1 M. Phil candidates. His research interests are in Wireless Sensor Networks, Reliability Theory and Artificial Neural Networks, etc. He has about 16 years experience in teaching. He is also a member of IEEE.

Dr. R. B. Patel received PhD from IIT Roorkee in Computer Science & Engineering, PDF from Highest Institute of Education, Science & Technology (HIEST), Athens, Greece, MS (Software Systems) from BITS Pilani and B. E. in Computer Engineering from M. M. M. Engineering College, Gorakhpur, UP. Dr. Patel is in teaching and Research & Development since 1991. He has supervised 30 M. Tech, 7 M. Phil and 2 PhD Thesis. He is currently supervising 3 M. Tech, and 8 PhD students. He has published more than 120 research papers in International/National Journals and Refereed International Conferences. He had been awarded for Best Research paper many times in India and abroad. He has written numbers books for engineering courses (These are "Fundamentals of Computing and Programming in C", "Theory of Automata and Formal Languages", "Expert Data Structures with C," "Expert Data Structures with C++," "Art and Craft of C" and "Go Through C". His research interests are in Mobile & Distributed Computing, Mobile Agent Security and Fault Tolerance, development infrastructure for mobile & Peer-To-Peer computing, Device and Computation Management, Cluster Computing, Sensor Networks, etc.

Dr. Manpreet Singh is working as Professor. & Head of computer science and Engineering department at MMEC, M. M. University Mullana, Ambala, India. He obtained his Ph.D. (Computer Science) from Kurukshetra University. He has number of publications in International journals/Conferences to his credit. His current research interest includes Grid Computing, Wireless communications, MANETs etc.

Rohit Vaid received his M. Tech. degree from Maharishi Markandeshwar University Mullana, Ambala, Haryana (India) respectively. He has number of publications in International journals/Conferences to his credit. He is currently supervising 6 M. Tech students. His research interests are in Mobile & Distributed Computing, Mobile Agent Security and Fault Tolerance, Cluster Computing, Wireless Sensor Networks, etc.

Deploying an Application on the Cloud

N. Ram Ganga Charan
Department of Computer Science & Engineering.
Geethanjali College of Engineering & Technology
Hyderabad, India-501 301

S. Tirupati Rao
Associate Professor of Computer Science & Engineering.
Geethanjali College of Engineering & Technology
Hyderabad, India-501 301

Dr .P.V.S Srinivas
Professor of Computer Science & Engineering.
Geethanjali College of Engineering & Technology
Hyderabad, India-501 301

Abstract— **Cloud Computing, the impending need of computing as an optimal utility, has the potential to take a gigantic leap in the IT industry, is structured and put to optimal use with regard to the contemporary trends. Developers with innovative ideas need not be apprehensive about non utility of costly resources for the service which does not cater to the need and anticipations. Cloud Computing is like a panacea to overcome the hurdles. It promises to increase the velocity with which the applications are deployed, increased creativity, innovation, lowers cost all the while increasing business acumen. It calls for less investment and a harvest of benefits. The end-users only pay for the amount of resources they use and can easily scale up as their needs grow. Service providers, on the other hand, can utilize virtualization technology to increase hardware utilization and simplify management. People want to move large scale grid computations that they used to run on traditional clusters into centrally managed environment, pay for use and be done with it .This paper deals at length with regard to the cloud, cloud computing and its myriad applications.**

Keywords-Cloud; Virtualization; EC2; IAAS; PAAS;SAAS;CAAS; DAAS; public cloud; private cloud; hybrid cloud; Community cloud.

I. Introduction

Cloud Computing can be defined as the novel style of computing where virtualized resources are provided as services on internet which are dynamically scalable[1].cloud computing represents a different way to architect and remotely managing computing resources[8]. It refers to both application delivered as the service over the internet and system software in the datacenters that provide those services .the data centre hardware and software is called cloud[2]. Cloud Computing is a major paradigm shift [3]. Most of the enterprises shifting their applications on to the cloud owing to its speed of implementation and deployment, improved customer experience, scalability, and cost control. Reliability, availability and security are the three greatest concerns for moving on to the cloud [3].

Businesses are running all kinds of applications in the cloud, like customer relationship management (CRM), HR, accounting, and much more. Some of the world's largest companies moved their applications to the cloud with salesforce.com after rigorously testing the security and reliability of infrastructure [17].Smart phones, laptops, PCS and PDAs can access programs, storage and application development platforms over the internet using cloud computing via services offered by the cloud providers. *Virtualization* is the key technology that enables Cloud Computing [3]. Remote hosting took its transformation from renting infrastructure to providing and maintaining Virtual servers sustaining the fluctuations in demand. The big players in cloud computing are Google, Amazon, and, of late, Microsoft and IBM.

The early adopter of this technology is Amazon. Amazon began providing Amazon Web Services in 2005, known only to the cognoscenti. Amazon's Web Services is the oldest and most mature of the public cloud service providers. Microsoft Azure represents a major evolution both of operating systems and of Microsoft's overall strategy. While written entirely from the ground up, it benefits from a long, mostly distinguished, and expensive pedigree. IBM was a very early proponent of both virtualization and cloud computing. IBM Smart Business cloud solutions support clouds built behind the enterprise firewall, or the IBM cloud. IBM's public cloud offering is still new, while its private cloud offerings are, for the cloud, very mature.

II. Phases of Computing Paradigms

- In phase 1, many users shared powerful mainframes using dummy terminals.
- In phase 2, stand-alone PCs became powerful enough to meet the majority of users' needs.
- In phase 3, PCs, laptops, and servers were connected together through local networks to share resources and increase performance.
- In phase 4, local networks were connected to other local networks forming a global network such as the Internet to utilize remote applications and resources [1].
- In phase 5, grid computing provided shared computing power and storage through a distributed computing system [6].
- In phase 6, cloud computing further provides shared resources on the Internet in a scalable and simple way. Mainframe computing offers finite computing power, while cloud computing provides almost infinite power and capacity. In addition, in mainframe computing dummy terminals acted as user interface devices, while in Cloud Computing

III. Cloud and Cloud Computing

A cloud is a pool of virtualized computer resources.

A cloud can: [10]

- Host a variety of different workloads, including batch-style back-end jobs and interactive ,User-facing applications
- Allow workloads to be deployed and scaled-out quickly through the rapid provisioning of Virtual machines or physical machines
- Support redundant, self-recovering, highly scalable programming models that allow Workloads to recover from many unavoidable hardware/software failures.
- Monitor resource use in real time to enable rebalancing of allocations when needed

A Cloud is a virtual space available to deploy the applications, whereas Cloud Computing is a general term for anything that involves delivering hosted services over the Internet. At its simplest, it is delivering the resources and capabilities of information technology dynamically as a service. Cloud Computing is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet [5]. It generally incorporates Infrastructure as a Service (IaaS), Platform as a service (PaaS), and Software as a Service (SaaS).

IV. Attributes and Benefits

A. Attributes

The attributes of cloud computing are: [13]

*1) Service Based***:** Consumer concerns are abstracted from provider concerns through service interfaces that are well-defined. The interfaces hide the implementation details and enable a completely automated response by the provider of the service to the consumer of the service.

*2) Scalable and Elastic***:** The service can scale capacity up or down as the consumer demands at the speed of full automation (which may be seconds for some services and hours for others). Elasticity is a trait of shared pools of resources.

*3) Shared***:** Services share a pool of resources to build economies of scale. IT resources are used with maximum efficiency. The underlying infrastructure, software or platforms are shared among the consumers of the service (usually unknown to the consumers). This enables unused resources to serve multiple needs for multiple consumers, all working at the same time.

*4) Metered by Use***:** Services are tracked with usage metrics to enable multiple payment models. The service provider has a usage accounting model for measuring the use of the services, which could then be used to create different pricing plans and models. These may include pay-as-you-go plans, subscriptions, fixed plans and even free plans. The implied payment plans will be based on usage, not on the cost of the equipment.

5)Uses Internet Technologies: The service is delivered using Internet identifiers, formats and protocols, such as URLs, HTTP, IP and representational state transfer Web-oriented architecture.

B. Benefits

The most frequently cited benefits of cloud computing are: [3]

- It is agile, with ease and speed of deployment
- Its cost is use-based, and will likely be reduced
- In-house IT costs are reduced
- Capital investment is reduced
- The latest technology is always delivered
- The use of standard technology is encouraged and facilitated

As an application moves to the cloud, the access to it becomes more simple and ubiquitous [12]. Low cost ultra light devices and inexpensive hand held devices build on latest operating systems such as android provide access to the internet, the number and types of tasks taking advantage of the new technology will increase by several orders of magnitude, going far beyond the comparatively modest list of things that we use computers and the Internet for today.

V. Cloud types

A common distinction is Public Clouds, Private Clouds, Hybrid Clouds and Community Clouds

A. Public Cloud

A public cloud, or external cloud, is the most common form of cloud computing, in which services are made available to the general public in a pay-as-you-go manner [13]. Customers – individual users or enterprises – access these services over the internet from a third-party provider who may share computing resources with many customers [16]. The public cloud model is widely accepted and adopted by many enterprises because ,the leading public cloud vendors as Amazon, Microsoft and Google, have equipped their infrastructure with a vast amount of data centers, enabling users to freely scale and shrink their rented resources with low cost and little management burden. Security and data governance are the main concern with this approach [5].

B.Private Cloud

A Private Cloud, or internal cloud, is used when the cloud infrastructure, proprietary network or data center, is operated solely for a business or organization, and serves customers within the business fire-wall [9]. Most of the private clouds are large company or government departments who prefer to keep their data in a more controlled and secure environment.

C. Hybrid Cloud

A composition of the two types (private and public) is called a Hybrid Cloud, where a private cloud is able to maintain high services availability by scaling up their system with externally provisioned resources from a public cloud when there are rapid workload fluctuations or hardware failures [11]. In the Hybrid cloud, an enterprise can keep their critical data and applications within their firewall, while hosting the less critical ones on a public cloud.

D. *Community Cloud*

The idea of a Community Cloud is derived from the Grid Computing and Volunteer Computing paradigms. In a community cloud, several enterprises with similar requirement can share their infrastructures, thus increasing their scale while sharing the cost. Another form of community cloud may be established by creating a virtual data center from virtual machines instances deployed on underutilized users machines [2].

VI. Cloud Architecture

From an engineering perspective the cloud is a computing architecture characterized by a large number of interconnected identical computing devices that can scale on demand and that communicate via an IP network. From a business perspective it is computing services that are scalable and billed on a usage basis [18].

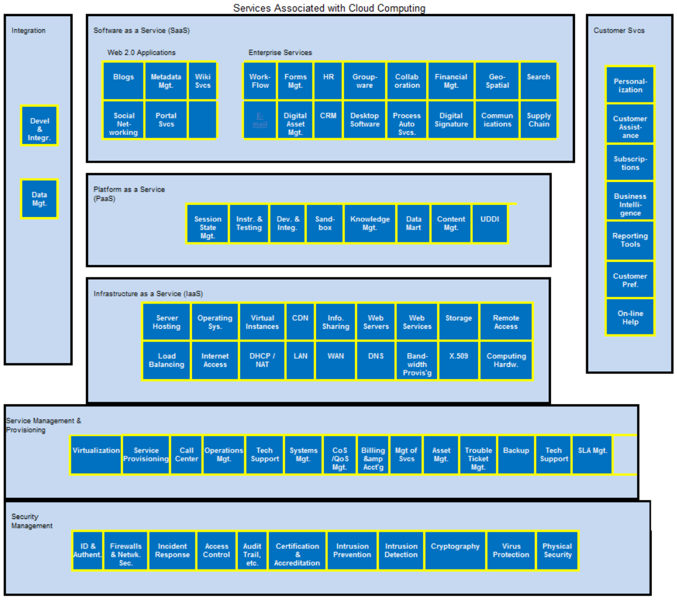


Figure 1. Architecture of a cloud

Like the OSI model, this framework is best read from the bottom up. Everything is dependent upon security. Provisioning is above security because it cannot occur reliably without security, and all of the services are based upon the ability to perform the tasks in the provisioning group. Infrastructure as a Service is the first grouping dependent upon (resting upon) provisioning services. (i.e.; once provisioning services are provided, Infrastructure as a Service can be delivered.

Likewise with PaaS(Platform as a service) and SaaS(software as a service). Integration and User Services are shown as vertical rectangles spanning IaaS, PaaS and SaaS because these regions represent services that are applicable to all three of these layers in the framework [18].

VII. Cloud Services

A Cloud is essentially a class of systems that deliver IT resources to remote users as a service. The resources encompass hardware, programming environments and applications [9]. The services provided through cloud systems can be classified into Infrastructure as a service (IaaS), Platform as a Service (PaaS) and Software as a service (SaaS).

A. *Infrastructure as a Service*

The IaaS is categorized into:

1) *Computation as a Service (CaaS),* in which virtual machine based servers are rented and charged per hour based on the virtual machine capacity – mainly CPU and RAM size, features of the virtual machine, OS and deployed software.

2) *Data as a Service (DaaS),* in which unlimited storage space is used to store the user's data regardless of its type, charged per GB for data size and data transfer.

Amazon has provided a popular universal and comprehensive solution to Cloud Computing, called the Amazon Elastic Compute Cloud (EC2). EC2 provides many useful features for customers, including a mature and inexpensive billing system able to charge for computing at a very fine-grained level (memory usage, CPU usage, data transfer, etc.), deployment between multiple locations, elastic IP addresses, connection to a customer's existing infrastructure through a Virtual Private Network, monitoring services by Amazon Cloud Watch, and elastic load balancing. EC2 has deployed such fine granularity and precision that it has become a benchmark and model in cloud computing [20].

Go Grid also provides Hybrid Hosting, which is a distinguishing feature. Many applications simply don't run well in a pure multi-tenant server environment [4]. Databases perform better on a dedicated server where they don't have to compete for input/output resources, and the situation is similar with web server applications. Go Grid provides these special applications with dedicated servers that also have high security assurance.

B. *Platform as a Service*

Platform as a Service (PaaS) cloud systems provide an execution environment that application services can run on. The environment is not just a pre-installed operating system but is also integrated with a programming-language-level platform, which users can be used to develop and build applications for the platform [1].

Microsoft's cloud strategy is to construct a cloud platform that users can move their applications to in a seamless way, and ensure its managed resources are accessible to both cloud services and on-premises applications. To achieve this, Microsoft introduced the Windows Azure Platform (WAP), which is composed of a cloud operating system named Windows Azure, and a set of supporting services. Windows Azure is the main part of the WAP. It employs virtual machines as its runtime environments.

C. Software as a Service

Software-as-a-Service (SaaS) is based on licensing software use on demand, which is already installed and running on a cloud platform. These on-demand applications may have been developed and deployed on the PaaS or IaaS layer of a cloud platform. SaaS replaces traditional software usage with a Subscribe/Rent model, reducing the user's physical equipment deployment and management costs. The SaaS clouds may also allow users to compose existing services to meet their requirements [5]. This section presents some SaaS clouds and applications.

VIII. Advantages of Cloud over Remote Hosting

The benefits of cloud hosting over the traditional remote hosting re as follows [6]:

- *Scalability*(also called elasticity), the ability to provision one or more servers quickly and to scale up or down quickly
- *Pre-configured operating system images,* such as a variety of popular Linux distributions: Ubuntu, Debian, Novell (SUSE), Gentoo, Centos, Fedora, Arch, and Red Hat Enterprise Linux and various versions of Windows-based servers.
- *Virtual servers or physical servers* that can be sized to different plans through a control panel, all the way up to 15.5 GB of RAM; servers can be configured with one to four processors and with one to four cores per processor, and disk drives are usually arranged in a fault-tolerant RAID configuration
- *Dedicated IP addresses* for cloud servers.
- *Communication* among servers in the same cloud (co-located), effectuated at high-speed and free of communications charges
- *Replication and/or distribution* over various geographical areas
- *Persistence* provided by a separate cloud storage facility, as storage in a *virtual* cloud server is not persistent when the server instance is shut down.

IX. Requirements for Deploying an Application over Public Cloud

A. Licensing

Application is made up of many components which are associated with some license agreements [9]. Analysis should be made about the effects of those license agreements on the deployment of application on cloud. Applications which are designed licensed for CPU, when we deploy it on the cloud increases the load by exciding the CPU license limit.

B. Processing requirements and memory locks

Application should be designed to work on the parallel architectures, because of the dynamic scalability of cloud. Multi threaded code which allows process to split in to small chunks suits for the cloud environment. A single threaded application cannot take the real advantage of clouds distributed nature [20].

C. Bandwidth requirements

Because a public cloud is accessed via the Internet, bandwidth is significantly limited when compared to a private cloud. Given the public cloud's bandwidth limitation, applications that have moderate client bandwidth requirements should only be considered. [9].

D. Communication protocol

The cloud is based on the Internet Protocol (IP), so for an application to be considered, it must use IP as its communication mechanism. While there are many protocols that can be run over IP, the use of Transport Control Protocol (TCP) is preferred.

E. Data security

The application will need to provide security at the data storage, processing and transmission stages. Three critical components of this are [5]:

- Data in transit needs to be protected either at the application or the transmission level.
- Data at rest must be protected by the application. The application must provide a mechanism to protect the data stored in the cloud. Encrypting data at rest is the best option at this time, and a future technical tip will delve into the specifics of this area [5].
- Servers to server communications are typically forgotten because they currently exist within the data center.

X. Application Deployement

In this section we present an example how the combination of virtualization and on of self service facilitate application deployment in this example we consider a two-tier Web application deployment using cloud.

A. Steps for deployment

The following steps comprise the deployment of the application [7]:

- A load balancer, Web server, and database server appliances should be selected from a library of preconfigured virtual machine images.
- Configuring each component to make a custom image should be made. Load balancer is configured accordingly; web server should be populated with the static contents by uploading them to the storage cloud where as the database servers are populated with the dynamic content of the site.
- The developer then feeds the custom code in to the new architecture making components meet their specific requirements.

- The developer chooses a pattern that takes the images for each layer and deploys them, handling networking, security, and scalability issues.

The secure, high-availability Web application is up and running. When the application needs to be updated, the virtual machine images can be updated, copied across the development chain, and the entire infrastructure can be redeployed. In this example, a standard set of components can be used to quickly deploy an application. With this model, enterprise business needs can be met quickly, without the need for the time-consuming, manual purchase, installation, cabling, and configuration of servers, storage, and network infrastructure [9].

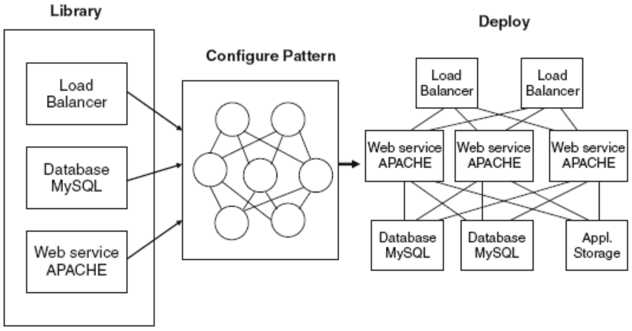


Figure 2. Deployment Strategy on Cloud for two tier architecture

B. *Deployment on azure cloud*

1) Step-1

Initially start visual studio in the administrator mode then go to file select new file. Select cloud service from project types and from template select web cloud service. In the solution explorer double click on default.aspx.Develop and press f5 to compile and debug the application.

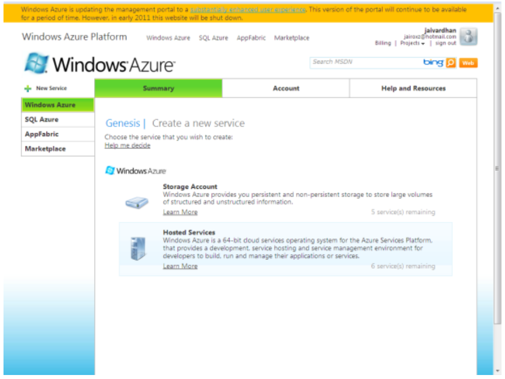
In the solution explorer, right click on the application and then click publish. A publish folder gets opened which contains service package file and cloud service configuration file.

2) Step-2

Log in to the windows azure portal using your windows live id to deploy the application on the cloud

3) Step-3

In the portal, click on the hosted services, storage accounts and CDN


Windows Azure is updating the management portal to a substantially enhanced user experience. This version of the portal will continue to be available for a period of time. However, in early 2011 this website will be shut down.

Genesis | Create a new service

Choose the service that you wish to create:
Help me decide

Windows Azure

Storage Account
Windows Azure provides you persistent and non-persistent storage to store large volumes of structured and unstructured information.
Learn More — 5 service(s) remaining

Hosted Services
Windows Azure is a 64-bit cloud services operating system for the Azure Services Platform, that provides a development, service hosting and service management environment for developers to build, run and manage their applications or services.
Learn More — 6 service(s) remaining

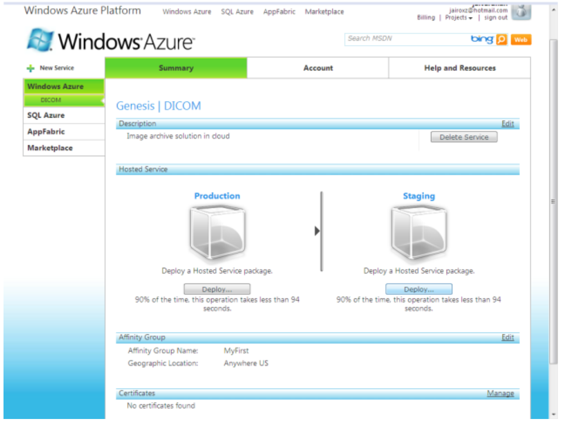
Figure 3. creation of a service on azure cloud

Click new hosted service. Select a subscription that will be used for application.

4) Step-4

Enter the name of the application, enter URL for your application, and then choose a region from the list of regions.

Select deploy to stage environment.



Genesis | DICOM

Description
Image archive solution in cloud

Hosted Service

Production: Deploy a Hosted Service package. 90% of the time, this operation takes less than 94 seconds.

Staging: Deploy a Hosted Service package. 90% of the time, this operation takes less than 94 seconds.

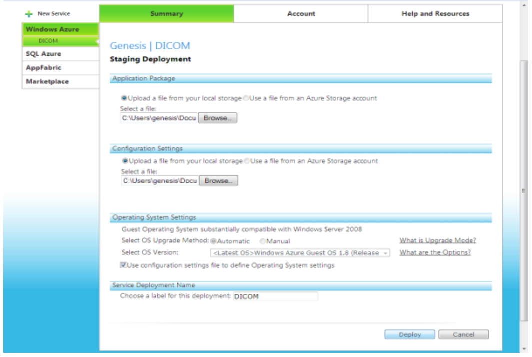
Affinity Group
Affinity Group Name: MyFirst
Geographic Location: Anywhere US

Certificates
No certificates found

Figure 4. Stages in Deployment

5) Step-5

Ensure that Start after successful deployment is checked. Specify a name for the deployment.



Genesis | DICOM
Staging Deployment

Application Package
● Upload a file from your local storage ○ Use a file from an Azure Storage account
Select a file: C:\Users\genesis\Docu

Configuration Settings
● Upload a file from your local storage ○ Use a file from an Azure Storage account
Select a file: C:\Users\genesis\Docu

Operating System Settings
Guest Operating System substantially compatible with Windows Server 2008
Select OS Upgrade Method: ● Automatic ○ Manual
Select OS Version: <Latest OS>Windows Azure Guest OS 1.8 (Release
☑ Use configuration settings file to define Operating System settings

Service Deployment Name
Choose a label for this deployment: DICOM

Figure 5. staging deployment

6) Step-6

For Package location**,** click the corresponding **Browse locally…** button, navigate to the folder where your <*Your Project Name*>.cspkg file is, and select the file.

For Configuration file, click the corresponding **Browse locally…** button, navigate to the folder where your Service Configuration.cscfg is, and select the file.

7) Step-7

Click **OK**. You will receive a warning after you click OK because there is only one instance of the web role defined for your application (this setting is contained in the Service Configuration. Cscfg file). For purposes of this walk-through, override the warning by clicking **yes**, but realize that you likely will want more than one instance of a web role for a robust application.

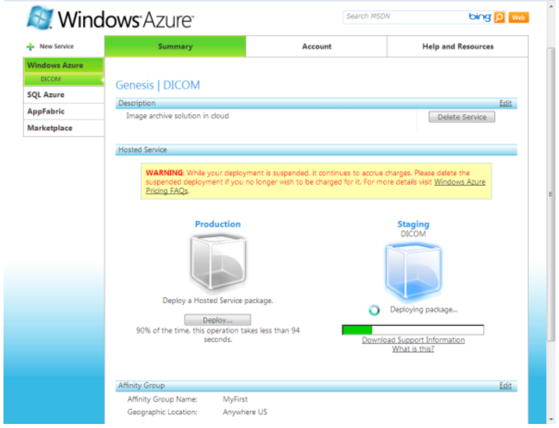


Image archive solution in cloud

WARNING: While your deployment is suspended, it continues to accrue charges. Please delete the suspended deployment if you no longer wish to be charged for it. For more details visit Windows Azure Pricing FAQs.

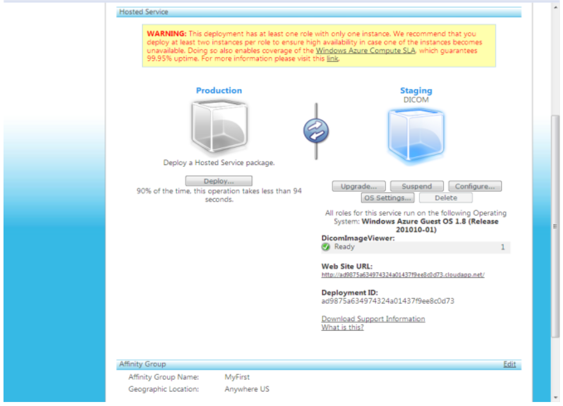
Deploy a Hosted Service package.

90% of the time, this operation takes less than 94 seconds.

Deploying package...

Affinity Group Name: MyFirst
Geographic Location: Anywhere US

Figure 6. Staging Of An Application



WARNING: This deployment has at least one role with only one instance. We recommend that you deploy at least two instances per role to ensure high availability in case one of the instances becomes unavailable. Doing so also enables coverage of the Windows Azure Compute SLA, which guarantees 99.95% uptime. For more information please visit this link.

Deploy a Hosted Service package.

90% of the time, this operation takes less than 94 seconds.

All roles for this service run on the following Operating System: **Windows Azure Guest OS 1.8 (Release 201010-01)**

DicomImageViewer:
Ready 1

Web Site URL:
http://ad9875a634974324a01437f9ee8c0d73.cloudapp.net/

Deployment ID:
ad9875a634974324a01437f9ee8c0d73

Affinity Group Name: MyFirst
Geographic Location: Anywhere US

Figure 7. Final deployement screen of application on azure

You can monitor the status of the deployment in the Windows Azure management portal by navigating to the **Hosted Services** section

XI. Cloud Computing Challenges

In summary, the new paradigm of cloud computing provides a number of benefits and advantages over the previous computing paradigms and many organizations are adopting it [5]. However, there are still a number of challenges, which are currently addressed by researchers and practitioners in the field. They are briefly presented below.

A. *performance*

Cloud may lack performance in some intensive transaction oriented and other intensive applications. High latency delays may be observed by the users who are at far long distance from cloud

B. *Security and Privacy*

Security is still a major criterion when coming to cloud computing, customers are worried about the attacks which are vulnerable, when information and other important resources are kept outside the firewall. Standard security practices should be done to overcome this problem [5].

C. *Control*

Some IT departments are concerned because cloud computing providers have a full control of the platforms. Cloud computing providers typically do not design platforms for specific companies and their business practices.

D. *Bandwidth costs*

With cloud computing, companies can save money on hardware and software; however they could incur higher network bandwidth charges. Bandwidth cost may be low for smaller Internet-based applications, which are not data intensive, but could significantly grow for data-intensive applications.

E. *Reliability*

Cloud computing still does not always offer round-the-clock reliability. There were cases where cloud computing services suffered few-hours outages. In the future, we can expect more cloud computing providers, richer services, established standards, and best practices. IBM has launched the Research Computing Cloud, which is an on-demand, globally accessible set of computing resources that support business processes.

XII. Conclusion

Cloud computing is a very flexible paradigm for delivering computational power. It means many things to many people. For some it means being able to set up a new start-up company knowing that initial resources will be inexpensive but a sudden increase in demand from users won't make the company a victim of its own success, as has happened in some cases in the past where servers have been unable to cope with demand, and the company loses clients as they become unhappy with poor response times.

For other people, cloud computing means easier administration, with issues such as licensing, backup and security being taken care of elsewhere. In other cases, cloud computing means having a powerful computational environment available anywhere that the user can access a web browser.

The various forms of service – infrastructure, platform, and software as a service provide exciting ways to deliver new products that innovators might come up with. Already there are examples of widely used products and web sites that have sustained remarkable growth because creative ideas could be implemented quickly, and because the subsequent demand could be met easily through the flexibility of cloud computing.

The future seems to be limited only by the imaginations of innovators who can think of applications that will help people communicate, store and process vast quantities of information, whether it is millions of individuals with small collections of personal information, or a single large organization with large collections of data to be processed.

REFERENCES

[1] Borko Furht Armando Escalante," Hand book of Cloud Computing "by Springer.

[2] A Technical report on: "Above the Clouds: Berkley view"

[3] David E.Y. Sarna: "Implementing and developing Cloud Applications".

[4] Shufen Zhang, Shuai Zhang, Xuebin Chen, Shangzhuo Wu. "Analysis and Research of Cloud Computing System Instance", 2010 Second International Conference on Future Networks, 978-0-7695-3940-9/10©IEEE.

[5] Jianfeng Yang, Zhibin Chen. "Cloud Computing Research and Security Issues", 978-1-4244-5392-4/10©2010 IEEE.

[6] Shuai Zhang, Shufen Zhang, Xuebin Chen, Xiuzhen Huo. "The Comparison between Cloud Computing and Grid Computing", 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), 978-1-4244-7237-6/ 2010 ©IEEE.

[7] Donald Robinson ,"Amazon web services made simple"

[8] Thomas B Winans and john seely brown,"cloud computing –a collection of working papers",May2009

[9] Introduction to the cloud computing architecture white paper 1st edition 2009 by sun Microsystems

[10] Greg Boss,Padma Malladi,Dennis Quan,Linda Legregni,Harold Hall-"cloud computing" 8 October 2007 IBM

[11] www.searchcloudcomputing.com

[12] www.wikipedia.com

[13] www.gigaspaces.com

[14] http://www.cioupdate.com/research/article.php/3827971/The-Five-Attributes-of-Cloud-Computing.htm

[15] http://techbasedmarketing.com/cutting-edge/the-amazon-business-cloud/2100/

[16] http://searchcloudcomputing.techtarget.com/definition/public-cloud

[17] http://www.salesforce.com/cloudcomputing/

[18] http://sites.google.com/site/cloudarchitecture/

[19] http://cloudslam09.com/

[20] http://www.cloudcomputingarchitecture.net/

[21] Rao, N. M. (2010). Cloud Computing Through Mobile-Learning. International Journal of Advanced Computer Science and Applications - IJACSA, 1(6).

AUTHORS PROFILE

N.Ram Ganga Charan, B.Tech. Graduate in computer science and Engineering from Geethanjali college of Engineering and Technology. He served as an Entrepreneurship leader at Geethanjali. He presented good number of papers and won prizes. He was an active member in workshop on cloud computing by Dr.Keith Maycock from National College of Ireland. Attended workshops on adobe flex, trends in cloud computing and Network programming. He worked in research under the guidance of Prof.Dr.P.V.S.Srinivas on deploying an application on cloud. His areas of research are cloud computing, inter planetary networks and Bionics.

S.Tirupati Rao is presently serving as an associate professor of computer science and Engineering at Geethanjali college of Engineering& Technology. He got his masters from Andhra University. He is perusing his Ph.D. from Andhra University presently. He worked as an associate professor in various colleges more than 9 years in the department of computer science. He joined as ISTE life member in 2005 with membership LM 47567.

Dr. P.V.S.Srinivas is presently serving as a Professor & Head, Department of Computer Science and Engineering, at Geethanjali College of Engineering and Technology, Hyderabad. He has got his Masters followed by PhD in Computer Science and Engineering in the area of Computer Networks from JNTU Hyderabad in the year 2003 and 2009 respectively. His main research interests are Wireless Communication, Mobile Computing and Mobile Ad hoc Networks. His research focus is on "Designing an Effective and Assured Communication in MANETs" and improving QoS in MANETs.
He has a rich experience of total 20 years, out of which 2 years of Industry and 18 years of academic. He is also serving as a Chief Panel Consultant in the area of wireless communications for a Hyderabad based company by name SCADA METER SOLUTIONS Pvt Ltd. He has published 28 research papers in different refereed International journals and conferences in India as well as Abroad. He is also serving as an Editor-in-Chief for an International Journal IJWNC and also a peer reviewer for 3 International Journals.

Radial Basis Function For Handwritten Devanagari Numeral Recognition

Prerna Singh
M.Tech Student
Shobhit University,
Meerut, India

Nidhi Tyagi
Assistant Professor
ISCEIT. Shobhit University,
Meerut, India

Abstract—**The task of recognizing handwritten numerals, using a classifier, has great importance. This paper applies the technique of Radial Basis Function for handwritten numeral recognition of Devanagari Script. Lot of work has been done on Devanagari numeral recognition using different techniques for increasing the accuracy of recognition. Since the database is not globally created, firstly we created the database by implementing pre-processing on the set of training data. Then by the use of Principal Component Analysis we have extracted the features of each image, some researchers have also used density feature extraction. Since different people have different writing style, so here we are trying to form a system where recognition of numeral becomes easy. Then at the hidden layer centers are determined and the weights between the hidden layer and the output layer of each neuron are determined to calculate the output, where output is the summing value of each neuron. In this paper we have proposed an algorithm for determining Devanagari numeral recognition using the above mentioned system.**

Keywords-Radial Basis Function; Devanagari Numeral Recognition; K-means clustering; Principal Component Analysis (PCA).

I. Introduction

Handwritten character recognition has been around since 1980’s. Till date many researches have been done. Automatic reading of numerical fields has been attempted in several application areas such as online handwritten recognition on computer tablets, recognize zip codes on mail for postal address sorting, processing bank check amounts, numeric entries in forms filled up by hand (for eg.Tax forms) and so on. While solving this domain of handwritten recognition many challenges are faced. As the handwritten digits are not always of same size, thickness, or orientation and position relative to the margins, many handwritten versions are even hard to recognize.

Handwritten recognition [1] is the ability of a computer to receive and interpret intelligible handwritten input from sources such as document, photographs, touch-screens & other devices. But in the past various schemes or algorithms are proposed in different languages like English [2] ,Chinese[3] , Arabian ,Persian , Bengali , Guajarati .Different classifiers have been used for handwritten digit recognition using neural network. Recently, signifi- cant contribution towards the improvement of recognition rates have been made by means of different combination strategies [4] and by the use of support vector machines [5] .

Machine learning is a artificial system[6] capable of autonomous acquisition & integration of knowle- dge.This capacity to learn from experience, analy- tical observation other means, results in system that can improve its own speed or performance and allows computer to “learn“. It can be categorized into supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning [7].

In this paper we proposed a method for recognition of Handwritten Devanagari numerals using Radial basis function. It is a feed forward neural network that computes activation at the hidden neuron in a way that is different from product between input vector and the weight vector. Hidden neuron activation in RBF are computed using an exponential of distance measure (usually the Euclidean distance or weighted norm) between the input vector and a prototype vector that character-izes the signal function at a hidden neuron.

Devanagari script was originally developed to write Sanskrit and after some time it was adapted by many other languages such as Rajasthani, Marathi [8], Nepali, Bhojpuri ,Konkani and Hindi [9], the mother tongue of majority of Indians.

The paper is organized in various sections. Further sections are like section 2, contains the related work on pattern recognition ,section 3 gives the overview of system, how recognition takes place using RBF, section 4 elaborates the proposed work and section 5 describes certain examples on the proposed system, section 6 gives the conclusion and future work.

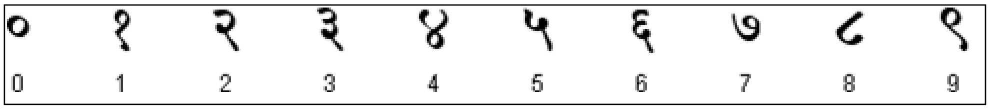


Fig.1 Devanagari Numerals

II. Related Work

Connell, R.M.K. Sinha and Anil K. Jain [10] descri- bes the combination of classifiers capturing both online and offline features using markov model, three nearest neighbour and clustering. A classifica- ion accuracy of 86.5% is obtained. Anita Pal & Dayashankar Singh has performed a work to recognize handwritten English character [2] using a multilayer preceptor with one hidden layer. Fourier descriptors with back propagation network yields good recognition accuracy of 94%. The skeleton- ized and normalized binary pixels of English characters were used as the inputs of the . M.Hanmandlu, J.Grover, S.Vasikarla, V.K.Madasu [9] used fuzzy modelling for recognition and produces 95% recognition rate. The criterion function of the relationship between entropy and energy of the fuzzy sets is used. G.G.Rajput, S. M. Mali [8] presented a method for recognition of isolated Marathi handwritten numerals using Fourier descriptors, invariant translation, scaling and rotation, and used as features of the images representing handwritten numerals. Support Vector Machine is used as pattern recognition and gives the accuracy rate of 97.85%. Shailedra Kumar Shrivastava and Sanjay S.Gharde [5] applies support vector machine for recognition and produces the average recognition rate of 99.85% using moment invariant and affine moment invariant.

The critical look at the available literature reveals that lot of efforts have been made to get higher accuracy but still tremendous amount of work need to be done for improving recognition accuracy by developing new feature extraction techniques or modifying the existing feature extraction technique. This paper proposed a technique for Devanagari numeral recognition using Radial basis function by forming the cluster of training data at the hidden neuron and a set of test data are input to recognize the exact digit.

III. Radial Basis Function (RBF)

RBF was originally introduced into the literature for the purpose of interpolation of data points on a finite training dataset [11]. This network typically have three layers input layers, hidden layers with a non-linear RBF [12] activation function and a linear output layer. Hidden layer performs mapping from the input space into a higher dimensional space in which the patterns become linearly separable.

The output ϕ of the network is thus for N number of neurons in the hidden layer is

$$\boldsymbol{\phi(x) = \sum a_i \sigma(\|x-C_i\|)} \ldots\ldots\ldots\ldots(1)$$

where C_i is the center vector for neurons i and a_i are the weight of the linear output neurons.

System Overview

The problem of recognition of handwritten numerals can be broadly classified into three blocks:-

I. Pre-processing
II. Feature extraction
III. RBF classifier

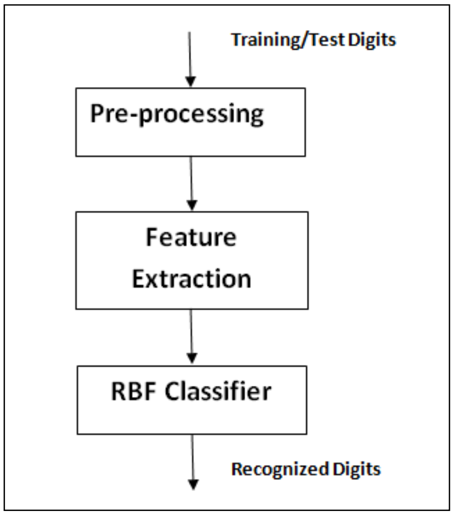


Fig 2.Block Diagram of System

A. *Creation of Data*

Our objective is to obtain a set of handwritten samples of Devanagari numerals that capture variation in hand writing between and within writers. Therefore we need numeral sample from multiple writers as well as multiples sample from each writer.

PRE-PROCESSING:-Pre-processing is the process to removing noise and distortion present in input text due to hardware and software limitations. Steps are performed in pre-processing are size normalization and noise removal.

Normalization-General problem which occurs is “how to standardize [13] the size of each image”.

Following steps took place:-

- Start the image with xsize0 * ysize0 pixels which fit the isolated numeral, by removing blank rows and columns.
- Rescale the size of the image to (Xsize*Ysize) pixels which is the maximum size according to xsixe0 or ysize0 i.e
  Xsize =max (xsize0, ysize0)
  Ysize=max (xsize0, ysize0)
- Find the center of gravity co-ordinates.
- Now put the character at the center of frame.
- Finally rescale the image to the size of 36*36 pixels.

Noise Removal:- It may be, while scanning some noise is intercepted. That noise could be because of some particles of dust on the scanner or because of the poor quality of paper on which numerals are written. Remove the noise from the image file.

B. *Feature Extraction*

When the input data to an algorithm is too large to be processed and is suspected to be notoriously redundant then the input data will be transformed into a reduced representation set of features. We can also consider density feature extraction [1,14] to encode style characteristics without being affected by variations in minute details of similarly formed characters.

Using Principal Component Analysis:- It is an effective technique for dimensionality reduction and extracting important information used in several areas such as machine learning, neural networks and signal processing. It maximizes the rate of decrease of variance and transforms a number of correlated variables into (smaller) number of uncorrelated variables called principal components [15]. The first principal component accounts for as much variabil- ity in the data as possible and each succeeding component accounts for as much of the remaining variability as possible. PCA can be done by eigenvalues decomposition of covariance matrix of a data matrix, usually after mean centering the data for each attribute.

IV. Proposed RBF Network

In addition to feature extraction, we need to calculate the centers of the basis function and the network is given in fig.3.The network divide its work in two phase, firstly it calculate the centers and then the weights between the hidden and the output neuron,

K-means clustering:-Now we want to calculate centers [16, 17] of basis function, the main idea is to define k centroids, one for each cluster. Steps for calculating centers are:-

1. Place K points in each cluster represent initial group center.
2. Assign each object to the group that has the closest centers
3. Re-calculate the position of the K centers , when all the objects are assigned
4. Repeat step 2,3 until the centers no longer move.

With the centers identified, now we need to deter- mine the weights [18,19,20] using LMS procedure from the hidden neurons to output summing neurons.

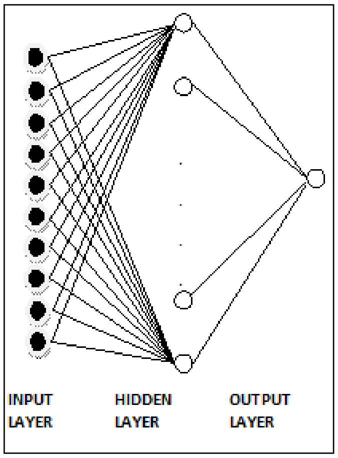


Fig.3 Radial Basis Function Network

Algorithm for the proposed network is given below:

Algorithm for pattern recognition using RBF

Input: image x of Devanagari handwritten digit from test set.
Initialize:
1.the training set patterns after normalizing the size and removal of noise.
2. Extract important information for each data to be stored in cluster.
3. Remove the redundant data items using PCA

Output: 1.Calculate the centers of basis function.
2. Determine weights between the hidden neuron 3.Input data matches with any of the cluster, for that particular cluster weight =1 otherwise 0.
4. Once values are calculated then we can determine the output by summing values of each neuron.

V. Example:

As per the proposed system, let’s consider an example when we insert an input, then what will be the values of each of the cluster represented by C.

I/P/	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9
०	1	0	0	0	0	0	0	0	0	0
१	0	1	0	0	0	0	0	0	0	0
२	0	0	1	0	0	0	0	0	0	0
३	0	0	0	1	0	0	0	0	0	0
४	0	0	0	0	1	0	0	0	0	0
५	0	0	0	0	0	1	0	0	0	0
६	0	0	0	0	0	0	1	0	0	0
७	0	0	0	0	0	0	0	1	0	0
८	0	0	0	0	0	0	0	0	1	0
९	0	0	0	0	0	0	0	0	0	1

C0…C9 represents the cluster having information regarding ० , १ numeral and so on.Suppose if input get matched with the data of the cluster then for that cluster value will be 1 and for rest of the cluster value will be 0.

VI. Conclusion & Future work

In this proposed work, we tried to recognize Devanagari numerals using RBF which have its own advantages and disadvantages over other systems. One of the advantages of RBF, is that the hidden layer is easier to interpret in RBF than the hidden layer in MLP. Training a RBF network is faster but once training is finished and it is being used, it is slower than a

MLP, so where speed is a factor a MLP may be more appropriate.

In future, we would like to implement this proposed work using MATLAB.

REFERENCES

[1] “Handwritten Character Recognition using Neural Network”,Sunith Bandaru.

[2] “Handwritten English Character Recognition Using Neural Networks”,Anita Pal and Dayashanker Singh,vol 1,no 2, July-December 2010.

[3] “Printed Chinese Character Recognition”,thesis by Yuan Lui.

[4] “Devanagari numeral recognition by combining decision of multiple connectionist classifier” Reena Bajaj,Lipika Dey and Santanu Chaudhary,Sadhana Vol-27,Part-I, February 2002.

[5] “SVM for handwritten Devnagri Numeral Recognition”, Shailendra Kr.Srivastava and Sanjay GHarde,vol 07.

[6] .Christopher M.Bishop, “Pattern Recognition and Machine learning”, Springer Publication,Singapore, 2006.

[7] “Introduction to Machine Learning” by Nils. J. Nilsson, 1997.

[8] G.G.Rajput, S.M.Mali “Fourier Descriptor based Isolated Marathi Handwritten Numeral Recognition”,vol 3-No.4,June 2010

[9] “Input Fuzzy Modeling for Recognition of Handwritten Hindi Numerals”, M.Hanmadhu, J.Grover, V.K.Madasu, S.Vasikarla.

[10] Recognition of unconstrained on-line Devanagari characters by “Scott D. Connell, R.M.K Sinha and Anil.K Jain.

[11] “Neural Networks”, Simon Haykins.

[12] “Off-line handwritten characters recognition using Radial Basis Function”, J.Ashok and Dr.E.G Rajan,vol 2,issue 04.

[13] ”Handwritten Digits Recognition” by Gaurav Jain and Jason Ko.

[14] “Statistical Pattern Recognition:A Review”, Anil K.Jain, IEEE, Robert and Jianchang, vol 22 no.1,2000.

[15] “Fast and robust scheme for recognition of handwritten Devanagari Numerals”, Vasantha,Ritu Jain and Patvardhan, NSC 2008.

[16] Improving the performance of Radial Basis Function by learning center location”, Dietrich and Thomas.

[17] Pattern Recognition using K-nearest neighbours”, Seiji,Kiyasu and Miyahara.

[18] Weighted Radial Basis Function for improved Pattern Recognition and signal processing”, Leonardo M. Reyneri.

[19] Adrian G. Bors, ”Introduction of Radial Basis Function”.

[20] R. M. K. Sinha and H. N. Mahabala, ”Machine recognition of Devanagari script", IEEE Transactions on Systems, Man and Cybernetics.

[21] Mohanty, S., Nandini, H., & Bebartta, D. (2011). Performance Comparison of SVM and K-NN for Oriya Character Recognition. International Journal of Advanced Computer Science and Applications - IJACSA, (Special Issue), 112-116.

[22] Jipeng, T. (2011). Skew correction for Chinese character using Hough transform. International Journal of Advanced Computer Science and Applications - IJACSA, (Special Issue), 45-48.

The Multi-Liaison Algorithm

Ms. Anjali Ganesh Jivani
Dept. of CSE
The M. S. University of Baroda
Vadodara, India

Ms.Amisha Hetal Shingala
Dept. of MCA,
SVIT, Gujarat Technological
University,
Vasad, India

Dr. Paresh. V. Virparia
Dept. of CS, SP University
V.V.Nagar, India

Abstract— **In this paper we present an approach for extracting multiple connections or links between subject and object from natural language input (English), which can have one or more than one subject, predicate and object. The parse tree visualization and the dependencies generated from the Stanford Parser are used to extract this information from the given sentence. Using the dependencies we have generated an output which displays which subject is related to which object and the connecting predicate. Finding the subjects and objects helps in determining the entities involved and the predicates determine the relationship that exists between the subject and the object. An algorithm has been developed to do so and this algorithm is elucidated in detail step-wise. We have named our algorithm 'The Multi-Liaison Algorithm' since the liaison between the subjects and objects would be displayed. The word 'liaison' has been used since we are displaying the relationship and association between the subjects and predicates. This output would be useful for natural language processing (NLP), information retrieval, information extraction and also text mining applications.**

Keywords- parse; liaison; relationship; extraction; information retrieval; text mining.

I. INTRODUCTION

An English sentence can have multiple subjects and objects and the Multi-Liaison Algorithm as proposed by us would display them with the connecting verbs/predicates. According to the approach presented in [3], a triplet in a sentence is defined as a relation between subject and object, the relation being the predicate. We have made an enhancement to it by including multiple connections or liaisons depending on the type of sentence that is given as input. The sentence is parsed with the help of the Stanford parser and then using the output of the parser as input to our program, all the subjects, objects and the predicates i.e. the multiple liaisons are displayed. The subjects can either be nouns or even pronouns. Moreover, one subject can be related to multiple objects and vice-versa.

The algorithm was developed in JAVA using Net Beans IDE 6.5 RC2. The performance of the application was measured using a system with CPU 1.99 GHz and 2.86 GB RAM. The algorithm was tested with a variety of input and some sample inputs and outputs are shown further. This algorithm is expected to be useful to researchers involved in NLP and Text Mining.

II. THE STANFORD PARSER

The Stanford Parser is a probabilistic parser which uses the knowledge of language gained from hand-parsed sentences to try to produce the most likely analysis of new sentences. This package is a Java implementation of probabilistic natural language parsers.

The Stanford dependencies provide a representation of grammatical relations between words in a sentence for any user who wants to extract textual relationships. The dependency obtained from Stanford parser can be mapped directly to graphical representation in which words in a sentence are nodes in graph and grammatical relationships are edge labels [10].

We have used them to extract the relation between multiple subjects and objects when the sentence to be parsed is a little complicated. Stanford dependencies (SD) are triplets: name of the relation, governor and dependent.

A. The Parse Tree and Dependencies

The parse tree generated by the Stanford Parser is represented by three divisions: A sentence (S) having a noun phrase (NP), a verbal phrase (VP) and the full stop (.). The root of the tree is S.

The Stanford typed dependencies representation was designed to provide a simple description of the grammatical relationships in a sentence that can easily be understood. The current representation contains approximately 52 grammatical relations [4]. The dependencies are all binary relations. The definitions make use of the Penn Treebank part-of-speech (POS) tags and phrasal labels.

To find the multiple subjects in a sentence our algorithm searches the NP sub tree. The predicate is found in the VP sub tree and the objects are found in three different sub trees, all siblings of the VP sub tree containing the predicate. The sub trees are: PP (prepositional phrase), NP (noun phrase) and ADJP (adjective phrase).

III. THE MULTI-LIAISON ALGORITHM

We explain our algorithm in detail followed by the output. In the first example we have displayed the Stanford output of the sentence followed by the output of our algorithm.

```
Function: CONVERT_ SENTENCE (Input_Str)
Returns: POS tagging, Parse tree, Typed Dependencies
Input_Str: Sentence to be parsed

[Run the Stanford parser with Input_Str as input]

Output_Str ← i) POS of each word
             ii) The parse tree generated
             iii) The typed dependencies
Return Output_Str

Function: MULTI_LIAISON (Output_Str)
Returns: Multiple liaisons or error message
         Function GET_TRIPLETS (Output_Str)
         Function GET_RELATIONSHIP (Output_Str)
Display the multiple liaisons
```

Figure 1. The Multi-Liaison Algorithm.

A. The Detailed Algorithm

We start with parsing a sentence by the Stanford parser and storing the result in some intermediate file so that it can be taken as input for our algorithm. The triplet extraction algorithm of [3] has also been considered before finding the liaisons.

As shown in Fig. 1, the Multi-Liaison Algorithm takes as input the POS of each word, the parse tree and the typed dependencies [9]. Two functions are then called, the first is the GET_TRIPLETS and the second is the GET_RELATIONSHIP.

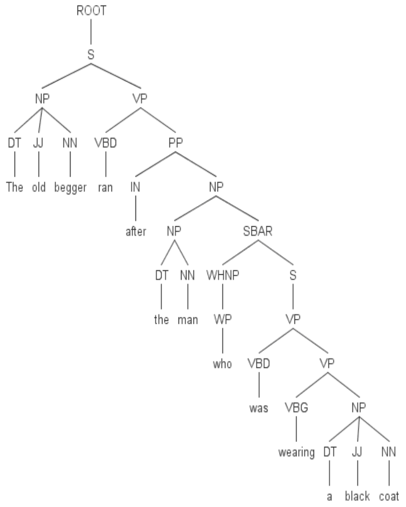


Figure 2. The Stanford Parse Tree.

```
Function: GET_TRIPLET (Output_Str)
Returns: Multiple subjects, objects and predicates
[Read level 1 of Parse Tree – refer Fig. 2]
If tree contains ‘NP’ or ‘NNP’ then
        Function GET_SUBJECT (NP sub tree)
Else
        Return error message
If tree contains ‘VP’ then
       Function GET_PREDICATE (VP sub tree)
       Function GET_OBJECT (VP sub tree)
Else
       Return error message

Function: GET_SUBJECT (NP sub tree)
Returns: Subject(s) and adjective(s)
For (all nodes of NP sub tree) do
    If NP sub tree contains ‘NN’ or ‘NNP’ or ‘NNS’
then
          Store POS as a subject
    If NP sub tree contains ‘JJ’ then
          Store POS as an adjective
Return the subject(s) and adjective(s)

Function: GET_PREDICATE (VP sub tree)
Returns: Predicate(s)
For (all nodes of VP sub tree) do
    If VP sub tree contains ‘VB?’ then
          Store POS as a predicate
    Else
          Return error message
Return the predicate(s)

Function: GET_OBJECT (VP sub tree)
Returns: Object(s)
For (all nodes of VP sub tree) do
    If VP sub tree contains ‘NP’ then
          For (all nodes of VP_NP sub tree) do
              If VP_NP sub tree contains ‘NP’ or ‘NN’
then
                     Store POS as an object
              Else
                     Return error message
    Else
          Return error message
Return the object(s)
```

Figure 3. The GET_TRIPLETS Function.

As shown in Fig. 3, the GET_TRIPLETS function takes as input the Stanford Parse Tree and by considering the nodes under the NP sub tree and the VP sub tree, finds all the subjects, objects and predicates.

The GET_RELATIONSHIP finds and displays the relationships between the subjects and objects. The algorithm is displayed in Fig. 4.

```
Function: GET_RELATIONSHIP (Output_Str)
Returns: Multiple liaisons / relations
[Read the Stanford typed dependencies from Output_Str]
For (all terms in typed dependencies) do
    If typed dependencies contain 'NSUBJ' then
        Store both words of NSUBJ as S1 and S2
        For each value of subject from GET_SUBJECT do
            If subject matches S2 then
                [Check for predicates]
                For each value of predicate from
                GET_PREDICATE do
                    If predicate matches S1 then
                        [Concatenate subject and predicate as
                         R1]
                        Store R1 in the relation
    If typed dependencies contain 'DOBJ' or 'PREP' then
        Store both the words as D1 and D2
        For each value of object in GET_OBJECT do
            If object matches D2 then
                Store value of object in the relation as R2
Return R1+R2
```

Figure 4. The GET_RELATIONSHIP Function.

B. *The Ouput of the Multi-Liaison Algorithm*

As per the algorithm discussed above, the output is shown below. In the first example, the outputs of the Stanford parse as well as the output of the Multi-Liaison both are displayed including the parse tree. In subsequent examples the parse tree is not displayed but the tagging, dependencies and the Multi-Liaison output is displayed. Fig. 2 displays the parse tree.

Example 1: The old beggar ran after the rich man who was wearing a black coat

The Stanford Parser output:

Tagging:

The/DT old/JJ beggar/NN ran/VBD after/IN the/DT rich/JJ man/NN who/WP was/VBD wearing/VBG a/DT black/JJ coat/NN

Parse Tree:

```
(ROOT
  (S
    (NP (DT The) (JJ old) (NN beggar))
    (VP (VBD ran)
      (PP (IN after)
        (NP
          (NP (DT the) (JJ rich) (NN man))
          (SBAR
            (WHNP (WP who))
            (S
              (VP (VBD was)
                (VP (VBG wearing)
                  (NP (DT a) (JJ black) (NN coat))))))))))
```

Typed Dependencies:

det(beggar-3, The-1)
amod(beggar-3, old-2)
nsubj(ran-4, beggar-3)
det(man-8, the-6)
amod(man-8, rich-7)
prep_after(ran-4, man-8)
nsubj(wearing-11, man-8)
aux(wearing-11, was-10)
rcmod(man-8, wearing-11)
det(coat-14, a-12)
amod(coat-14, black-13)
dobj(wearing-11, coat-14)

The Multi-Liaison Output:

Subject: 1
NN beggar
Predicate: 3
VBD ran
VBD was
VBG wearing
Object: 2
NN man   JJ rich
NN coat JJ black

Relationship:

beggar - ran - man
man - wearing - coat

Figure 5. Example 1.

As shown above, the Multi-Liaison Algorithm displays the relationship between the subject and object (beggar and man) as well as the relationship between the two objects (man and coat).

Example 2: The dog and the cat ran after the mouse and the mongoose

Tagging:

The/DT dog/NN and/CC the/DT cat/NN ran/VBD after/IN the/DT mouse/NN and/CC the/DT mongoose/NN

Typed Dependencies:

det(dog-2, The-1)
nsubj(ran-6, dog-2)
det(cat-5, the-4)
conj_and(dog-2, cat-5)
nsubj(ran-6, cat-5)
det(mouse-9, the-8)
prep_after(ran-6, mouse-9)
det(mongoose-12, the-11)
prep_after(ran-6, mongoose-12)
conj_and(mouse-9, mongoose-12)

The Multi-Liaison Output:
Subject: 2
NN dog
NN cat
Predicate: 1
VBD ran
Object: 2
NN mouse
NN mongoose

Relationship:
dog - ran – mouse - mongoose
cat - ran - mouse - mongoose

Figure 6. Example 2.

Example 3: Jack and I visited the zoo with our children

We have also considered pronoun as a subject and therefore we have got the relationship with 2 subjects in terms of noun and pronoun.

Tagging:
Jack/NNP and/CC I/PRP visited/VBD the/DT zoo/NN with/IN our/PRP$ children/NNS

Typed Dependencies:
nsubj(visited-4, Jack-1)
conj_and(Jack-1, I-3)
nsubj(visited-4, I-3)
det(zoo-6, the-5)
dobj(visited-4, zoo-6)
poss(children-9, our-8)
prep_with(visited-4, children-9)

The Multi-Liaison Output:
Subject: 2
NNP Jack
PRP I
Predicate: 1
VBD visited
Object: 2
NN zoo
NNS children
PRP$ our

Relationship:
Jack - visited - zoo - children
I - visited - zoo - children

Figure 7. Example 3.

All the three examples shown in Fig. 5, Fig. 6 and Fig. 7 have different number of subjects and objects and the relationship between them is also not similar. The Multi-Liaison Algorithm output in this way can be very useful for text mining applications where a variety of sentences are to be mined.

IV. Performance of Multi-Liaison Algorithm

The application was written in JAVA using Net Beans IDE 6.5 RC2. It parsed a single sentence of 12 words in 8.35 seconds and displayed the output as shown in the examples above.

This algorithm works equally well with simple as well as complex sentences and the output is very clear and precise.

V. Conclusion

In this paper we have presented an algorithm which displays the relationships between subjects and objects in sentences where there are multiple subjects and objects. The Stanford parser output was used to generate this result.

This algorithm would be usable not only by text mining experts and computational linguists but also by the computer science community more generally and by all sorts of professionals like biologists, medical researchers, political scientists, business and market analysts, etc. [10]. In fact it would be easy for users not necessarily versed in linguistics to see how to use and to get value from the simple relationship that is displayed so effectively.

This work can be carried forward for different text mining and natural language processing applications. In continuation to this we aim to use the Multi-Liaison Algorithm in text summarization and Information Retrieval.

Acknowledgment

We are grateful to our Departmental Heads and Research Guides for their timely suggestions and technical support. Moreover we are extremely thankful to all the authors mentioned below whose research papers we have referred. Their work has formed the foundation and backbone for us in developing this algorithm.

References

[1] D. Klein, C. D. Manning, "Fast exact inference with a factored model for natural language parsing" in Advances in Neural Information Processing Systems 15 (NIPS 2002), Cambridge, MA: MIT Press, pp. 3-10, 2003.

[2] D. Klein, C. D. Manning, "Accurate unlexicalized parsing" in Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430, 2003.

[3] Delia Rusu, Lorand Dali, Blaz Fortuna, Marko Grobelnik, Dunja Mladenic, "Triplet extraction from sentences" in Artificial Intelligence Laboratory, Jožef Stefan Institute, Slovenia, Nov. 7, 2008.

[4] D. Lin, P. Pantel, "DIRT - Discovery of inference rules from text" in Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001. pp. 323-328, 2001.

[5] J. Leskovec, M. Grobelnik, N. Milic-Frayling, "Learning sub-structures of document semantic graphs for document summarization" in Proceedings of the 7th International Multi-Conference Information Society IS 2004, Volume B. pp. 18-25, 2004.

[6] J. Leskovec, N. Milic-Frayling, M. Grobelnik, "Impact of linguistic analysis on the semantic graph coverage and learning of document extracts" in National Conference on Artificial Intelligence, 2005.

[7] O. Etzioni, M. Cafarella, D. Downey, A. M. Popescu, T. Shaked, S. Soderland, D. S. Weld, A.Yates. Unsupervised named-entity extraction

from the Web: An experimental study. Artificial Intelligence, Volume 165, Issue 1, June 2005, Pages 91-134.

[8] Marie-Catherine de Marneffe, Bill MacCartney, Christopher D. Manning, "Generating typed dependency parses from phrase structure parses" in LREC 2006.

[9] Marie-Catherine de Marneffe, Christopher D. Manning, "The Stanford typed dependencies representation" in COLING Workshop on Cross-framework and Cross-domain Parser Evaluation, 2008.

[10] Marie-Catherine de Marneffe, Christopher D. Manning, "The Stanford typed dependencies manual" in Revised for Stanford Parser v1.6.2, February, 2010.

[11] Daniel Cer, Marie-Catherine de Marneffe, Daniel Jurafsky, Christopher D. Manning, "Parsing to Stanford dependencies: Trade-offs between speed and accuracy" in 7th International Conference on Language Resources and Evaluation (LREC 2010).

[12] Dick Grune and Ceriel Jacobs, "Parsing Techniques – A Practical Guide," in Proceedings of the 8th International Conference, CICLing 2007, Mexico City, A. Gelbukh (Ed), pp. 311-324, Springer, Germany, 2007.

[13] Takale, S. A. (2010). Measuring Semantic Similarity between Words Using Web Documents. International Journal of Advanced Computer Science and Applications - IJACSA, 1(4), 78-85.

[14] Firdhous, M. F. M. (2010). Automating Legal Research through Data Mining. International Journal of Advanced Computer Science and Applications - IJACSA, 1(6), 9-16.

[15] Satapathy, S. K., & Mishra, S. (2010). Search Technique Using Wildcards or Truncation : A Tolerance Rough Set Clustering Approach. International Journal of Advanced Computer Science and Applications - IJACSA, 1(4), 73-77.

AUTHORS PROFILE

Ms. Anjali Ganesh Jivani is an Associate Professor in the Department of Computer Science and Engineering, The Maharaja Sayajirao University of Baroda. She is pursuing her doctorate in Computer Science and her research area is Text Mining. She has published a number of research papers related to Text Mining. Her paper titled "'The Shared Nearest Neighbor Algorithm with Enclosures (SNNAE)" has been published by IEEE Computer Society and World Research Organization, ISBN 978-0-7695-3507-4, DOI 10.1109/CSIE2009.997, pg. 436. The paper is available on the ACM DL portal. She has co-authored a book titled 'SQL & PL/SQL Practice Book', ISBN 978-81-910796-0-9.

Ms. Amisha Hetal Shingala is an Assitant Professor in the Department of MCA, Gujarat Technological University. She is pursuing her doctorate in Information Retrievel. She has published a number of papers. She has co-authored a book titled 'SQL & PL/SQL Practice Book', ISBN 978-81-910796-0-9.

Dr. Paresh V. Virparia is an Associate Professor in the Dept. of Computer Science, SP Universiy. He has done doctorater in Simulation and Modeling. He has guided 9 Ph.D. students.

A Performance Study of Some Sophisticated Partitioning Algorithms

D.Abhyankar
School of Computer Science
Devi Ahilya University
Indore, M.P.

M.Ingle
School of Computer Science
Devi Ahilya University
Indore, M.P.

***Abstract*— Partitioning is a central component of the Quicksort which is an intriguing sorting algorithm, and is a part of C, C++ and Java libraries. Partitioning is a key component of Quicksort, on which the performance of Quicksort ultimately depends. There have been some elegant partitioning algorithms; Profound understanding of prior may be needed if one has to choose among those partitioning algorithms. In this paper we undertake a careful study of these algorithms on modern machines with the help of state of the art performance analyzers, choose the best partitioning algorithm on the basis of some crucial performance indicators.**

Keywords- Quicksort; Hoare Partition; Lomuto Partition; AQTime.

I. INTRODUCTION

Partitioning is undoubtably a core part of the Quicksort on which the performance ultimately depends. Quicksort is a leading and widely used sorting algorithm. For instance C, C++ and Java libraries use Quicksort as their sorting routine. The Partitioning is a key component of the Quicksort and selection algorithm. There are several partitioning algorithms that accomplish the task, but only a few deserve special attention. Hoare, Lomuto, Modified Lomuto and Modified Hoare are those few selected partition algorithms. This paper carries out an in depth study of the selected partitioning algorithms. The important question is as to which partitioning algorithm is superior so that we can call the superior algorithm in sorting routine. This study attempts to answer the same question. In past Scientists studied and compared these algorithms; the comparisons however were theoretical and were made on old architectures. An algorithm effective on old architectures may not be effective on modern machines. A study valid on old architectures may not be so on modern architectures. Moreover in past researchers did not have advanced performance analyzers to study cache miss and page faults. Consequently researchers relied on cache simulations. Therefore their results may be inaccurate. Hence it is beneficial to compare the algorithms on contemporary architectures using state of the art performance analyzers.

It has not escaped our notice that state of the art machines are Multicore and if an algorithm has to be effective it should be Multicore ready [13]. Future lies in parallel/multithreaded algorithms, but even then one should not forget that parallel algorithms or multithreaded algorithms will need sequential algorithms at lower level. The basic question is which sequential sorting algorithm to call at lower level. Calling a slow sequential algorithm at lower level will neutralize the advantage of parallel sorting gained by multiple cores. So the question which sequential sorting is the best option at lower level is of paramount importance. Literature suggests that Quicksort offers the most effective answer at least today. If the Quicksort is lower level sequential sorting algorithm, then the very next question is which Partitioning algorithm we should choose. This study is going to solve the same question.

To study the performance of selected partitioning algorithms on contemporary machines is the central idea of the paper. A fair test of the algorithm's performance is its execution time; however the drawback of this approach is that no intuition is provided as to why the execution time performance was good or bad. The reason(s) may be high instruction count, high cache miss count and high branch misprediction count. Even high page fault count affects the performance. Earlier researchers studied the impact of these factors using cache simulation and similar techniques. Fortunately today researchers have performance analyzing softwares which are not merely effective in capturing execution time but also acquire accurate data about cache miss, branch mispredictions and page faults.

II. LITERATURE REVIEW

In the past researchers did not enjoy the luxury of sophisticated profilers which we enjoy now. Instead they relied heavily on theoretical models and cache simulations. Majority of algorithm researchers compare the algorithmic performance on the basis of unit cost model. The RAM model is a most commonly used unit cost model in which all basic operations involve unit cost. The advantage of unit cost model is that it is simple and easy to use. Moreover it produces results which are easily comparable. However, this model does not reflect the memory hierarchy present in modern machine. It has been observed that main memory has grown slower relative to processor cycle times, consequently Cache miss penalty has grown significantly [12]. Thus good overall performance cannot be achieved without keeping cache miss count as low as possible. Since RAM model does not count cache miss, it is no longer a useful model.

Usually algorithm researchers in sorting area only count particular expensive operations. Analysis of sorting and searching algorithms, for instance, only counts the number of comparisons and swaps. There was exquisite logic behind only counting comparison operation which was expensive in the past. That simplified the analysis and still retained accuracy since the bulk of the costs was captured, but this is no longer true because the shift in the technology renders the "expensive operations" inexpensive and vice versa. Same happened with comparison operation which is not expensive anymore. Indeed it is no more expensive than addition or copy. Hence the study favours a practical approach and is not biased towards a single performance indicator. The idea is to have a fairly objective view and goal of good overall performance rather than concentrating on a single performance indicator.

Literature reveals that every partitioning algorithm incurs (n-1) comparisons, where n is total number of elements in the array[1, 2, 3, 4, 5, 6, 7, 8, 9]. Partitioning algorithms differ in swap count or data transfer operations. Hoare partition & Modified Hoare partition algorithms lead to adaptiveness of swap count / data transfer operation count. In the worst case, for Hoare and Modified Hoare algorithms swap count/ data transfer count is approximately(n/2), whereas for Lomuto and Modified Lomuto swap count/data transfer count is approximately (n)[14].

III. PERFORMANCE STUDY ON MODERN ARCHITECTURES

This paper studies the performance of Hoare partition, Lomuto partition, Modified Hoare partition and modified Lomuto partition on contemporary computers. Thus to study algorithms were tested on Pseudorandom numbers using state of the art Machines. Experiments were performed on state of the art COMPAQ PC which was equipped with Windows Ultimate operating system. Following tables and figure present the average case statistics generated by the tests on 3 important performance indicators: elapsed time, CPU Cache Miss, Branch mispredictions. AQtime software was instrumental in gathering the reliable profiling data. Elapsed time given in the table is in milliseconds.

TABLE I: STATISTICS OF LOMUTO PARTITION

Lomuto Partition			
N	***Elapsed Time***	***CPU MisPredicted Branches***	***CPU Cache Misses***
10000	7.52	280360	435
20000	16.83	596571	857
30000	28.49	998545	1778
40000	30.43	1372734	3651
50000	50.65	1708752	3132
60000	52.82	2064278	7865
70000	68.02	2660145	3406
80000	72.52	2990947	5161
90000	86.64	3270575	6328
100000	103.15	3815563	6324

TABLE II: STATISTICS OF MODIFIED LOMUTO PARTITION

Modified Lomuto Partition			
N	***Elapsed Time***	***CPU MisPredicted Branches***	***CPU Cache Misses***
10000	1.57	66588	82
20000	2.99	148860	60
30000	4.65	222592	171
40000	6.46	312973	269
50000	8.2	387050	447
60000	10.27	477689	422
70000	11.75	569962	590
80000	13.32	637354	1206
90000	15.32	741244	206
100000	17.62	806429	798

TABLE III: STATISTICS OF HOARE PARTITION

Hoare Partition			
N	***Elapsed Time***	***CPU MisPredicted Branches***	***CPU Cache Misses***
10000	3.52	165923	745
20000	7.55	351172	650
30000	12.26	530738	615
40000	16.04	708813	1139
50000	20.53	890882	982
60000	24.95	1141577	1566
70000	31.46	1303499	4242
80000	33.96	1561376	1686
90000	38.36	1680148	3812
100000	44.58	1862329	2100

TABLE: IV: STATISTICS OF MODIFIED HOARE PARTITION

Modified Hoare Partition			
N	***Elapsed Time***	***CPU MisPredicted Branches***	***CPU Cache Misses***
10000	1.46	83287	45
20000	3.08	173563	520
30000	4.71	267342	319
40000	6.28	369235	234
50000	7.87	463831	245
60000	9.55	575380	1014
70000	11.17	725573	714
80000	12.92	724582	785
90000	14.22	829880	1489
100000	16.15	908749	1091

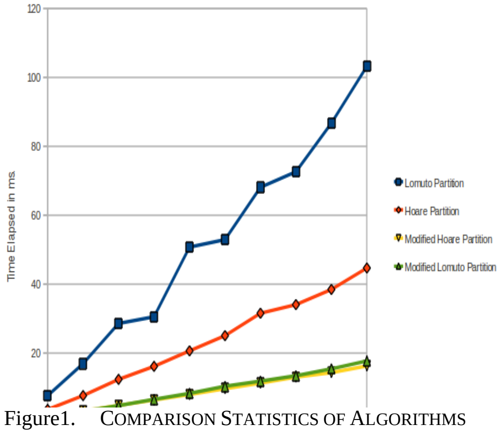


Figure1. COMPARISON STATISTICS OF ALGORITHMS

IV. Analysis, Results and Conclusion

Tables and Figure 1, show the results based on random input, depict the performance on 3 crucial performance indicators. Since Page fault count was 0 for each one of the algorithms, it was not shown explicitly in the tables. Zero page fault count is due to large main memory size which was not feasible earlier. Modified Hoare partition outperforms the other algorithms in almost all entries in the table. Modified Lomuto is the second one to finish and is not too behind. Modified Lomuto is followed by Hoare partition which in turn is followed by Lomuto which is the last one to complete. It is easy to see that among the studied algorithms the one with the better cache miss count is usually the first one to complete the partitioning. Lomuto algorithm and Hoare algorithm are slow because of their higher instruction count, poor cache miss count and fairly high branch misprediction count. The interesting question that emerges is why Modified Hoare and Modified Lomuto have lower cache miss count whereas others have cache miss count on higher side. The intuitive reason is that instruction cache miss count is likely to go down as overall instruction count and code size goes down. If we can keep data cache miss count in check then overall cache miss count will be low. Same seems to have happened with Modified Hoare and Modified Lomuto partitioning algorithms.

References

[1] J. L. Bentley and M. D. Mcilroy "Engineering a sort function," Software—practice and experience, VOL. 23(11), 1249–1265 (NOVEMBER 1993).

[2] R. Sedgewick, ‘Quicksort’, PhD Thesis, Stanford University (1975).

[3] C. A. R. Hoare, "Partition: Algorithm 63, " "Quicksort: Algorithm 64," Comm. ACM 4(7), 321-322, 1961.

[4] D. E. Knuth, The Art of Computer Programming, Vol. 3, Pearson Education, 1998.

[5] C. A. R. Hoare, "Quicksort," Computer Journal5 (1), 1962, pp. 10-15.

[6] S. Baase and A. Gelder, Computer Algorithms:Introduction to Design and Analysis, Addison-Wesley, 2000.

[7] J. L. Bentley, "Programming Pearls: how to sort," Communications of the ACM, Vol. Issue 4, 1986, pp. 287-ff.

[8] R. Sedgewick, "Implementing quicksort Programs," Communications of the ACM, Vol. 21, Issue10, 1978, pp. 847-857.

[9] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, Introduction to Algorithms, Second Edition. MIT Press and McGraw-Hill, 2001.

[10] G. S. Brodal, R. Fagerberg and G. Moruz, "On the adaptiveness of Quicksort," Journal of Experimental AlgorithmsACM, Vol. 12, Article 3.2, 2008.

[11] S.Carlsson, “A variant of HEAPSORT with almost optimal number of comparisons,” Information Processing Letters Modified 24:247-250,1987.

A. G. LaMarca, “Caches and Algorithms,” PhD theses University of Washington, 1996.

[12] M. Edahiro, “Parallelizing fundamental algorithms such as sorting on multi-core processors for EDA acceleration, “ASP-DAC '09 Proceedings of the 2009 Asia and South Pacific Design Automation Conference, 2009.

[13] D. Abhyankar and M. Ingle, “Engineering of a Quicksort Partitioning Algorithm,” Journal of Global Research in Computer Science Vol. 2, No. 2, 2011.

[14] V. Aho, J. E. Hopcroft, J.D. Ulman, “The Design and Analysis of Computer Algorithms, “Addison-Wesley, 1974.

[15] V. Aho, J. E. Hopcroft, J.D. Ulman, “Data Structures and Algorithms, “Addison-Wesley, 1983.

[16] J. L. Bentley, “Writing Efficient Programs, “ Prentice-Hall, 1982.

[17] G. Brassard and P. Bratley, “Fundamentals of Algorithmics, “Prentice Hall, 1996.

[18] R. Mansi, “Enhanced Quicksort Algorithm, “ The International Arab Journal of Information Technology, Vol. 7, No. 2, April 2010