

Natarajan Meghanathan
Jan Zizka (Eds)

Computer Science & Information Technology

The Sixth International Conference on Advances in Computing and
Information Technology (ACITY 2016)
Chennai, India, July 23~24, 2016



AIRCC Publishing Corporation

Volume Editors

Natarajan Meghanathan,
Jackson State University, USA
E-mail: nmeghanathan@jsums.edu

Jan Zizka,
Mendel University in Brno, Czech Republic
E-mail: zizka.jan@gmail.com

ISSN: 2231 - 5403
ISBN: 978-1-921987-53-3
DOI : 10.5121/csit.2016.60901 - 10.5121/csit.2016.60911

This work is subject to copyright. All rights are reserved, whether whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the International Copyright Law and permission for use must always be obtained from Academy & Industry Research Collaboration Center. Violations are liable to prosecution under the International Copyright Law.

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The Sixth International Conference on Advances in Computing and Information Technology (ACITY 2016) was held in Chennai, India, during July 23~24, 2016. The Seventh International Conference on VLSI (VLSI 2016), The Sixth International Conference on Artificial Intelligence, Soft Computing and Application (AIAA 2016) and The Third International Conference on Computer Networks & Data Communications (CNDC 2016) were collocated with the ACITY-2016. The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The ACITY-2016, VLSI-2016, AIAA-2016, CNDC-2016 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, ACITY-2016, VLSI-2016, AIAA-2016, CNDC-2016 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the ACITY-2016, VLSI-2016, AIAA-2016, CNDC-2016.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Natarajan Meghanathan
Jan Zizka

Organization

General Chair

Natarajan Meghanathan
Dhinaharan Nagamalai

Jackson State University, USA
Wireilla Net Solutions, Australia

Program Committee Members

Abdelkrim Haqiq	Hassan 1st University, Morocco
Ali Asghar Safaei	Tarbiat Modares University, Iran
Ali Elkateeb	University of Michigan-Dearborn, USA
Amelia Zafra Gomez	University of Cordoba, Spain
Ankit Chaudhary	Truman State University, USA
Balasubramanian K	European University of Lefke, Turkey
Chen ZhiQiang	University of Missouri-Kansas City, USA
Chiranjib Sur	University of Florida, US
Dana Petcu	West University of Timisoara, Romania
Daniel Dacuma Dasig	Jose Rizal University, Philippines
Djouani	University Paris-Est Créteil, France
Ehsan Heidari	Islamic Azad University, Iran
Ercan Oztemel	Marmara University, Turkey
Erritali Mohammed	Sultan Moulay Slimane University, Morocco
Farzad Kiani	Istanbul S.Zaim University, Turkey
Fatih Korkmaz	Çankiri Karatekin University, Turkey
Francisco Macia Perez	University of Alicante, Spain
Hamid Mcheick	University of Quebec at Chicoutimi, Canada
Hao-En Chueh	Yuanpei University, Taiwan, Republic of China
Hassan Chizari	University Technology Malaysia, Malaysia
Hossein Iranmanesh	University of Tehran, Iran
Houcine Hassan	Universidad Politecnica de Valencia, Spain
Intisar Al-Mejibli	University of Essex, United Kingdom
Isa Maleki	Islamic Azad University, Iran
Ishthaq Ahamed	G Pulla Reddy Engineering College, India
Jianfeng Wang	Xidian University, China
Jin-Whan Kim	Youngsan University, South Korea
Jose A.R.Vargas	University of Brasilia, Brazil
Jose Vicente Berna	University of Alicante, Spain
Kannammal	Coimbatore Institute of Technology, India
Khattab Alheeti	Anbar University College of Computer, Iraq
Kwangjin Park	Wonkwang University, South Korea
Labed Said	University of Constantine, Algeria
Li Zheng	University of Bridgeport, USA
Lin Wang	University of Jinan, China
Luigi PATRONO	University of Salento, Italy
Luis Fernando de Mingo Lapez	Technical University of Madrid, Spain

Lynne Grewe	California State University East Bay, USA
Mohamed Ashik	Salalah College of Technology, Oman
Mahdi Mazinani	IAU Shahreqods, Iran
Maher Ben Jemma	University of Sfax, Tunisia
Mahi Lohi	University of Westminster, UK
Manish Mishra	Haramaya University, Ethiopia
Marc Sevaux	Universite de Bretagne, France
Marta Beltran Pardo	Universidad Rey Juan Carlos, Spain
Mayyash	The California State University(CSU), USA
Michal Wozniak	Wroclaw University of Technology, Poland
Mohamed Waleed Fakhr	University of Bahrain, Bahrain
Mohammad Talib	University of Botswana, Botswana
Mohammad Yamin	King Abdulaziz University, Saudi Arabia
Mohammed Amin	Higher Colleges of Technology, UAE
Mohammed Ghanbari	University of Essex, United Kingdom
Moses Ekpenyong	University of Edinburgh, Nigeria
Mujiono Sadikin	Universitas Mercu Buana, Indonesia
Nabila Labraoui	University of Tlemcen, Algeria
Nadia Qadri	University of Essex, United Kingdom
Nazmus Saquib	University of Manitoba, Canada
Neda Darvish	Islamic Azad University, Iran
Neetesh Saxena	The State University of New York, USA
Nourddine Bouhmala	Buskerud and Vestfold University, Norway
Othmane Alaoui Fdili	Mohammed V University, Morocco
Peter Ogedebe	BAZE University, Nigeria
Rahil Hosseini	Islamic Azad University, Iran
Raveendra Rao	University of Western Ontario, Canada
Saad M.Darwish	Alexandria University, Egypt
Selwyn Piramuthu	University of Florida, Florida
Sergio Kurokawa	Universidade Estadual Paulista, Brazil
Sergio Pastrana	University Carlos III of Madrid, Spain
Seyyed Reza Khaze	Islamic Azad University, Iran
Shahid Siddiqui	Integral University, India
Stefano Berretti	University of Florence, Italy
SugamSharma	Ames Laboratory Information System, USA
Thuc-Nguyen	University of Science, Vietnam
Tri Kurniawan Wijaya	Technische Universitat Dresden, Germany
Weifa Liang	Australian National University, Australia
Yahya Slimani	Faculty of Sciences of Tunis, Tunisia
Yassine Boukal	University of Lorraine, France
Yong-Jin Lee	Korea National University of Education, Korea
Youssef Fakhri	University Ibn Tofail, Morocco
Yuhanis binti Yusof	Universiti Utara Malaysia, Malaysia
Zaher Al Aghbari	University of Sharjah, UAE
Zaw Zaw Htike	International Islamic University, Malaysia
Zivic Natasa	University of Siegen, Germany

Technically Sponsored by

Networks & Communications Community (NCC)



Computer Science & Information Technology Community (CSITC)



Information Technology Management Community (ITMC)



Organized By



Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

The Sixth International Conference on Advances in Computing and Information Technology (ACITY 2016)

A Survey on Security Risk Management Frameworks in Cloud Computing	01 - 11
<i>Rana Alosaimi and Mohammad Alnuem</i>	
Automated Short Answer Grader Using Friendship Graphs	13 - 22
<i>Soumajit Adhya and S. K. Setua</i>	
Automatic Generation and Optimization of Test Data Using Harmony Search Algorithm	23 - 32
<i>Rajesh Kumar Sahoo, Deeptimanta Ojha, Durga Prasad Mohapatra and Manas Ranjan Patra</i>	
Big Graph : Tools, Techniques, Issues, Challenges and Future Directions	119 - 128
<i>Dhananjay Kumar Singh and Ripon Patgiri</i>	

The Seventh International Conference on VLSI (VLSI 2016)

Glitch Analysis and Reduction in Combinational Circuits	33 - 40
<i>Ronak Shah</i>	
Analysis of CMOS and MTCMOS Circuits Using 250 Nano Meter Technology	41 - 51
<i>M Suresh, A K Panda, Mukesh Sukla, Marakonda Patnaikuni Vasanthi and Sowpati Santhi</i>	
IP Core Design of Hight Lightweight Cipher and its Implementation	97 - 105
<i>Sruthi.N, R.Nandakumar and Rajkumar.P</i>	
Design of IEEE 1149.1 Tap Controller IP Core	107 - 118
<i>Shelja A S, Nandakumar R and Muruganantham C</i>	

**The Sixth International Conference on Artificial Intelligence, Soft
Computing and Application (AIAA 2016)**

**Dengue Detection and Prediction System Using Data Mining with Frequency
Analysis..... 53 - 67**

Nandini V, Sriranjitha R and Yazhini T.P

OBIA on Coastal Landform Based on Structure Tensor..... 69 - 80

Sun Shuting, Liu Jianqiang and Zou Bin

**The Third International Conference on Computer Networks & Data
Communications (CNDC 2016)**

**Concealed Data Aggregation with Dynamic Intrusion Detection System to
Remove Vulnerabilities in Wireless Sensor Networks..... 81 - 96**

Bharat Bhushan, Keshav Kaushik and G Sahoo

A SURVEY ON SECURITY RISK MANAGEMENT FRAMEWORKS IN CLOUD COMPUTING

Rana Alosaimi¹ and Mohammad Alnuem²

Department of Information Systems,
King Saud University, Riyadh, Saudi Arabia

¹Rana.io@hotmail.com

²malnuem@ksu.edu.sa

ABSTRACT

Cloud computing technology has experienced exponential growth over the past few years. It provides many advantages for both individuals and organizations. However, at the same time, many issues have arisen due to the vast growth of cloud computing. Organizations often have concerns about the migration and utilization of cloud computing due to the loss of control over their outsourced resources and cloud computing is vulnerable to risks. Thus, a cloud provider needs to manage the cloud computing environment risks in order to identify, assess, and prioritize the risks in order to decrease those risks, improve security, increase confidence in cloud services, and relieve organizations' concerns on the issue of using a cloud environment. Considering that a conventional risk management framework does not fit well with cloud computing due to the complexity of its environment, research in this area has become widespread. The aim of this paper is to review the previously proposed risk management frameworks for cloud computing and to make a comparison between them in order to determine the strengths and weaknesses of each of them. The review will consider the extent of the involvement and participation of consumers in cloud computing and other issues.

KEYWORDS

Cloud Computing; Risk Management & Information Security

1. INTRODUCTION

Cloud computing is a new paradigm shift in the technological industry which will continue to grow and develop in the next few years. The rate of organizations migrating to a cloud computing environment is increasing daily due to its advantages [1, 2]. The major cloud computing advantages which benefit organizations are: high scalability and flexibility in organizations' resources in order to meet peak time demand, excellent reliability and availability in that resources can be accessed from anywhere and at any time, and there is no upfront cost for installing and managing the software and hardware infrastructure [3, 4, 5, 6].

On the other hand, cloud computing also has brought many risks to organizations due to the fact that they outsource IT resources which make services completely managed and delivered by a third party. Therefore, such organizations might lose control over how they secure their environment and they might be concerned with privacy and security as the new technology is a

major source of new vulnerabilities in these areas [7, 8, 9, 10]. Therefore, it is important to establish several controls which will work together to decrease the risks, provide layered security, increase confidence in cloud services, and relieve the fear of using a cloud computing environment. Risk management is one of the cloud computing environment controls which aims to assess and manage risks related to cloud computing and to prevent those risks from impacting business goals.

This paper will provide a systematic review of the previously proposed risk management frameworks for cloud computing environments. The paper will be organized as follows. In section II, an overview of the two main subjects – cloud computing and risk management – will be provided. Section III will include reviews of relevant work on previously proposed cloud computing risk management frameworks. Section IV will present the advantages and disadvantages of each of them. Section V will include a discussion of the results of the review and a comparison of the frameworks. Section VI will conclude the paper and will include recommendations for future work in this area.

2. BACKGROUND

2.1. Cloud Computing

Cloud computing is a new type of computing model extended from distributed computing, parallel computing, and grid computing. It provides various additional features to users such as secure, quick, and convenient data storage and a net computing service centred on the Internet. The factors that have propelled the frequency of occurrence and development of cloud computing include the development of grid computing, the appearance of high-quality technology in storage and data transportation, and the appearance of Web 2.0 – especially the development of virtualization [11]. Cloud computing consists of five essential characteristics: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service [12, 13]. Cloud computing generally provides services on three main diverse levels [14, 15]. These models are:

- Software as a Service (SaaS) – cloud computing delivers an application which is already customized with all of the required hardware, software, operating system, and network to be accessible by various consumers (regardless of their location) by using the Internet without the need to install software on the servers [16].
- Platform as a Service (PaaS) – cloud computing offers a developmental environment and all the developers' requirements (such as software tools, libraries, programming languages, and services for cloud consumers to develop or install their own software and applications). The applications are then delivered to the users via the Internet [17].
- Infrastructure as a Service (IaaS) – cloud computing offers fundamental computing resources (such as processing, servers, storage, and networks) and virtualization technology for consumers to install and run their own operating systems and applications [18].

Furthermore, cloud computing services can be deployed in four ways dependent on consumers' requirements [19]. These four deployment models are:

- Public Cloud – the cloud infrastructure and computing resources are made available to public consumers over a public network. Multiple organizations can work on and access the provided infrastructure at the same time. The public cloud model is controlled and managed by a third party: a cloud provider [5, 20].

- Private Cloud – cloud services are dedicated only to a specific consumer (organization) and offer the highest security of client data and greater control over the cloud infrastructure. A private cloud may be managed and owned by the organization itself or a cloud provider [21].
- Community Cloud – in a community cloud, the cloud infrastructure is provisioned for a group of consumers (organizations) which have the same shared demands. They can share resources by using the connections between the associated organizations. The community cloud is similar to a private cloud in that it can be managed and owned either by the relevant community organizations or a cloud provider [5, 20].
- Hybrid Cloud – A hybrid cloud is a mixture of two or more types of cloud deployment models (public, private, or community) which are connected together to allow for the transfer of data and application between them but without affecting each other [21].

2.2. Risk Management

Organizations usually face and are exposed to several types of risk (e.g. policy, programme, operational, project, financial, human resources, technological, health, safety, and political risks) [22]. The International Organization for Standardization (ISO) defined risk as “the effect of uncertainty on objectives”. Otherwise, the risk is expressed as a combination of the consequences of an event and the associated probability of occurrence [23].

Risk management is a systematic mechanism for managing the risks or threats facing an organization in order to enable it to recognize the events that may result in unfortunate or damaging consequences and to establish the best course of action for identifying, assessing, understanding, acting on, and communicating risk issues [22, 24]. Risk management adds value and offers many objectives to an organization. Some of these objectives are: increasing system security, protecting and enhancing the organization’s assets, making well-informed decisions, and optimizing operational efficiency [25, 26].

3. RISK MANAGEMENT FRAMEWORKS IN CLOUD COMPUTING

The researchers’ efforts are based on three different perspectives on cloud computing risk assessment. The first proposed risk assessment frameworks can be used only by a cloud computing consumer. It was suggested that, in some cases, risk should be transferred to the cloud provider or a trusted third party [27, 28]. However, these researchers ignore the fact that the cloud provider owns and manages the infrastructure of the cloud environment and cannot disclose their security models and procedures to anyone who might be a malicious user. On the other hand, other researchers have proposed risks should only be assessed by the cloud provider, without taking into account the importance of involving the cloud consumers in the process because the cloud provider is the real owner of the data and the only party who knows the real value of the assets in the cloud environment [29, 30]. Thereafter, some researchers believed in the importance of involving the consumers in the risk assessment process [31, 32, 33]. When and to what extent the consumers are involved in this process was considered differently by the different proposed risk assessment frameworks.

3.1. Risk Management by Cloud Consumers

Saripalli and Walters [27] presented a quantitative framework by which the impact and risk of cloud security (based on the Federal Information Processing Standards (FIPS)) can be assessed [25]. Thus, in addition to the security objectives already defined by the US Federal Information

Security Management Act (FISMA) for information and information systems – confidentiality, integrity, and availability – Saripalli and Walters added three more security objectives in the context of cloud platforms – multiparty trust, mutual auditability, and usability. These six security objectives of cloud platforms are referred to as the CIAMAU framework. The typical threats and events are mapped into one or more of these six categories. Saripalli and Walters assessed risk as a combination of the probability of a security threat event and its measured consequence or impact. The probability of a threat event is assessed by using statistical data and the impact of a threat event is evaluated based on expert opinions by using the modified Wideband Delphi method to collect the necessary information [34].

Tanimoto, et al. [28] considered the risk factors in cloud computing from a consumer's viewpoint based on the risk breakdown structure (RBS) method. Their study aimed to address three main security subjects in terms of the cloud environment: the security guarantees in a disclosure environment, the existence of two or more stakeholders, and mission critical data problems. They classified the risk factors into three main divisions: risks for a company which is introducing cloud computing, risks for a cloud service provider, and risks for others. Therefore, in terms of risk analysis, a hybrid method has been proposed based on a quantitative decision tree analysis and the qualitative risk matrix method. Thus, the risk matrix method categorizes risk into four classifications in accordance with the degree of incidence and generation frequency: risk avoidance, risk mitigation, risk acceptance, and risk transference. Consequently, the analyzed and evaluated risks and a detailed countermeasure and proposal have been produced based on these results.

3.2. Risk Management by a Cloud Provider

Fito, et al. [29] proposed a semi-quantitative BLO-driven cloud risk assessment (SEBCRA) approach which is based on the Federation of European Risk Management Association's standards [35]. Fito, et al.'s risk management procedure aims to determine the risk impacts (either positive or negative) on the level of the business objectives (BLOs) of a given cloud organization instead of its impact on the whole cloud environment. Therefore, their risk management framework involves these steps: SEBCRA (the overall process of risk analysis and evaluation), risk reporting and communication, risk treatment, and risk monitoring. SEBCRA was proposed as it is a core sub-process by which cloud risks can be prioritized according to their impact and consequences on different BLOs. In the risk analysis process, Fito, et al. used a standard level matrix to extract the risk level estimation (RLE) as the output for each of the affected BLOs, which is the product of the risk probability and its impact on the BLO. Thus, any risks in which RLE is within unacceptable ranges and has a negative impact on the BLOs are avoided and this has the benefit of leading to an improvement in achieving the BLOs.

Zhang, et al. [30] presented an information risk management framework for cloud computing which covers all cloud service models and cloud deployment models. The framework is based on the evolving ISO/IEC 27001 standards [36], the NIST risk management guide for information technology systems [25], and the Booz Allen Hamilton information security governance [37]. Furthermore, it is similar to the traditional standard quality management (Plan, Do, Check, Act) cycle of continuous improvement and involves seven processes: selection of the relevant critical areas, strategy and planning, risk analysis, risk assessment, risk mitigation, the assessment and monitoring programme, and risk management review. Thus, Zhang, et al. focused on critical areas in cloud computing which must be protected and designed in order to protect the security objectives of information assets: confidentiality, integrity, and availability. In addition, risk analysis and assessment processes, threats, vulnerability identification, and assessment of the output from the identification can ascertain the risk levels (High, Medium, and Low) of relevant

critical areas which were selected previously from 12 critical areas which address both tactical and strategic security.

3.3. Risk Management by Cloud Providers and Consumers

Almorsy, et al. [31] proposed a security management framework aimed at improving collaboration between cloud providers, service providers, and service consumers in terms of managing the security of the cloud platform and the hosted services. Cloud consumers are advised to participate in every step of the risk assessment processes in order to extend their security management process (SMP) to include cloud-hosted assets. The framework has been introduced based on aligning the NIST-FISMA standard with the cloud computing model [38]. Almorsy, et al.'s framework consists of three main layers: a management layer, an enforcement layer, and a feedback layer. The framework includes six main phases: service security categorization, security controls selection, security controls implementation, security controls assessment, service authorization, and security monitoring. Their security management framework can be applied to each developed and deployed service in cloud computing and it is considered the overall security categorization for each service.

Xie, et al. [32] presented a risk management framework which includes users, providers, and third party agencies. The main aim of their framework is for cloud providers to ascertain a user's requirements clearly and to enhance the trust between them. The framework is composed of five basic processes: user requirement self-assessment, cloud service providers' desktop assessment, risk assessment, third party agencies review, and continuous monitoring. In the user requirement self-assessment phase, the user should determine the required cloud computing service and deployment model and the security level of authentication, access control, auditing, data integrity, etc. in order to determine potential cloud providers based on these selections. Thus, the aims of the cloud providers' desktop assessment phase are to evaluate the plans of those candidates and to review their past security levels. The risk assessment phase consists of seven stages: preparation of the risk assessment, asset identification, threat identification, vulnerability identification, existing security measures, risk analysis, and risk assessment documentation. The third party agencies review phase involves authoritative security evaluation institutions (including the review group and expert group) which review the procedures of the security assessment plan of the user. Albakri, et al. [33] proposed a security risk assessment framework based on the SaaS model with a public deployment model based on the ISO/IEC 27005 standard. The framework considered both the cloud provider and the cloud consumer in the risk assessment process by providing a dynamic relationship between them. They aim to balance the realistic results which will derive from the participation of consumers and the potential complexity which may occur due to their involvement. Thus, cloud computing consumer participation in risk management processes is limited to only three tasks: determining the regulatory and legal requirements, determining the security risk factors, and getting feedback from the cloud provider and applying the required security tasks. Therefore, their framework consists of six phases: context establishment, risk assessment, risk treatment, risk acceptance, risk communication and consultation, and risk monitoring and review. In the context establishment process, each consumer should start its own context establishment for its data which will move to the cloud environment to define legal compliance. Furthermore, the same applies to the risk assessment process. Each consumer should identify the risk for its own data which will move to the cloud. Thereafter, the cloud provider will be able to perform a risk analysis and the rest of the processes for its entire infrastructure and consumers' data.

4. RISK MANAGEMENT FRAMEWORKS PROS AND CONS

There is no perfect risk management framework and, due to the complexity of the cloud computing environment, there are many reasons which could make a framework more effective

or which could reduce its effectiveness. Table 1 below represents the strengths and weaknesses of each of the above mentioned security risk management frameworks which have been proposed for use in a cloud computing environment.

Table 1. Risk Management Frameworks Advantages and Disadvantages.

Research paper	Advantages	Disadvantages
Saripalli and Walters	<ul style="list-style-type: none"> • The approach is fully iterative convergence and enables a comparative assessment of the relative robustness of different cloud vendor offerings in a defensible manner. • It proposes three additional specific security objectives for a cloud environment to be appropriate for a cloud security risk assessment. 	<ul style="list-style-type: none"> • It requires the careful and precise collection of input data for a probability calculation of threat events, which needs to be used to assess cloud computing risks. • It only focuses on risk assessment, which is only one step in the risk management process. The remaining steps are still required. • A quantitative risk assessment method has been used; thus, the results may be confusing and even imprecise. In addition, the method is expensive and requires solid experience with advanced tools. • The risk assessment has focused only on the cloud consumer and has overlooked that the cloud provider is the manager and owner of the cloud infrastructure.
Tanimoto, et al.	<ul style="list-style-type: none"> • This approach analyses and ascertains the risk factors of cloud computing and gives detailed countermeasures. • It uses a combination of quantitative and qualitative methods for risk analysis and achieved the advantages of both. It has avoided bias and inaccuracy in the assessment results. 	<ul style="list-style-type: none"> • It lacks a risk identification process for the threats, vulnerabilities, and assets of a cloud computing environment. • The risk factors were ascertained only from the consumers' viewpoints and the approach overlooked that the cloud provider is the manager and owner of the cloud infrastructure.
Fito, et al.	<ul style="list-style-type: none"> • This approach evaluates the impact of cloud risks on the BLOs of a cloud organization, instead of considering the impacts on the whole cloud environment. It therefore has strong focus and precision. • It uses a combination of quantitative and qualitative methods for risk analysis and achieves the advantages of both. It has avoided bias and inaccuracy in the assessment results. 	<ul style="list-style-type: none"> • There is no explanation for the risk identification method, which is an important and critical process in the risk assessment of cloud environment. • The impact of risks has been evaluated based only on the BLOs of a cloud provider and has overlooked consumers' objectives and the fact that the cloud consumer is the real owner of the data assets.

Zhang, et al.	<ul style="list-style-type: none"> The risk management was based on selecting critical areas in a cloud computing environment, which makes the risk assessment process strongly focused. 	<ul style="list-style-type: none"> The risk management was semi-static because the list of critical areas was fixed. This may make the risk assessment of the cloud environment inflexible and some of the risks may be ignored. A qualitative risk assessment method was followed. This makes the costs and benefits analysis during the selection of recommended controls difficult. The risk management has focused only on the cloud provider and has overlooked that the cloud consumer is the real owner of the data assets.
Almorsy, et al.	<ul style="list-style-type: none"> This approach tackles the loss of trust and security control problems by enabling cloud consumers to extend their SMP to include cloud-hosted assets. It mitigates the loss of control for cloud providers in terms of the hosted services developed by other parties. The security management framework was undertaken separately for each of the provided services. This is where the problem of multi-tenancy lies. 	<ul style="list-style-type: none"> Cloud consumers were involved in every step of the risk assessment processes. This complicates the risk assessment processes, particularly when the number of consumers increases. A qualitative risk assessment method was followed. This makes the costs and benefits analysis during the selection of recommended controls difficult.
Xie, et al.	<ul style="list-style-type: none"> This approach analyses the security status of cloud service providers by reviewing historical incidents. It introduces third party assessment agency to ensure the effectiveness and safety of cloud computing applications. 	<ul style="list-style-type: none"> Consumer involvement was not really considered to be active in the risk assessment process, which is only able to decide the security level in general and to select a cloud computing service and deployment model. Consumers are only involved in determining the appropriate cloud providers based on their requirements. A qualitative risk assessment method was followed. This makes the costs and benefits analysis during the selection of recommended controls difficult. There is a lack of risk treatment or acceptance in terms of the appropriate action to be taken for each risk.

Albakri, et al.	<ul style="list-style-type: none"> • This approach activates the involvement of consumers in the risk management process. • It tries to balance between the benefits of the participation of consumers and the complexity caused thereby. 	<ul style="list-style-type: none"> • The involvement of consumers involves notifying them at each phase that their participation is needed and completion of their responses must be awaited. This could disrupt or delay the process. • The cloud computing consumer does not participate in risk treatment and acceptance. It is the consumer who experiences the risks to its own assets and, therefore, they should make the decision. • A qualitative risk assessment method has been followed. This makes the costs and benefits analysis during the selection of recommended controls difficult.
-----------------	---	--

5. RESULTS AND DISCUSSION

Based on the review of several frameworks proposed previously by different authors, it can be confirmed that the traditional risk management may fail and may not fit well with a cloud computing environment. Thus, the strengths and weaknesses of those frameworks lead to a conclusion that some of key issues should be taken into account when applying a risk management framework to a cloud computing environment. These issues are outlined as follows:

- The involvement of consumers in the risk management process is important because they are the only ones who know the value of their assets.
- Consumer participation should not be limited to the extent of inactivity and consumers should not be involved in each step to the extent of complicating the process.
- Context establishment and risk identification (sub-processes of risk assessment) are critical processes in risk management.
- The participation of cloud consumers in the risk treatment process is significant due to the fact that they are part of the problem; therefore, they must be a part of the solution.
- It is preferable for the risk assessment process to be performed for each of the provided services separately in order to handle conflicts in the consumers' security requirements, due to the multi-tenancy feature of cloud computing.
- The conflict between consumers which appears in the risk identification process should be handled well in order to implement their security requirements and to achieve consumer satisfaction.
- It is advisable to use a combination of quantitative and qualitative methods in the risk analysis process in order to benefit therefrom and to avert their disadvantages.

6. CONCLUSION AND FUTURE WORK

The conventional risk management framework does not fit well with a cloud computing environment due to its specific characteristics. Therefore, several studies have been conducted on risk management in cloud computing. There are no specific criteria by which a risk management framework can be considered very bad or very good. A perfect risk management framework cannot be achieved. On the other hand, a particular framework can be decided on the basis of whether it is appropriate and effective or not based on a cloud environment.

This paper has reviewed the previously proposed risk management frameworks in cloud computing with specific regards to the participation of consumers therein and has determined the strengths and weaknesses of each. It was concluded that some specific issues are important when proposing a risk management framework for a cloud computing environment.

In future work, the researchers hereof propose a new risk management framework which takes the advantages of the previously proposed risk management frameworks and averts their disadvantages.

REFERENCES

- [1] R. Charanya, M. Aramudhan, K. Mohan, S. Nithya, "Levels of Security Issues in Cloud Computing," *International Journal of Engineering and Technology*, 2013.
- [2] M. Alzain, B. Soh, E. Pardede, "A Survey on Data Security Issues in Cloud Computing: From Single to Multi-Clouds," *Journal of Software*, 2013.
- [3] L. Qian, Z. Luo, Y. Du, and L. Guo, "Cloud Computing: An Overview," M. Jaatun, G. Zhao, & C. Rong, *Cloud Computing*, pp. 626-631. Berlin: Springer Berlin Heidelberg, 2009.
- [4] R. Bhadauria, and S. Sanyal, "Survey on Security Issues in Cloud Computing and Associated Mitigation Techniques," *International Journal of Computer Applications*, 2012.
- [5] A. Apostu, F. Puican, G. Ularu, G. Suciuc, and G. Todoran, "Study on advantages and disadvantages of Cloud Computing – the advantages of Telemetry Applications in the Cloud," *Recent Advances in Applied Computer Science and Digital Services*, 2013.
- [6] A. Apostu, F. Puican, G. Ularu, G. Suciuc, G. Todoran, "Study on advantages and disadvantages of Cloud Computing – the advantages of Telemetry Applications in the Cloud," *Recent Advances in Applied Computer Science and Digital Services*, 2013.
- [7] M. Hölbl, "Cloud Computing Security and Privacy Issues," *The Council of European Professional Informatics Societies*, 2011.
- [8] G. Tucker, and C. Li, "Cloud Computing Risks," *Proceedings on the International Conference on Internet Computing*, 2012.
- [9] T. Chou, "Security Threats on Cloud Computing Vulnerabilities," *International Journal of Computer Science & Information Technology*, 2013.
- [10] M. Ryan, "Cloud computing security: the scientific challenge, and a survey of solutions," *Journal of Systems and Software*, 2013.
- [11] S. Zhang, S. Zhang, X. Chen, and X. Huo, "Cloud Computing Research and Development Trend," *Second International Conference on Future Networks*, 2010.

- [12] M. Ali, S. Khan, A. Vasilakos, "Security in cloud computing: Opportunities and challenges," *Informatics and Computer Science Intelligent Systems Applications*, 2015.
- [13] F. Ahamed, S. Shahrestani, A. Ginige, "Cloud Computing: Security and Reliability Issues," *IBIMA*, 2013.
- [14] P. Sareen, "Cloud Computing: Types, Architecture, Applications, Concerns, Virtualization and Role of IT Governance in Cloud," *International Journal of Advanced Research in Computer Science and Software Engineering*, 2013.
- [15] I. Ashraf, "An Overview of Service Models of Cloud Computing," *International Journal of Multidisciplinary and Current Research*, 2014.
- [16] G. Kulkarni, P. Chavan, H. Bankar, K. Koli, and V. Waykule, "A new approach to Software as Service Cloud," *7th International Conference on Telecommunication Systems, Services, and Applications*, 2012.
- [17] J. Gibson, D. Eveleigh, R. Rondeau, and Q. Tan, "Benefits and Challenges of Three Cloud Computing Service Models," *Fourth International Conference on Computational Aspects of Social Networks*, 2012.
- [18] W. Hsu, "Conceptual Framework of Cloud Computing Governance Model - An Education Perspective," *IEEE Technology and Engineering Education*, 2012.
- [19] R. Sharma, R. Trivedi, "Literature review: Cloud Computing –Security Issues, Solution and Technologies," *International Journal of Engineering Research*, 2014.
- [20] F. Liu, J. Tong, J. Mao, R. Bohn, J. Messina, L. Badger, and D. Leaf, "NIST Cloud Computing Reference Architecture," *National Institute of Standards and Technology*, 2011.
- [21] A. Gajbhiye, and K. Shrivastva, "Cloud Computing: Need, Enabling Technology, Architecture, Advantages and Challenges," *Confluence The Next Generation Information Technology Summit*, 2014.
- [22] H. Berg, "Risk Management: Procedures, Methods and Experiences," *Bundesamt für Strahlenschutz, Salzgitter, Germany*, 2010.
- [23] ISO/Guide 73, "Risk Management-Vocabulary," *International Organization for Standardisation*, 2009.
- [24] G. Dickson, "Principles of Risk Management," *Glasgow Caledonian University*, 1995.
- [25] G. Stoneburner, A. Goguen, and A. Feringa, "NIST SP 800-30 Risk Management Guide for Information Technology Systems," pp. 8-26, *NIST*, 2002.
- [26] "A Risk Management Standard," *The Institute of Risk Management (AIRMIC) and The Public Risk Management Association (Alarm)*, 2002.
- [27] P. Saripalli, and B. Walters, "A Quantitative Impact and Risk Assessment Framework for Cloud Security," *IEEE 3rd International Conference on Cloud Computing*, pp. 280-288, *IEEE*, 2010.
- [28] S. Tanimoto, M. Hiramoto, M. Iwashita, H. Sato, and A. Kanai, "Risk Management on the Security Problem in Cloud Computing," *First ACIS/JNU International Conference on Computers, Networks, Systems, and Industrial Engineering*, pp. 147-152, *IEEE*, 2011.
- [29] J. Fito, M. Macias, and J. Guitart, "Toward Business-driven Risk Management for Cloud Computing," *Network and Service Management (CNSM)*, pp. 238-241, *IEEE*, 2010.

- [30] X. Zhang, N. Wuwong, H. Li, and X. Zhang, "Information Security Risk Management Framework for the Cloud Computing Environments," IEEE International Conference on Computer and Information Technology, pp. 1328-1334, IEEE, 2010.
- [31] M. Almorsy, J. Grundy, and A. Ibrahim, "Collaboration-Based Cloud Computing Security Management Framework," IEEE 4th International Conference on Cloud Computing, pp. 364-371, IEEE, 2011.
- [32] F. Xie, Y. Peng, W. Zhao, D. Chen, X. Wang, and X. Huo, "A Risk Management Framework For Cloud Computing," IEEE 2nd International Conference, pp. 476-480, IEEE, 2012.
- [33] S. Albakri, B. Shanmugam, G. Samy, N. Idris, and A. Ahmed, "Security risk assessment framework for cloud computing environments," Security and Communication Networks, Wiley Online Library, 2014.
- [34] H. Linstone, and M. Turoff, "The Delphi Method: Techniques and Applications," Addison-Wesley, 1975.
- [35] FERMA, "FERMA's Risk Management Standard," 2003, Retrieved from [http://www.ferma.eu/Portals/2/documents/RMS/RMS-UK\(2\).pdf](http://www.ferma.eu/Portals/2/documents/RMS/RMS-UK(2).pdf)
- [36] E. Humphreys, "Implementing the ISO/IEC 27001 Information Security Management System Standard," Artech Print on Demand, 2007.
- [37] J. Miller, L. Candler, and H. Wald, "Information Security Governance Government Considerations for the Cloud Computing Environment," Booz Allen Hamilton, pp. 4-11, 2009.
- [38] NIST, "Standards for Security Categorization of Federal Information and Information Systems," FIPS-199, 2002, Retrieved from <http://csrc.nist.gov/publications/fips/fips199/FIPS-PUB-199-final.pdf>

AUTHOR

Rana Musaad Alosaimi is a Master student of Information Systems at the College of Computer and Information Sciences of King Saud University (KSU), Saudi Arabia. She received her bachelor's degree in Information Technology from the same university. Her fields of research interests include Information Systems, Cloud Computing, and Information Security.

Mohammed Abdullah Alnuem is an Assistant Professor in the field of Computer Science, in the college of Computer and Information Sciences, King Saud University, Saudi Arabia. Further, he is also serving as Vice Dean for Development and Quality in E-Transactions and Communications. He received his PhD in Mobile Computing and Networks from the school of Informatics, University of Bradford, UK and M.S. in Distributed Systems and Networks from the same university. His fields of research interests include Computer Networks, Distributed Systems, Wired and Wireless Networks and Software Engineering.

INTENTIONAL BLANK

AUTOMATED SHORT ANSWER GRADER USING FRIENDSHIP GRAPHS

Soumajit Adhya¹ and S.K. Setua²

¹Department of Management, J.D. Birla Institute, Kolkata, India
smjt.adhya@gmail.com

²Dept. of Computer Science, University of Calcutta, Kolkata, India
sksetua@gmail.com

ABSTRACT

The paper proposes a method to assess short answer written by student using friendship matrix, representation of friendship graph. The Short Answer is a type of answer which is based on facts. These answers are quite different from long answers and Multiple Choice Question (MCQ) type answers. The friendship graph is a graph which is based on friendship condition i.e. the nodes have only one common neighbor. Friendship matrix is the matrix form of the friendship graph. The student answer is stored in a friendship matrix and the teacher answer is stored in another friendship matrix and both the matrices are compared. Based on the number of errors encountered from student answer an error marks is calculated and that number is subtracted from full marks to get student grade.

KEYWORDS

Friendship Graph, Friendship Matrix, Short Answer, Triplets

1. INTRODUCTION

Now-a-days various types of students' answers are evaluated by the examination systems. One is multiple choice questions answering which checks whether the student has marked the correct option or not. Another type is long answers which grades the student based on the content and writing style. The third type is the Short Answers (SA). SA are the type of answers, based on facts and concepts. Computer Assisted Assessment is common term for assessing student grades. This has become quite popular today because of the errors that humans make while evaluating an answer. To use computers for checking Multiple Choice Question (MCQ) type answers is very popular nowadays because of its speed and reliability. The development of Optical Mark Recognition (OMR) sheets have made it possible for examiners/examination system to use MCQ questions for conducting their examination. But MCQ questions do not test the true knowledge of the students. So there is need to check the student answers both short and long answers with the help of computer system. This paper proposes a system which can be used for marking SA by using computer system.

Generally there are 3 categories by which a SA can be marked. First, using statistical technique a SA can be marked. This is done by matching the student answer to the percentage of matching to the model answer by using various statistical techniques. Secondly using information extraction technique where the information is extracted from the text and those concepts are matched with the model answer. Third is the natural language technique which find the semantic meaning of

Natarajan Meghanathan et al. (Eds) : ACITY, VLSI, AIAA, CNDC - 2016
pp. 13–22, 2016. © CS & IT-CSCP 2016

DOI : 10.5121/csit.2016.60902

answer through parsing and finally compare it with instructors answer and assign the final scores[5].

This paper proposes a method by which the SA can be graded. It proposes a system which compare friendship matrices of given answers and model answers and accordingly provides the grade. It also provides a frame work by which the subject, predicate and object of each sentence is extracted teraining the semantic meaning of each sentence. So, this grading procedure takes into account semantic meaning of each and every statement. In this procedure the semantic meaning of the student answer is checked with semantic meaning of the model answer. This algorithm is a variation from C-Rater as it enters concepts from teacher answer one sentence at a time and stores it in a friendship matrix. Similarly the student answer is stored in another friendship matrix and is compared with the teacher answer to check how many concepts exactly matches and then it is graded accordingly.

In this paper Section 2 deals with related work in automated SA grader, Section 3 deals with terminologies associated with this paper, Section 4 deals with the problem definition, Section 5 deals with the proposed method for SA grading and Section 6 is the Conclusion.

2. RELATED WORK

Normally automated SA grading can be graded into 5 eras.

Era 1 is the concept mapping technology where the following systems like ATM, C-Rater, Wang'08. The idea of concept mapping is to check whether a particular concept is present in the student answer or not and then grade accordingly. Concept mapping is based on the sentence level. Another technique is the facet mapping which is derived from concept mapping. Facets are the words and triples and any concept can be broken into facets.

The ATM (Automated Text Marker) breaks the student answer and teacher answer into concepts which are of few words and matches them to find the common concept and then it is assessed accordingly. The C Rater is aimed at to match the sentence level concepts between teacher and student answers. The two answers are compared by representing the texts using variation of syntax, anaphora, morphology, synonyms, and spelling correction. The teacher answers are entered as separate sentence for each every concept. Only one concept is assessed at a time. It gives a higher accuracy than other conceptual raters.

Era II is the Information Extraction technology where the following technologies like AutoMark, e max were developed. Since SA are usually expected to provide specific ideas and hence can be modeled by templates. Information extraction uses pattern matching, and also can extract structured data from unstructured sources and represent structured data as tuples.

Era III is using corpus based technology where the following technologies like Willow, Mohler-09 were developed. They use the statistical properties of large document corpora. Normally these methods are used for large texts but also can be effectively applied for SA also. But the correctness of student answer is limited only to the teacher answer vocabulary and can be increased by including synonyms and bilingual parallel corpora.

Era IV is the use of Machine Intelligence Learning technology where systems like e examiner and CAM was developed. These techniques uses the measurements extracted from natural language processing and combine them to form a single grade using classification or regression model.

Era V is the evolution era. This era is independent of methods. In this case the researchers all over the world compete in various competitions and tournaments [5].

3. TERMINOLOGY

3.1. Friendship Graph and Friendship Matrix

A graph is called a friendship graph if every pair of its nodes has exactly one common neighbor. This condition is called the friendship condition [2]. This graph is used to model the subject-predicate-object structure of a sentence. A friendship matrix is the relational or tabular structure of the friendship graph. In this case the common node (neighbor) is associated with the other nodes i.e. subjects and objects. The relation or table name is the common node and the subjects & objects are the nodes that are associated with it.

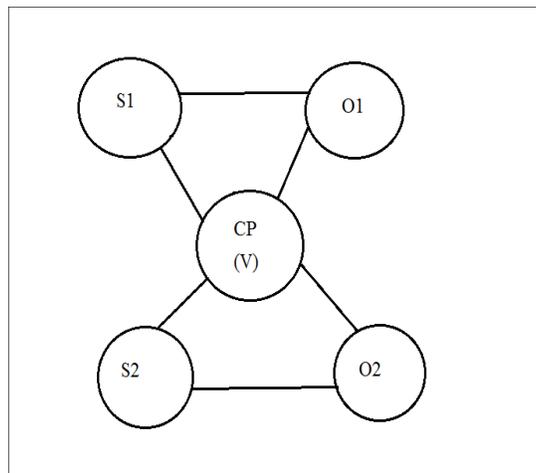


Figure 1: A friendship graph

Table Name: V (the common node)

SUBJECT	PREDICATE
S1	O1
S2	O2
S3	O3
S4	O4

Table 1: The Friendship Matrix which is derived from the Friendship Graph

The friendship graph structure is an ideal structure to store the RDF triplets. The common node can be used as the common MAIN PREDICATE, with the neighboring connected nodes as Subjects and Objects. Now this friendship graph structure can be saved in a matrix form by making the common MAIN PREDICATE as the TABLE NAME with PRE PREDICATE, SUBJECT and OBJECT as the column names.

PRE PREDICATE	SUBJECT		OBJECT		
	PRE SUBJECT	MAIN SUBJECT	PRE OBJECT	MAIN OBJECT	POST OBJECT

Table 2: A friendship matrix to represent the above friendship graph

Since the number of friendship matrix tables is going to be large so there will be a lot of overhead costs. So instead of maintaining individual friendship matrices we can save it in a single table with the following structure:

Common main Predicate	Corresponding Table					
Predicate 1		Subject		Object		
	Pre Predicate	Pre Subject	Main Subject	Pre Object	Main Object	Post Object

Table 3: Final Friendship Matrix to store all the subject-predicate-object triplets derived from teacher student answer

3.2. Sentence Extraction

The sentence extraction algorithm extracts individual sentences from the document. We assume that the individual sentences are separated by full stops. By analyzing the full stops the sentence is extracted and each individual sentence is passed to the Co Reference Resolution pass one after another [4].

3.3. Natural Language Processing Techniques [4]

- Co Reference Resolution (CRR) Pass: In this pass all the entities are recognized which are single words or a block of sequential words. Entities refer to any name, place etc. CRR attempts to find the words in a sentence that refers to an entity and replaces these references with the target entity. Then this modified sentence is passed to tokenization and parts of speech tagger.
- Tokenization and Part of Speech Tagger: Each Sentence is tokenized and part of speech is tagged for each and every word. Then this tokenized sentence is sent to Full parsing phase.
- Full parsing phase: In this case the sentence is written in Penn Tree Bank style which shows the phrasal structure and attachments. The nesting level is denoted by using tabs. Then this parse tree is sent for split coordinating conjunction phase.
- Split Coordinating Conjunction Phase: Complex sentences are broken into simple sentences based on conjunction..
- Extract Dependent Clauses: Sentences with dependent clauses, known as complex sentences in linguistics—as opposed to simple sentences with a single clause—are common in text. A dependent clause is introduced by either a subordinate conjunction

(for adverbial clauses) or a relative pronoun (for relative clauses), so those two cases have to be handled differently. This pass also extracts parenthesized phrases and clauses as they can be handled similarly, although not all are technically dependent clauses. Adverbial clauses are extracted into modifiers, whereas relative or parenthesized clauses are broken off into separate sentences.

- f. Extract Adjective Phrases: Adjective phrases typically appear in sentences between one or two commas, and appear in the parse tree as nested under their subject.
- g. Extract Prepositional Clauses: Prepositional Phrases are the main type of adjunct that is converted into a triple modifier. Because the attachments of modifiers are ignored by this system, attachments don't need to be captured.
- h. Lemmatization: Reducing the verbs to their base form.
- i. Synonym Conversion: All synonyms are checked from synonym table and converted into base word.

4. PROBLEM DEFINITION

SA do not have a huge length. There are no marks given for writing style for SA. This property distinguishes between SA and long answer. Today most of the commercial software only marks MCQ type questions. This will assess student's depth of knowledge only at lower level of Blooms taxonomy of educational objectives. They fail to assess student's performance at higher level of taxonomy of educational objective. So, writing SA for factual and conceptual answers have become a necessity now [5]. This paper proposes a method by which the SA can be graded. It also provides a frame work by which the subject, predicate and object of each sentence can be extracted so that the semantic meaning of each sentence is not lost. So, this grading procedure takes into account semantic meaning of each and every statement. In this procedure the semantic meaning of the student answer is checked with semantic meaning of the teacher answer.

5. PROPOSED ALGORITHM

This paper proposes a grader system for automatically marking SA. The student written factual answer is compared with the teacher answer. Both the answers which are written in paragraph forms are converted into friendship matrix form and then the two matrices are compared. Based on number of matches of tuples the student is given a grade.

Each sentence of teacher answer is extracted which is generally a complex sentence and is sent to NLP Converter. NLP converter converts the complex sentence which is extracted from the teacher answer into simple sentence(s). Then each simple sentence is passed to TTripletExtractor to create the Teacher Friendship Matrix.

Each sentence of student answer is extracted which is generally a complex sentence and is sent to NLP Converter. NLP converter converts the complex sentence, extracted from the student answer into simple sentence(s). Then each simple sentence is passed to STripletExtractor to create the Student Friendship Matrix.

The Grader will be applied to compare and to find the number of matches of tuples between the teacher friendship matrix and student friendship matrix. Based on number of matches the student

answer is graded. Every unmatched tuples or part of tuples of teacher friendship matrix and student friendship matrix is treated as errors. There are 4 types of errors i.e.

Error1: Error due to missing words in pre subject, pre object, and post object for matching subject/object

Error2: Error due to object of teacher answer not found in student answer for a matching main subject.

Error3: Error due to main subject of teacher answer not found in student answer for a matching common predicate.

Error4: Error due to predicate of teacher answer not found in student

To convert a complex sentence to simple sentence the following NLP techniques are used in order:

CRR, Tokenization and Parts of speech tagger, Full parsing, Split Coordinating conjunction, Extract Dependent Clauses, Extract Adjective Clauses, Extract Prepositional Clauses, lemmatization and Synonym Conversion[1][3][4].

The overall method is formalized as below:

Sentence Extraction (*Answer*)

```
{
  Extract the sentences from model answer one by one.
}
```

NLP Converter (*A_Complex_Sentence*)

```
{
  Use the existing NLP techniques to convert the complex sentences to simple sentences for each and every sentence of model answer.
}
```

TTripletExtractor (*A_Simple_Sentence*)

```
{
```

Step 1: Find the deepest verb from the Verb Phrase (VP) sub tree of the parse tree and match it in the predicate field. If the matching predicate is not found then add that predicate to the friendship matrix and go to Step 2. If the matching predicate is found then go to Step 2.

Step 2: While finding the deepest verb all the nodes that are encountered from the parse tree in the VP sub tree of the parse tree are combined to form a string and store it in the pre predicate field with corresponding to that common predicate which was found in Step 1.

Step 3: Find the first noun from the NP sub tree of the parse tree and store it in the main subject field with corresponding to that common predicate which was found in Step 1. While finding the first noun all the nodes that are encountered from the parse tree are combined to form a string and stored in the pre subject column.

Step 4: Find the first adjective, noun or pronoun from the VP sub tree of the parse tree and stored as object with corresponding to that common predicate which was found in Step 1. While finding the first noun/adjective/pronoun all the nodes that are encountered from the parse tree are

combined to form a string and stored in the pre object column and other nodes which followed object are to form a string and stored in the post object column.

}

STripletExtractor (*A_Simple_Sentence*)

{

Step 1: Find the deepest verb from the Verb Phrase (VP) sub tree of the parse tree and match it in the predicate field. If the matching predicate is not found then add that predicate to the friendship matrix and go to Step 2. If the matching predicate is found then go to Step 2.

Step 2: While finding the deepest verb all the nodes that are encountered from the parse tree in the VP sub tree of the parse tree are combined to form a string and store it in the pre predicate field with corresponding to that common predicate which was found in Step 1.

Step 3: Find the first noun from the NP sub tree of the parse tree and store it in the main subject field with corresponding to that common predicate which was found in Step 1. While finding the first noun all the nodes that are encountered from the parse tree are combined to form a string and stored in the pre subject column.

Step 4: Find the first adjective, noun or pronoun from the VP sub tree of the parse tree and stored as object with corresponding to that common predicate which was found in Step 1. While finding the first noun/adjective/pronoun all the nodes that are encountered from the parse tree are combined to form a string and stored in the pre object column and other nodes which followed object are to form a string and stored in the post object column.

}

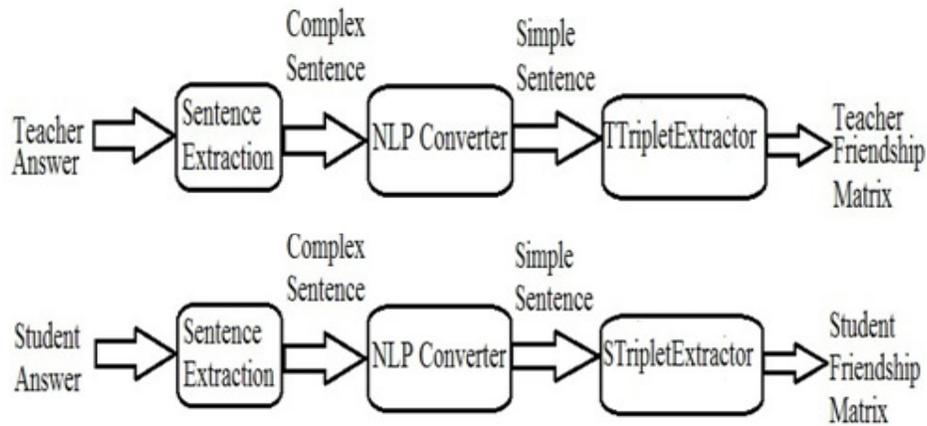


Figure 5: Process of Converting Teacher Answer and Student Answer into respective friendship matrix

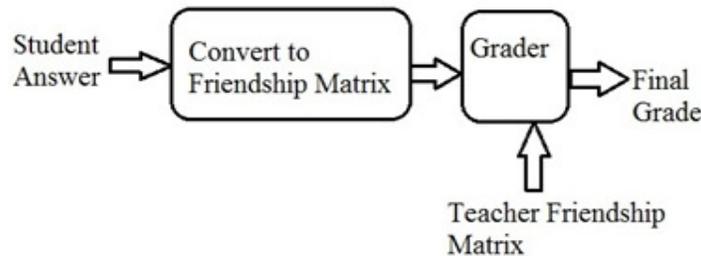


Figure 6: Grading the Student answers with help of teacher answers

Grader (*S_Friendship_Matrix*, *T_Friendship_Matrix*)

{

Step 1: Initialization of variables *error_col_count*, *error_row_count*, *semantic_error* and *predicate_error*.

Step 2: Take a common predicate from teacher friendship matrix and match with the common predicate of the student friendship matrix. If the common predicate is not found then go to Step 4 else go to Step 3.

Step 3: For each main subject for the matching common predicate from teacher friendship matrix repeat from Step 3.1 to Step 3.6

Step 3.1: Take a main subject from the teacher friendship matrix for the common matching predicate.

Step 3.2: Match it with the main subject of the student matrix for the corresponding matching predicate one at a time. If a match is found then go to Step 3.3 else go to Step 3.7

Step 3.3: Match the words present in the pre subject field to that corresponding matching main subject of the teacher friendship matrix with the pre subject field to that corresponding matching main subject of the student friendship matrix. For each unmatched word increase the *error_col_count* by 1 and go to Step 3.4

Step 3.4: Take the main object of the teacher friendship matrix to that corresponding main subject and match it with the main object of the student friendship matrix to that corresponding main subject. If match is found then go to Step 3.5 else increase the *semantic_error* by 1 and go to Step 3.7

Step 3.5: Match the words present in the pre object field to that corresponding matching main object of the teacher friendship matrix with the pre object field to that corresponding matching main object of the student friendship matrix. For each unmatched word increase the *error_col_count* by 1 and go to Step 3.6

Step 3.6: Match the words present in the post object field to that corresponding matching main object of the teacher friendship matrix with the post object field to that corresponding matching main object of the student friendship matrix. For each unmatched word increase the *error_col_count* by 1 and go to Step 3.1

Step 3.7: If no match is found increase the *error_row_count* by 1 and go to Step 2

Step 4: Increase the *predicate_error* count by 1. If all the common predicates are exhausted then Go to Step 5 else go to Step 2

Step 5: Calculation based on errors which are subtracted from the full marks i.e.

$$M = FM - EM$$

where,

$$MC = FM / N$$

$$ER1 = [MC/W]*ECC$$

$$ER2 = [MC/2]*SE$$

$$ER3 = MC*ERC$$

$$ER4 = MC*PE$$

$$EM = ER1 + ER2 + ER3 + ER4$$

Abbreviations:

M = Final marks of answer

FM = Full marks of an answer

EM = Total Error marks

MC = Marks of each concept
 W = Average number of words present in pre subject, pre object, post object
 SE = Semantic Error
 ECC = Error Column Count
 ERC = Error Row Count
 ER1 = Total Error Marks occurred due to missing words in pre subject, pre object,
 post object for matching subject/object
 ER2 = Total Error Marks occurred due to object of teacher answer not found in student
 answer for a matching main subject.
 ER3 = Total Error Marks occurred due to main subject of teacher answer not found in
 student answer for a matching common predicate.
 ER4 = Total Error Marks occurred due to predicate of teacher answer not found in
 student
 }

6. CONCLUSIONS

This paper proposes a technique which uses facet mapping era. The NLP techniques is used to convert the text into base text from where the concept mapping technique is used to store the triplets in a friendship matrix, and to grade the answers. In SA concepts are checked rather than the writing style. In this case the concepts are compared between teacher answer and student answer and the number of errors is computed. These errors are then converted to a number which is subtracted from the full marks of the student answer. This technique can be expanded in future to mark long essays which depend on writing style. Also the future work relating this paper is to devise a faster information matching algorithm which can match student answer and teacher answer to grade them.

REFERENCES

- [1] Delia Rusu, Lorand Dali, Blaž Fortuna, Marko Grobelnik & Dunja Mladenić, (2007) "Triplet Extraction from Sentences," Proceedings of the Conference on Data Mining and Data Warehouse (SiKDD 2007) held at 10th International Multi conference on Information Society
- [2] Endre Boros, Vladimir A. Gurvich & Igor E. Zverovich, (2008) DIMACS Technical Report, 1 RUTCOR, Rutgers Center for Operations Research Rutgers, The State University of New Jersey
- [3] Jonathan Hayes and Claudio Gutierrez, (2004) "Bipartite Graphs as Intermediate Model for RDF", The Semantic Web – ISWC 2004, Springer Berlin Heidelberg, Vol. 3298, pp 47 – 61.
- [4] Aaron De Fazio, (2009) "Natural Question Answering over Triple Knowledge Bases", Australian National University
- [5] Steven Burrows, Iryna Gurevych & Benno Stein, (2015) "The Eras and Trends of Automatic Short Answer Grading," International Journal of Artificial Intelligence in Education25, IOS Press, p. 60-117

AUTHORS

Soumajit Adhya holds a M.Sc degree in Computer and Information Science from University of Calcutta and currently employed as a IT faculty in JDBI, Department of Management.



S. K. Setua is an associate professor in the Department of Computer Science & Engineering at University of Calcutta. His research interest includes distributed computing, information & network security, big data analytics, SDN. He has more than 50 research publications in international journals and conferences.



AUTOMATIC GENERATION AND OPTIMIZATION OF TEST DATA USING HARMONY SEARCH ALGORITHM

Rajesh Kumar Sahoo¹, Deeptimanta Ojha¹, Durga Prasad Mohapatra² and Manas Ranjan Patra³

¹Department of Computer Engineering, A.B.I.T, Cuttack
rajesh_sahoo@rediffmail.com
deeptimantaojha@gmail.com

²Department of Computer Engineering, NIT, Rourkela
durga@nitrkl.ac.in

³Department of Computer Engineering, Berhampur University, Berhampur
mrpatra12@gmail.com

ABSTRACT

Software testing is the primary phase, which is performed during software development and it is carried by a sequence of instructions of test inputs followed by expected output. The Harmony Search (HS) algorithm is based on the improvisation process of music. In comparison to other algorithms, the HSA has gain popularity and superiority in the field of evolutionary computation. When musicians compose the harmony through different possible combinations of the music, at that time the pitches are stored in the harmony memory and the optimization can be done by adjusting the input pitches and generate the perfect harmony. The test case generation process is used to identify test cases with resources and also identifies critical domain requirements. In this paper, the role of Harmony search meta-heuristic search technique is analyzed in generating random test data and optimized those test data. Test data are generated and optimized by applying in a case study i.e. a withdrawal task in Bank ATM through Harmony search. It is observed that this algorithm generates suitable test cases as well as test data and gives brief details about the Harmony search method. It is used for test data generation and optimization.

KEYWORDS

Harmony search algorithm, meta-heuristics, test case generation, test case optimization, test data.

1. INTRODUCTION

The test case generation is based on the requirements. It completely ignores the aspect of system execution. Apart from this, the test case design from program code may cause difficult to imbrute. Test cases may not expose the missing functionalities. The proposed approach focuses the redundancy, test cases, and test case optimization challenges. It uses HS optimization

algorithm to optimize the random test cases. Moreover, this proposed method inspires the developers to generate random test cases to improve the design quality. This paper is intended to present the result of the outcome of HS to find the optimum solution in the software construct. Optimization can be represented through the process of finding the best result under the given circumstances which might be used for maximizing or minimizing the local or overall optimum value of a function. The evolutionary algorithms like Harmony search is discussed in this paper. Z. W. Geem [6] introduced Harmony search method in 2009. By this technique, a flawless state of harmony is observed through the musical process. It is correspondent to generate the optimum value through the musical optimization process. The improvisation of music is a technique where the musician plays various musical notes with divergent types of musical instruments and generates the best amalgamation of tunes with frequency.

The rest of the paper is organized as follows. Section 2 discusses basics of test data automation, Section 3 is for literature survey, Section 4 represents the fundamentals of Harmony Search (HS) algorithm, proposed systems, and methodology, Section 5 focuses the simulation results, Section 6 represents discussion and future scope and Section 7 concludes the paper.

2. BASIC CONCEPTS

2.1 Automated Test data generation

Testing is the phenomenon of finding errors after executing the programs. Software testing can be defined by many processes designed sequentially and does not do anything unintended [19]. The objective of software testing is to finalize the application software against the user requirements. It must have good test coverage to test the application software and perform as per the specifications. For generating list of coverage's, the test cases should be designed with maximum possibilities of finding various errors or bugs [20]. The test cases should be very effective and is measured through the number of defects or errors reported. Generation of test data is a method for identifying the data set which satisfies the criteria. Automated generation of test data plays a key factor in software testing. Most of the researchers are used the heuristics approaches for automated generation of test data. Automatic test data generation data helps to minimizing the time and cost in developing test cases.

2.2 Overview of Harmony Search

The harmony search metaheuristic optimization algorithm is based on the music. The harmony search technique is inspired by the musician when he composes the music. Usually harmony consists of different possible amalgamations of music pitches saved in the memory. In this technique first, random solutions are directly stored in the harmony memory based on memory considering rate and pitch adjustment rate, and then pitch adjustment distance will be calculated between different selected random solutions. The best solution is stored in the memory of harmony by discarding the worst solution.

3. LITERATURE SURVEY

Biswal et al.[1] described how the activity diagram is derived from the test scenarios by converting into a control flow graph (CFG) where every node explains an activity and the edges of the node with the activities of control flow. According to Geem et al. [6, 7] harmony search a

metaheuristic population-based algorithm where multiple harmonies are used in parallel which gives better performance with high efficiency. Geem [7] focused the Pitch Adjustment Rate (PAR) function which is used in simulated annealing and it also increases the robustness of the algorithm. It is highly reliable. Geem et al. [8] described how the parameter like Pitch adjustment rate (PAR) increases in a linear way with the number of generations while the bandwidth (BW) decreases exponentially in Harmony Search algorithm. Manjarres et al. [5] focused described various applications of HS like industry, power systems, construction design, and information technology. Das et al. [3] described the background of the power of Harmony Search (HS) which gives the better solutions. The major drawback is user must specify the minimum and maximum bandwidth values. It is difficult to deduce the program which is dependent. Chen et al. [2] proposed a concept which is having the relation between nodes of control flow graph (CFG). It also defined the fitness function for evaluating the generation of test data by using Genetic Algorithm (GA) approach which performed the random technique. Ghiduk et al. [10] described how the test data are generated by using a search based technique but in case of random search techniques the optimal test cases are not generated effectively and efficiently. Kirkpatrick et al. [11] focused on a key factor like the Pitch adjustment which is diversified randomly in Harmony Search method. Mahdavi et al. [12] explained how HS algorithm is used in clustering problems of web pages through continuous and discrete representation of data. Mira et al. [13] focused on the parameters which required for improving the efficiency of an HS algorithm in a particular problem where diversification and intensification are available. Omran et al. [14] focused on the how the Global Harmony Search (GHS) is simulated through Particle Swarm Optimization (PSO). This paper also explained how the pitch adjustment of Harmony Search technique can imitate by harmony memory through the harmony vector. It also replaces bandwidth (BW) and adds to the dimension of HS. Peng et al. [15] proposed a novel algorithm which upgrades the search based test case generation by using an adaptive genetic algorithm (AGA) and it gives better results in the software testing tool. Suresh et al. [16] described a genetic algorithm used for test data generation and it also generates the basic path. In this case the genetic algorithm combines the features of local and global test data optimization. According to this paper sequence diagram is converted into control flow graph and generates the optimal test cases using genetic algorithm. Yang et al. [17] focused on HS algorithm which is diversified with the randomization operation of pitch adjustment rate through the probability. Existing pitch (or solution) is saved in the Harmony Memory. It also gives the global search solutions. Yang et al. [18] proposed an approach to generate the test data in a single specific path based genetic algorithm (GA). In this case the effectiveness of the fitness function and performance are evaluated.

4. PROPOSED SYSTEM

We have proposed a methodology for generating test cases for withdrawal system of an ATM machine and test cases are optimized by Harmony search algorithm (HSA). This method is used for evaluating its efficiency and effectiveness for generating the test cases and to maximize to achieve the goal.

4.1. Necessity of Proposed System:

The proposed system is intended to generate an automatic and optimized test case with existing approaches of Harmony Search. Optimized test cases may not be helpful in the testing process. It may be required to differentiate between the various test cases. First of all the system may be initialized with harmony memories. Each harmony memory maintains its own prevailing location

on the basis of which the test cases may be generated. Harmony Search has a memory that helps to maintain the solutions by harmonies. This paper also finds out the effectiveness of the proposed approach through the number of test cases or test data.

4.1.1. Harmony Search Algorithm (HSA)

The main principle and working of proposed approach:

Harmony search is a meta heuristic population-based optimization technique. Through this technique, the problem is represented through different test cases. The quality of each test case is calculated through the fitness value of the problem. The working principle of harmony search technique is inspired by the musician when he composes the music; a musician usually tries different combinations of music pitches which are stored in the harmony memory. Perfect harmony search needs a correspondent for generating the process to find the solution which is optimal. The main steps of a harmony search method are:

- Generate random solutions which will be stored in the Harmony Memory (HM).
- Select a random solution from Harmony Memory based factors that are Harmony Memory Considering Rate (HMCR) and Pitching Adjustment Rate (PAR).
- Apply adjustments pitch distance to the selected random solution.
- Compare the fitness function values of the mutated solution or newly formed solution with the worst solution.
- If better solution is found then the worst solution is substituted with the mutated solution available in the harmony memory otherwise the solution is discarded.
- Remember the solution which is best so far.

In this case, mainly two parameters are used. They are HMCR and PAR. Harmony Memory Considering Rate (HMCR) is described as the probability of selecting a particular component or solution available in HM. Pitching Adjust Rate (PAR) describes the probability distribution of the candidate solution for mutation purpose through HM.

The value of Pitching Adjust Rate (PAR) can be calculated as follows.

$$PAR = (PAR_{max} - PAR_{min}) / (it_{max}) * it + PAR_{min} \quad (1)$$

Where PAR_{max} → maximum pitch adjusting rate

PAR_{min} → minimum pitch adjusting rate

it_{max} → given maximum iterations

it → current iteration number

The bandwidth factor 'bw' can be calculated as.

$$bw = bw_{max} * \exp(\text{coef} * it) \quad (2)$$

Here the value of 'coef' is evaluated as follows:

$$\text{coef} = \log (b w_{\min} / b w_{\max} / i t_{\max}) \quad (3)$$

Where $b w$ = bandwidth
 $b w_{\min}$ = minimum bandwidth
 $b w_{\max}$ = maximum bandwidth

The bandwidth factor $b w$ is used to alter the value of existing solution available in Harmony Memory. This can be done as follows:

$$x_{\text{new}} = x_{\text{old}} + \text{rand} (0,1) * b w \quad (4)$$

Where x_{new} = new solution
 x_{old} = old existing solution
 $\text{rand} (0,1)$ = a random value ranging in between 0 and 1.

Harmony Search is a meta heuristic search method which is generally used to optimize problems whose solution may be analyzed in an n-dimensional space. According to Harmony Search Algorithm, each musician plays a musical note to generate the best possible harmony altogether. A new solution is generated for a musical note with the help of harmony factors like HMCR (Harmony Memory Considering Rate), PAR (Pitch Adjustment Rate) and $b w$ (Bandwidth). Each musical note usually possesses its current position. After iteration, the worst harmony is substituted by a new better solution. The flow chart of Harmony Search is depicted in Figure-1.

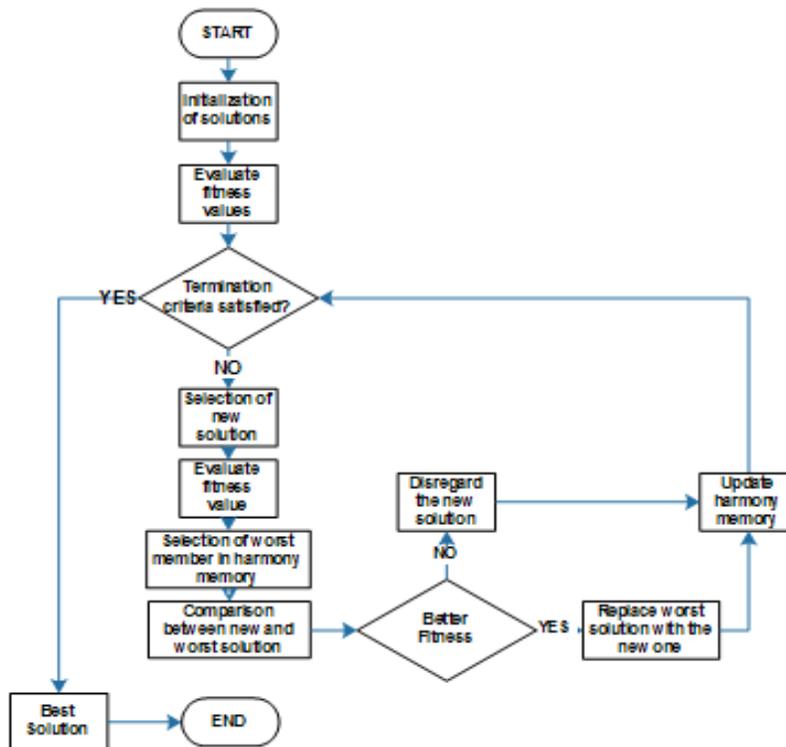


Figure 1: Flow chart of Harmony Search

4.2. Pseudo code for Harmony Searches for generating test data

The number of generation is initialized.

The population size is initialized

Generate Initial Population and store it in the Harmony Memory

Evaluate the initial fitness function value 'fx' of all candidate solutions

$$f_x = 1 / (\text{abs}(\text{net_bal} - \text{wd_amt}) + \epsilon)^2$$

where ϵ varies from 0.1 to 0.9

Find the initial best and worst solution

Initialize different HARMONY parameters

i.e., PAR_{\min} , PAR_{\max} , bw_{\min} , bw_{\max} , HMCR

While Generation < max_iteration do

Evaluate the values of PAR, bw

If $\text{rand}(0, 1) \leq \text{HMCR}$ then

Choose a random value from Harmony Memory i.e., x

If $\text{rand}(0, 1) \leq \text{PAR}$ then

Adjusting the value of x by using the following equation:

$$x_{\text{new}} = x + \text{rand}(0, 1) * \text{bw}$$

End if

else

Choose a random value with the help of following equation:

$$x_{\text{new}} = \text{min_value} + \text{rand}(0, 1) * (\text{max_value} - \text{min_value})$$

End if

Checking the boundary condition of the new solution

Evaluate the fitness function value of the new solution

If $\text{fitness}(\text{new solution}) \geq \text{fitness}(\text{worst solution})$

Accept the new solution

Replace the worst solution with the new solution available in the Harmony Memory.

End if

Update current best solution and worst solution

Generation (t) = Generation (t) + 1

End While

Select the best fitness with function value

4.3. Methodology:

For Mathematical function

$$f(x) = 1 / (\text{abs}(\text{suc_bal}) + \epsilon)^2 \quad (5)$$

Where $0.1 \leq \epsilon < 0.9$ (taking ϵ -value because overflow condition due to infinity).

Here Successive Amount (suc_amt) is defined as :

$$\text{suc_bal} = \text{net_bal} - (\text{wtd_amt} - \text{min_bal}) \quad (6)$$

Where net_bal = current account balance
min_bal= minimum bank balance limit

Initially, each solution is initialized with a musical note. The musician will search for optimal solutions by changing the pitch and bandwidth of their musical note. It will keep track of best note in the population and upgrade its solution. The optimal solution is used to maximize the mathematical function $f(x)$ which may be implemented in Harmony Search using MATLAB-7.0.as shown in Table-1.This table primarily focuses on the musician's attempt to generate the best note in the search space.

Table 1: Fitness Function Value for each sample space or test data

Iteration Number	Test Cases/Test Data	Fitness Function Value
0	4000	5.9488e-010
10	5700	6.4746e-010
20	7300	7.0357e-010
40	10800	8.5496e-010
60	10800	8.5496e-010
80	14300	1.061e-009
100	16400	1.2225e-009
120	18700	1.4457e-009
140	20800	1.7075e-009
160	24300	2.3338e-009
180	26600	2.9537e-009
200	29800	4.3282e-009
220	31000	5.1018e-009
240	33700	7.8313e-009
260	35400	1.085e-008
280	39900	3.8445e-008
300	43500	4.4438e-007
320	43800	6.943e-007
340	44000	9.9976e-007
350	44000	9.9975e-007

In this case, 20 numbers of sample test cases are considered. The function value depends on upon the parametric values of the input variables. It was found that the solution reaches its optimum value after 325 iterations.

5. SIMULATION RESULTS

The proposed approach generates the automated test cases through test data for Bank ATM using Harmony Search Algorithm. The figure-2 shows the relation between two variable quantities which are fitness function value range and test data measured along one of a pair of axis represented in table-1.

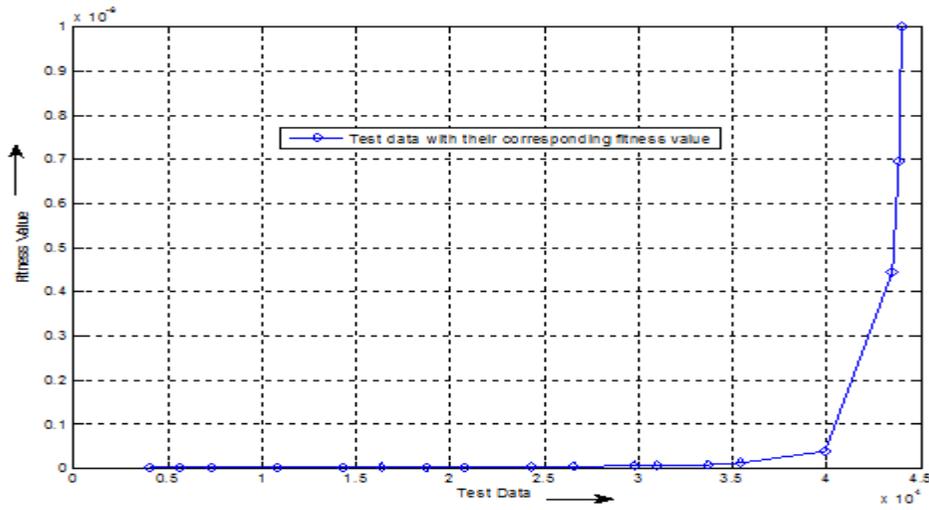


Figure 2: Graphical representation of test data and fitness function value for table-1

The proposed approach generates the test data for Bank ATM’s withdrawal operation using Harmony Search Algorithm. Table-2 represents the range of fitness value with different test data and also it gives the individual candidate solution according to the range of fitness value in terms of percentage.

Table 2: %of test data in terms of maximum fitness value

Fitness Value Range	% of Test data
$0 \leq f(x) < 0.3$	25
$0.3 \leq f(x) < 0.7$	55
$0.7 \leq f(x) < 1.0$	20

The above table shows that around 10% test cases or test data are having the higher fitness function $f(x)$ value and lies in between 0.7 and 1.0. Figure-2 shows a pictorial representation of the relation of two variable quantities like percentage of test data and fitness value range.

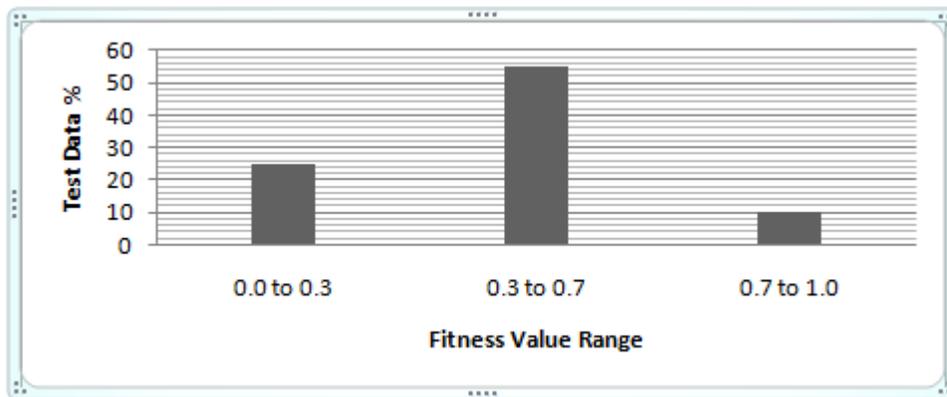


Figure 2: Graphical representation of % test data and fitness value range for table-2

6. DISCUSSION AND FUTURE SCOPE

While considering the mathematical function $f_x=1/(\text{abs}(\text{net_bal-wd_amt})+\varepsilon)^2$, where ε varies from 0.1 to 0.9, along with each member is initialized with a harmony. It has been found that optimality of solution keeps track of best and worst member in the harmony memory and updates its solution accordingly. By considering some sample test cases it has been observed that the function value depends upon the parametric values of the input variables like Harmony Memory Considering Rate, Pitching Adjusting Rate and the Bandwidth. The proposed approach generates the test data for small applications. The future approach to this work could enhance the test data generation for large programs automatically. The different parameters could be added to this approach which gives more optimized test cases and also increases the efficiency of Harmony Search (HS) technique. Another perspective area could be the randomly generated test data by using various paths according to the control flow graph (CFG). Test Cases can be generated by using various kinds of meta heuristic algorithms like GA, FA, PSO, BCO etc. The test data generated by using HS algorithm is compared with test data generated by PSO and it was found that HS produces optimal result in very less time and with more accuracy.

7. CONCLUSION

Harmony search algorithm (HSA) is a very important tool for optimization of test cases or test data. It has been diversified the problems in a very effective manner for generating the test data automatically. In this paper, HS algorithm has been discussed to generate the test cases which are optimized by taking an example of withdrawal operation of an ATM machine. This paper also describes the fundamental notions of HSA, how the random test cases are generated and finding the optimal solution to maximize the problem. This paper will inspire researchers to work on HSA by applying in computer science engineering area to generate the effective automated test cases.

REFERENCES

- [1] Biswal.B.N., Nanda.P., Mohapatra.D.P.,(2008),"A Novel Approach for Scenario-Based Test Case Generation", International Conference on Information Technology,pp.244-247.
- [2] Chen Yong, Zhong Yong, Tingting Shi, Liu Jingyong,(2009), "Comparison of Two Fitness Functions for GA-based Path-Oriented Test Data Generation", Fifth International Conference on Natural Computation, Vol.4,pp.177-18.
- [3] Swagatam Das, Arpan Mukhopadhyay, Anwit Roy, Ajith Abraham, and Bijaya K. Panigrahi,(2011),"Exploratory Power of the Harmony Search Algorithm: Analysis and Improvements for Global Numerical Optimization", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics. Vol.41, No.1, pp. 89-106.
- [4] Desikan.S. and Ramesh.G,Software testing principles & practices, Pearson Education,2007.
- [5] D. Manjarres, I. Landa-Torres, S. Gil-Lopez et al., (2013)," A survey on applications of the harmony search algorithm", Engineering Application in Artificial Intelligence. 26(8), pp.1818–1831.
- [6] Geem, Z.W.,(2009),"Harmony search algorithms for structural design optimization, in Studies in computational intelligence", Springer Berlin Heidelberg: Berlin, Heidelberg. 23(9), pp. 228-242.

- [7] Geem, Z.W.,(2007),“Harmony search algorithm for solving Sudoku”, in Proceedings of the 11th international conference, KES 2007 and XVII Italian workshop on neural networks conference on Knowledge-based intelligent information and engineering systems: Part I.,Springer-Verilog: Vietri Sul Mare, Italy.
- [8] Geem, Z,(2010),“State-of-the-Art in the Structure of Harmony Search Algorithm”, in Recent Advances In Harmony Search Algorithm, Springer Berlin / Heidelberg, pp. 1-10.
- [9] Geem, Z.W., and K.-B. Sim,(2010),” Parameter-setting-free harmony search algorithm”, Applied Mathematics and Computation, Vol.217 (8):, pp. 3881-3889.
- [10] Ghiduk, Ahmed S, and Girgis, Moheb R.,(2010), “Using Genetic Algorithms and Dominance Concepts for Generating Reduced Test Data”, Informatica (Slovenia), Volume 34, pp.377-385.
- [11] Kirkpatrick, S., “Optimization by simulated annealing: Quantitative studies”, Journal of Statistical Physics, 34(5-6):1984, pp. 12.
- [12] Mahdavi, M. and H. Abolhassani,(2009), “Harmony K-means algorithm for document clustering”, Data Mining and Knowledge Discovery, 18(3): pp. 370-39.
- [13] Mira, J., J. Álvarez, and X.-S. Yang,(2005),” Engineering Optimizations via Nature-Inspired Virtual Bee Algorithms, in Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach”, SpringerBerlin / Heidelberg, pp. 317-323.
- [14] Omran, M.G.H. and M. Mahdavi,(2008),” Global-best harmony search”,Applied Mathematics and Computation, 198(2): pp. 643-656.
- [15] Peng Lin, Xiaolu Bao, Zhiyong Shu, Xiaojuan Wang, Jingmin Liu,(2012),” Test Case Generation Based on Adaptive Genetic Algorithm”,IEEE,978-1-4673-0788-8/12/\$31.00.
- [16] Y.Suresh and S.Rath,(2013), “A genetic Algorithm based Approach for Test Data Generation in Basis Path Testing”, The International Journal of Soft computing and software Engineering, vol.3, issue.3.
- [17] Yang, X.-S.,(2008),” Nature-Inspired Metaheuristic Algorithms”,Springer-Verilog, pp. 73-80.
- [18] Yang Cao, Chunhua Hu and Luming Li,(2009),” An Approach to Generate Software Test Data for a Specific Path automatically with Genetic Algorithm”, International Conference on Reliability, Maintainability and Safety,pp.888-892.
- [19] Glenford J.Myers,(2004), The art of software testing, 2nd ed.: Wiley.
- [20] S.Kuppuraj and S.Priya,(2012),”Search-Based Optimization for Test Data Generation Using Genetic Algorithms,” in Proc of the 2nd International Conference on Computer Applications, pp.201-205.

GLITCH ANALYSIS AND REDUCTION IN COMBINATIONAL CIRCUITS

Ronak Shah

B.Tech Student, Electronics and Communication,
Faculty of Technology, Dharmsinh Desai University, Nadiad, Gujarat, India
ronakrameshbhaishah@gmail.com

ABSTRACT

Hazard in digital circuits is unnecessary transitions due to gate propagation delay in that circuit. Hazards occur due to uneven delay offered in the path of the various ongoing signals. One of the important reasons for power dissipation in CMOS circuits is the switching activity. This include activities such as spurious pulses, called glitches. Power optimization techniques that concentrate on the reduction of switching power dissipation of a given circuit are called glitch reduction techniques. In this paper, we analyse various Glitch reduction techniques such as Hazard filtering Technique, Balanced Path Technique, Multiple Threshold Technique and Gate Freezing Technique. Using simulation, we also measure the parameters such as noise and delay of the circuits on application of various techniques to check the reliability of different circuits in various situations.

KEYWORDS

Glitch, Power dissipation, Gate freezing, balanced path delay, multiple threshold transistor, Hazard filtering, Noise, Delay and switching activity.

1. INTRODUCTION

Power dissipation is an increasingly critical issue in modern VLSI design and testing .Hence, Low Power Circuit Design has become very crucial in today's era of modern portable consumer gadgets. For CMOS combinational circuits, the reduction of dynamic power dissipation is very important. A signal transition can be of two types: a functional transition and glitch. Before reaching the steady state, a signal might go through several state changes which are called glitches. As they dissipate 20-70% of total power dissipation, glitch is needed to be eliminated for low power design.

$$P_{\text{Total}}=P_{\text{Static}} +P_{\text{dynamic}} \quad (1)$$

$$P_{\text{Total}}=P_{\text{Switching}}+P_{\text{Short-Circuit}}+ P_{\text{leakage}} \quad (2)$$

Total Power dissipation consists of mainly dynamic power dissipation and static power dissipation, further these are divided into switching power dissipation, leakage power dissipation, short circuit power dissipation. Dynamic power dissipation is a major source of leakage power,

which is directly proportional to the number of signal transitions(1-0 and 0-1) in a digital circuit. Switching power dissipation ($P_{\text{switching}}$) is directly proportional to switching activity(a), load capacitance(C_{load}), Voltage supply (V_{dd}) and clock frequency(f_{clk}) as shown in equation(3).

$$P_{\text{switching}} = a \cdot C_{\text{load}} \cdot V_{\text{dd}}^2 \cdot f_{\text{clk}} \quad (3)$$

In this paper we have chosen the best available techniques to reduce the glitch power. We have selected highly glitchfull circuits in our analysis and tried to reduce glitch power, delay and noise using these techniques. Using Tanner tool simulation, we have also put up the statistics to make the analysis simpler to understand.

2. TECHNIQUES FOR GLITCH REDUCTION

2.1 Gate Freezing

This method is useful for minimization of glitches. In this method, glitchfull and high power dissipating circuits are selected and replaced by a modified library cell called 'F-gate' with a control signal(CS) as shown in Fig.1 where Vdd is supply voltage ,I is input ,O is output CS is control signal to n-type library cell and Gnd is ground. This gate is controlled in order to freeze the cell's output for reducing the amount of glitch from the circuit. Basic CMOS gate and Gate frozen CMOS layout is shown in Fig.1.The control signal(CS) drives the gate input of this n-type cell.This method transforms some of the gates that are more glitchfull into modified devices that are able to filter out unnecessary output transitions when a control signal(CS) is activated.

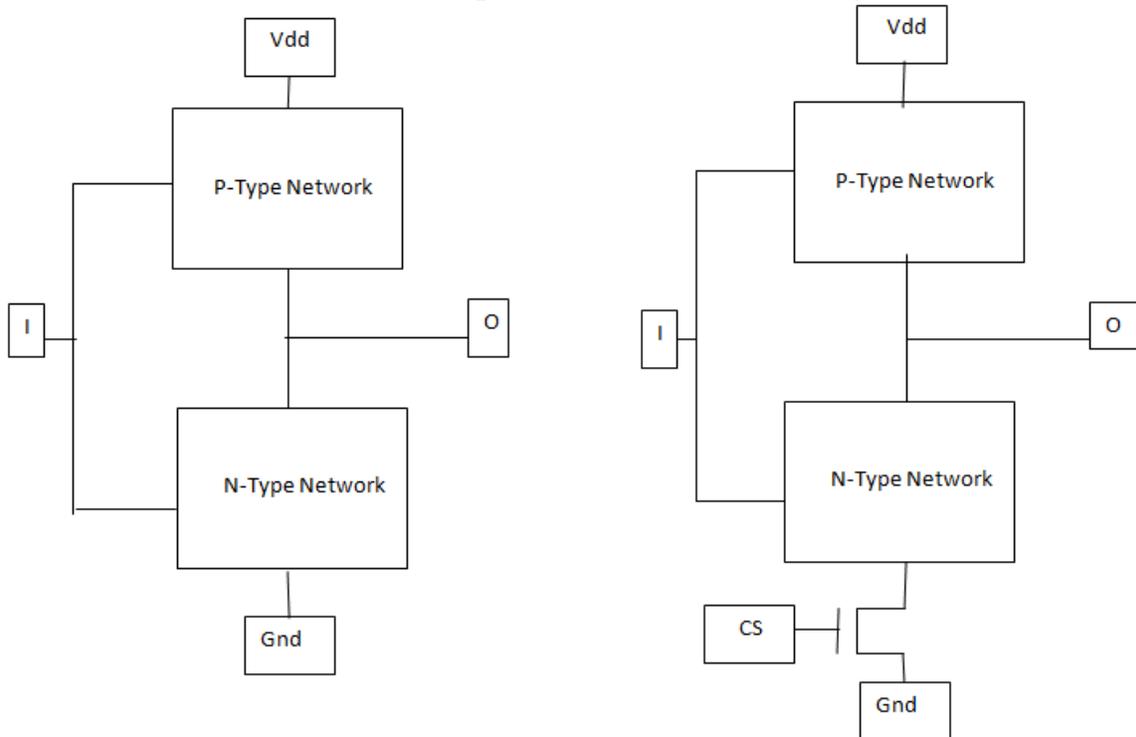


Figure 1 CMOS logic and CMOS logic with library cell

2.2 Balanced Path Technique

Balanced path delay technique is used for resolving differing path delays. To make path delays equal, buffer insertion is done on the faster paths. Balanced path delay will avoid glitches in the output. This technique is not considered efficient in terms of power consumption due to addition of buffers. Hence the more innovative method is hazard filtering discussed next.

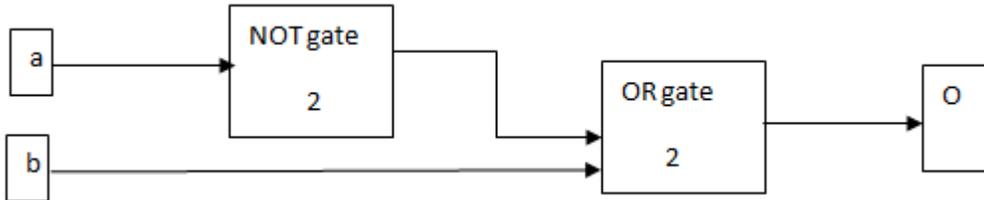


Figure 2 Original Circuit with glitch output

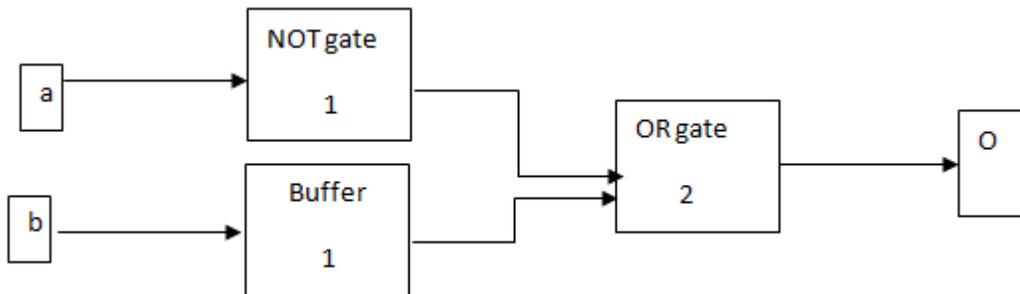


Figure 3 After applying balanced path delay technique

2.3 Hazard Filtering Technique

Hazard in digital circuits is unnecessary transitions as in case of glitch due to gate propagation delay in that circuit.

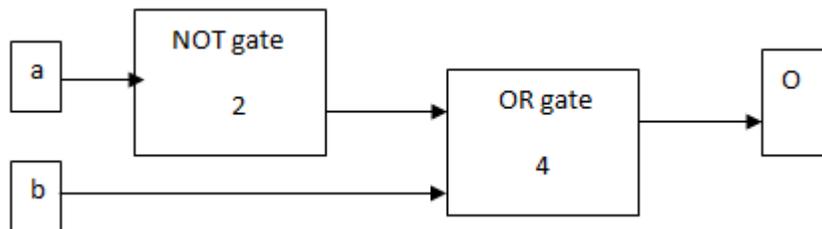


Figure 4 After applying hazard filtering technique

Hazards occur due to uneven delay offered in the path of the various ongoing signals. So apart from balanced path delay technique and using buffers to balance the delays in path, we use the hazard filtering technique in which we increase the delay of receiving hardware to such an extent so that spurious transitions are eliminated and hence the glitch is eliminated. This is shown in the figure and also verified using simulation.

2.4 Multiple Threshold Technique

This is a technique to reduce power dissipation and reducing glitch in digital circuits. As delay of each gate is a function of threshold voltage (V_{th}), gates that are in non critical paths were selected and their threshold voltages were rised manually, then the propagation delays along different paths can be balanced so that unnecessary transition will be minimized. Therefore, it is a new efficient technique for minimizing glitch in digital circuits that lead to low power dissipation.

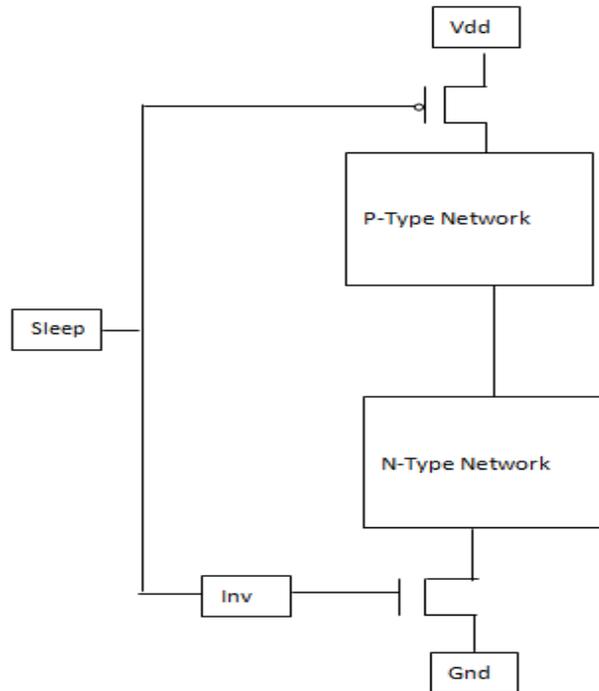


Figure 5 Multiple Threshold Implementation

3. SIMULATION AND RESULTS

We present the analysis of various parameters after applying various techniques on original glitchfull circuits.

Analysis of Circuit1(a'+b)

Name	Avg. Power V_1(watts)	Max.Power V_1(watts)	Delay(s)	Noise (Sq V/Hz)	Avg. Power V_1(watts)
Original Circuit	3.048279e-013	2.642508e-002	2.9497e-007	13.02006a	3.048279e-013
Hazard Filtering Technique	1.384576e-013	1.968104e-002	2.9513e-007	8.69371a	1.384576e-013
Balanced Path Delay Technique	4.538634e-013	6.502845e-002	2.9515e-007	4.36361a	4.538634e-013
Multiple Threshold	2.208378e-013	1.084178e-002	9.5208e-008	13.20418a	2.208378e-013
Gate Freezing	5.211884e-013	5.674504e-002	2.0005e-007	8.69361a	5.211884e-013

Analysis of Circuit 2 (AB'+BC)

Name	Avg. Power V_1(watts)	Max.Power V_1(watts)	Delay(s)	Noise (Sq V/Hz)
Original Circuit	5.786543e-013	5.488896e-002	2.9443e-007	4.41438a
Hazard Filtering Technique	8.379241e-013	1.790674e-001	2.9508e-007	4.41438a
Balanced Path Delay Technique	8.639895e-013	8.481847e-002	2.9511e-007	4.41438a
Multiple Threshold	3.182036e-013	5.595556e-002	2.9462e-007	4.41438a
Gate Freezing	5.191436e-013	5.683199e-002	2.9508e-007	8.86631a

Analysis of Circuit3 ((a'b)'c)'

Name	Avg. Power V_1(watts)	Max.Power V_1(watts)	Delay(s)	Noise (Sq V/Hz)
Original Circuit	2.425840e-013	3.211329e-002	1.9991e-007	4.36533a
Hazard Filtering Technique	1.192698e-013	1.766607e-002	1.9986e-007	17.28354a
Balanced Path Delay Technique	1.008878e-012	9.210829e-002	2.9510e-007	4.36537a
Multiple Threshold	3.604858e-014	8.604082e-003	1.9956e-007	30.15881a
Gate Freezing	1.024181e-013	1.775080e-002	1.9987e-007	17.28324a

3.1 Average Power Analysis:

Results show that Hazard Filtering Technique and Multiple Threshold Technique are better than other techniques where it is reduced upto 54.58% and 85.13% respectively. Also in multiple threshold technique the output voltage is reduced as compared to maximum voltage. So in this situation, we may prefer Hazard Filtering Technique to fulfill this criterion. The power consumption is highest for balanced path technique as expected.

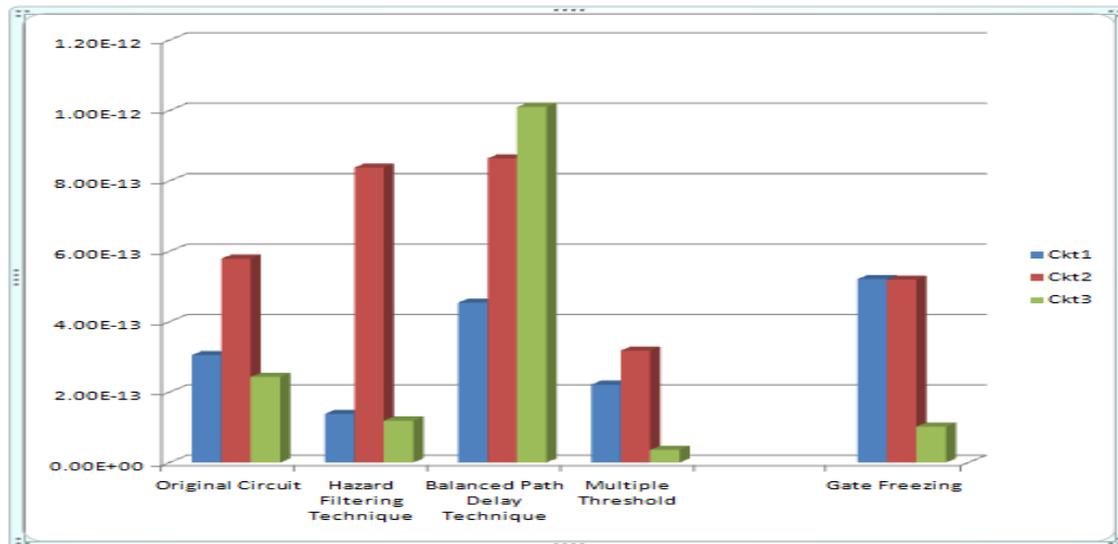


Figure 6 Circuit wise Average power consumption in watts

3.2 Max. Power Analysis

Similar to Average Power Analysis, we see that Maximum Power dissipation follows the same trend except for ckt2 Hazard filtering technique. Otherwise, Multiple threshold Technique and Hazard Filtering Techniques are advisable. The power consumption in balanced path technique as expected is higher.

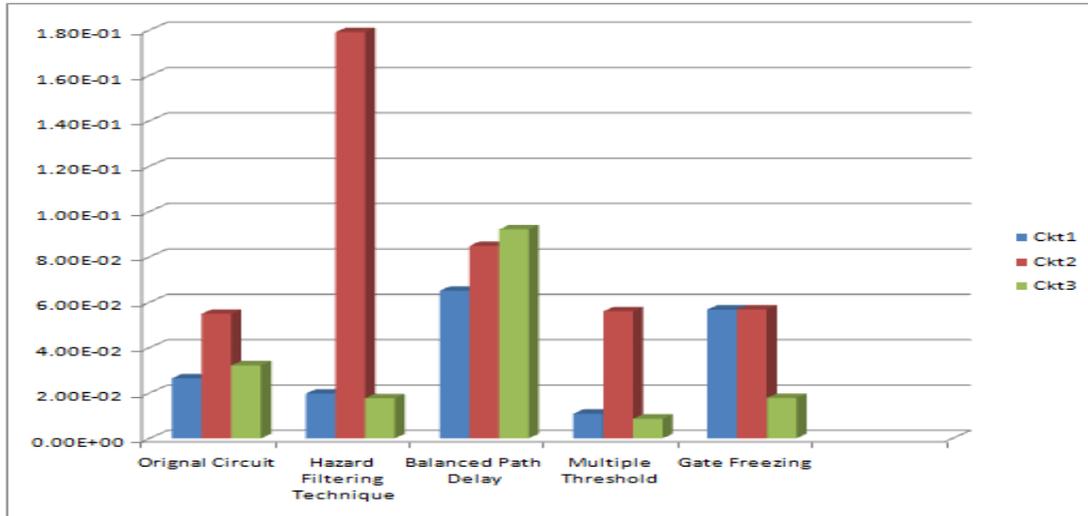


Figure 7 Circuit wise Maximum power consumption in watts

3.3 Delay Analysis

All techniques perform in similar manner for delay analysis of various circuits with a smaller lead to Gate Freezing upto 67.72% and Multiple Threshold Technique upto 32.18%. Hence these techniques may be used with proper analysis for reducing delay of the circuit.

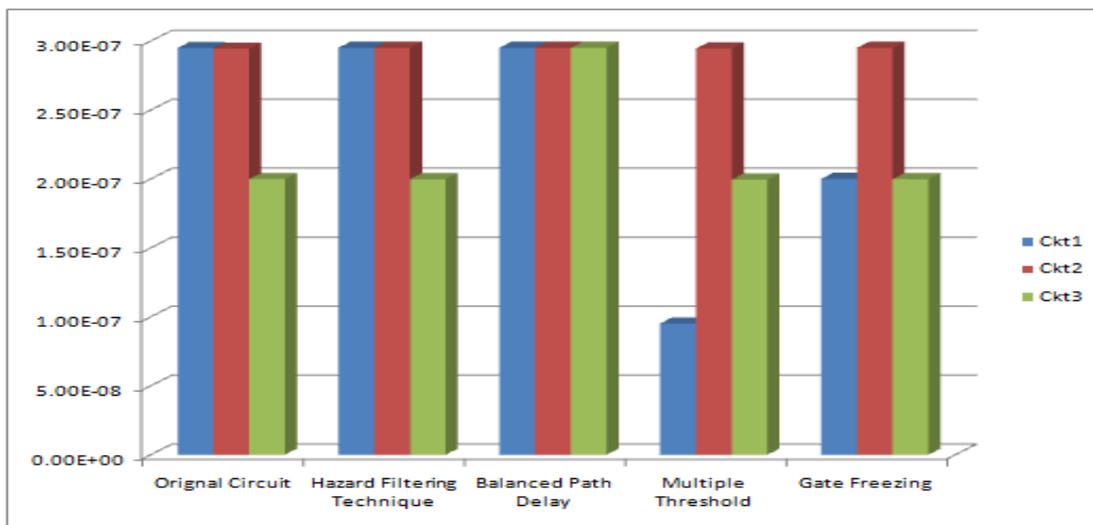


Figure 8 Circuit wise delay analysis in sec

3.4 Noise Analysis

As seen from the chart, balanced path technique is the best when analyzed for least noise in the output waveform upto 66.48%. The second best technique is the hazard filtering technique which also offers lesser noise in the output.

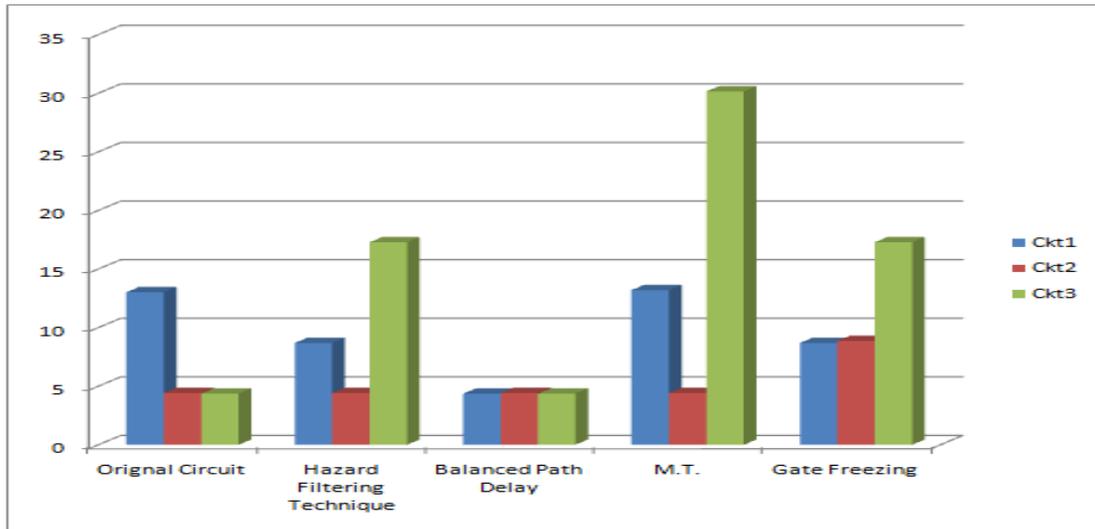


Figure 9 Circuit wise noise analysis (in Sq V/Hz)

4. CONCLUSION

For applications such as mobile computational systems, low power design is a criterion which should be satisfied. In this paper, we try to reduce the glitch power in the circuits and analyze the various available techniques such as gate freezing, hazard filtering, balanced path delay and Multiple threshold technique for noise, delay and power using tanner tool. We ascertain the various techniques according to required specification in terms of these parameters. We show that Hazard Filtering Technique and Multiple Threshold Technique are better than other techniques to reduce power consumption in the circuit upto 54.58% and 85.13% respectively. To reduce delay and speed up the circuit, we can use Multiple Threshold Techniques where it is reduced upto 67.72%, Gate freezing technique is the second best approach to speed up the circuit upto 32.18%. Noise is best reduced in balanced path delay technique upto 66.48%.

ACKNOWLEDGMENT

This research was supported by National Institute of Science and Technology under summer undergraduate research fellowship.

REFERENCES

- [1] Vikas, Deepak "A REVIEW ON GLITCH REDUCTION TECHNIQUES "IJRET: International Journal of Research in Engineering and Technology, Volume : 03 Issue: 02,2014

- [2] Masanori Hashimoto, Hidetoshi Onodera and Keikichi Tamaru” A Practical Gate Resizing Technique Considering Glitch Reduction for Low Power Design”,1999
- [3] Warren Shum and Jason H. Anderson, “FPGA Glitch Power Analysis and Reduction”, International Symposium on Low power electronics and design (ISLPED) 2011, page no. 27-32.
- [4] J. Lamoureux, G. Lemieux, and S. Wilton, “GlitchLess: Dynamic power minimization in FPGAs through edge alignment and glitch filtering”. IEEE TVLSI,16(11):1521– 1534, Nov. 2008.
- [5] Henrik Eriksson and Per Larsson-Edefors, “Impact of Voltage Scaling on Glitch Power Consumption”, Integrated Circuit Design Lecture Notes in Computer Science Volume 1918, 2000, pp 139-148.
- [6] Masanori Hashimoto, Hidetoshi Onodera and Keikichi Tamaru, “A Practical Gate Resizing Technique Considering Glitch Reduction for Low Power Design” .
- [7] Luca Benini, Giovanni De Micheli, Fellow, Alberto Macii, Enrico Macii, Massimo Poncino and Riccardo Scarsi, “Glitch Power Minimization by Selective Gate Freezing”, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 8, No. 3, June 2000.
- [8] Zhanping Chen, Liqiong Wei and Kaushik Roy, “Reducing Glitching And Leakage Power In Low Voltage CMOS Circuits”, Purdue University Purdue e Pubs ECE Technical Report, march 1997.
- [9] Masanori Hashimoto, Hidetoshi Onodera and Keikichi Tamaru, “A Power Optimization Method Considering Glitch Reduction by Gate Sizing”, International Symposium on Low Power Electronics and Design,1998, page no. 221- 226.
- [10] J. P. Uyemura, Circuit Design for CMOS VLSI. Norwell, MA: Kluwer,1992.
- [11] N. H. E. Weste and K. Eshraghian, Principles of CMOS VLSI Design a Systems Perspective , 2d ed. Reading, MA: Addison-Wesley, 1993.
- [12] P. E. Allen and D. R. Holberg, CMOS Analog Circuit Design, 2d ed.New York: Oxford Univ. Press, 2002.
- [13] A. Hastings, The Art of Analog Layout. Upper Saddle River, NJ: Pren-tice-Hall, 2001.
- [14] N. Ranganathan and Ashok K. Murugavel “A Microeconomic Model for Simultaneous Gate Sizing and Voltage Scaling for Power Optimization, Department of Computer Science and Engineering,Nanomaterial and Nanomanufacturing Research Cente 21st International Conference on Computer Design ,2003 IEEE
- [15] Luca Benini , Member, IEEE , Giovanni De Micheli, Fellow, IEEE , Alberto Macii, Student Member, IEEE ,Enrico Macii, Member, IEEE , Massimo Poncino , Member, IEEE , and Riccardo Scarsi “Glitch Power Minimization by Selective Gate Freezing”, IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, VOL. 8, NO. 3, JUNE 2000 287
- [16] Ki-Seok Chung Taewhan Kim C. L. Liu Design Technology Dept. of EECS and Dept. of Computer Science Intel Corporation National Tsing Hua Univ. Santa Clara, 95052 USA Hsinchu, Taiwan R.O.C. Adv. Information Tech. Research Center Korea Adv. Institute of Sci.&Tech. Taejon, Korea,“A Non-Zero Delay Model for Glitch Analysis in Logic Circuits”, 43rd IEEE Midwest Symp. on Circuits and Systems, Lansing MI, Aug 8- 11,2000

ANALYSIS OF CMOS AND MTCMOS CIRCUITS USING 250 NANO METER TECHNOLOGY

M Suresh¹, A K Panda¹, Mukesh Sukla¹, Marakonda Patnaikuni Vasanthi²
and Sowpati Santhi³

¹Department of ECE, NIST, Berhampur, Odisha
msuresh73@hotmail.com
akpanda62@hotmail.com
Mukesh_ele02@yahoo.co.in

²Department of ECE, RGUKT, Basar, Telangana
vasanthimpiit@gmail.com

³Department of ECE, RGUKT, Nuzvid, Andhra Pradesh
santhisowpati555@gmail.com

ABSTRACT

The low-power consumption with less delay time has become an important issue in the recent trends of VLSI. In these days, the low power systems with high speed are highly preferable everywhere. Designers need to understand how low-power techniques affect performance attributes, and have to choose a set of techniques that are consistent with these attributes. The main objective of this paper is to describe, how to achieve low power consumption with approximately same delay time in a single circuit. In this paper, we make circuits with CMOS and MTCMOS techniques and check out its power and delay characteristics. The circuits designed using MTCMOS technique gives least power consumption.

All the pre-layout simulations have been performed at 250nm technology on tanner EDA tool.

KEYWORDS

MTCMOS, sleep mode, leakage current, header switch, footer switch

1. INTRODUCTION

In the earlier days VLSI designers mainly concentrated on area, performance, speed, cost and reliability. But this performance improvement has lead to the increase in power dissipation. Reducing this power dissipation and achieving low power consumption has become a challenging task to the current day designers as cooling technology and packing are very expensive and also now a days because of the battery life time, the electronic circuit designers are worried about decreasing the total power consumption to increase the battery life time[11], especially for portable embedded systems and decrease the battery's size which is reflected on the portability of the devices. Power is very much concerned due to the remarkable growth and success of fields

like personal computing devices and wireless communication system which demand high speed computation and complex functionality with low power consumption. As the technology continues to scale down a significant portion of the total power consumption in high performance digital circuits is due to leakage current because of reduced threshold voltage [1]. MOSFETs are fabricated with high overall doping concentration, lowered source/drain junction depths, halo doping, high mobility channel materials, etc. Furthermore the reduction of the gate oxide thickness causes drastic increase in the gate tunnelling leakage current due to carriers tunnelling through the gate oxide, which is a strong exponential function of the voltage magnitude across the gate oxide [2],[7] to minimize the leakage current. Here our main aim is to decrease the leakage current using MTCMOS technique.

2. CMOS

Complementary metal-oxide semiconductor is the most leading semiconductor technology used in the transistors that are manufactured into most of today's computer microchips. CMOS logic is well known for its extremely low static power dissipation and high noise immunity. CMOS is sometimes referred to as complementary-symmetry metal oxide semiconductor. Complementary-symmetry refers that a typical CMOS design style uses complementary and symmetrical pairs of p-type and n-type metal oxide field effect transistors for logic functions [8]. In CMOS technology, both kinds of transistors are used in a complementary way to form a current gate that forms an effective means of electrical control.

In this, all the PMOS devices will be together called as pull-up network and substrates are connected to the VDD, all the NMOS devices will be together called as pull-down network and substrates are connected to the VSS. The output is taken at the centre as a function of inputs.

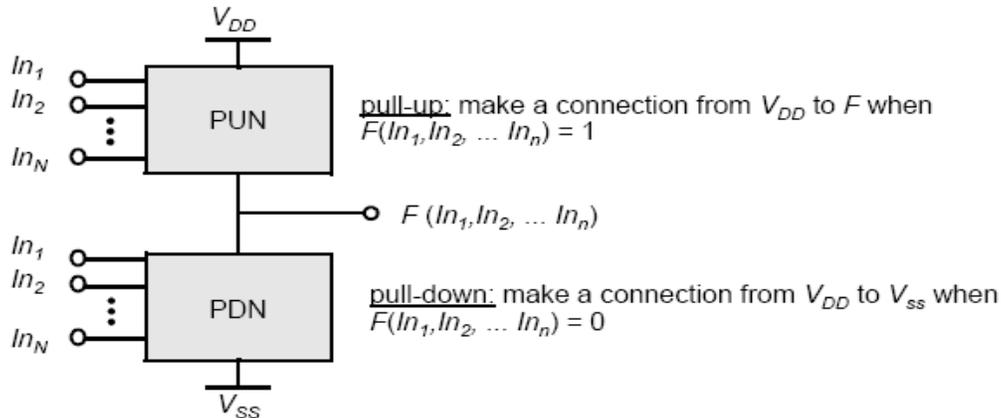


Figure 1. CMOS basic structure

3. MTCMOS

Multi-threshold CMOS is a power reduction technique, widely used in today's industry to lower the gate leakage current. The multi threshold CMOS technology has two main parts. First, "Active" and "sleep" operational modes [13] are associated with MTCMOS technology, for efficient power management. Second, two different threshold voltages are used for N channel and P channel MOSFET in a single chip [9]. The low threshold voltage transistor is able to switch

faster but has a high leakage current when turned off compared to the high threshold voltage transistor which is slower to switch but has a low leakage current when off. In this we use low threshold transistor for logic and to separate it from power /ground with high threshold transistors and also the circuit is operated at high performances because of low threshold voltage transistor.

When a logic circuit is active, the sleep signals are de-asserted which turn on high threshold voltage transistors and create virtual VDD and GND around the logic. In inactive mode the sleep signals are asserted which separate the logic from the power/ground, there by lowers the leakage current. The MTCMOS technique shows no impact over circuit parameters such as output impedance, gain, threshold voltage, fluctuations and frequency response [10].

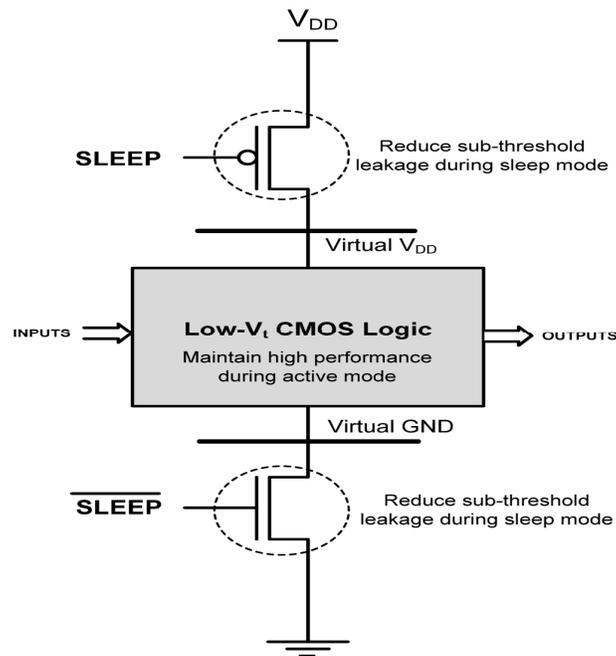


Figure 2. MTCMOS basic structure.

4. POWER CONSUMPTION OF CMOS AND MTCMOS CIRCUITS

The Complementary-Metal Oxide Semiconductor technology is well-known for its low power consumption. CMOS gates in older technologies were very efficient. In newer technologies, it has been skyrocketed due to transistor scaling, chip transistor counts and clock frequencies.

The average power over the time interval is

$$P_{avg} = E/T = (1/T) \int_0^T i_{DD}(t) V_{DD} dt$$

Where $i_{DD}(t)$ is supply current and V_{DD} is the supply voltage.

MTCMOS is a power gating technique in which a power gating transistor will be placed between the logic transistors and either powered or grounded, thus creating a virtual supply and virtual

ground, respectively. Power gating is a technique used to reduce power consumption by shutting off the current to blocks of the circuit that are not in use. Lowering the threshold voltage results in an exponential increase in sub-threshold current [6]. As the circuit spends more time in the ideal (stand-by) mode, so it is practical to reduce the leakage current to minimize the static power which represents the dominant part of the total power consumption. Multi-threshold CMOS (MTCMOS) technology is one of the most effective techniques to reduce the leakage current during the standby mode by using a low threshold voltage transistor in the critical paths of the circuit to improve the performance while the high threshold voltage one is in uncritical paths and is used as an isolation switch between the virtual supply lines (VDD, GND) and the real one [14]. The high threshold voltage transistor is used for power consumption in the shortest path [3],[4]. Both active mode and sleep mode are associated for efficient power management.

5. IMPLEMENTATION OF 2-BIT SERIAL IN SERIAL OUT SHIFT REGISTER

A Shift Register is a sequential logic circuit that is used to store a sequence of data and this data is shifted by one clock pulse for every clock period at its output. They are a group of flip-flops that are connected in chain so that the output of one flip-flop will be given as the input to another flip-flop connected to it. All flip-flops are driven by a common clock and also they are set and reset simultaneously. Here we are taking both input and output in a serial manner so we call it as serial in serial out (SISO) shift register.

5.1. Conventional 2-Bit SISO Shift register using CMOS

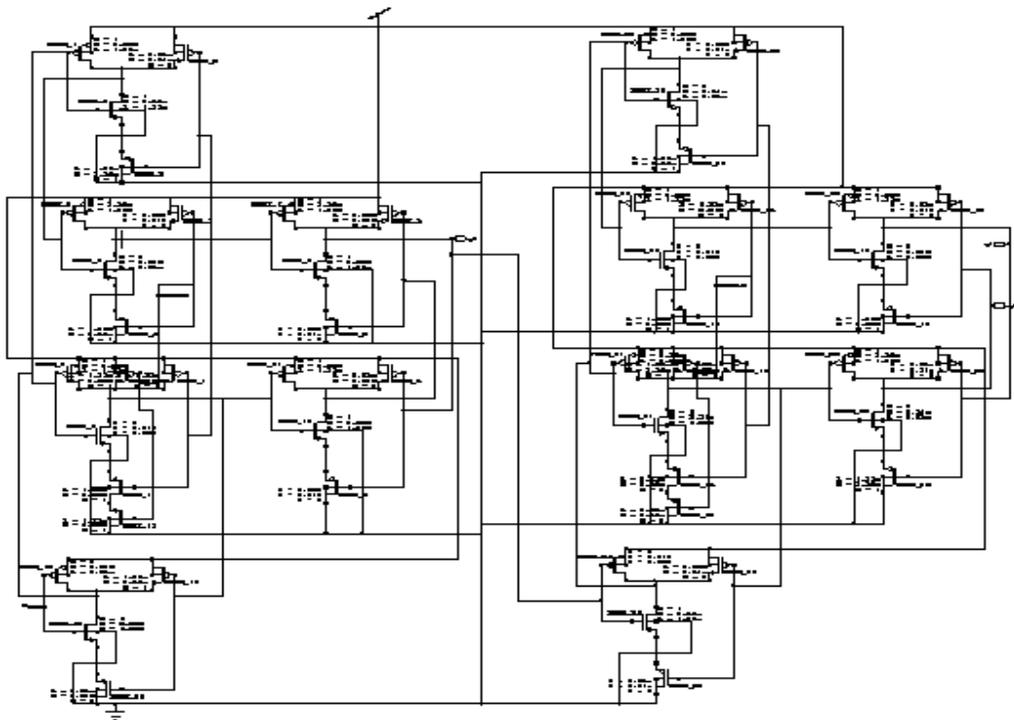


Figure 3. Schematic diagram of Conventional 2-Bit SISO Shift Register using CMOS

5.2. Simulation Result

Taking d as the input, clock as the clock input, q1 as MSB and q0 as LSB we get the following output waveform for 2-Bit SISO Shift Register using normal CMOS technique.



Figure 4. Output waveforms of conventional 2-Bit SISO Shift Register

5.3 Proposed 2-Bit SISO Shift Register using MTCMOS technique

To get the proposed design, we add a PMOS transistor that connects VDD and the circuit and forms a virtual power supply and a NMOS transistor that connects VSS and circuit and forms a virtual VSS. An inverter is designed, using it we give the sleep signal directly to PMOS and its inverted output to the NMOS.

The sleep transistor is controlled by a sleep signal that can be used to switch on and off the device. The PMOS sleep transistor can be called as “header switch” as it connects VDD supply to the circuit and the NMOS sleep transistor can be called as “footer switch” as it connects VSS supply to the circuit.

5.4. Simulation Result

Taking d as the input, clk as the clock input, q1 as the MSB and q0 as the LSB we get the following output waveform for 2-Bit SISO Shift Register using MTCMOS technique.

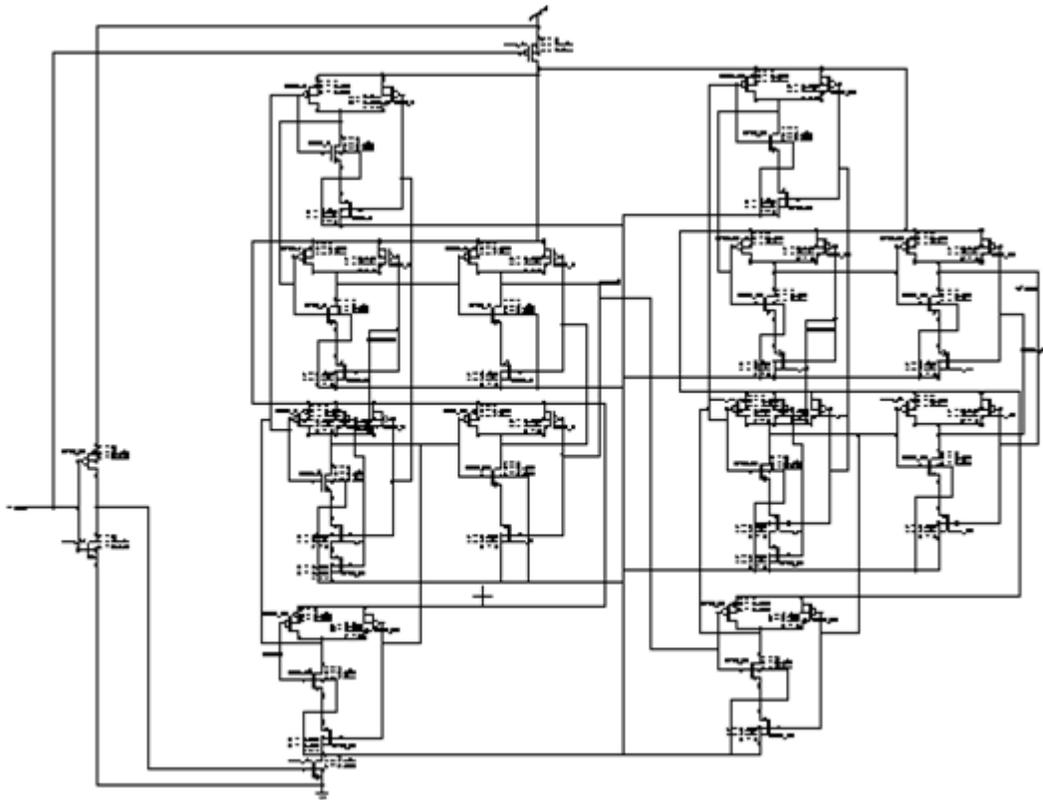


Figure 5. Schematic Diagram of Proposed 2-Bit SISO Shift Register using MTCMOS.

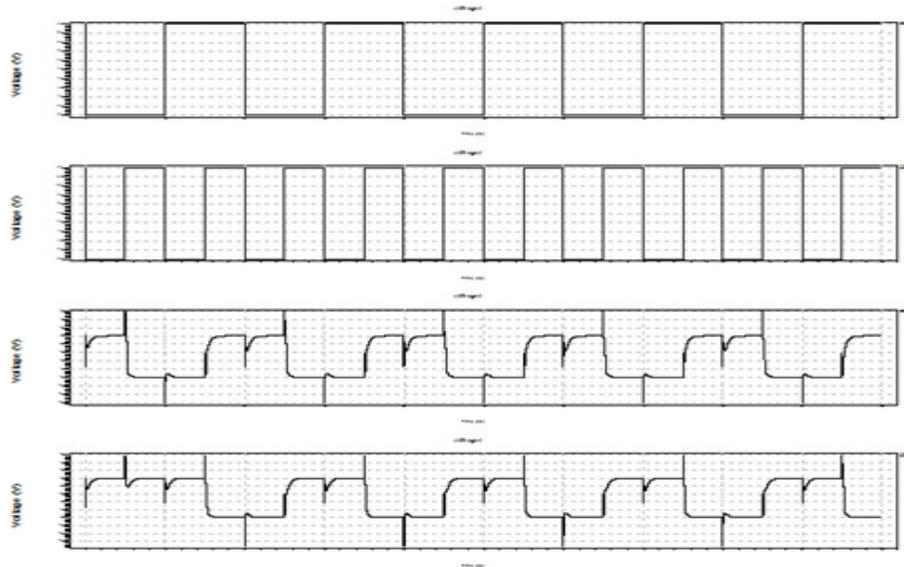


Figure 6. Output waveforms of proposed 2-Bit Shift Register.

6. IMPLEMENTATION OF 2-BIT BINARY INCREMENTER

The Binary Incrementer increases the value stored in the register by '1'. For this implementation it simply adds '1' to the existing value stored in the register. For this implementation we need 'n'

half adders to add 'n' number of bits. The storage capacity of the register is to be incremented. Here in the below example we are using two half adder to get 2-bit incrementer. The carry of the first half adder is used as input for the second half adder.

6.1. Conventional 2-bit incrementer design using CMOS

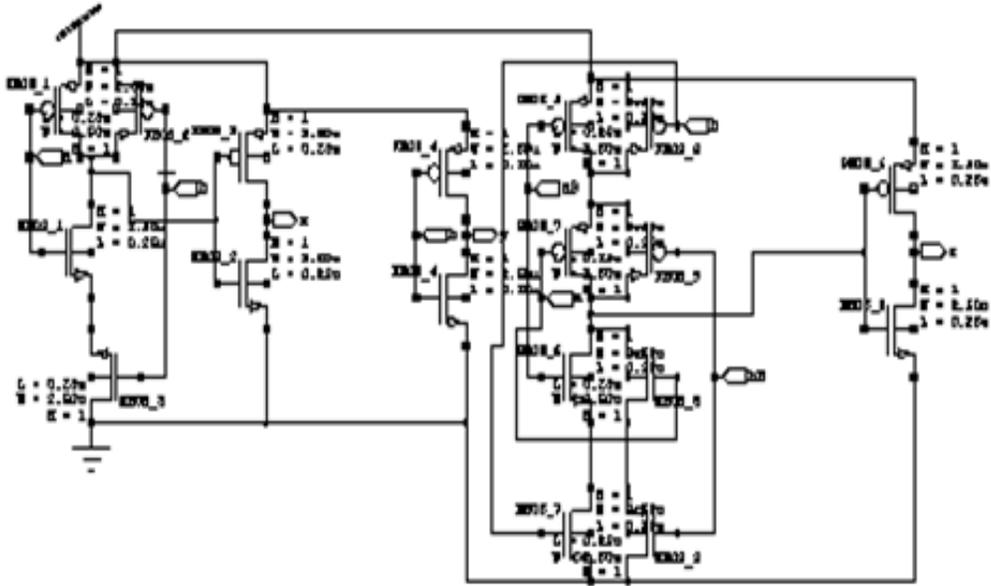


Figure 7. Schematic diagram of Conventional 2-bit binary incrementer using CMOS.

6.2. Simulation Result

Considering a, b as the inputs and x, z, y as the outputs we get the following output waveforms for 2-bit binary incrementer using CMOS.

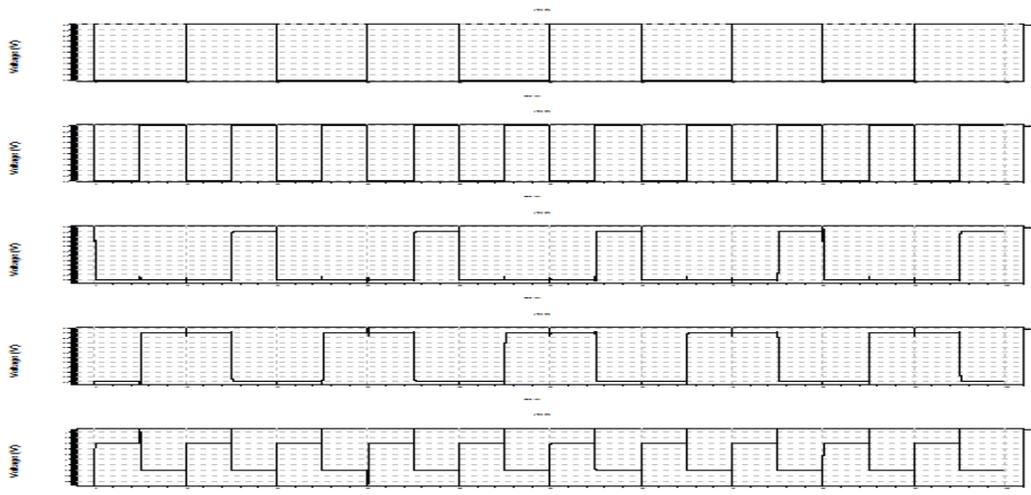


Figure 8. Output waveforms of conventional 2-bit binary incrementer using CMOS

6.3. Implementation of Proposed 2-bit binary incremter Using MTCMOS

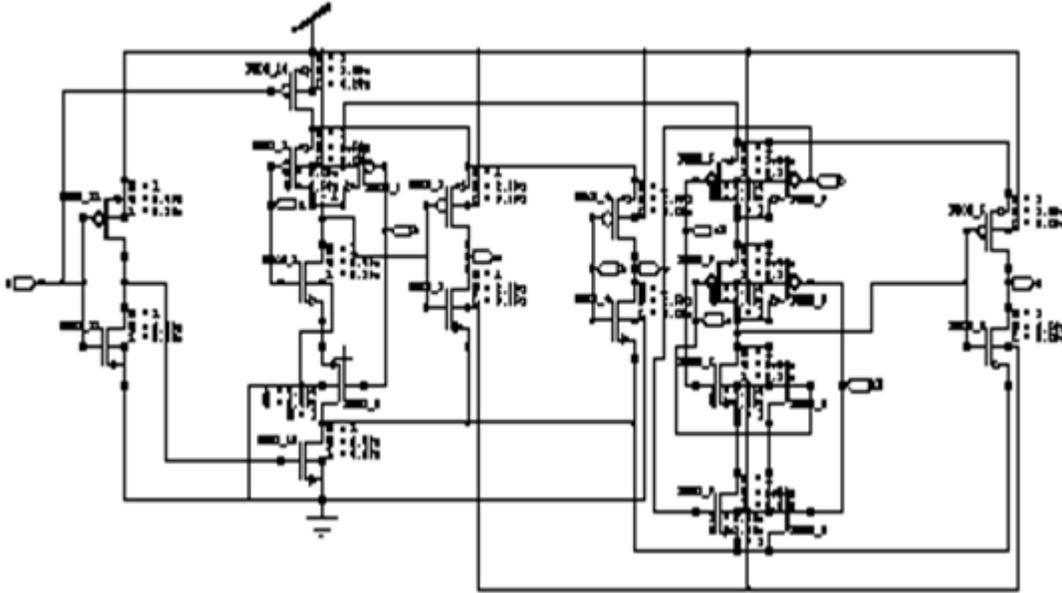


Figure 9. Schematic diagram of proposed 2-bit binary incremter using MTCMOS

6.4. Simulation Result

Considering a, b as the inputs x, z, y as the outputs we get the following output waveforms using MTCMOS technique.

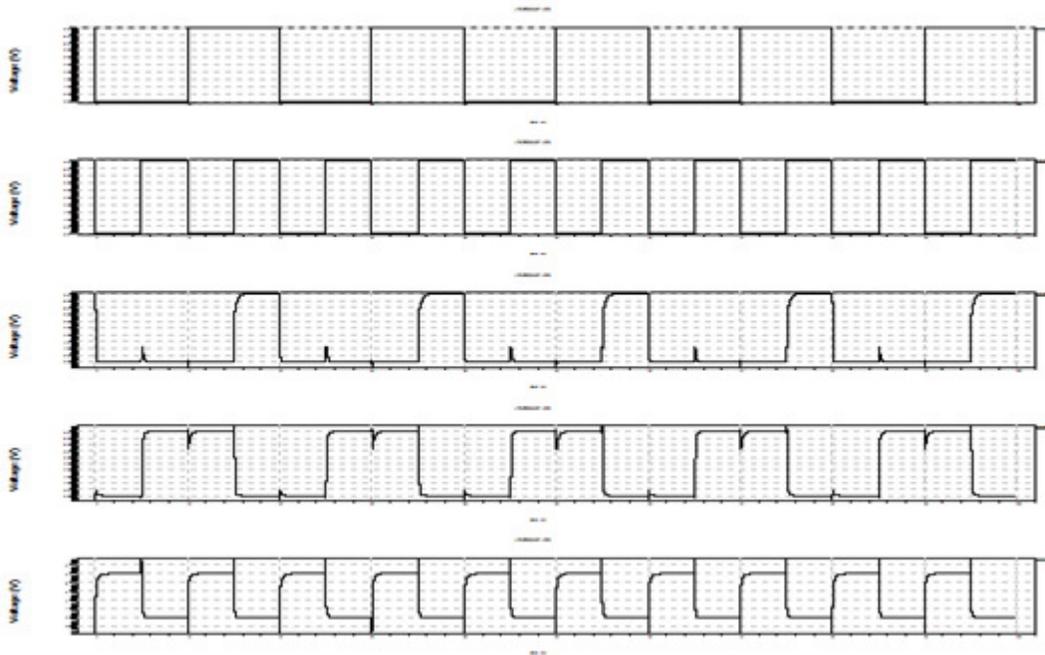


Figure 10. Output waveforms of proposed 2-bit binary incremter

7. POWER CONSUMPTIONS OF 2-BIT INCREMENTER AND SHIFT REGISTER USING CMOS AND MTCMOS

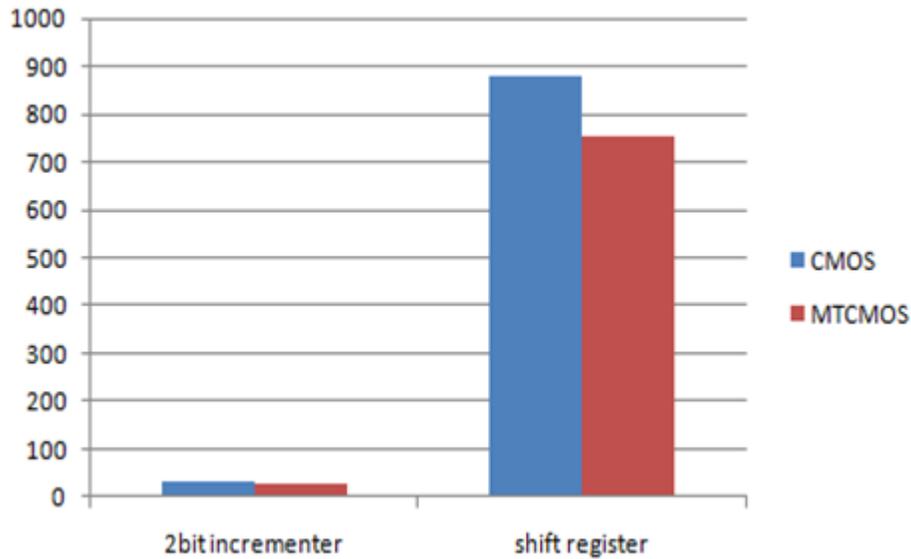


Figure 11. Graph of power consumptions of 2-bit incrementer and shift register

7.1 Generalizing MTCMOS technique to different circuits

Now apply the proposed MTCMOS technique to different kinds of gates and circuits and observe it's reduction in power consumption when compared to normal CMOS.

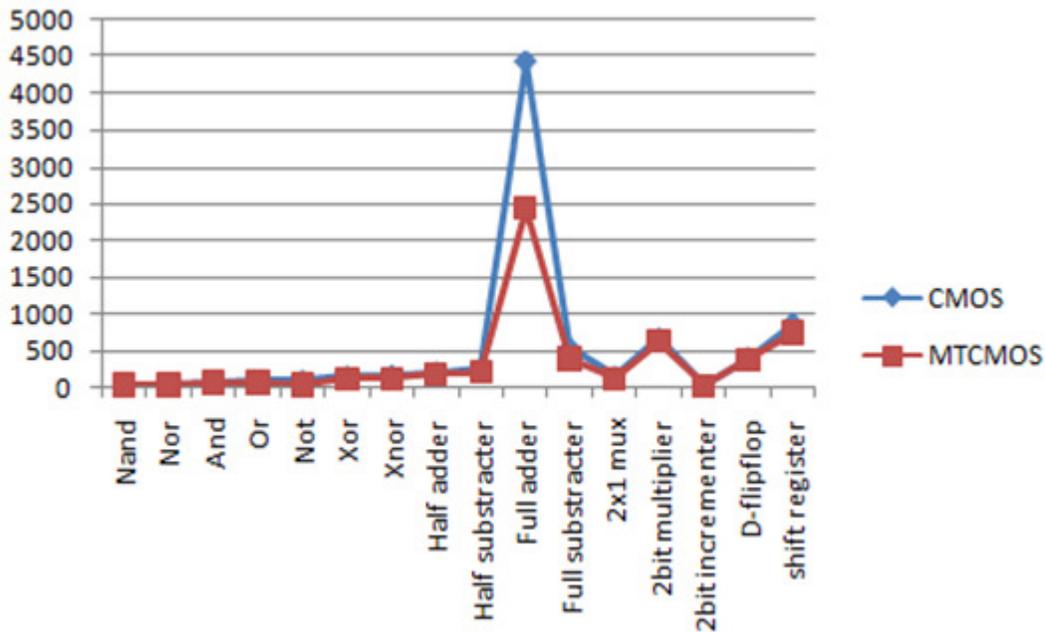


Figure 12. Difference in power consumptions between CMOS and MTCMOS Circuits

8. CONCLUSIONS

In this paper we concentrated on the leakage current analysis and it can be reduced using the MTCMOS technique. The proposed technique is associated with different threshold transistors to build the CMOS circuits. All the circuits are designed using 250nm technology and operated with supply voltage. In this the total average power is decreased because of its reduction in leakage currents using MTCMOS. As this technique deals with leakage current, we should always take care of the temperature.

ACKNOWLEDGEMENTS

The authors would like to thank Rajiv Gandhi University of Knowledge Technologies and National Institute of Science and Technology, Berhampur for providing their esteemed guidance and support to carry out their research work successfully.

REFERENCES

- [1] Pawar Chander, Pokala Santhosh, Prasad Kurhe, "VLSI DESIGN OF FULL SUBTRACTOR USING MULTI-THRESHOLD CMOS TO REDUCE LEAKAGE CURRENT AND GROUND BOUNCE NOISE", ISSN, Volume-2, Issue-2, 2015.
- [2] D. A. Antoniadis, I. Aberg, C. N. Chleirigh, O. M. Nayfeh, A. Khakifirooz and J. L.Hoyt, "Continuous MOSFET performance increase with device scaling: The role of strain and channel materialinnovations," IBM. J. Res.Develop., vol. 50, no. 4, pp 363-376,Jul,2006.
- [3] Dong Whee Kim, Jeong Beom Kee, "Low-Power Carry Look-Ahead Adder With Multi-Threshold Voltage CMOS Technology", in Proceeding of ICSICT International Conference on Solid-State and Integrated-Circuit Technolgy, pp.2160-2163,2008.
- [4] H. Thapiliyal and N. Ranganathan, "ConservativeQCAGate (CQCA) for Designing Concurrently Testable Molecular QCA Circuits", Proc. Of the 22nd Intl. Conf.on VLSI Design, New Delhi, India, pp. 511-516, 2009.
- [5] H. Thapiyal, M.B Srinivas and H.R.Arabnia, "Reversible Logic Synthesis of Half, Full and ParallelSubtractors", Proc. Of the 2005 Intl. Conf. on Embedded Systems and Applications, Las Vegas,pp. 165-181,2005.
- [6] Hematha S , Dhawan A and Kar H , "Multithreshold CMOS Design for low power digital circuits" ,TENCON 2008-2008 IEEE Region 10 Conference,pp.1-5,2008.
- [7] K. Roy, S. Mukhopadhyay , and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deepsubmicrometerCMOS circuits, "Proc. IEEE, vol.91, no. 2,pp. 305-327, Feb. 2003.
- [8] "what is CMOS memory?" Wicked Sago. Retrieved 3 March. 2013.
- [9] H.Thapliyal, M.B Srinivas, H.R Arabnia, "Reversible Logic Synthesis of Half, Full and ParallelSubtractors", proc. of the 2005 Intl. Conf. on Embedded Systems and Applications, Las Vegas, pp. 165-181.

- [10] Anis, M.H.; and Elmarsy, M.I. (2002). Power reduction via an MTCMOS implementation of MOS current mode logic. Proceedings of IEEE ASIC/SOC conference, 193-197.
- [11] Itziar Marin, Eduardo Arceredillo, Jagoba Arias, Aitzol Zuloaga, Iker Losada, "low-power aware design: Topics on low battery consumption", proceedings of the 4th WSEAS Int. Conf. on Information Security, Communications and computers , Tenerife, Spain, December 16-18,2005(pp47-52).
- [12] A. Chilambuchelvan,S. Saravanan, B. Chidhambarar Ajan, J.Raja Paul Perinbam, "Certain Investigations on Energy saving techniques using DVS for low power embedded system", Proceedings of the 6th WSEAS International Conference on Applied Informatics and communications, Elounda, Greece, August 18-20,2006(pp298-305) .
- [13] Alice Wang , Benton H. Calhoun, Anantha P. Chandrakasan, "subthreshold Design for Ultra Low-Power Systems", Springer US, 2006.
- [14] J.T. Kao, A. P. Chandrakasan, "Dual-threshold voltage techniques for low-power digital circuits", IEEE Journal of Solid-State Circuits, Volume 35, ISSUE 7, July 2000,pp. 1009-1018.
- [15] K. Roy, "Leakage power reduction in low-voltage CMOS designs", IEEE International Conference on Circuits and Systems, Volume 2, Sept.1998, pp.167-173.
- [16] B.H. Calhoun, F. A. Honore, A. p. Chandrakasan , "A leakage reduction methodology for distributed MTCMOS", IEEE Journal of Solid-State Circuits ISSUE 5, May 2004, pp.818-826.
- [17] Jan M.Rabaey, Anantha Chandrakasan, Borivoje Nikolic "Digital integrated circuits", a design perspective, second edition, 2003.
- [18] Electronic Publication: Digital Object Identifiers(DOIs): Article in a journal:
- [19] j. Clerk Maxwell, A. Treatise on Electricity and Magnetism, 3rd ed., vol.2. oxford: clarendon, 1892,pp.68-73.
- [20] M. Young The Technical writer's Handbook. Mill Valley, CA: University Science, 1989.

INTENTIONAL BLANK

DENGUE DETECTION AND PREDICTION SYSTEM USING DATA MINING WITH FREQUENCY ANALYSIS

Nandini. V¹ and Sriranjitha. R² and Yazhini. T. P³

Department of Computer Science and Engineering,
SSN College of Engineering, Kalavakkam

¹nandini.vishwa94@gmail.com

²sriranjitha.raghuraman@gmail.com

³tp.yazhini@gmail.com

ABSTRACT

Clinical documents are a repository of information about patients' conditions. However, this wealth of data is not properly tapped by the existing analysis tools. Dengue is one of the most widespread water borne diseases known today. Every year, dengue has been threatening lives the world over. Systems already developed have concentrated on extracting disorder mentions using dictionary look-up, or supervised learning methods. This project aims at performing Named Entity Recognition to extract disorder mentions, time expressions and other relevant features from clinical data. These can be used to build a model, which can in turn be used to predict the presence or absence of the disease, dengue. Further, we perform a frequency analysis which correlates the occurrence of dengue and the manifestation of its symptoms over the months. The system produces appreciable accuracy and serves as a valuable tool for medical experts.

KEYWORDS

Named Entity Recognition, Part of Speech tagging, Classification, Prediction, SMO

1. INTRODUCTION

Mining unstructured data is a very pressing issue in the field of text mining. This is especially a major subject in the area of medicine. Clinical decisions are often made based on doctor's intuition and experience rather than on the knowledge-rich data hidden in the database. Dengue is attracting global concern from researchers and health care professionals over the world. Statistics reveal that almost 25,000 people die from dengue every year. Timely detection of symptoms associated with this deadly disease, and apt prevention measures will go a long way in bringing down its effects on the world populace. Hence, we need a system that will first learn the characteristics of people with dengue, and use this knowledge to predict dengue in new patients. Over the years several NLP systems like cTakes, MetaMap, etc. [3] [2] were used to extract medical concepts from clinical text. They focused on rule based, medical knowledge driven dictionary lookup approaches. While some researchers have contributed to disease prediction,

they have concentrated primarily around heart attacks [6] [7] [10] [11] [12]. Inspiration drawn from such work, combined with the increasing rate of dengue cases around the world motivates us to develop a system to model, predict and analyze dengue instances. The inability to extract useful information from clinical documents may hamper the health care experts' efforts from understanding the relationship between the prevalence of diseases and the associated factors. The frequency of diseases can also be allied with its time frame. This is especially true in the case of water-borne and air-borne diseases. Addressing this task will be a major help to doctors, experts and patients. This relation will enable health care connoisseurs to take preventive measures and reduce the prevalence of these diseases.

2. PROBLEM STATEMENT

The knowledge available in medical repositories is effectively mined and analyzed using the proposed system. The input is a set of annotated discharge summaries containing data pertaining to the disease dengue. Disorder names are extracted from these summaries and looked up in a summarized UMLS (Unified Medical Language System). The output produced in this step is supplied to classifiers which then perform detection and prediction. Further, frequency correlation is performed with the time frame.

3. OVERVIEW OF PROPOSED SYSTEM

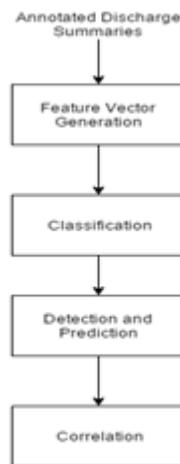


Figure 1. Overview of the system

The annotated discharge summaries are supplied to feature extraction algorithms and the extracted features are in turn used to generate a feature vector. This is supplied as input to a classification algorithm and a prediction model is developed. The model generated can then be used to detect and predict the presence of dengue. Finally, a correlation analysis is performed to determine how the disease is spread over the months.

4. RELATED WORK

N. Aditya Sundar et al [5] use regular factors contributing to heart diseases, including age, sex, blood sugar and blood pressure, to predict the likelihood of a patient getting a heart disease. Data mining techniques of Naïve Bayesian classification and WAC (Weighted Associative Classifier)

are used to train a model on existing data. Subsequently, patients and nurses can use this model to supply features and get a prediction on a possible heart attack. Oona Frunza et al [6] present a machine learning approach that identifies semantic relations between treatments and diseases and focuses on three semantic relations (prevent, cure and side effect). Later, features were extracted from unstructured clinical text, and were used to classify the relationship between diseases and associated treatments. Jyoti Soni et al [7] have developed a predictive data mining algorithm to predict the presence of heart disease. Fifteen attributes were selected to perform the prediction and Decision Tree was found to produce the best results. Classification based on clustering algorithms was found to not perform well. [12] Proposes a Medical Diagnosis System for predicting the risk of cardiovascular disease. It uses genetic algorithm to determine the weights for a neural network. This feed forward neural network is subsequently used for classification and prediction of heart diseases. A data set of 303 instances of heart disease with 14 attributes each is used for training the system. Devendra Ratnaparkhi, Tushar Mahajan and Vishal Jadhav in their paper [10] describe a system for prediction of heart disease using Naïve Bayes algorithms. They further propose a web interface to help healthcare practitioners assess the possibility of a heart problem in patients. A similar attempt proposes a heart disease prediction system [11] using Decision Tree and Naïve Bayes and its implementation in .NET platform by I.S.Jenzi et al in their paper. Some data mining techniques used for modeling and prediction of dengue include SVM [13], decision tree [14] and neural network [15].

5. SYSTEM DESIGN

The system design is divided into 2 parts.

5.1. Feature Vector Generation

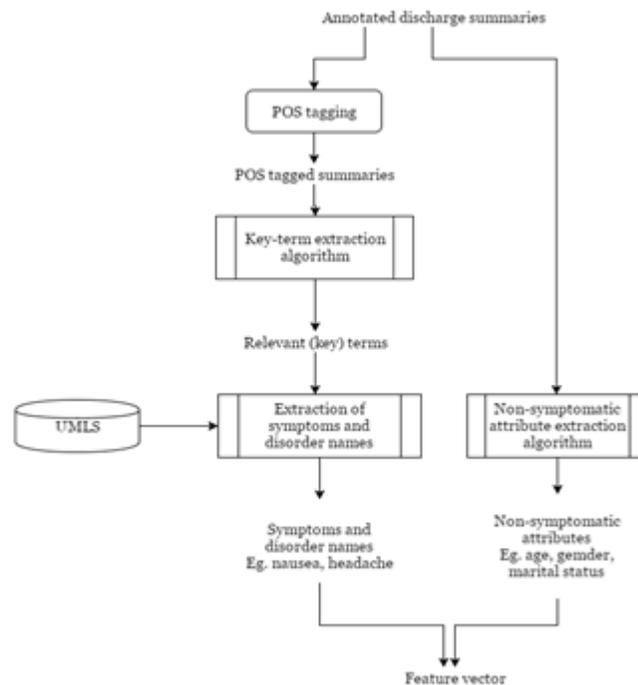


Figure 2. Feature Vector Generation

5.1.1. POS Tagging

The Stanford POS Tagger is used to tag the discharge summaries. An instance of the tagger class is created. The input data is stored in a folder. The program iterates through the folder's files and tags all the input files using the tagger instance created. The tagged data is stored in a file.

```
POSTagger tag = new POSTagger();
File f = new File("DischargeSummary.txt");
while str=f.read() != NULL do
| tag.tagSample(str);
end
```

Figure 2. POS tagging pseudo code

5.1.2. Key Term Extraction

The key terms such as nouns and adjectives (specified by the tags NN NNP NNS NNPS JJ JJS etc) are extracted from the tagged data and stored in a file.

```
for all word sequences of the form (adjective,noun),(noun),(gerund preposition
noun) do
| Add Sequence to keyterms file;
end
```

Figure 3. Key term extraction pseudo code

5.1.3. Duplicates Removal

The file generated might contain redundant attributes. To avoid this, the duplicates are removed. The pseudo code for the same is given below

```
Load keyterms file;
Create or open distinctKeyterms file;
Create an empty ArrayList of Strings;
for each word sequence in keyterms file do
| if sequence is not already in ArrayList then
| | Add sequence to ArrayList
| end
end
```

Figure 4. Duplicates removal pseudo code

5.1.4. Dictionary Look Up

UMLS (Unified Medical Language System) serves as a repository of mentions. The UMLS is used to extract the relevant symptoms from the tagged file. FileSearcher Class is imported and the FindWordInFile method is used to search for a word in a given file.

```

Load UMLS dictionary;
Load ArrayList of distinct keyterms file;
Create or open distinctFilteredKeyterms file;
Import FileSearcher class ;
Create new FileSeacher object;
for each keyterm in distinctKeyterms file do
    Call findWordInList(word,file) function of FileSearcher object as
    findWordInList(keyterm,uMLS) if returned value is true, keyterm is present in
    UMLS then
    | Add keyterm to distinctFilteredKeyterms file
    else
    end
end

```

Figure 5. Dictionary lookup pseudo code

5.1.5. Temporal Data Extraction

The discharge summaries are fed as input to the temporal data extraction algorithm. The admission months are extracted using regular expressions.

```

String pattern= Expression;
Pattern p = new Pattern(pattern);
Matcher m = p.matcher(Line to be processed);
if m.find() then
    | return month;
else
    | return null;
end

```

Figure 6. Temporal data extraction pseudo code

5.1.6. Non-Symptomatic Feature Extraction

Non-Symptomatic features such as age, gender, marital status, family history and past medical history are extracted using regular expressions from the annotated discharge summaries.

- The age is computed using the date of birth of the patient.
- The gender can be either M or F (Male or Female)
- The marital status can be Y or N (Yes or No)
- The family history can be Y or N (Yes or No)
- The past medical history can be Y or N (Yes or No)
- The disease can be Y or N (Yes or No)

5.1.7. Feature Vector Generation

The feature vector is the input supplied to the classifier. The features extracted are combined in a comma separated format and a feature vector is generated. The vector can be represented using frequency value representation or using binary representation. The frequency value format implies that, the frequency of occurrence of the feature in the document is considered. The binary representation on the other hand only considers the presence or absence of the feature in concern. Dengue has very few prominent symptoms and therefore it is not advisable to use the frequency value representation to retrieve them from clinical text. Binary representation of symptoms is therefore preferred in this case. Non-symptomatic features are represented as nominal attributes.

```

Load distinctFilteredKeyterms file;
Create or open featureVector file;
Import FileSearcher class ;
Create new FileSeacher object;
for each file in input data folder do
| Add input file name to featureVector file
end
for each keyterm in distinctFilteredKeyterms file do
| Call findWordInFile(keyterm,input file) on FileSearcher object
end
if returned value is true then
| Print 1 ;
else
end
Print 0 ;
Print newline;

```

Figure 7. Feature Vector Generation pseudo code

The feature vector generated is supplied to a set of classifiers. To identify the best classifier an analysis is performed.

5.2. Classification and Analysis

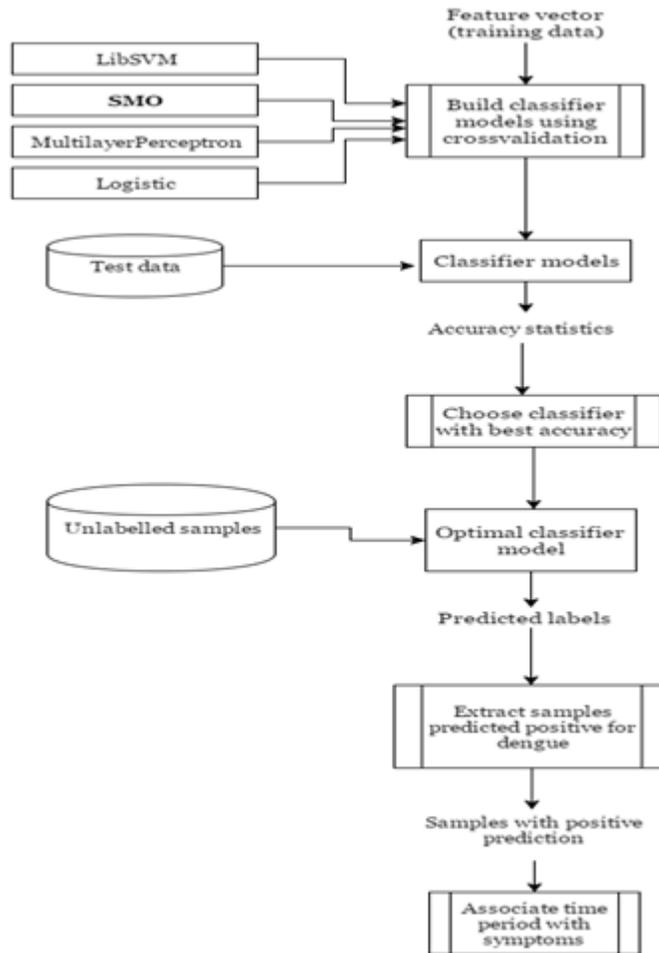


Figure 8. Classification and analysis

5.2.1. Classification

The following is the gist of steps followed during classification process:

1. Prepare Training set
2. Supply training set to the classifiers
3. Build the classification models
4. Save the models that have been built
5. Prepare the test set
6. Evaluate the test set on the saved models

7. Analyze the performance of the classifiers
8. Choose the best classifier
9. Supply unlabeled dataset to the best classifier
10. Obtain the prediction

5.2.2. Frequency Analysis

Frequency analysis aims at correlating the frequency of occurrence of the disease over the months. Eight most common and highly contributing symptoms for dengue have been chosen. The occurrences of these symptoms over the months is represented using graphs to give a better understanding of which symptom contributes the most to the presence of dengue.

6. IMPLEMENTATION

6.1. Data set used

We have used 100 samples of annotated discharge summaries as input to this system. The personal details of the patients are already preprocessed to ensure patient confidentiality. They contain details like age, date of birth, date of admission, patient's medical history, medication administered to the patient during the period of stay in the hospital. And the final diagnosis of the patient is also mentioned.

6.2. Tagged file

The above dataset is sent to a POS tagger to perform the part of speech tagging. An instance of the tagger is created and its TagFile method is used to tag the data. This tagged file is sent to a key term extraction algorithm and the relevant features are extracted. The duplicate terms are removed from using the duplicates removal algorithm. These terms are stored in a file.

6.3. UMLS Look up

A subset of the UMLS containing terms relevant to the disease are used as basis to perform the dictionary look up. The file containing the key terms is then compared with the thesaurus and symptoms that contribute to dengue are stored in another file.

6.4. Feature Extraction and Vector Generation

6.4.1. Symptomatic features

To extract the symptomatic features, the following steps are performed:

1. A file reader object is created
2. The discharge summaries are read line by line
 - Each line is split into words
 - The words are compared with the file containing filtered output

- If there is a match , 1 is written to the feature to the feature vector

3. If there is no match, 0 is written to the vector

6.4.2. Non- Symptomatic features

The non-symptomatic features are extracted using regular expressions. The features are extracted and written to the feature vector file. The feature vector is saved as an arff file.

Snapshot of the generated vector is as shown:

```
@relation DENGUE
@attribute age numeric
@attribute gender (M/F) {M,F}
@attribute ad_mon (01-12) numeric
@attribute marital_stat (M/N) {M,N}
@attribute med_his (Y/N) {Y,N}
@attribute fam_his (Y/N) {Y,N}
@attribute fever (0/1) numeric
@attribute headache (0/1) numeric
@attribute nausea (0/1) numeric
@attribute vomit (0/1) numeric
@attribute Pain_behind_eye (0/1) numeric
@attribute Fatigue (0/1) numeric
@attribute muscle_pain (0/1) numeric
@attribute skin_rash (0/1) numeric
@attribute disease (Y/N) {Y,N}
@data
35, M, 7, M, Y, N, 1, 0, 1, 1, 0, 1, 0, 0, Y
45, M, 8, M, N, N, 1, 1, 0, 0, 0, 0, 1, 1, Y
39, M, 7, N, N, Y, 0, 0, 1, 1, 0, 0, 0, 0, N
57, M, 9, M, N, Y, 1, 1, 0, 0, 0, 0, 0, 1, Y
29, M, 9, M, N, N, 1, 1, 0, 1, 1, 0, 0, 0, Y
22, M, 12, N, Y, N, 1, 1, 0, 1, 0, 0, 1, 1, Y
31, F, 1, N, N, Y, 1, 0, 0, 1, 0, 0, 0, 0, N
37, F, 5, M, Y, Y, 1, 1, 1, 0, 0, 1, 1, 1, Y
78, M, 2, M, Y, N, 1, 1, 0, 1, 1, 0, 0, 1, Y
62, M, 8, N, Y, Y, 1, 0, 0, 0, 0, 0, 1, 1, 1, Y
```

Figure 9. Feature vector

6.5. Classification

The training set is supplied as input to 6 classifiers. Classification analysis was performed on the classifiers. The steps involved in this analysis are:

- Import the weka and java packages
- Call function useClassifier with the data to be classified as parameter
- Create the classifier object
- Build the classifier model
- Save the model
- Create an Evaluation object
- Cross validate using 10 fold cross validation
- Print the confusion matrix

The results of the analysis are discussed in the Results and Discussions section of the paper.

6.5.1. Prediction on Test Set

The test set contains the samples that aren't known to the classification model yet. The saved model is then evaluated on the test set and the accuracy is obtained.

6.5.2. Prediction on Unlabeled Dataset

Unlabeled dataset is fed to the saved model. The disease label is a "?" in this case. The model then predicts the labels for these samples.

6.5.3. Graphical User Interface

A GUI was developed to simplify access to the dengue detection system. Separate panels, one for researchers and another for common users were developed. Researchers can upload a folder consisting of discharge summaries which will be used as the training set. Common users can indicate which symptoms they are experiencing and get a prediction from the system.

The screenshot shows a window titled 'Patient GUI' with two tabs: 'Researchers' and 'Patients'. The 'Patients' tab is active. Under the heading 'Symptoms:', there are two columns of input fields. The first column includes: Age (text box with '45'), Gender (radio buttons for Male and Female, with Female selected), Medical history (radio buttons for Yes and No, with Yes selected), Fever (radio buttons for Yes and No, with Yes selected), Nausea (radio buttons for Yes and No, with Yes selected), Pain behind the eye (radio buttons for Yes and No, with No selected), and Fatigue (radio buttons for Yes and No, with Yes selected). The second column includes: Admission month (text box with '7'), Marital status (radio buttons for Married and Not married, with Married selected), Family history (radio buttons for Yes and No, with No selected), Headache (radio buttons for Yes and No, with Yes selected), Vomiting (radio buttons for Yes and No, with No selected), Muscle pain (radio buttons for Yes and No, with No selected), and Skin rash (radio buttons for Yes and No, with No selected). Below the symptoms is a 'Get Prediction' button. Under the heading 'Evaluations:', there is a text box containing the text 'PREDICTED POSITIVE FOR DENGUE'.

Figure 10. Patient GUI

The screenshot shows a window titled 'Researcher GUI' with two tabs: 'Researchers' and 'Patients'. The 'Researchers' tab is active. Under the heading 'Enter the training set file path:', there is a text box containing the path 'ISecond_Review\input_output\Dischargesummary' and an 'Attach' button. Below this is a 'Start' button. Under the heading 'Evaluations:', there is a scrollable text box containing the following text:

```

1. Meta-classifier
Evaluation
Correctly Classified Instances  91    91 %
Incorrectly Classified Instances  9    9 %
Kappa statistic                0.7634

```

Figure 11. Researcher GUI

6.6. Frequency Analysis

To perform frequency analysis, we have used bar charts. The bar charts are generated using JFreeCharts. The correlations of the spread of the symptoms and in turn the disease over the months are reported briefly to give a clear picture to the researchers. This feature is only available to the researchers.

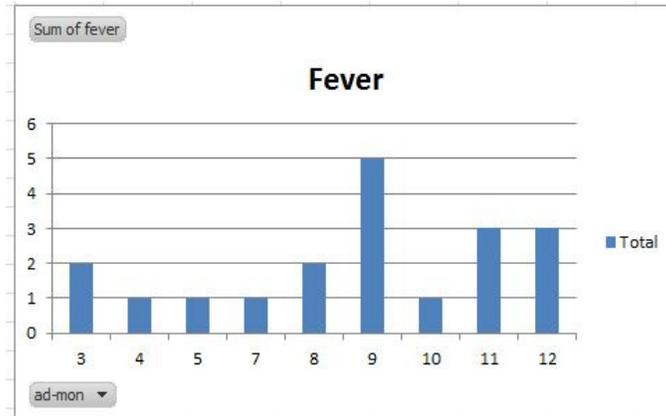


Figure 12. Fever vs month

7. RESULTS AND DISCUSSIONS

The feature vector is supplied to various supervised learning algorithms and classifier models are generated. LibSVM is integrated software for support vector classification, regression and distribution estimation. It supports multi-class classification. Logistic regression classifier uses a sigmoid function to perform the classification. Multilayer perceptron is a classifier based on Artificial Neural Networks. Each layer is completely connected to the next layer in the network. Naïve Bayes methods are a set of supervised learning methods based on applying Bayes theorem with the naïve assumption of independence between every pair of features. The Sequential Minimal Optimizer uses John Platt's sequential minimal optimization algorithm for training a support vector classifier. It also normalizes all attributes by default. The Simple Logistic Classifier is used for building linear logistic regression models. These classifiers are subject to two types of classifications – 10-fold cross-validation and percentage split (2/3rd training and 1/3rd test). Accuracies obtained from the 2 methods are compared. In addition, accuracy of the various classifiers are analyzed based on five performance metrics (Accuracy, Kappa statistics, Mean absolute error, Root mean squared error, Relative absolute error) [16] and the best model is chosen.

- **Accuracy:** The number of samples that are correctly classified from the given 100 input samples.
- **Kappa Statistic:** The Kappa Statistic can be defined as measuring degree of agreement between two sets of categorized data. Kappa result varies between 0 to 1 intervals. Higher the value of Kappa means stronger the agreement/ bonding. If Kappa = 1, then there is perfect agreement. If Kappa = 0, then there is no agreement. If values of Kappa statistics are varying in the range of 0.40 to 0.59 considered as moderate, 0.60 to 0.79 considered as substantial, and above 0.80 considered as outstanding.

- Mean Absolute Error: Mean absolute error can be defined as sum of absolute errors divided by number of predictions. It is measure set of predicted value to actual value i.e. how close a predicted model to actual model. The lower the value of MAE the better the classification.
- Root Mean Squared Error : Root mean square error is defined as square root of sum of squares error divided by number of predictions. It is measure the differences between values predicted by a model and the values actually observed. Small value of RMSE means better accuracy of model. Lower the value of RMSE, better the prediction and accuracy.
- Relative Absolute Error: Relative error is the ratio of the absolute error of the measurement to the accepted measurement. A lower percentage indicated better prediction and accuracy.

Classifier Analysis					
Classifier	Accuracy	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error
LibSVM	73%	0.0927	0.27	0.5196	68.0175%
Logistic	85%	0.6239	0.1439	0.3735	36.2592%
MultiLayer Perceptron	80%	0.3812	0.2412	0.3806	64.8464%
Naive Bayes'	83%	0.542	0.2261	0.342	56.9706%
SMO	91%	0.7634	0.09	0.3	22.6725%
Simple Logistic	86%	0.6449	0.2004	0.3197	50.4718%

Figure 13. Classifier analysis using 10-fold cross validation

Based on the above analysis, SMO is identified to be the most optimal classifier.

7.1 Analysis and correlation

The predicted results are visualized in graphical form subsequent to prediction. Counts of occurrences of various symptoms over the months are depicted using bar charts, and these values are compared with the graphs generated for the original training dataset. The month with maximum manifestation of all symptoms was found to be September. This was also the month with maximum cases of dengue, according to the prediction. This inference was also corroborated by the graph generated from the initial training dataset, and we gather from these graphs that August, September and October are the months most vulnerable to dengue.

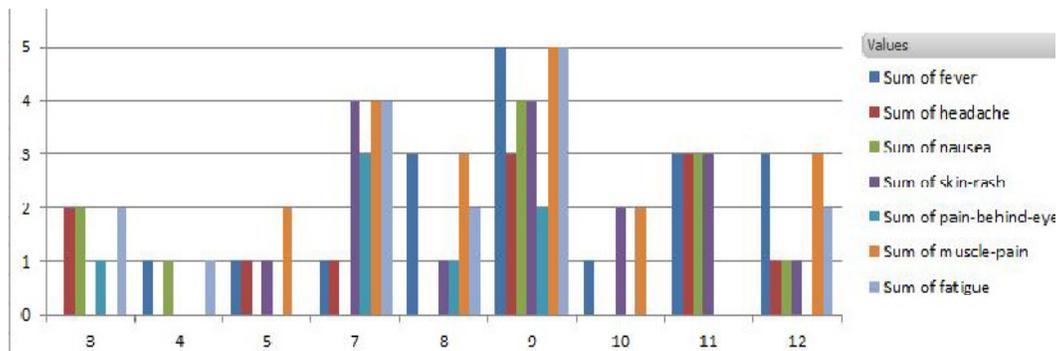


Figure 14. Overview of all symptoms spread over the months

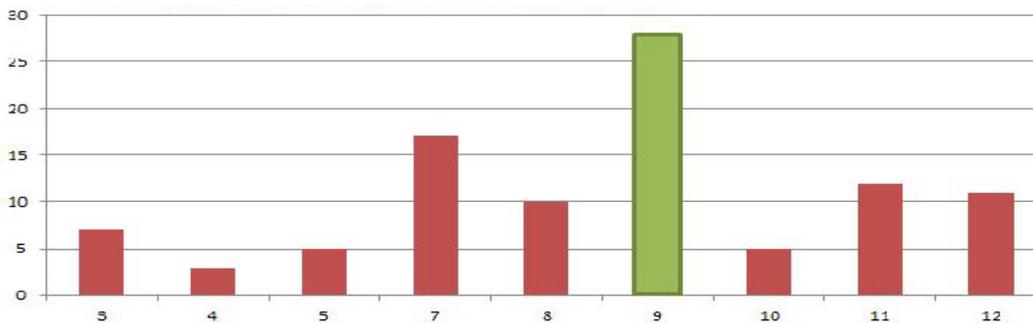


Figure 15. Occurrences of all symptoms over the months

8. CONCLUSION

To conclude, we have discussed, in this report, the detailed design and related algorithms for a system to identify disorder mentions from clinical text and correlate its frequency with the time frame. The annotated discharge summaries are tagged and feature extraction algorithms are used to obtain the features relevant to the disease, Dengue. This is followed by the generation of a feature vector (Binary representation). This vector is then used to train and build various classification models and SMO is found to produce the best results. The model generated further aids in the prediction of the disease. Bar graphs are then used to succinctly represent this correlation. Additionally the correlation of training samples with time frame was compared with the correlation obtained from predicted results and the disease occurrence was found to concentrate in the months of August, September and October in both the cases.

9. LIMITATIONS

Our system uses only 15 features. Extracting more features might increase the accuracy of the model. The feature vector is depicted using the binary representation. Using the frequency value representation might improve overall classification.

10. FUTURE WORK

As a part of our future work, we intend to write an implementation to produce bag of words and extract more features to produce an extensive analysis. Further, we also intend to implement

tagging of the discharge summaries using BIOS tagging [5]. Whenever hospitals receive new samples showing a tendency for dengue, those samples must be integrated with the existing training set. This was, the training and predictive capacity of the model will grow, possibly giving better results in the future. To provide and up-to date analysis, we could extend the project to be used as a desktop app or browser plugin which will automatically synchronize with new data received from the hospitals' end.

REFERENCES

- [1] Sameer Pradhan, Noemie Elhadad, Wendy Chapman, Suresh Manandhar & Guergana Savova, (July 2014) "Analysis of Clinical Text", SemEval-2014 Task 7.
- [2] Melinda Katona & Richard Farkas, (June 2014) "SZTE-NLP: Clinical Text Analysis with Named Entity Recognition", SemEval-2014.
- [3] Koldo Gojenola, Maite Oronoz, Alicia Perez & Arantza Casillas, (December 2014) "IxaMed: Applying Freeling and a Perceptron Sequential Tagger at the Shared Task on Analyzing Clinical Texts", SemEval-2014.
- [4] Parth Pathak, Pinal Patel, Vishal Panchal, Narayan Choudhary, Amrith Patel & Gautam Joshi, (July 2014) "ezDI: A Hybrid CRF and SVM based Model for Detecting and Encoding Disorder Mentions in Clinical Notes", SemEval-2014.
- [5] Oana Frunza, Diana Inkpen & Thomas Tran, (June 2011) "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts", IEEE transactions on knowledge and data engineering, vol. 23, Issue no. 6.
- [6] Deepali Chandna, (2014) "Diagnosis of Heart Disease Using Data Mining Algorithm", International Journal of Computer Science and Information Technologies, Vol. 5 (2), pp1678-1680.
- [7] Jyoti Soni, Ujma Ansari, Dipesh Sharma & Sunita Soni, (March 2011) "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications, vol 17.
- [8] Smitha T & Dr.V Sundaram, (2012) "Knowledge Discovery from Real Time Database using Data Mining Technique", International Journal of Scientific and Research Publications, Volume 2, Issue 4.
- [9] M.A.Nishara Banu & B Gomathy, (Nov-Dec 2013) "Disease Predicting System Using Data Mining Techniques", International Journal of Technical Research and Applications, Volume 1, Issue 5, PP. 41-45
- [10] Devendra Ratnaparkhi, Tushar Mahajan & Vishal Jadhav, (November 2015) "Heart Disease Prediction System Using Data Mining Technique", International Research Journal of Engineering and Technology, Volume: 02 Issue: 08.
- [11] I.S.Jenzi, P.Priyanka & Dr.P.Alli, (March 2013) "A Reliable Classifier Model Using Data Mining Approach for Heart Disease Prediction", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3.
- [12] N. G. Bhuvaneshwari Amma, (February 2012) "Cardiovascular Disease Prediction System using Genetic algorithm and Neural Network", International Conference on Computing, Communication and Applications, IEEE, pp1-5.

- [13] A.Shameem Fathima & D.Manimeglai, (March 2012) “Predictive Analysis for the Arbovirus-Dengue using SVM Classification”, International Journal of Engineering and Technology, Volume 2 No. 3
- [14] Daranee Thitipryoonwongse, Prapat Suriyaphol & Nuanwan Soonthornphisaj, (2012) “A Data Mining Framework for Building Dengue Infection Disease Model”, 26th Annual Conference of the Japanese Society for Artificial Intelligence
- [15] N.Subitha & Dr.A.Padmapriya, (August 2013) “Diagnosis for Dengue Fever Using Spatial Data Mining”, International Journal of Computer Trends and Technology, Volume 4 Issue 8
- [16] Yugal kumar & G. Sahoo, (July 2012) “Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA”, I.J. Information Technology and Computer Science, Volume 7, pp43-49.

AUTHORS

Nandini V is currently pursuing her final year, Computer Science and Engineering in SSN College of Engineering. She has published a paper on Machine Vision in the ARPN Journal of Engineering and Applied Sciences. Her research interests include Artificial Intelligence, Robotics, Machine Learning, Machine Vision and Data Mining.



Sriranjitha R is currently pursuing her final year, Computer Science and Engineering in SSN College of Engineering. She is a member of CSI (Computer Society of India). Her research interests include Machine Learning, Artificial Intelligence, Data Mining and Data Structures.



Yazhini T P is currently pursuing her final year, Computer Science and Engineering in SSN College of Engineering. Her research interests include Computer Networks, Data Mining, Artificial Intelligence and Web Technology.



INTENTIONAL BLANK

OBIA ON COASTAL LANDFORM BASED ON STRUCTURE TENSOR

Sun Shuting, Liu Jianqiang and Zou Bin

Key laboratory of Space Ocean Remote Sensing and Application,
SOA, Beijing, China

National Satellite Ocean Application Service, Beijing, China

sst19910323@gmail.com

jqliu@mail.nsoas.org.cn

zoubin@mail.nsoas.org.cn

ABSTRACT

This paper presents the OBIA method based on structure tensor to identify complex coastal landforms. That is, develop Hessian matrix by Gabor filtering and calculate multiscale structure tensor. Extract edge information of image from the trace of structure tensor and conduct watershed segment of the image. Then, develop texons and create texon histogram. Finally, obtain the final results by means of maximum likelihood classification with KL divergence as the similarity measurement. The study findings show that structure tensor could obtain multiscale and all-direction information with small data redundancy. Moreover, the method described in the current paper has high classification accuracy.

KEYWORDS

OBIA, Structure Tensor, Texon & Coastal Zone

1. INTRODUCTION

Remote sensing with its large amount of information, access to information faster, shorter cycle, less restricted and so on provides a new method for the investigation of the coastal zone. In view of the characteristics of the coastal zone, the methods of land use classification are proposed. The original classification method of the coastal zone is based on pixel-based classification, such as Maximum Likelihood or ISODATA. These methods are suitable for low and medium spatial resolution remote sensing image; as for high-resolution images, information, such as texture information single pixel able to contain has its limitation. Thus, object oriented classification is widely used. That method firstly segment image to object, which is a series of homogeneous regions adjacent to each other. After that, classify these objects.

In recent years, some scholars have applied the object-oriented classification to the classification of coastal landform. In the year 2009, Wang Changying[1] has proposed a coastline extraction methods based on Apriori Algorithm, which can extract 5 different kind of coastline: artificial coastline, bedrock coastline, sandy coastline, muddy coastline and biological coastline. Using such method, the extraction accuracy can reach 3 pixels by using Landsat image(30m).

Proposed by Bigün[2] in 1987, structure tensor indicates the consistency of the matrix which was originally used for detection of linear symmetry. In recent years, structure tensor method is improved from 2 aspects:

In practice, the distance between objects and the viewpoint is different, in addition, objects have specific spatial scale. Therefore, single scale structure tensor is unable to describe texture information of different scales. Thus, scientists expand single scale structure tensor to multiscale structure tensor to enhance the capacity in describing image information, Sensitive to noise, structure tensor usually requires filtering. The most common filter is the Gaussian filter. Nonlinear structure tensor was proposed by M Rousson et al [3] in 2003, that is, Gaussian filter is replaced by nonlinear filter, which could preserve more information and restrain noise at the same time.

Multiscale structure tensor is able to extract multiscale and all-direction information with small data redundancy. In the past 30 years, it has been successfully applied in many fields such as image recovery [4], frequency and ridge estimation [5], texture analysis [6] and image processing in neural biology field [7]. In this paper, in order to develop multiscale structure tensor, Hessian matrix is constructed using Gabor wavelet and nonlinear diffusion filter is used to reduce the noise. Multiscale structure tensor serves as texture information about coastal areas.

It is difficult to directly compare the similarity of 2 objects due to different amount of pixels in different objects. In this paper, texton histogram is used to describe the information of the objects.

Texton is the fundamental microstructure of natural image. [8] A plurality of similar texture and spectral features in the same texture area can compose such microstructure. These measurement vectors have certain similarity. Typical vectors can be used to represent all similar vectors in the region and discard the others. Which can reduce amount of computation and to multivariable decoupling in order to reduce the correlation of vectors. The texton histogram is a typical application of this theory. In this method, New measurement vectors are obtained by clustering the old measurement vectors.

2. STUDY AREA AND RESEARCH DATA

The study area is located in the southeast of Hainan Island. As it's shown in figure 1.



Figure 1. Location of Study Area

The research data is the image of WorldView2, a high-resolution commercial remote sensing satellite launched by Globe Digital Company in October 2009. WorldView2 contains 8 multi spectral bands and 1 Panchromatic band. The Panchromatic resolution is 0.46m, and multi spectral resolution is 1.84 m. In this paper, the image product has 4 bands (RGB and NIR) with resolution at 2 m (after image fusion).

3. TECHNOLOGICAL PROCESS

The technical process is shown in Figure 2.

First, the satellite image is filtered by Gabor to develop Hessian matrix, employ nonlinear diffusion filtering to develop structure tensor. Then the edge information of the image is obtained by analyzing the trace and conduct watershed segment. After that, the author utilizes the trace and eigenvalues of structure tensor, structure tensor and spectral information to develop textons and calculate texton histogram. Finally, the author carries out the maximum likelihood classification based on KL divergence to get the final classification result.

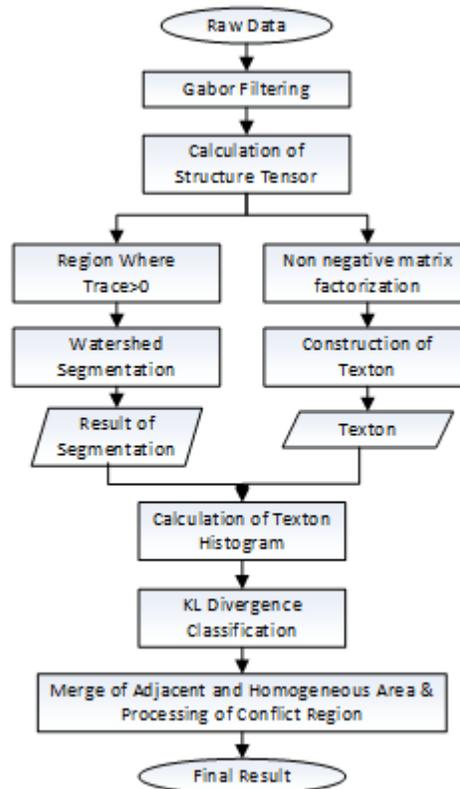


Figure 2. Technology Process

4. STRUCTURE TENSOR

For grey scale image I , the structure tensor is shown as Formula (4.1):

$$T = k * \begin{vmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{vmatrix} \quad (4.1)$$

Where I_x is the partial derivative of the image in x direction and I_y is the partial derivative in y direction. It is easy to extend Formula (4.1) to higher order as shown in Formula (4.2):

$$T = k * \begin{vmatrix} \sum(I_x^2) & \sum(I_x I_y) \\ \sum(I_x I_y) & \sum(I_y^2) \end{vmatrix} \quad (4.2)$$

For multi spectral data, the structure tensor is the sum of the tensor product of each band.

Scheunders[9] has proposed the multiscale feature description method based on wavelet and the wavelet basis function γ satisfies the Formula (4.3):

$$\begin{cases} \lim_{x,y \rightarrow \infty} \gamma(x,y) = 0 \\ \int \gamma(x,y) dx dy = 1 \end{cases} \quad (4.3)$$

Where γ is a kind of two dimensional differentiable function. Therefore, wavelet basis function can be used as multi scale convolution kernel. The elements in a Hessian matrix are obtained by convolving the image and calculating the partial derivative.

$$\begin{pmatrix} D_x \\ D_y \end{pmatrix} = a^s \begin{pmatrix} \frac{\partial(I*\gamma_s)}{\partial x} \\ \frac{\partial(I*\gamma_s)}{\partial y} \end{pmatrix} \quad (4.4)$$

Where a is the wavelet base (constant). γ_s is the wavelet basis function in scale s . Gabor wavelet is used as wavelet basis function based on the method of Shoudong Han et al [7].

4.1. Gabor Filters

Gabor filters is to divided signal into many small time intervals. Using Fourier transform to analyse each time intervals. The specific process is to add a Gaussian function to $f(t)$ as window function and then doing Fourier transform. As for the two-dimensional data, its window function is as follows [10]:

$$g(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\left(\frac{x+y}{\sigma}\right)^2} \quad (4.5)$$

Put it into Gabor transformation function:

$$G_f = \int_{-\infty}^{\infty} f(t)g(t-b)e^{-i\omega t} dt \quad (4.6)$$

Obtain Gabor basis function of two-dimensional discrete data:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\left(\frac{x+y}{\sigma}\right)^2} \cos(2\pi f(x\cos\theta + y\sin\theta)) \quad (4.7)$$

f is the frequency, determined by formula (5.4); θ is the angle, make it as $0^\circ, 45^\circ, 90^\circ, 135^\circ$

$$f = \frac{1}{2^n \cdot 2\sqrt{2}} \quad (n = 0,1,2) \quad (4.8)$$

Therefore, the calculation of the multiscale structure tensor is as Formula (4.9):

$$T = k * \begin{vmatrix} \sum(D_x^2) & \sum(D_x D_y) \\ \sum(D_x D_y) & \sum(D_y^2) \end{vmatrix} \quad (4.9)$$

4.2. Nonlinear Diffusion Filter

PM diffusion is a nonlinear diffusion model proposed by Perona and Malik [11] in 1990. It is a kind of nonlinear adaptive PDE algorithm. The calculation process contains detection of edge and direction of the image. In non-edge area, the filter is isotropic of high smooth degree. In edge area, the filter is anisotropic with low smooth degree. The application of it to images could reduce noise and maintain or even enhance image characteristics at the same time.

Discrete nonlinear diffusion PM equation is shown in Formula (4.10):

$$\partial T(x, y) = \text{div}(g(\sum_{i=1}^2 \sum_{j=1}^2 |\nabla T(i, j)|^2) \nabla T(x, y)) \quad (4.10)$$

Where $T(x,y)$ is the structure tensor of pixel in the image with a coordinates of (x, y) , g is the edge detection function shown in Formula (4.11):

$$g(|\nabla x|) = \frac{1}{|\nabla x|^{\varepsilon+\delta}} \quad (4.11)$$

Where ε is a positive constant to prevent the denominator as 0; and δ is the harmonic parameter of diffusion filter. This paper takes δ as 0.7.

5. IMAGE SEGMENTATION

In this paper, watershed segmentation is applied, which is a region-based mathematical morphology segmentation method. It could obtain single pixel width, connection, enclosure and accurate outline.

Watershed algorithm can only segment single band image. In case of processing multi-bands image, certain algorithm should be used to reduce the dimension.

Structure tensor with higher order (4.2) can effectively process multi-band data. In addition, the Hessian matrix is developed by gradient, which contains the structure information of local neighbourhood areas without mutual offset of the gradients of rising edge and falling edge.

The trace of structure tensor of any pixel (x,y) of an image is calculated by Formula (5.1):

$$\text{trace}(x, y) = \text{trace} \left(\begin{vmatrix} T(x, y) & T(x, y + n) \\ T(x + m, y) & T(x + m, y + n) \end{vmatrix} \right) \quad (5.1)$$

Where m and n is the length and width respectively of the image. In this paper, the part of trace >0 is selected as the gradient image.

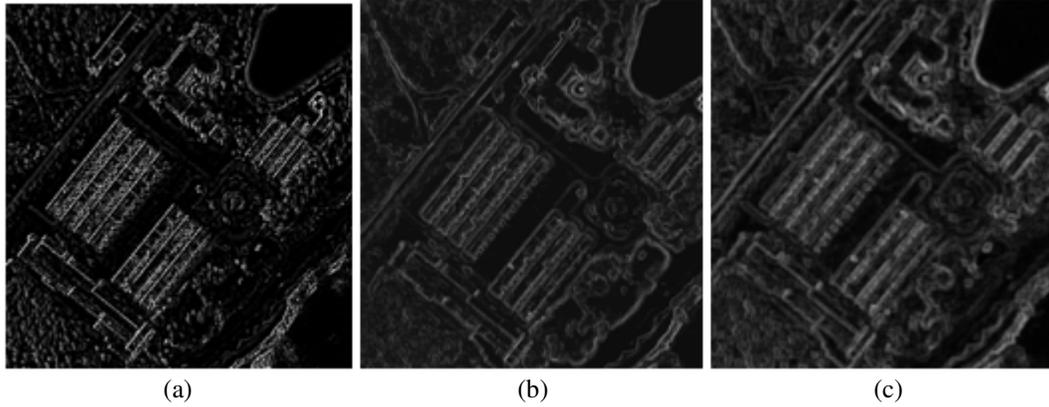


Figure 3. Comparison of the Results of CIE Transformation

In Figure 3, a is the gradient obtained by the method in this paper; b is multidimensional gradient and c is morphological gradient. It can be seen that the trace of structure tensor can more effectively use the information of each band and obtain more accurate detection of the edge.

6. CONSTRUCTION OF TEXTON

OBIA requires describing objects with complex spectral and texture information and of different size after image segmentation. In this paper, the author develops texton to describe the information of object. Firstly, texture operator is selected; 4 bands spectral data is combined to develop textons. Then, the non-negative matrix factorization is employed to reduce the dimension.

6.1 Texture Feature

Texture operator selected is shown in Table 1. A total of 20 texture feature vectors are chosen as texture information of the object. Among them, the texture operator based on eigenvalues of the structure tensor is estimated by Formula (6.1)

$$E(x, y) = \frac{1}{1+B(\lambda_1-\lambda_2)^2} \quad (6.1)$$

Where B is a constant, taking 1 in the current paper. λ is the eigenvalue of structure tensor at the point (x, y) .

Table 1. List of Texture Feature

Texture Name	Dimension	Notification
Structure Tensor	12	3 scales, 4 directions a total of 12 dimensions
Trace	3	As formula(5.1), 3 scales a total of 3 dimensions
Eigenvalue	3	As formula(6.1), 3 scales a total of 3 dimensions
NDWI	1	
ARVI	1	

6.2 Non negative matrix factorization

In this paper, texture is reduced from 20 bands to 6-dimensions using non negative matrix factorization. Reducing computation load and reducing the correlation between bands. In Figure

4, a shows the clustering results by using this method, b shows the result without using nnmf, c shows the clustering result by only using Spectral information.

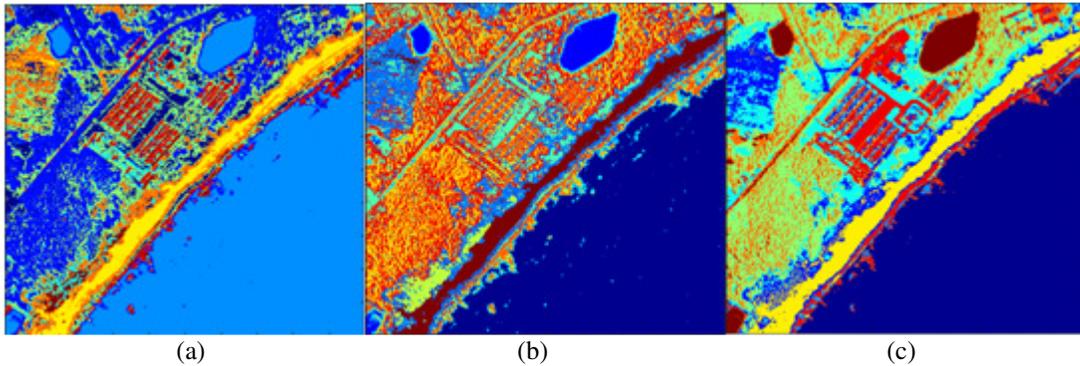


Figure 4. Results of Clustering of the Texton

The comparison of the results of the above three methods shows that texton can effectively identify most objects and is an effective method. NNMF not only reduces computation load, but also reduces the correlation between bands. It maintains more information without reducing classification accuracy after dimensionality reduction.

At a result, a 10 dimensional texton is obtained.

7. CLASSIFICATION OF THE OBJECTS

The author divides the images into the following 7 categories: seawater, land water area, forest land, farm land, building, bare land and coastal (tidal zone) coastal (intertidal zone). based on characteristics of the image and Szuster B W's [12] method.

The vegetation is classified based on density degree in line with Ecker's standards [13] as well as the Ethnography of Hainan. Forest in the study area in the south-eastern end of Hainan Island is the forest land and other terrestrial vegetation is farm land. The coastal type is the muddy coast, which can be further divided into tidal and intertidal zones.

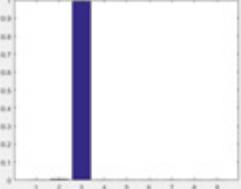
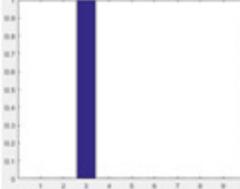
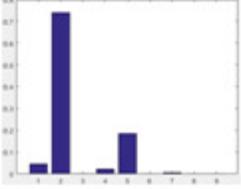
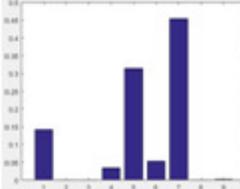
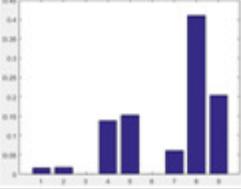
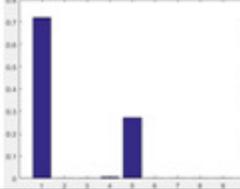
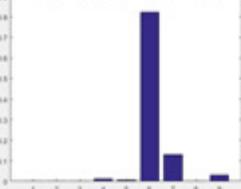
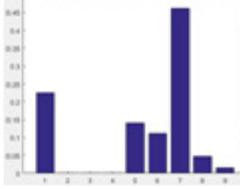
7.1. Texton Histogram of the Training Sample

With the calculation in Chapter 6, the author employs 10 dimension textons to describe the characteristics of the object. However, the author cannot directly compare two objects due to different amount of textons (i.e. pixels) in different objects. Therefore, Texton histogram is used in the current paper to describe the information of object so as to compare the similarity of two objects with different number of pixels. The process of generating texton histogram is as the followings:

- a) Classify these textons into 9 groups by ISODATA and record the group of each texton.
- b) Obtain texton histogram by count the frequency of the group of all textons in an object.
- c) Turns different histogram in the same dimension by histogram normalization.

Texton histogram of each class is shown in Table 2.

Table 2. Texton Histogram of the Training Sample

Class	Texton Histogram	Class	Texton Histogram
Seawater		Land Water Area	
Forest Land		Farm land	
Buildings		Bare land	
Tidal		Intertidal	

7.2. Maximum Likelihood Classification

The maximum likelihood is used in the current paper to classify the objects. Texton histogram reflects the distribution of probability of various kinds of textons because it is the probability statistics of texton types contained in the object. Therefore, this paper employs KL divergence as similarity measurement as shown in the following Formula:

$$D(p||q) = \sum p(x) \log \frac{p(x)}{q(x)+a} \quad (7.1)$$

Where p and q are the probability distribution of x. In this paper, p is the texton histogram of subject sample and q is the texton histogram of control sample. To avoid denominator is 0, constant a is added. This paper takes 0.01 as a value. The classification of an object can be identified until the KL divergence is the minimum among them and it's smaller than the threshold value t; and t is considered as 0.1.

7.3. Processing of Classification Result

Watershed algorithm cannot avoid the over-segmentation, so the phenomenon exists that 2 areas adjacent to each other are of the same class. The border width of Watershed algorithm is fixed as 1. Kang, T. Yung's [14] Bridge algorithm can eliminate such boundary. Bridge algorithm is defined as a connection of connected points. In the binary image, for a point whose value is 0, if there is more than 2 points with value of 1 as its neighbourhood, value of the point is 1. As it's shown in Formula (7.2)

$$\begin{array}{ccc}
 1 & 0 & 0 \\
 1 & 0 & 1 \\
 0 & 0 & 1
 \end{array}
 \text{Change into}
 \begin{array}{ccc}
 1 & 0 & 0 \\
 1 & 1 & 1 \\
 0 & 0 & 1
 \end{array}
 \quad (7.2)$$

Texture histogram of seawater and land water area is similar, water areas are combined after extraction, mainly, the largest area is seawater.

8. RESULT OF THE EXPERIMENT AND ASSESSMENT OF THE ACCURACY

The study area is shown in Figure 6. The image is taken by Worldview2 on June 27th, 2011, located at 18°46'39"N, 110°29'45"E with the size at 1 000×1 000 pixels.



Figure 5 Original Image of the Study Area

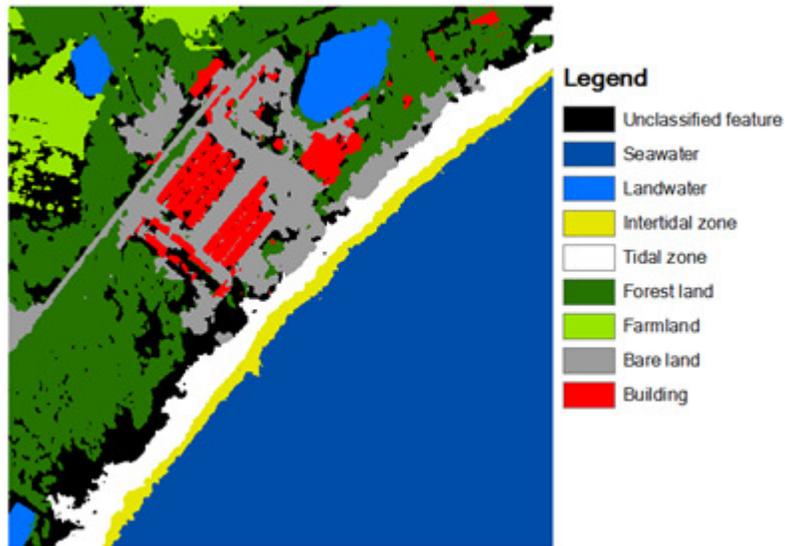


Figure 6. the Result of Classification

Table 3. Confusion Matrix

Landform	Method in this Paper		Maximum Likelihood		Method Without NIR	
	PA(%)	UA(%)	PA(%)	UA(%)	PA(%)	UA(%)
Seawater	100	100	90.5	95	90.9	100
Land water	100	100	100	95	100	95
Forest Land	90.0	90	80.0	80	81.8	90
Farm Land	84.2	80	73.7	70	75	75
Bare Land	73.9	85	62.5	75	65	65
Building	90.0	90	85.7	90	80	80
Tidal	89.5	85	76.2	80	77.3	85
Intertidal	94.7	90	93.3	70	88.2	75
OA(%)	90.0		81.9		83.1	
Kappa	0.886		0.793		0.807	

In table 3, first column shows the result of method in this paper, second column is the result using Maximum Likelihood on pixel. Third column shows the result without NIR data using structure tensor method (without NDWI and ARVI texture). Training sample of all methods are of the same. It is found by comparison of the results of the above three methods:

Firstly, overall accuracy of Maximum Likelihood is higher than 80%, similar to structure tensor method without NIR data. With NIR information, quality of the texture has been significantly improved. Besides, all methods have a high accuracy among landform with obvious characteristics like seawater which represent spectral information of the image is fundamental.

Secondly, method of this paper have an obvious improvement on classification of bare land and farmland and that of method without NIR data is passable. Therefore, structure tensor is an effective texture algorithm on representation of image local information.

Thirdly, overall accuracy of the method described in this paper can reach 90.0%, 8.9% higher than that of the maximum likelihood method with 0.093 rise of Kappa coefficient. In conclusion, method of this paper has high classification accuracy.

8. CONCLUSIONS

The findings of the current study demonstrate that structure tensor is an excellent texture operator because of its capacity in obtaining all-direction and multiscale information with small data redundancy. The overall accuracy of the OBIA method based on structure tensor adopted in this paper is above 90% and this method can be effectively applied in classifying coastal landforms.

Meanwhile, the author finds that there is quite big room for improvement of the current method as structure tensor texture can be further optimized by energy function of wavelet.

FUNDING

Project supported by Welfare Special Industry Fund of Marine Public Welfare, SOA of China (Grant No. 201505014).

REFERENCES

- [1] 王常颖. 基于数据挖掘的遥感影像海岸带地物分类方法研究 [D]. 青岛: 中国海洋大学, 2009.
- [2] Bigün J, Granlund G H, Wiklund J. Multidimensional orientation estimation with applications to texture analysis and optical flow[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1991 (8): 775-790.
- [3] Rousson M, Brox T, Deriche R. Active unsupervised texture segmentation on a diffusion based feature space[C]//*Computer vision and pattern recognition, 2003. Proceedings. 2003 IEEE computer society conference on. IEEE, 2003, 2: II-699-704 vol. 2.*
- [4] Chierchia G, Pustelnik N, Pesquet-Popescu B, et al. A nonlocal structure tensor-based approach for multicomponent image recovery problems[J]. *Image Processing, IEEE Transactions on*, 2014, 23(12): 5531-5544.
- [5] Mikaelyan A, Bigun J. Frequency and ridge estimation using structure tensor[C]//*Biometric Technologies in Forensic Science (BTFS 2013)*, Nijmegen, Netherlands, October 14-15, 2013. Radboud University Nijmegen, 2013: 58-59.
- [6] Han S, Wang X. Texture Segmentation Using Graph Cuts in Spectral Decomposition Based Riemannian Multi-Scale Nonlinear Structure Tensor Space[J]. *International Journal of Computer Theory and Engineering*, 2015, 7(4): 259.
- [7] Budde M D, Frank J A. Examining brain microstructure using structure tensor analysis of histological sections[J]. *Neuroimage*, 2012, 63(1): 1-10.
- [8] Tuceryan M, Jain A K. Texture analysis[J]. *The handbook of pattern recognition and computer vision*, 1998, 2: 207-248.
- [9] Scheunders P. A multivalued image wavelet representation based on multiscale fundamental forms[J]. *IEEE Transactions on Image Processing*, 2002, 11(5):568-575.

- [10] Lee T S. Image representation using 2D Gabor wavelets[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1996, 18(10): 959-971.
- [11] Lindeberg T. Scale-space theory in computer vision[M]. Springer Science & Business Media, 2013.
- [12] Szuster B W, Chen Q, Borger M. A comparison of classification techniques to support land cover and land use analysis in tropical coastal zones[J]. Applied Geography, 2011, 31(2): 525-532.
- [13] Eckert S. Improved Forest Biomass and Carbon Estimations Using Texture Measures from WorldView-2 Satellite Data[J]. Remote Sensing, 2012, 4(4):810-829.
- [14] Kong T Y, Rosenfeld A. Digital topology: introduction and survey[J]. Computer Vision, Graphics, and Image Processing, 1989, 48(3): 357-393.

AUTHORS

Sun Shuting (1991.3~)

Master student of National Satellite Ocean Application Service, China.

Research direction: OBIA and Antarctic Sea Ice.



Liu Jianqiang (1964.9~)

Research Fellow,

Deputy director of National Satellite Ocean Application Service, China.

Research direction: Construction of Ocean Satellite and Ocean Remote Sensing



Zou Bin (1969.2~)

Research Fellow of National Satellite Ocean Application Service, China.

Research direction: Ocean Remote Sensing



CONCEALED DATA AGGREGATION WITH DYNAMIC INTRUSION DETECTION SYSTEM TO REMOVE VULNERABILITIES IN WIRELESS SENSOR NETWORKS

Bharat Bhushan¹, Keshav Kaushik² and G Sahoo³

^{1,3}Department of Computer Engineering, BIT Mesra, Ranchi, India

Bharat_bhushan1989@yahoo.com

gsahoo@bitmesra.ac.in

²Department of Computer Engineering, HMRITM, New Delhi, India

Keshavkaushik96@gmail.com

ABSTRACT

Data Aggregation is a vital aspect in WSNs (Wireless Sensor Networks) and this is because it reduces the quantity of data to be transmitted over the complex network. In earlier studies authors used homomorphic encryption properties for concealing statement during aggregation such that encrypted data can be aggregated algebraically without decrypting them. These schemes are not applicable for multi applications which lead to proposal of Concealed Data Aggregation for Multi Applications (CDAMA). It is designed for multi applications, as it provides secure counting ability. In wireless sensor networks SN are unarmed and are susceptible to attacks. Considering the defence aspect of wireless environment we have used DYDOG (Dynamic Intrusion Detection Protocol Model) and a customized key generation procedure that uses Digital Signatures and also Two Fish Algorithms along with CDAMA for augmentation of security and throughput. To prove our proposed scheme's robustness and effectiveness, we conducted the simulations, inclusive analysis and comparisons at the ending.

KEYWORDS

Concealed Data Aggregation, Digital Signatures, DYDOG, Wireless Sensor Networks.

1. INTRODUCTION

WIRELESS sensor networks (WSNs) have gained much significance in past few years because of its huge number of applications and areas of use. The application domain ranges from military investigations to environment monitoring and ecological monitoring. The sensor networks generally comprises of several sensor nodes gathered from deployed environments in a large scale [1]. Sensor nodes in sensor networks face a major problem as sensor nodes are energy constrained and these have limited power, storage, communication, and processing capabilities. Thus the major problem in wireless sensor network is energy consumption. Thus to conserve energy and power sensor networks brings forth the concept of data aggregation [2]. This means

converting many values sensed from different environments into one single value and aggregated value is computed at sink by the use of some mathematical functions [3]. The technique for aggregation is used mainly for the reduction in amount of data to be sent in the sensor environments. As a result of reduction of amount of data communicated within WSNs, there is energy conservation of battery [4]. Sensor nodes send their readings to a special type of node for performing aggregation of data i.e., aggregators, that sends only the condensed or aggregated reading further [5]-[6]. These aggregators may be some kind of special nodes or normal sensor nodes also.

Sensor nodes requires high security as it prompts many security issues like confidentiality, data integrity, data authentication, key management, etc. High security is required in wireless sensor networks so it is one of the most popular research topics and much advancement have been reported on in recent years[7]. In this paper we mainly focus on security aspect of data transmission in WSNs and we propose a method of secure transmission of encrypted data across sensor nodes in sensing environments as well as secure key generation methods involved in attack detection and prevention in wireless sensor networks.

Encryption of data being transmitted in WSNs is necessary as this type of sensors can be subject to many different types of attacks. The attacker can either listen secretly the data being transmitted in WSNs (attacker may deduce the secret key) or send forged or duplicated data to sensor nodes, aggregators or base station (attacker may send forged data to cheat BS without knowing the secret key) or even compromise secrets of components of WSNs by capturing them. so as encryption is necessary sensor nodes must encrypt data on hop-by-hop basis. [8]-[9]-[10]-[11]. The mechanism of key generation involves an overhead activity making this an expensive and complicated operation [13]-[14]. Different key generation schemes have been proposed but they involve high computations for encryption of data and require more CPU, bandwidth and memory. For this reason we in this paper are using digital signatures along with two fish algorithm as a procedure for key generation[12].

2. DESIGN ISSUES AND ROUTING CHALLENGES IN WSN

Though the wireless sensor network have numerous applications, they have some restrictions like limited energy supply, low bandwidth of links connecting sensor nodes or limited computational power. The major goal of WSNs is to prolong the lifetime of the communication network and also prevent the problems in connectivity by implementation of aggressive and efficient energy management techniques. There are some challenging factors that influence the design of different routing protocols in WSNs. Here in the following we present an overview of some of the routing challenges as well as certain design issues that may affect the routing process in WSNs.

2.1. Node Deployment

This process of node deployment in WSNs is basically application dependent and it also affects the routing performance. There are two ways of Deploying Nodes, namely deterministic and randomized node deployment. The techniques in which sensors can be placed manually and data routing are done through predefined paths are called deterministic deployment. However in the case of random deployment the nodes are randomly scattered giving an impression of infrastructure of ad-hoc distribution. Optimal clustering becomes necessary if the overall

distribution of nodes is non-uniform simply to allow energy efficient routing or to allow connectivity.

2.2. Energy Consumption

While performing computation or transmitting information's, the sensor nodes use their energy supply which is always limited in wireless environment. Thus the life of a sensor nodes shows a heavy dependence on the network battery lifetime. Every node in a multihop WSN plays dual role, both as a data sender and data router. Significant topological changes can be caused by power failures which may be a result of malfunctioning of certain sensor nodes. This may require the reorganization and rerouting of the packets being permitted in the network.

2.3. Node or Link Heterogeneity

Generally the sensor nodes are considered to be homogenous that is they all have some capacity in terms of communication, computation and power. But sometimes sensor nodes may have different capabilities or roles depending on their application as some of the applications may require a blend of sensors for monitoring pressure, temperature and humidity of the environment capturing the image or tracking of mobile objects. These special types of sensor nodes can either be independently deployed or all the different functionalities could be included in some sensor nodes. For example some hierarchical protocols can designate a cluster head to be different from normal sensor nodes. The cluster heads are chosen from among deployed sensors and are more powerful than the normal sensor nodes in terms of memory, energy and bandwidth. This handles the burden of transmission.

2.4. Fault Tolerance and Scalability

The failure of certain sensor nodes due to physical damage or lack of power or environment interface should not affect the overall performance of the sensor network. There may be a problem when many nodes may fail, the MAC and the routing protocols should accommodate the formation of new links or new routers to the data collecting BS. Thus multilevel redundancy is required in a fault tolerant sensor network. Scalability is another issue in WSN as the number of SN deployed in any sensing area they may be in the order of thousands or even more. The routing scheme used must be capable of working with the huge number of sensor nodes. Also the routing protocols being used must be scalable enough to respond to any event or operation in the environment. Most of the sensors remain in sleep state until an event occurs.

2.5. Connectivity and Coverage

Higher node density in any sensor network prevent them from being isolated from each other. Thus the sensor nodes are expected to be well connected. This may not prevent different network topologies from being variable and also from the network Size shrinking due to failure of some sensor nodes. Thus the connectivity depends on the random distribution of sensor nodes. In WSNs all the sensor nodes obtain their own view of sensing environment. This view of sensor is limited both in accuracy as well as in range. Thus only a limited physical sensing area can be covered. Thus Area coverage also becomes an important design parameter in WSN.

3. PRELIMINARIES

3.1. Privacy Homomorphism Encryption

An encryption scheme with homomorphic property is privacy homomorphism encryption. The homomorphic property means that the algebraic operations on PT can be executed with the manipulation of the corresponding CT with the help of a key.

$$D_k(E_k(m_1) \circ E_k(m_2)) = m_1 @ m_2,$$

Where $D_k()$ is decryption with key K , $E_k()$ is encryption with key K , \circ denote operations on cipher text and $@$ denote operations on plaintext.

PH schemes are of two types, similar to conventional encryption schemes. First one is Symmetric cryptosystems where keys are identical and second one is Asymmetric cryptosystem where keys are different. Symmetric PH schemes have greater efficiency as compared to Asymmetric PH schemes. The best known Asymmetric schemes are the one based on ECC (Elliptic Curve Cryptography) which provides the same security as RSA cryptosystem and that too with a smaller key size and cipher text. A 160-bit ECC cryptosystem provides the same security as provided by a 1,024-bit RSA cryptosystem [24].

3.2. Data Aggregation and Encryption

There is a major problem of aggregation of encrypted data in WSNs [23] which was firstly introduced by Gira et al. in [10] and it was further refined in [15]. Homomorphic encryption schemes were used to enable arithmetic operations over cipher texts that is to be transmitted on a multi-hop basis. Secure aggregation also involves some problems with public-key encryption mechanisms [16]-17]. Solution to public key encryption mechanism is to equip nodes with private keys for increasing the security level. This limits the effect of attacker that compromises some of the nodes but this is not deployed yet because of certain reasons mainly being the high computational cost involved in encryption and decryption of plaintext and cipher texts. Also the expansion in bit size during plaintext to cipher text conversion involves high overhead hence depleting the sensors energy.

3.3. Routing Protocols

The efficiency of a sensor networks heavily depends on the routing protocols used. Energy Efficient & Secure Pattern Based Data Aggregation protocol (ESPDA) was proposed that considered data aggregation and security together for wireless sensor networks [18]. In ESPDA cluster heads prevent transmission of redundant data from sensor nodes making ESPDA as energy and bandwidth efficient. Next concept was Secure Reference Based Data Aggregation (SRDA) in which the raw sensed data by sensor nodes is compared with referenced data values and the only the differential data is transmitted rather than the raw data [19]. Hein Zelman, et al. [20] proposed a hierarchical clustering algorithm for sensor networks. This was Low Energy Adaptive Cluster Hierarchy (LEACH) based protocol. Here the operations were divided into rounds and during each round another set of nodes acts as CHs. Main advantage of this was that energy consumption is uniformly distributed among all the nodes and the main disadvantage was that it uses scheduling criteria based on (TDMA) time division multiple access which makes it

inclined to long delays when it is applied to large sensor networks. An enhancement over LEACH protocol was published in [21]. This protocol was PEGASIS (Power Efficient Gathering in Sensor Information Systems). It was a chain based protocol designed for extending the lifetime of the network which elects a leader from the chain, based on residual energy level which results in average energy spent by each node being reduced. Virtual Grid Architecture (VGA) was another energy efficient routing paradigm proposed in [22]. This protocol used data aggregation and also in network processing to maximize the lifetime of the network as it performs data aggregation at two levels: local and global. PEGASIS greatly prolongs the lifetime of network when transmission range is limited and VGA saves more energy when transmission range is more.

3.4. CDA Based Privacy Homomorphism Schemes

Our work focusses on the solution for confidential data exchanges in WSNs that incorporates data aggregation. To the best of our knowledge, CDA (Concealed Data Aggregation) was the first concept that proposed a solution for end-to-end encryption along with the data aggregation model. In [8], the basic idea of CDA was introduced and it also showed the way to apply privacy homomorphism in WSNs. CDA provides end-to-end security along with providing in-network processing. They use algebraic properties of the applied PH: additive and multiplicative PH. In recent years, Castellucia, et al. Introduced an efficient data aggregation of encrypted data in WSNs and this is also based on additive homomorphism of encryption scheme [15]. Next concept introduced was CDAMA where the private keys is kept secret and it is only known by the base station. There is same public key for SNs in same group and no one outside knows the public key of the group. Also here BS extracts individual aggregated results from aggregated CT by performing individual decryption.

3.5. Digital Signatures

In some hostile environments, like a military battlefield, or environment monitoring, the broadcast authentication is an essential requirement to ensure the security and privacy of a Wireless Sensor Networks. The best way to provide broadcast authentication is to use digital signatures in sensor networks [25]. Public-key cryptography is considered to be too expensive for small sensor nodes, because it (like RSA) requires extensive computations and generally is not suitable for tiny sensors. In order to achieve Message integrity, Message Authentication Codes (MACs) are used. It allows for the application of digital Signatures that provide Data integrity and repudiation. Only the party that has the private key can create a particular signature. When a message along with a signature is received, only then the corresponding public key is used to verify the signature and once the signature verification is done, the receiver knows message integrity is still preserved. Digital signature is the most critical security services that cryptography offers. Digital Signature is an authentication mechanism which enables the sender to attach a unique code (signature). The signature is generated by taking the hash of the message and then encrypting the message with the sender's private key. It is an NIST standard that uses secure hash algorithm. The plaintext message, the message signature and the Public Key of sender are bundled together and transformed into signed and encrypted message using the Public Key of the receiver. The receiver unbundles received message and computes the message digest of the received message that is compared to the decrypted signature.

4. MODULE DESIGN

4.1. WSN set-up Model

In this module we set up a WSN environment in which network is divided into static clusters containing SN. Sensor nodes having limited energy and secure communication among them are necessary. Aggregator nodes are chosen based on residual energy level of nodes. Each sensor nodes sends the sensed data to corresponding aggregators which aggregates the received value and transfers the aggregated result to Base Station BS. We assume the Base Station to have immense computational power so it generates two types of keys, both public and private keys for sensor nodes using CDAMA scheme. All sensors have common public key but different private keys. Now the generated key is assigned to all the sensor nodes.

4.2. Aggregation Model

In WSN information is collected by sensor nodes from deployed environments and this collected information is forwarded to base station via multi-hop transmission based on cluster topology. This accumulated transmission results in high energy consumption for the intermediate nodes. Thus to increase the lifetime of the sensor networks cluster topology enables the intermediate nodes to perform data aggregation(AG).After performing aggregation AGs forwards the aggregated result to next hop. Aggregation of data takes place by two methods i.e., algebraic operations (e.g., adding or multiplying) or statistical operations (e.g., mean, median, mode, max, min). AG forwards only the aggregated result instead of forwarding the entire raw data.

4.3. Attack Model

Here in this model, we create two unauthorized sensor nodes called the attacker nodes which have more energy and threshold as compared to the normal nodes. There are different types of possible attacks on WSNs. Here we in this paper are considering the DOS attack Denial-of-Service attack which causes Black hole attack, Wormhole attack, Sybil attacks, Selective forwarding attacks etc. DOS attack is based on node-id. The attacker node behaves as normal nodes with its changed node id and receives data packets and drops them causing loss of data. Attacker nodes also change the threshold of the normal nodes thus drying the energy of the normal nodes. There are two methods followed by the attacker nodes here. Firstly, it traces the node id and changes the node id (based on node id) and secondly changes the threshold value (based on energy level).

4.4. Security Model

To ensure Data integrity and Data Confidentiality homomorphic encryption is used, which allows operations on encrypted data without decrypting them at the intermediate nodes thus preventing the access to plaintext. Considering the security issues in WSNs, we use Dynamic Intrusion Detection Protocol model (DYDOG) where Dynamic Intrusion Detection nodes are deployed which acts as both forwarding node and also intrusion monitoring and detection nodes with respect to data flow through sensor network. It uses secure session key management technique without deploying separate intrusion monitoring nodes. This technique makes network more dynamic and flexible against various kinds of attacks. Here in our work, we have used DYDOG

technique along with a secure key generation scheme that is based on Digital Signatures and Two Fish Algorithms.

5. SYSTEM DESCRIPTION AND SECURITY OBJECTIVES

5.1. Network Architecture and Operating Mechanism

In this paper, we consider a wireless sensor network system consisting of a fixed base station and large number of sensor nodes. These sensor nodes are homogenous in functionalities as well as capabilities. We suppose, the sink as reliable always, but the sensor nodes are subject to be compromised by the attackers. In this wireless system, the data are sensed by the sensor nodes and are transmitted to a base station with the help of CHs that performs data aggregation. We also assume that, all sensor nodes and the BS use the symmetric radio channel, sensor nodes are distributed randomly, and are energy constrained. The protocol used is CDAMA that elects CHs, and a sensor node transmits the data to its CH.

5.2. Security Vulnerabilities and Objectives

Like all other type routing protocols in WSNs, CDAMA are vulnerable to number of security attacks e.g., jamming, spoofing, replay attacks, etc. Because of depending on the CHs for data aggregation and routing procedure, attacks involving CHs could be the serious to the network system. If an attacker compromises the secrets of a CH, it can provoke attacks such as sinkhole or selective forwarding attacks, thereby degrading the network performance. Also the attacker may intend to inject any forged sensing data into the network towards the CHs. If we use DYDOG and Digital Signatures along with the normal CDAMA protocol then we notice lesser attacks and improved network performance on the basis of throughput, packet delivery ratio, and throughput.

5.3. Protection from Vulnerabilities Mechanism

Wireless sensor networks are vulnerable to attacks like Denial Of Service (DOS) which includes mainly Black hole attack, Wormhole attack, Sybil attack, Jamming attacks and selective forwarding attacks. This is a serious problem in WSN. A packet drop attack or black hole attack is a type of DOS attack in which a node that is supposed to relay packets actually discards them instead. This occurs when a node is compromised by the attacker. Because the packets are dropped from a lossy network, it is very hard to detect and prevent these packet drops. The adversary can make several compromised nodes in Black hole intercepted region. Intruder can also sense or read the secret data from. Similarly wormhole attack records and also uses the secret data in unauthorized manner, Sybil attack causes faulty identification and the Selective forwarding attack data loss in wireless sensor networks. Against these various types of attacks Dydog model provides flexible and resilient solution by Dynamic Intrusion Detection Nodes for High-Data rate. Dynamic Intrusion Detection Protocol model (DYDOG) Design is based on data flow in Wireless Sensor Networks (WSNs). In this the Dynamic Intrusion Detection nodes are deployed which will act as forwarding node as well as an Intrusion Monitoring Node. The selection of dynamic intrusion detection nodes are from the neighbor's non-forwarding node list using Secure Session Key Management approach without deploying any separate Intrusion Monitoring Nodes. Because of this the network becomes more dynamic and flexible against

various types of attacks and provides availability of maximum monitoring node's with high error rate Wireless Networks.

Within the session itself depending on mobility the monitoring nodes dynamically change its behavior. For an attacker it creates problem to identify and attack these nodes within the limited session. By this technique the attacks and the Compromised nodes can be effectively and easily identified at runtime even in high data rate dynamic or static Wireless Sensor Networks (WSNs).

6. OUR SCHEME

We have used DYDOG mechanism for security enhancement of CDAMA. First we performed the Wormhole attack on CDAMA technique. Wormhole acts by three procedures.

- 1) Message duplication
- 2) Compromising the node-id of normal nodes
- 3) Packet dropping

To overcome this wormhole attack we use DYDOG mechanism, Digital signatures and Two fish algorithm.

Steps involved are as follows:

6.1. Key Generation Procedure

1. If source is transmitting data
2. Count the number of requests
3. Evaluate $N = \lfloor \text{expr}(\text{len}_{q1}) * (\text{len}_{q2}) * (\text{len}_{q3}) \rfloor$
4. Initialize E
5. Randomize GEN as value of index.
6. Evaluate $H = \lfloor (q1) * (q2) * \text{GEN} \rfloor$
7. Evaluate $T_{\max} = \lfloor (T) / (x) \rfloor$
8. Evaluate $P = \lfloor (q2) * (q3) * (\text{GEN}) \rfloor$
9. Find Public key $[((N) * (E) * (P) * (H) * (T_{\max}))]$
10. Return Public key

6.2. Encryption Procedure

1. If data is received at destination
2. Count the number of reply
3. If request= message_id then randomize the value of R
4. Calculate cipher text as per expression $C = \lfloor (M) * (P) + (R) * (H) \rfloor$
5. Return the value of ciphertext as C.
6. Calculate aggregation count $\text{AGG_C} = \lfloor (\text{Message_id} * P) + (\text{Message_id} * Q) + (\text{Message_id} * H) \rfloor$
7. Return the value of AGG_C

6.3. Aggregation Procedure

1. Compute the aggregated result as cipher text $C' = C1 + C2$. It also includes the randomness of both groups.
2. Return C'

6.4. Decryption Procedure

1. Compute $M, M = \log_p (q_2 q_3 * C)$
2. Return M

6.5. DyDog Two Fish and Digital Signature Procedure

In this the Dynamic Intrusion Detection nodes are deployed which will act as forwarding node as well as an Intrusion Monitoring Node.

We define the following apps for using DyDOG along with Digital Signatures and Two Fish algorithm.

6.5.1. Cryptographic App

1. Initialize Crypto App with new app.
2. Use the catch variable in a procedure, if catch has crypto app and it should load the path of library where cryptography is available else it should return the error code.
3. Using the class namespace, eval, tcl drop, two fish and encryption is accessed. All will be accessed in the name space.
4. Variable of two fish version is specified i.e., version 0.1.
5. Package is provided for accessing the two fish and tcl drop and for calling the version initialized above
6. Package is provided for accessing the encryption and tcl drop
7. If the information is available for accessing the numversion of tcl return the same
8. Call the check module function for encryption
9. Define three procedures with keywords (Encrypt, Decrypt and Encrypt Password).

6.5.2. Cipher Text App

1. initialize cipher app with value new_app
2. if twofish.tcl file is available (file will be generated when running the code and disappears once executed) take it as source
3. check the packages required for Tcl, Tcl 8.4, tcltest, itwofish
4. proc h2b {hex}
return [binary format H* \$hex] //return the expression
5. proc b2h {bin}
binary scan \$bin H* dummy //scan binary value and make it as dummy value
return \$dummy //returning dummy value

6. If length of argument value is equal to 0 for each statement with digital signature number and cipher, then increment the testnum
7. Initialize the engine with two fish and digital signature number
8. Initialize encrypted value of engine in the encrypted block with the value of h2b and clear
9. Initialize the decrypted value of engine in the decrypted block with the value of encrypted which is obtained above.
10. if no string is encrypted using b2h or if no string is decrypted using b2h then reject the increment else approve the increment
11. Initialize a file in open status and trim the string and perform Monte Carlo test.
12. If the value of input vector and Monte Carlo value is equal create another engine using two fish and digital signature number.

6.5.3. Tests

We perform three tests here.

i. CBC mode test:

1. Initialize the engine with two fish and digital signature number.
2. Encrypt h2b (two fish object value) and pt (variable name) and make it encrypted with engine value.
3. Configure the engine with h2b (two fish object value).
4. Decrypt the encrypted value with engine and store it in decrypted app.
5. Access the itcl and delete the engine else call the value and check the below test

ii. Monte Carlo test

1. Initialize the engine with two fish and digital signature number.
2. Initialize the data with h2b (two fish object value) and pt (variable name).
3. Initialize the data with engine and encrypt block with data and make data encrypted.
4. Initialize the data with engine and decrypt block with data and make data decrypted.
5. Access the itcl and delete the engine else check the below test.

iii. ECB mode test

1. Initialize the engine with two fish and digital signature number.
2. Initialized encrypted value with engine and encrypted block with h2b (twofish object).
3. Initialized decrypted value with engine and decrypted block with called encrypted value.
4. Access the itcl and delete the engine.
5. If encrypted text does not match expectation or decrypted text does not match original plaintext then increment is failed otherwise approve the increment.
6. Initialize t with some message and encrypt the same with e and the decryption of block should be done while calling e.

6.5.4 Message App

1. Calling self to view the messages by node

2. Globally declaring ns with digital signature without encryption
3. Merge the messages viewed with respect to the message id's
4. Trace the node and node address
5. Calling self to send the size of message id and data with encrypt port.
6. Check the number of nodes and assign the node values to i.
7. Append the node id and message and assign Digital Signatures to all the nodes.

6.5.5 Check and Channel App

1. Check the value of packet size and if the expression of node boundary values is perfect with the data node, it will start transmitting the request otherwise node transmit stop
2. Assign each packet to channel and call the watch dog procedure and send it to all the channels.

6.5.6 Request

1. Creating new application called request.
2. Set flag as 0 and if source transmits data and reply from the destination is equal to source received reply then start the data transmission.
3. Else update the request and transmit the request till it reached the maximum value and initialize till request is maximum

6.5.7 Transmit

1. Create a new application called transmit
2. If source has the data received and the data received is equal to request, assign flag as 0
3. If flag is 0, drop the request, else transmit the request
4. If destinations have the received data and if node id request and node id reply are same then initialize the reply again

6.5.8 Reply

1. Create a new application called reply.
2. If reply is received and the received reply is not equal to source, assign flag as 1.
3. If flag is 0, transmit reply or acknowledgement or data and assign flag as 1 else drop the request.
4. Else assign flag as 1 and if reply received to node is not equal to source, transmit new reply or acknowledgement or data and assign flag as 1.

Count the values if node transmits are equal to 0 and if count is less than 2 then transmit count else stop the reply.

7. SIMULATION PARAMETER

This model is implemented using a network simulator 2.34. The simulation parameters are 500 X 500 sq. area, and consisted of 50 to 60 number of nodes with flat-grid topology, two ray ground ground radio propagation model and 802.15.4 MAC layer .AODV, CDAMA, AODV under Attack, CDAMA under attack and CDAMA along with DYDOG and Digital Signatures from different perspectives such as Average-delay, Packet Delivery ratio, Energy Spent and

Throughput. The network simulator set up is shown below in the table.

TABLE I SIMULATION PARAMETERS

SI. No.	PARAMETERS	Values
1	Simulation area	500 X 500 square meters
2	Propagation	Two ray ground propagation
3	Queue type	Drop tail
4	Antenna type	Omni antenna
5	Number of nodes	50 to 60 nodes
6	Topology	Flat grid topology
7	Routing protocol	CDAMA
8	Maximum packets in interface queue length	200
9	Network interface type	Phy/wireless
10	MAC type	802.11

8. SIMULATION RESULTS AND DISCUSSIONS

8.1. Average Delay

Average delay includes all the possible types of delays that may be either due to buffering during route discovery latency, or queuing at the interface queue or may be the transfer times of the data packets. The figure shows the end to end delay incurred in transferring the data from source node to sink node by different routing schemes. The maximum delay is in AODV with attack, Sybill attack followed by the CDAMA with attack. In an efficient network the average delay should be less and when CDAMA is compared to AODV under attack and CDAMA under attack, it has lesser delay but when we incorporate DYDOG along with Digital signatures the delay is further reduced.

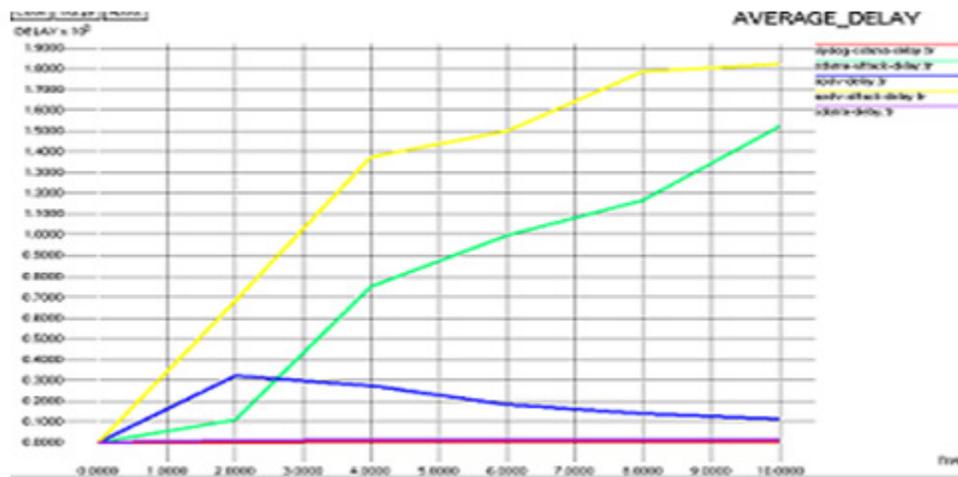


Fig.1. Variation of average delay with time

8.2. Packet Delivery Ratio

Packet Delivery Ratio is the ratio of the data packets that has been delivered to destinations to those that has been generated by constant bit rate (CBR) sources. The figure shows the packet delivery ration achieved by different routing techniques. The packet delivery ratio is highest for CDAMA enhanced technique followed by the normal CDAMA technique. In case of CDAMA under attack the number of packets sent is high but received number of packets is less so the PDR decreases for the AODV and CDAMA under attack.

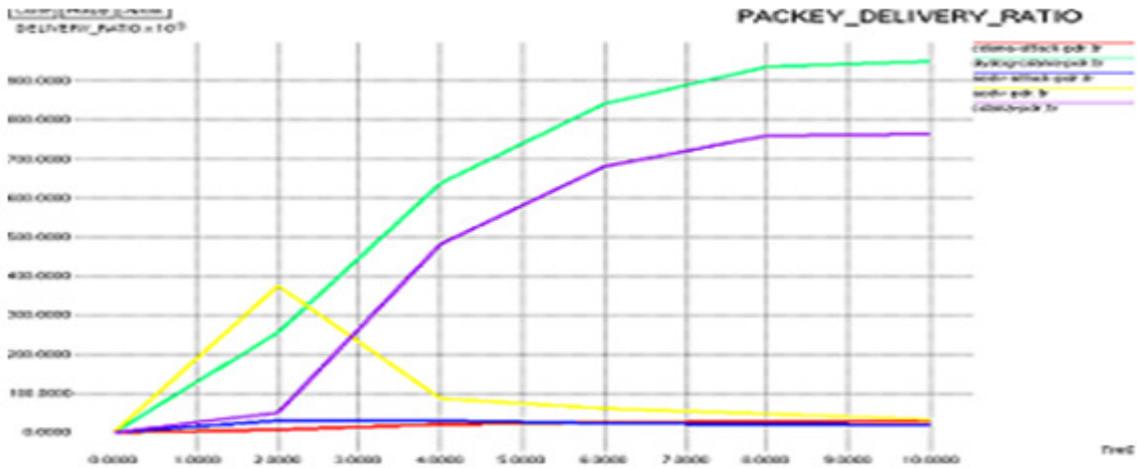


Fig.2. Variation of Packet delivery Ratio with Time

8.3. Energy Consumption

Average Energy Consumption by the sensor nodes in the network is one of the most important metrics to evaluate energy efficiency of the routing protocol that has been proposed. The figure shows the energy spent by nodes in the sensor network. Energy consumption for CDAMA technique is lesser than enhanced CDAMA because of some additional procedures like Two fish algorithms, Digital signatures and Dydog mechanisms. These extra procedures result in more energy consumption for enhanced cdama. Though DYDOG consumes little more energy but it also increases the throughput

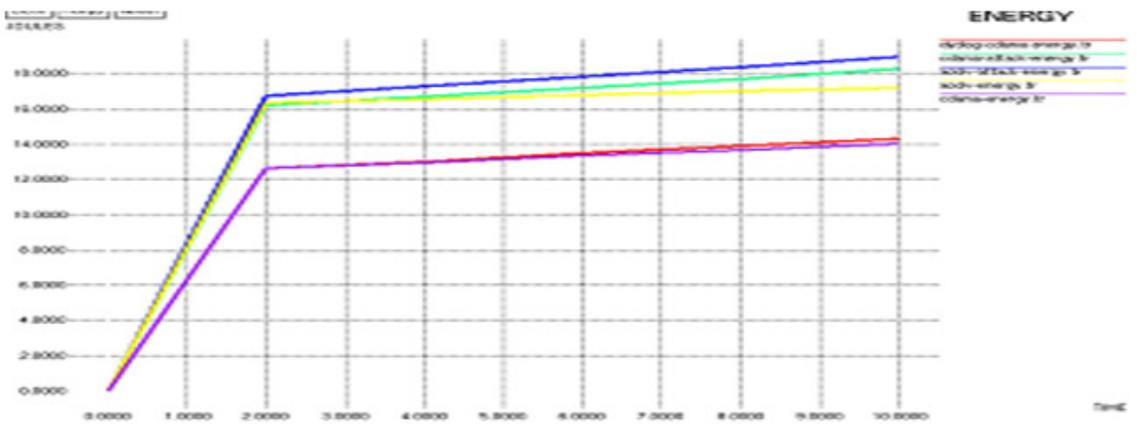


Fig.3. Variation of energy consumption with time

8.4. Throughput

Throughput is the total number of routing packets transmitted per data packets that has been delivered at destination. The throughput of CDAMA with DYDOG mechanism is more because of digital signatures along with Two fish algorithm. Though energy consumption is more but as throughput of overall transmission increases and hence the security also increases.

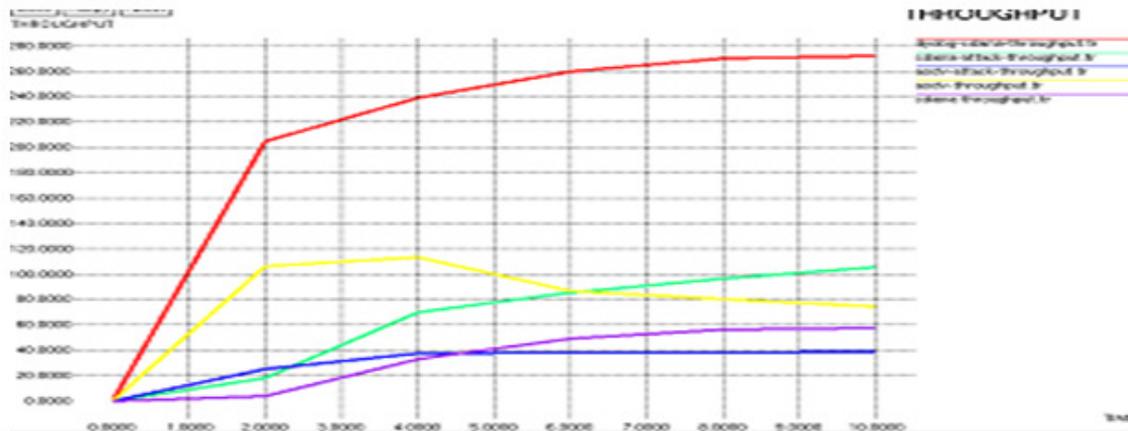


Fig.4. variation of Throughput with time.

9. CONCLUSION

The work proposes a secure, increased throughput and a better packet delivery ration scheme than normal CDAMA technique. Here the Dynamic Intrusion Detection Protocol (DYDOG) model is used along with Digital Signatures along with Two Fish Algorithms for CDAMA technique where cipher text of different applications can be aggregated together. While using these algorithms and protocols we have enhanced the working of CDAMA technique that mitigates the impact and reduces the overall damage to acceptable condition. CDAMA performs better than the traditional AODV routing protocol but the proposed technique provides higher security using Digital Signatures along with two Fish algorithm. The proposed technique defends the altered routing, selective forwarding and wormhole attacks. In this paper the energy consumption is still a issue as this technique leads to more energy consumption by the sensor nodes In our future work we will be proposing a technique for lesser energy consumption than the proposed technique.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

REFERENCES

- [1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A Survey on Sensor Networks" IEEE Comm. Magazine, vol. 40, no. 8, pp. 102-114, Aug. 2002.
- [2] R. Min, A. Chandrakasan, "Energy-Efficient Communication for Ad-Hoc Wireless Sensor Networks," Proc. Conference Record of the 35th Asilomar Conference Signals, Systems and Computers, vol. 1, 2001.

- [3] B. Przydatek, D. Song, A. Perrig, "SIA: Secure Informations Aggregation in Sensor Networks," Proc. First International Conf. Embedded Network Sensor Systems, pp. 255-265, 2003.
- [4] R.Chandramouli, S.Bapatla, and K.P.Subbalakshmi, "Battery power-aware encryption.ACM transactions on information and system security," pp. 162-180, 2006.
- [5] K.Akkaya,M.Demirbas, RS.Aygun, "The Impact Of Data Aggregation on the performance of Wireless Sensor Networks," wiley wireless Communication Mobile Computing (WCMC), J(8), 171-193, 2008.
- [6] Alzaid,H, Foo,E, and Nieto J.G, " Secure data aggregation in wireless sensor network-:a survey," In the proceedings of the sixth Australasian conference on information security, Volume-81, pp.93-105, 2008.
- [7] J.Girao, M. Schneider, and D.Westhoff, "CDA:Concealed Data Aggregation in wireless sensor networks,"Proceedings of the ACM workshop on Wireless Security, 2004.
- [8] D. Westhoff, J. Girao, and M. Acharya, "Concealed Data Aggregation for Reverse Multicast Traffic in Sensor Networks: Encryption, Key Distribution, and Routing Adaptation," IEEE Trans. Mobile Computing, vol. 5, no. 10, pp. 1417-1431, Oct. 2006.
- [9] E. Mykletun, J. Girao, and D. Westhoff, "Public Key Based Crypto schemes for Data Concealment in Wireless Sensor Networks," Proc. IEEE Int'l Conf. Comm. (ICC '06), vol. 5, 2006.
- [10] J. Girao, D. Westhoff, M. Schneider, N. Ltd, and G. Heidelberg, "CDA: Concealed Data Aggregation for Reverse Multicast Traffic in Wireless Sensor Networks," Proc. IEEE Int'l Conf. Comm. (ICC '05), vol. 5, 2005.
- [11] J. Girao, D. Westhoff, E. Mykletun, and T. Araki, "Tinypeds:Tiny Persistent Encrypted Data Storage in Asynchronous Wireless Sensor Networks,"Ad Hoc Networks, vol. 5, no. 7, pp. 1073-1089, 2007.
- [12] C.d. Westhoff,B.Lamparter, and A.Weimerskirch,"on digital signaturesin ad hoc networks," J.Eur.Trans. telecom, vol.16, no. 5, pp. 411-425, 2005.
- [13] R.Watro, D.Kong, S.Cuti, C.Gardiner, C.lynn, and P.kruus, "Sensor networks with public key technology", in proc. 2nd ACM Workshop Security ad hoc sensor network, pp.59-64, 2004.
- [14] C.karlof ,D.Wagner, "Secure routing in wireless sensor networks: Attacks and countermeasures,"in the proc IEEE Int. Workshop Sensor Netw. Protocols Appl., May 2003, pp. 113-127, May 2003.
- [15] C.castelluccia, E.Mykletun and G.Tsudik, "Efficient Aggregation of encrypted data in wireless sensor networks," Mobile and Ubiquitous Systems: Networking and Services, 2005.
- [16] E. Mykletun, J. Girao, and D. Westhoff, "Public Key Based Cryptoschemes for Data Concealment in Wireless Sensor Networks", IEEE Int'l Conf. Communications (ICC '06), pp. 2288-2295, 2006.
- [17] T.ELGamal, "A Public Key Cryptosystem and a Signature Scheme based on Discrete Logarithms," CRYPTO,IT-31(4): pp. 469-472, 1985.
- [18] H. Cam, S. O'zdemir, P. Nair, D. Muthuavinashiappan, and H.O. Sanli, "Energy-Efficient Secure Pattern Based Data Aggregation for WSNs," Computer Comm., vol. - 29, no. - 4, pp. 446-455, 2006.

- [19] H. Sanli, S. Ozdemir, and H. Cam, "SRDA: Secure Reference based Data Aggregation Protocol for WSNs," Proc. IEEE 60th Vehicular Technology Conf. (VTC '04-fall), vol. 7, 2004.
- [20] M. Younis, M. Youssef and K. Arisha, "Energy Aware Routing in Cluster Based Sensor Networks", in the Proceedings of the 10th IEEE/ACM (MASCOTS2002), Fort Worth, TX, October 2002.
- [21] S. Lindsay and C. Raghavendra, "PEGASIS: Power Efficient gathering in Sensor Info. Systems", international conference on communications, 2001.
- [22] J.N. Al-Karaki, et al., "data Aggregation in Wireless Sensor Networks-Exact and approximate algorithms," Proc IEEE Wks. High Perf. Switching and Routing 2004, phoenix, AZ, Apr. 18-21, 2004.
- [23] Sanjeev Setia, a. Sankardas Roy and Sushil Jajodi "Secure Data Aggregation in Wireless Sensor Networks" Proc. of 33rd STOC, pp. 266-275, 2001.
- [24] N. Koblitz, A. Menezes, S., Vanstone, "State of Elliptic Curve Cryptography," Designs, Codes & Cryptography, vol. 19, no. 2, pp. 173-193, 2000.
- [25] Cryptography and Network Security Principles and practices, William Stallings, Pearson Education, Fifth Edition.

AUTHORS

Bharat Bhushan (M'26). Date Of Birth-17th Dec 1989. Phd Scholar (Dept. of Computer Sc. & Engg.) student at Birla Institute of Technology, Ranchi, Jharkhand-835215, India. He has worked as Network Engineer for 1 years in HCL Infosystems Ltd., Noida



Keshav Kaushik (M'20). Date Of birth- 21/05/1996 BTech Computer Science & engineering student at HMR institute of technology & management.



Gadadhar Sahoo (M'60) Professor and Dean (Admissions and Academic Coordination), Dept. Of Computer Science & Engg., Ph.D., IIT Kharagpur. He has working experience of teaching field of 26 years and research experience of 31 years. He is now with Birla Institute Of Technology, Mesra, Ranchi. He has 167 publications in National/International journals and 80 publications in National/International Conferences. Dr. Sahoo is also a member of ISTE.



IP CORE DESIGN OF HIGHT LIGHTWEIGHT CIPHER AND ITS IMPLEMENTATION

Sruthi.N¹, R.Nandakumar² and Rajkumar.P³

¹Student, VLSI DESIGN, Department of Electronics and Communication
Engineering, NCERC, University of Calicut
shruthishanti@gmail.com

²Scientist 'C', NIELIT, Calicut
nanda@nielit.gov.in

³Faculty, Department of Electronics and Communication Engineering, NCERC,
University of Calicut
rajkumar1073@ncerc.ac.in

ABSTRACT

In the present era of e-world where security has got a larger weightage, cryptography has its role to play. Nowadays, the devices available in the market are of resource constrained type. Hence we need lightweight ciphers for the efficient encryption of data thereby increasing the performance. In this project a detailed study of HIGHT cryptographic algorithm is done which outperforms standard algorithms. HIGHT is an ISO Standard block cipher which has 64-bit block length and 128-bit key length. HIGHT was designed to be proper for the implementation in the low resource environment such as WSN, WBN, RFID tag or tiny ubiquitous devices. It is implemented on Spartan 6 FPGA evaluation kit and performance metrics are found out. A HIGHT cryptocore is being designed, characterized and implemented which will be a reference platform for hardware design engineers to model devices which require lightweight characteristics.

KEYWORDS

HIGHT, Lightweight cryptography, low resource devices, FPGA

1. INTRODUCTION

For secret communication there is a need of hidden writing and this part of science is called cryptography. With the help of cryptography we are able to achieve data integrity, data confidentiality and authentication. In such cases, certain protocols are created and analyzed and they are known as ciphers. These ciphers are the combination of mathematics, computer science and electrical science. They are mainly used in e-commerce, ATM passwords and other applications where there is a need of privacy. In today's world everyone needs privacy for communication hence cryptography has a major role to play. Ciphers are basically classified into Symmetric ciphers and Asymmetric ciphers. There is a common key for encryption and

decryption in symmetric ciphers whereas in asymmetric ciphers there is a public key to encrypt and private key to decrypt. Hence data manipulation is done. Symmetric ciphers are further classified as block ciphers and stream ciphers. In block ciphers, data is being divided into blocks of particular size and whereas in stream ciphers bit by bit manipulation of the data is being done. The block ciphers can be transformed into stream ciphers by operating in OFR and CTR modes. In stream ciphers, hidden internal state changes as the cipher operates. Block ciphers are better analyzed and has got broader range of applications. The basic 2 properties of the ciphers are diffusion and confusion. Diffusion dissipates statistical structure of plaintext over ciphertext (redundancies are dissipated) whereas confusion property gives the relationship between cipher text and key as complex as possible. The basic design elements of a cipher include block size, key size, number of rounds, subkey generation algorithm, round function, fast software en/decryption and ease of analysis. Block ciphers are iterated ones i.e they transform fixed size blocks of plaintext into identical size ciphertext through the repeated application of an invertible transformation known as round function. Round functions take different round keys k as second input which are derived from the original key. The design criteria for ciphers are efficiency. In block ciphers usage of Sbox leads to larger hardware footprint. Memory expense is the major constraint of designing a block cipher. Based on the structure of algorithm, the block ciphers are classified into SP networks and Feistel networks. The main advantages of using feistel network are that en/decryption operations are very similar i.e only reversal of key schedule is required. The cryptographic algorithms developed before 1990s was mainly focused to work on standard devices which consume larger area and power.[2] But gradually the devices were made to work in the resource constrained environment. For securing such devices, lightweight ciphers were invented. These ciphers are developed bit away from industry demands. The design criteria of lightweight ciphers are efficiency, simplicity and security. The block size can be 32,48 or 64 bits and key size can be 80 or 128 bits. The power, area consumption of lightweight ciphers is minimum.

In this paper, HIGHT cryptographic algorithm is implemented in both software and hardware platform. The results and the resource utilized by the design is also given.

2. HIGHT

The block cipher HIGHT was developed in Korea. HIGHT is the shortform of HIGH security and lightweight. HIGHT is a ISO/IEC 18033-3:2010 which has 64 bit input /output data block with no Sbox, 32 round with XOR, modular addition and circular shift operations. The HIGHT algorithm is defined below,

The entire plain text is divided into 8 subtexts, each 8 bit each. From the 128 master keys are being divided into 16 keys, 8 bit each. 8 whitening keys are generated from the master keys and the 128 subkeys from the constant generation algorithm. Out of these 8 whitening keys, first 4 are used in the initial transformation of the plain text and last 4 are used in the final transformation. Constant generation algorithm is based on a 7-bit LFSR. The 7 bits i.e initial state of the LFSR is '0101101' and from this basic constant, by doing the XOR operation of last 2 bits next constant is being generated and the process is continued to generate further 127 constants.[1] These 128 constants along with the master keys are used to generate 128 subkeys. In the 32 rounds of HIGHT, each round uses 4 subkeys for the operations.

The plain text P

$$(1) P = P_7 || P_6 || P_5 || P_4 || P_3 || P_2 || P_1 || P_0$$

Master Key K

$$K = K_{15} || K_{14} || K_{13} || K_{12} || K_{11} || K_{10} || K_9 || K_8 || K_7 || K_6 || K_5 || K_4 || K_3 || K_2 || K_1 || K_0$$

Whitening and Subkey generation

a) The generation of whitening keys is defined as follows

for $i = 0, 1, 2, 3$:

$$WK_i = K(i+12)$$

For $i = 4, 5, 6, 7$:

$$WK_i = K(i-4)$$

b) The 128 subkeys are used for encryption and decryption, 4 subkeys per round .
The generation of sub keys is defined as follows.

$$(1) s_0 = 0, s_1 = 1, s_2 = 0, s_3 = 1, s_4 = 1, s_5 = 0, s_6 = 1$$

$$d_0 = s_6 || s_5 || s_4 || s_3 || s_2 || s_1 || s_0$$

(2) for $i = 1$ to 127 ;

$$s(i+6) = s(i+2) \wedge s(i-1)$$

$$d_i = s(i+6) || s(i+5) || s(i+4) || s(i+3) || s(i+2) || s(i+1) || s(i)$$

(3) for $i = 0$ to 7 ;

for $j = 0$ to 7 ;

$$SK(16 * i + j) = K(j - i \bmod 8) [+] d(16 * i + j)$$

for $j = 0$ to 7 ;

$$SK(16 * i + j + 8) = K((j - i \bmod 8) + 8) [+] d(16 * i + j + 8)$$

Initial transformation

$$(2) X_{0,0} = P_0 [+] WK_0$$

$$X_{0,2} = P_2 \wedge WK_1$$

$$X_{0,4} = P_4 [+] WK_2$$

$$X_{0,6} = P_6 [+] WK_3$$

$$X_{0,1} = P_1$$

$$X_{0,3} = P_3$$

$$X_{0,5} = P_5$$

$$X_{0,7} = P_7$$

32 rounds

(3) For $i = 0$ to 30:

$$X(i+1),0 = X_{i,7} \wedge (F_0(X_{i,6}) [+] SK(4*i + 3))$$

$$X(i+1),2 = X_{i,1} [+] (F_1(X_{i,0}) \wedge SK(4*i))$$

$$\begin{aligned}
 X(i+1),4 &= X_i,3 \text{ [^] (F0}(X_i,2) \text{ [+] SK}(4*i + 1)) \\
 X(i+1),6 &= X_i,5 \text{ [+] (F1}(X_i,4) \text{ [^] SK}(4*i + 2)) \\
 X(i+1),1 &= X_i,0 \\
 X(i+1),3 &= X_i,2 \\
 X(i+1),5 &= X_i,4 \\
 X(i+1),7 &= X_i,6
 \end{aligned}$$

For $i=31$:

$$\begin{aligned}
 X(i+1),1 &= X_i,1 \text{ [+] (F0}(X_i,0) \text{ [^] SK124)} \\
 X(i+1),3 &= X_i,3 \text{ [^] (F1}(X_i,2) \text{ [+] SK125)} \\
 X(i+1),5 &= X_i,5 \text{ [+] (F0}(X_i,4) \text{ [^] SK126)} \\
 X(i+1),7 &= X_i,7 \text{ [^] (F1}(X_i,6) \text{ [+] SK127)} \\
 X(i+1),0 &= X_i,0 \\
 X(i+1),2 &= X_i,2 \\
 X(i+1),4 &= X_i,4 \\
 X(i+1),6 &= X_i,6
 \end{aligned}$$

Final transformation

$$\begin{aligned}
 (4) \quad C_0 &= X_{32,0} \text{ [+] WK4} \\
 C_2 &= X_{32,2} \text{ [^] WK5} \\
 C_4 &= X_{32,4} \text{ [+] WK6} \\
 C_6 &= X_{32,6} \text{ [^] WK7} \\
 C_1 &= X_{32,1} \\
 C_3 &= X_{32,3} \\
 C_5 &= X_{32,5} \\
 C_7 &= X_{32,7}
 \end{aligned}$$

Final Cipher Text

$$5) C=C7||C6||C5||C4||C3||C2||C1||C0$$

The F0 and F1 round functions are:

$$\begin{aligned}
 F_0(x) &= (x \lll 1) \text{ [^] } (x \lll 2) \text{ [^] } (x \lll 7) \\
 F_1(x) &= (x \lll 3) \text{ [^] } (x \lll 4) \text{ [^] } (x \lll 6)
 \end{aligned}$$

The decryption operation is identical in operation to encryption apart from the following two modifications

(1) All [+] operations are replaced by [-] operations except for the [+] operations connecting SK_i and outputs of F₀

(2) The order in which the keys WK_i and SK_i are applied is reversed.

The toplevel block diagram of HIGHT is shown below

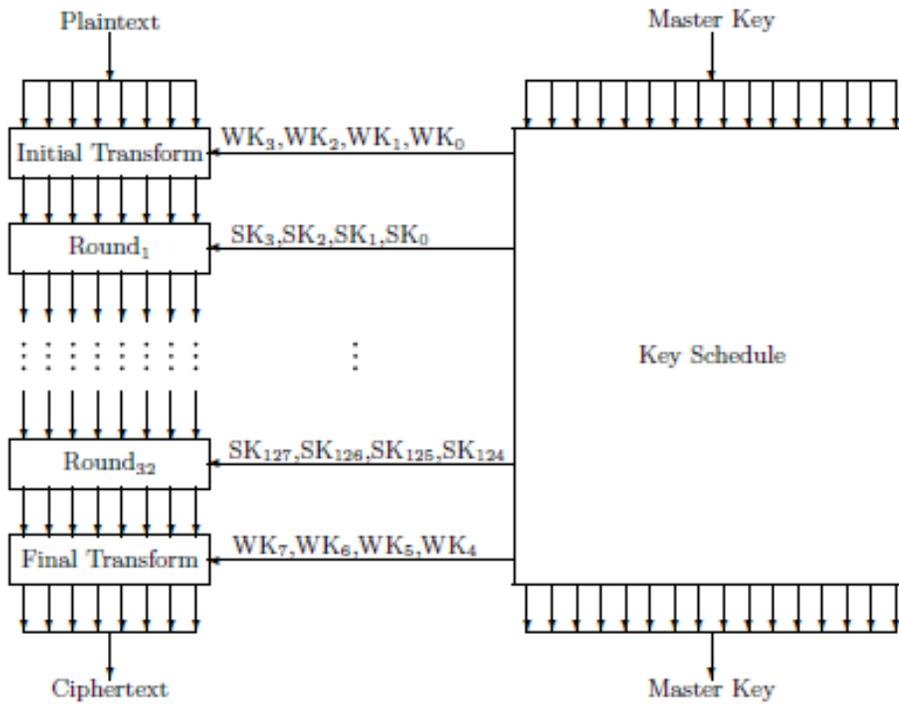


Figure 1. Toplevel Diagram

3. SOFTWARE PLATFORM IMPLEMENTATION

All the cryptographic algorithms are implemented on a software platform so that their behavior in such an environment is found out. The software platform implementation mainly aims at optimization of speed, memory size, power or energy.

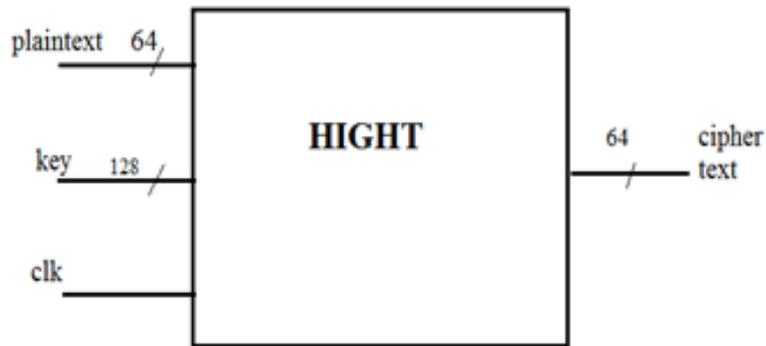


Figure 2:Input Output Diagram

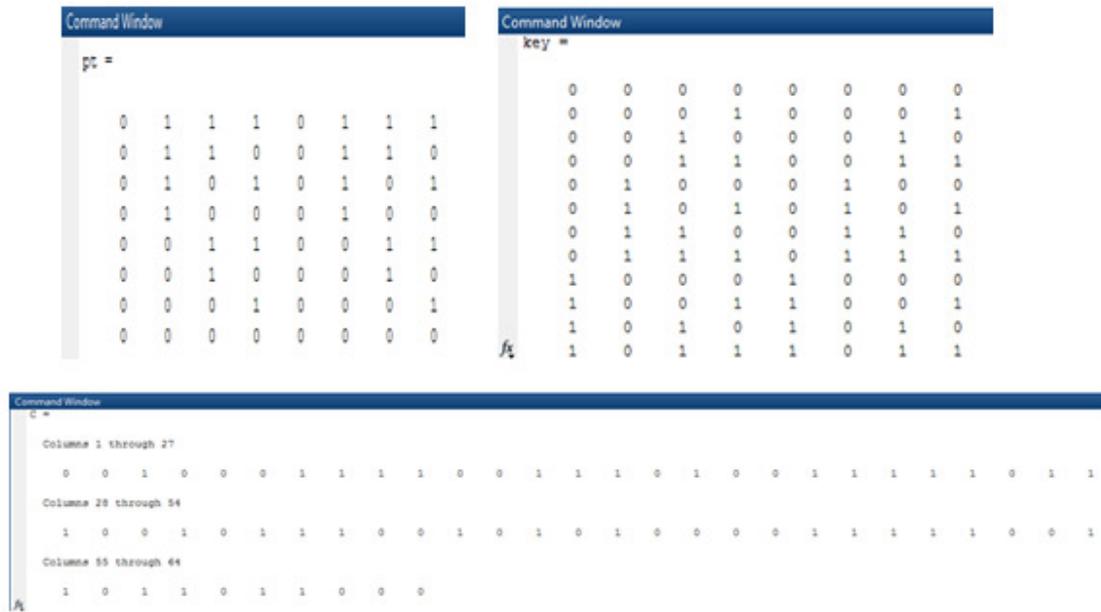


Figure 3: MATLAB Result

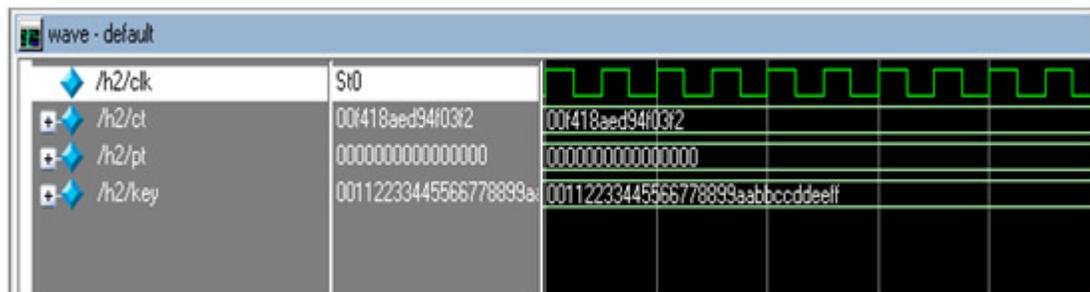


Figure 4: ModelSim Result

The HIGHT algorithm was implemented in MATLAB and MODELSIM 6.2c. Based on the input- output diagram of HIGHT, the Verilog code was created and was implemented in the software environment. The MATLAB calculator for the HIGHT was created and it was verified. To simulate the Verilog code, the code was run on the ModelSim and the results were found out. The results obtained are shown.

Table 1:Function Table

PLAINTEXT	MASTERKEY	CIPHERTEXT
0011223344556677	ffeeddccbbaa99887766554433221100	23ce9f72e543e6d8
0000000000000000	00112233445566778899aabbccddeeff	00f418aed94f03f2
0123456789abcdef	00112233445566778899aabbccddeeff	73aa299327a22684
0123456789abcdef	ffeeddccbbaa99887766554433221100	8181e2a70f8346f7
0000000000000000	ffeeddccbbaa99887766554433221100	3181ff9102b64cca

4. HARDWARE PLATFORM IMPLEMENTATION

Hardware implementations are mainly done on FPGA and ASIC technology. In ASIC, main aim is to reduce the design time. Comparing to ASIC implementation, FPGA is more advantageous because it provides flexibility, agility of algorithms and modifications are made easier. The Verilog code was run on Xilinx 14.3 and the synthesis results were obtained. The code was implemented on a Spartan -6 evaluation kit XC6SLX45T-3FGG484.

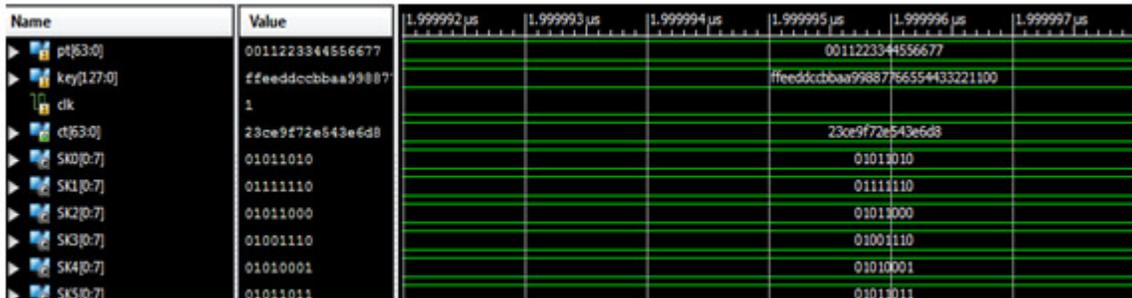


Figure 5: Xilinx Result

5. ONCHIP DEBUGGING AND PROTOTYPING RESULTS

Finally design is being analyzed using the ChipScope Pro Analyzer and on chip results were obtained. These results were used to compare with the simulation and synthesis results .The results obtained are shown below

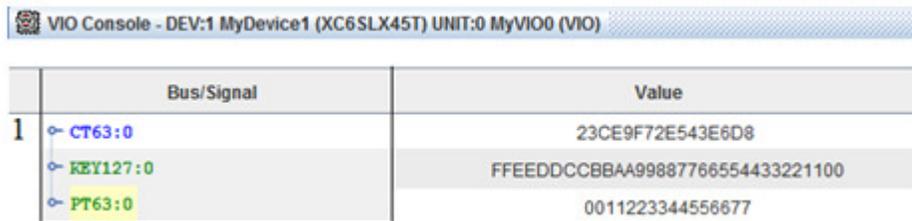


Figure 6 : On Chip Debugging Results

Power Analysis report is obtained on XPA tool on the Spartan-6 kit and from the report the power consumed by the design is equal to 0.037 W. After implementing on the FPGA kit, the design was implemented on ASIC platform and the area, power and timing details were obtained. The area consumed was found to be 0.22µm² and power consumed was found to be 0.06 nW .And the maximum frequency of operation is found to be 119.847 MHz . These results obtained from the ASIC implementation are used to calculate the performance metrics of the HIGHT cryptographic algorithm. Calculated throughput is 767.018Mbps.

Table 2 :Resource Utilization Summary

Slice Logic Utilization	Used	Available	Utilization
Number of Slice Registers	2,409	54,576	4%
Number used as Flip Flops	2,409		
Number of Slice LUTs	2,689	27,288	9%
Number used as logic	2,453	27,288	8%
Number used exclusively as route-thrus	236		
Number of occupied Slices	991	6,822	14%
Number of MUXCYs used	568	13,644	1%
Number of bonded IOBs	1	296	1%
Number of BSCANs	1	4	25%
Average Fan-out of Non-Clock Nets	3.85		

6. CONCLUSIONS

This paper focuses on the characterization of HIGHT algorithm and has developed an IP Core of HIGHT which will be reference one for the design engineers. A detailed study on HIGHT block cipher was done and carried out its algorithm validation. HIGHT block cipher is a lightweight block cipher of block size 64 bit and key size 128 bit targeted to provide cryptographic security for resource constrained applications e.g. RFID, sensor networks etc. The behavioural description of the design is written in Verilog HDL and simulated using XilinxISE 14.3 and ModelSim 6.2 c software platforms. Then the design is successfully implemented on Xilinx Spartan6 FPGA. The performance metrics were found out and the results are presented. A detailed analysis of HIGHT cryptographic algorithm was done. In-depth analysis of linear and differential attacks needs to be carried out.

ACKNOWLEDGEMENTS

The gratification of this Project will be incomplete without mentioning all the people who helped me to make it possible, whose gratitude and encouragement were valuable to me. I would like to thank my guides for their whole hearted support. I would also like to thank my parents and friends who encouraged me and gave me the motivation to complete the work. Above all I would like to thank God for his abundant grace.

REFERENCES

- [1] Deukjo Hong, Jaechul Sung, Seokhie Hong, Jongin Lim, "HIGHT: A New Block Cipher Suitable for Low-Resource Device," Springer 2006.
- [2] Mohd Bj, et al, "A survey of lightweight block ciphers for low-resource devices-Comparative studies and open issues," Journal of Network and Computer Application , 2015
- [3] B.Han, H.Lee, H.Jeong, "The HIGHT Encryption Algorithm," Internet Working group, June 24, 2011

- [4] Woo Kwon Koo, Hwa seong Lee, Yong Ho Kim, Dong Hoon Lee, "Implementation and Analysis of New lightweight cryptographic algorithm suitable for wireless sensor networks ," International Conference on Information Security and Assurance, IEEE, April 2008
- [5] Panasayya Yalla and Jens-Peter Kaps, "Lightweight Cryptographys for FPGA ," International Conference on Reconfigurable Computing and FPGAs, IEEE, Dec 2009, pp. 225–230.
- [6] Fernando Melo Nascimento , Fernando Messias dos Santos , Edward David Moreno, "A VHDL Implementation of the Lightweight Cryptographic Algorithm HIGHT ,"

AUTHORS

Sruthi. N did BTech from NSS College of Engineering in Electronics and Communication during 2010-2014 under University of Calicut. Doing MTech in VLSI DESIGN at Nehru College of Engineering (2014-2016) under University of Calicut.

R. Nandakumar working as Scientist 'C' at NIELIT, Calicut. ME in Communication Systems and MBA in Project Management. Area of specialisation includes VLSI DESIGN and Communication Engineering. Coordinator for PG Diploma VLSI Design, Coordinator for PG Diploma ESDM Resource Person for STTPs, Corporate & Industrial Training & Collaborative Workshops, IEEE NIELIT SB Counselor. In charge of NBA Accreditation

P.RajKumar has been working as Senior Assistant Professor at Nehru College of Engineering and Research Centre in the Electronics and Communication Engineering Department since June 2013. His educational qualifications include Master of Engineering (M.E) in the specialization Communication Systems and Bachelor of Engineering (B.E) in Electronics and Communication Engineering. His areas of research interest comprise Image Processing, Networks and VLSI Design.

INTENTIONAL BLANK

DESIGN OF IEEE 1149.1 TAP CONTROLLER IP CORE

Shelja A S¹, Nandakumar R² and Muruganantham C³

¹Department of Electronics and Communication Engineering, NCERC.
sheljaas@gmail.com

²Assistant scientist/engineer, N.I.E.L.I.T, Calicut
nanda@nielit.go.in

³Department of Electronics and Communication Engineering, NCERC
murugananthamc994@ncerc.ac.in

ABSTRACT

An implementation of IEEE 1149.1 TAP controller is presented in this paper. JTAG is an established technology and industry standard for on-chip boundary scan testing of SoCs. JTAG TAP controllers are becoming a delivery and control mechanism for Design For Test. The objective of this work is to design and implement a TAP controller IP core compatible with IEEE 1149.1-2013 revision of the standard. The test logic architecture also includes the Test Mode Persistence controller and its associated logic. This work is expected to serve as a ready to use module that can be directly inserted in to a new digital IC designs with little modifications.

KEYWORDS

IP core; IEEE 1149.1, TAP; TMP controller; JTAG; boundary scan; DFT

1. INTRODUCTION

A TAPC is probably the most common part used in support of on-chip testing. With the emergence of JTAG many SOC design will be using multiple TAPC for accessing internal logic during test. TAPC is a part of the IEEE 1149.1 standard. IEEE 1149.1, the standard for test access port and boundary scan architecture is a common platform for device, board and system level testing. The standard, popularly known as JTAG was originally introduced in the year 1990 and it is now a well established technology in the industry. A TAPC is probably the most common part used in support of on-chip testing. The original motivation for JTAG was boundary scan testing. Boundary scan is a method for gaining direct control of IO pins of a circuit at boundary of chip during test. This enables efficient testing on interconnection between devices that are mounted on a circuit board. The JTAG control is not limited at just the boundary of device it can also be used to gain access to internal structures during the test of device itself. It provides low cost technique for functional, and in- circuit testing that does not requires the test system to have direct access to each node.

The architecture of IEEE 1149.1 boundary scan includes a Test Access Port (TAP) interface, TAP controller (TAPC) logic, Boundary Scan Registers (BSR), Instruction Register (IR) and Test Data Registers (TDR). The IR and TDR form separate scan paths arranged between the Test Data Input (TDI) pin and Test Data Output (TDO) pin. This architecture allows the TAP to select and shift data through one of the path without accessing the other. When the test logic is active only one register is connected between the TDI and TDO interface depending on the value at Test Mode Select (TMS) signal. TCK is the test clock and is not synchronized with the system clock. A significant advantage of JTAG is that it requires only a minimum set of test access pins as it uses a serial interface. It facilitates design reuse and provides a standard protocol on-chip testing.

But when this standard is used for boundary-Scan chained device is put into test mode where their I/Os are completely under control of the Boundary register content. But when the chained TAPs pass through the TLR state; these instructions are replaced with not test mode instructions. The I/O pins then revert to being connected to the internal device logic which will be in an unknown state. The results of this reconnection are unpredictable [3].

The 2013 revision of the standard consider this issue and suggest an optional controller for avoiding disruptions caused to the device. An attempt is made to study the effect of the new controller to the test logic of the standard.

2. TAP CONTROLLER

TAP controller is a synchronous machine which provides access to the device under test and controls the behaviour of test logic using its 4-wired interface. It is a 16 state FSM which generates clocks and control signals to the associated test logic. Figure 1 shows the top level architecture of TAPC. Test clock TCK and mode select signal TMS controls the operation of TAPC. An optional TRSTN pin may be used to asynchronously reset the test logic if required and it is active low. A reset of the test logic can also be achieved within five TCKs or less by setting the TMS input high.

The TAPC will be initialized to test logic reset state at the power up. State transitions occur on the rising edge of TCK based on the value of TMS. The FSM has two scan paths for data transmission: one for instruction scan and other for data scan. The state diagram includes six steady states: Test-Logic-Reset, Run-Test/Idle, Shift-DR, Pause-DR, Shift-IR, and Pause-IR. To load and execute a new instruction FSM control is moved to the Select IR-Scan state, from where, it moves through the various states, Capture-IR, Shift-IR, and Update-IR, as required. The last operation is the Update-IR operation and the instruction loaded into the shift section of the Instruction register is latched to the Instruction register to become the new current instruction. This causes the Instruction register to be deselected as the register connected between TDI and TDO and the Data register identified by the new current instruction to be selected as the Data register between TDI and TDO. From now, one can manipulate the data register with the generic signals; Capture-DR, Shift-DR, and Update-DR control signals. TCK can be stopped in either a high or low state without loss of data.

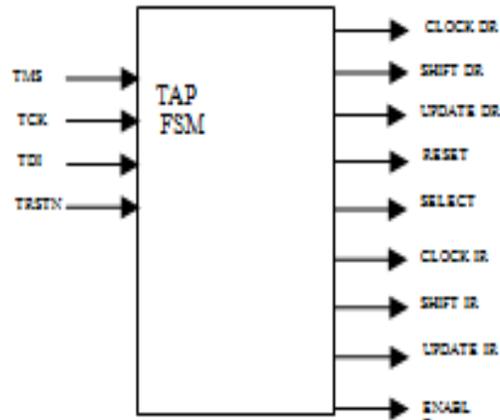


Fig 1. TAP top level

The functional table for the implemented example is given in table 1 with the encoding. The encoding used here is just an example. Different encoding schemes may be used.

Table 1. Functional table of TAP

STATE OF TAP	STATE ENCODING	PRESENT STATE	NETSTATE		OUTPUT								
			X=0	X=1	CLK DR	SHIFT DR	UPDATE DR	CLK IR	SHIFT IR	UPDATE IR	RESET	SELECT	ENABLE
TEST LOGIC RESET	F	1111	1100	1111							0	1	0
RUN TESTIDLE	C	1100	1100	0111							1	1	0
SELECT DR SCAN	7	0111	0110	0100							1	0	0
CAPTURE DR	6	0110	0010	0001	1						1	0	0
SHIFT DR	2	0010	0010	0001	1	1					1	0	1
EXIT1 DR	1	0001	0011	0101							1	0	0
PAUSE DR	3	0011	0011	0000							1	0	0
EXIT2 DR	0	0000	0010	0110							1	0	0
UPDATE DR	5	0101	1100	0111			1				1	0	0
SELECT IR SCAN	4	0100	1110	1111							1	0	0
CAPTURE IR	E	1110	1010	1001				1			1	1	0
SHIFT IR	A	1010	1010	1001				1	1		1	1	1
EXIT1 IR	9	1001	1011	1110							1	1	0
PAUSE IR	B	1011	1011	1000							1	1	0
EXIT2 IR	8	1000	1010	1101							1	1	0
UPDATE IR	D	1101	1100	0111						1	1	1	0

2.1. Simulation Results

The design is simulated using Modelsim simulator and the waveform obtained is shown in figure 2. The state transition and output signals of TAPC can be verified by applying the following exhaustive test pattern to the TMS input:

1011000100010000110001000010001000110011

IC Reset (optional): Provide test control of system reset and related inputs through TAP.

Power domain control (optional): to support multiple power domains in a system having a single TDR, an optional standard TAP to TDR interface is recommended that allows for segmentation of test data registers. The concept of register segments allows for segments that may be excluded or include.

Procedural Description Language (PDL) (optional): a new executable description language to document test procedures unique to a component.

4. IP CORE ARCHITECTURE

The demand for more powerful products and the increasing capacity of today's silicon technology have moved the design methodology to the system abstraction level. The integration technology supports the integration of a complete system in silicon (System-on-chip) and design methodologies are more and more based on pre-defined and pre-designed Intellectual Property blocks (IP-core). The reusing of IP-cores has been an alternative to reduce the increasing gap between design productivity and chip complexity of emerging SoC designs [7]. [7] suggests a structured method for IP core design.

Figure 3 shows the revised test logic architecture. It includes the optional TMP controller (TMPC) and the associated TMP status register as per the specification in IEEE 1149.1-2013 revision. The IR used here is a four bit shift register. Only bypass register and TMP status register is the implemented data registers. Boundary scan registers and other optional registers are not implemented in this architecture.

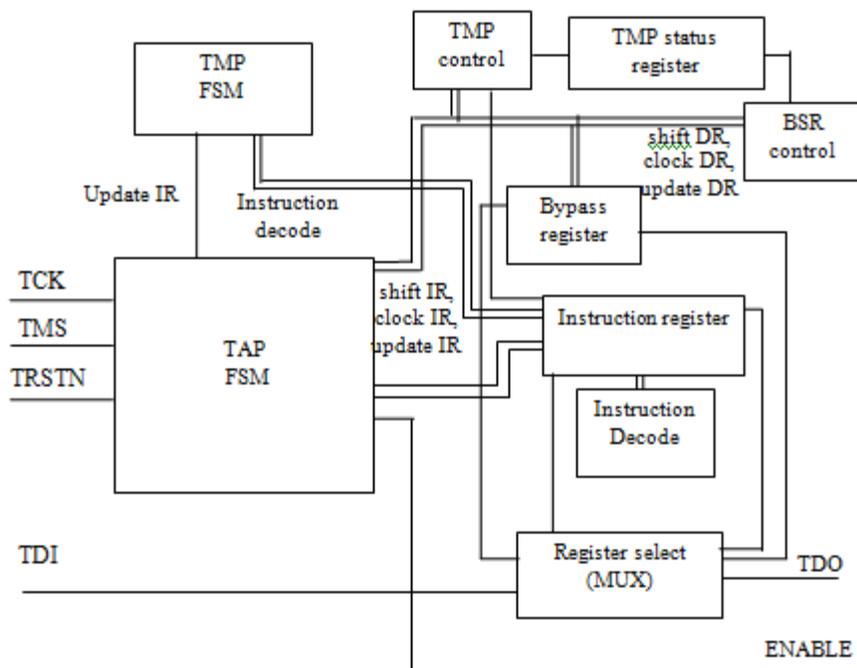


Fig .3 Conceptual schematic of the IP core

4.1. TMP Controller

TMPC is a synchronous state machine which keeps a device in test mode persistently Irrespective of the state of the remaining test logic. During board or system test TMP controller provides control over which components are in test mode and which are not, independent of the active instruction [1]

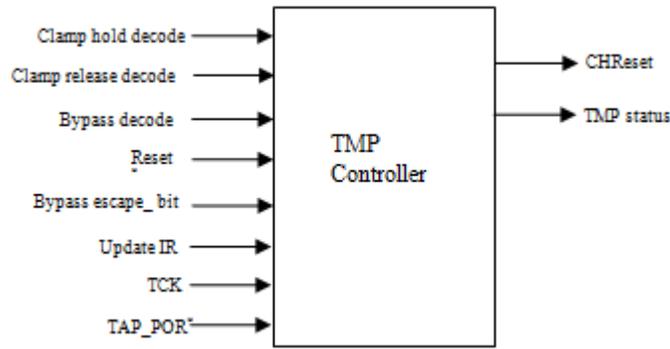


Fig 4. TMP controller top module

The TMP controller is an FSM, as shown in figure 5, with two states: persistence on and persistence off. State of the controller is determined by decode signals from instruction register and TAP controller output signals. Asynchronous reset options, either TRSTN or POR*, must be provided to reset the controller.

Table 2. TMP controller signals

Pin name	Type	Signal origin	Function
Clamp hold decode	In	From IR	To set TMP controller to persistence on mode
Clamp release decode	In	From IR	To set TMP controller to persistence off mode
Bypass decode	In	From IR	Sets the bypass escape bit to 1 together with the update signal
Update IR	In	From TAP	State of TAP which sets the bypass escape bit
Reset*	In	From TAP	Reset signal from TAP. Generate chreset
Bypass_escape	In	From TMP status register	Allows a component to escape test mode
TAP_por*	In	Input pin	Asynchronous on-chip reset at Power up
Tck	In	Input pin	Clock
Chreset*	Out	To boundary register	Reset signal to boundary registers
TMP_status	Out	To boundary register	Indicates State of TMP controller

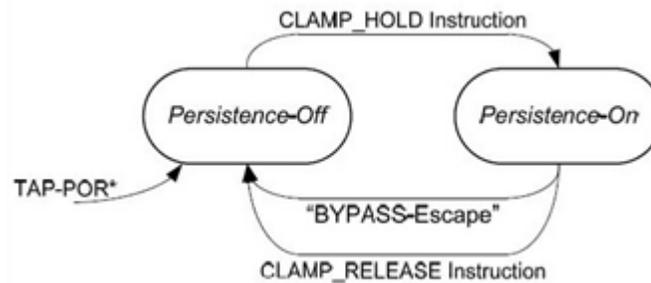


Fig 5. TMP controller state diagram

A possible implementation of TMPC is shown in figure 6. TMPC generates reset signals to control the boundary registers and other design specific registers. The TMPC together with the initialisation instruction allows safe switching between mission mode and test mode of system being tested.

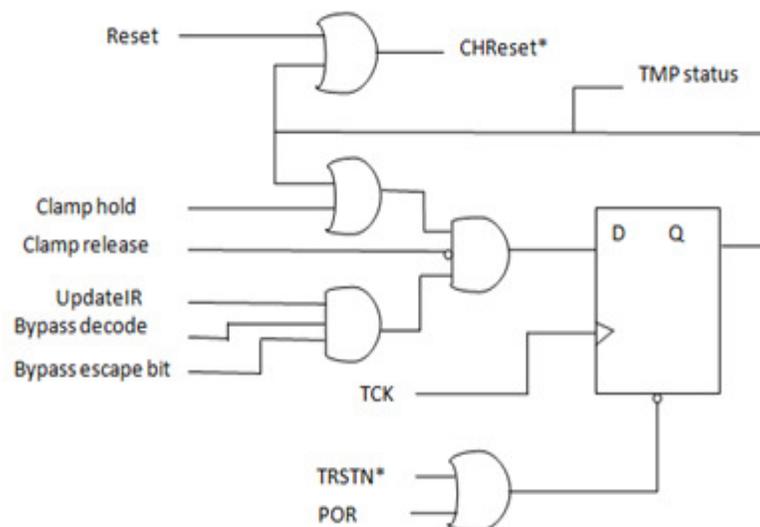
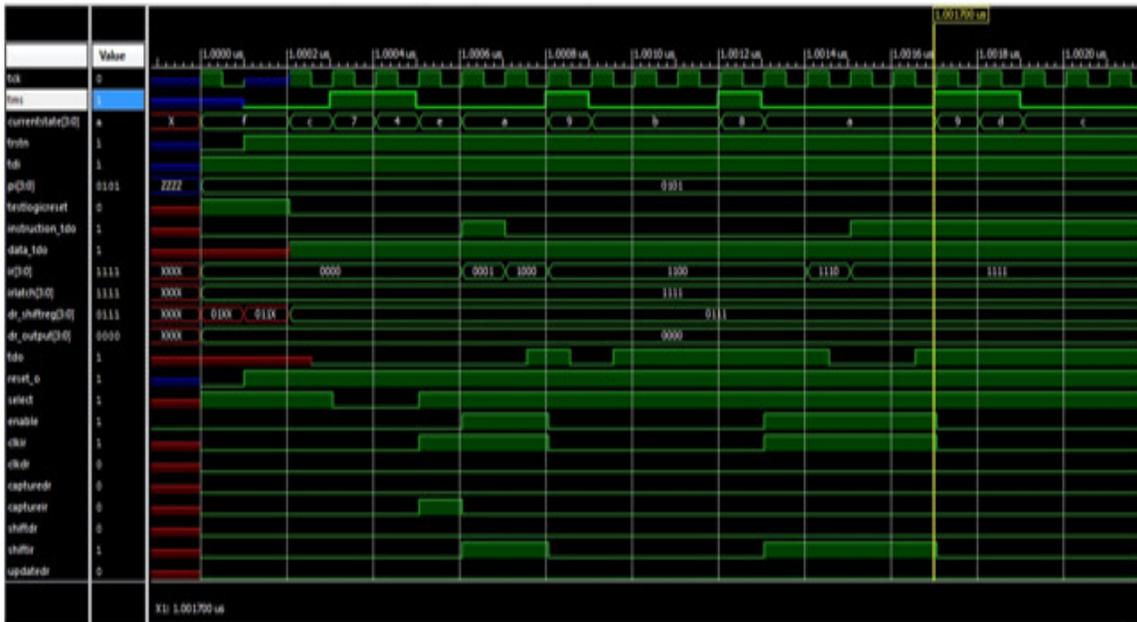


Fig 6. TMP controller example implementation

4.2. TAP Instruction Scan

Instructions may be loaded in to TAP controller's IR by traversing the TAP state machine down to Update-IR state. At this state, the current contents of the register are shifted out (by TDO) as new value is shifted in. However, the value shifted out will always be fixed with 01 as last two bits, as mandated by the IEEE 1149.1 standard for use in testing the functionality of the JTAG interface. The procedure for shifting a bit in and out of either shift register (IR or DR) is identical; the TAP state machine must be in the Shift-IR or Shift-DR state, respectively. The input bit (at TDI) is sampled by the TAP controller on the rising edge of each TCK cycle, while the output bit is driven out on the falling edge of the TCK cycle.

For each bit in the transfer, except the final bit, TMS must remain low. This is important so that, as each bit is shifted in and shifted out, the state machine remains in the Shift-IR



Instruction shift operation at IR shift register (ir [3:0]) currentstate [3:0] is at shift_ir state (b)

FIG 7. TAP IR scan

4.3. TAP Data Scan

Shifting values into and out of the DR of the TAP controller is performed in a similar manner to that of the IR. The simulation result of data scan is given in figure 8.

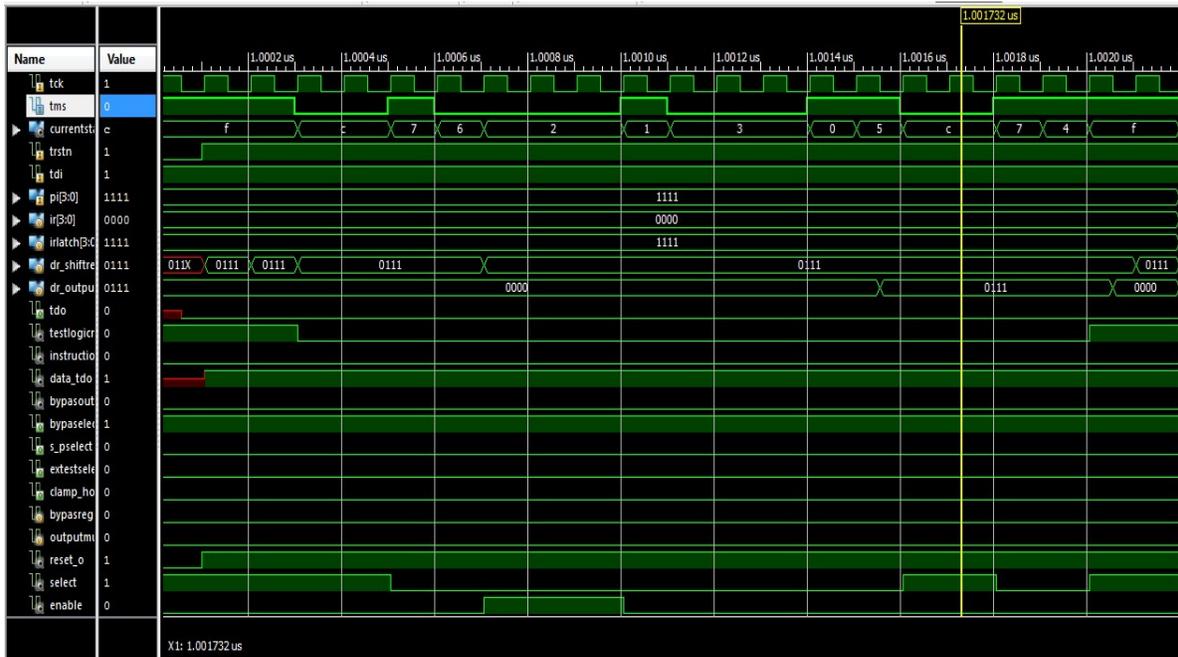


Fig 8. TAP DR scan

4.4. Simulation Results of Test Logic

The test logic is simulated using Xilinx ISE design suite and the result is shown in figure 9.

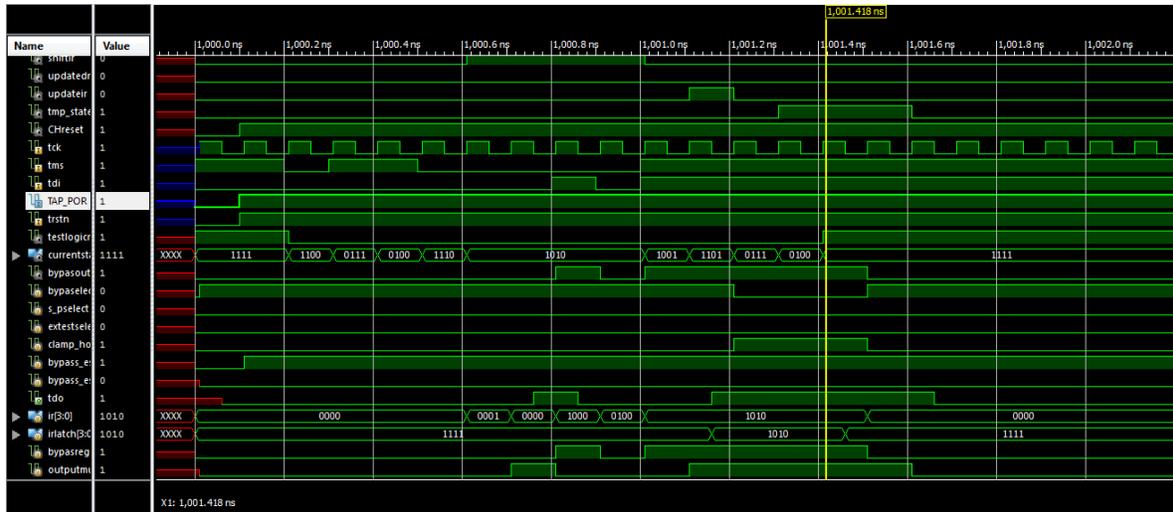


Fig 9. Test logic simulation waveform

5. IMPLEMENTATION RESULTS

Design is implemented in Xilinx XC6LX16-CS324 evaluation board and results are analyzed. The Xilinx nexys evaluation board is used here for physical design automation (floor planning, placement and routing). The RTL design is also analyzed using Cadence Encounter™ RTL Compiler. Cadence RTL Compiler is a powerful tool for logic synthesis and analysis for digital designs. The FPGA utilization of TAPC is shown in table 3.

Table 3. Resource utilization of TAP controller

Logic Utilization	Used	available	utilization
Number of slice flip flops	4	1536	1%
Number of 4 input LUTs	13	11536	1%
Number of occupied slices	7	768	1%
Number of bonded IOBs	18	63	28%
IOB latches	7		
Number of BUFGMUXs	1	8	12%
Average fan-out of Non-clock nets	7.00		

Table 4. Resource utilization of IP core

Logic Utilization	Used	available	utilization
Number of slice flip flops	17	8224	1%
Number of 4 input LUTs	19	9112	1%
Number of occupied slices	10	2278	1%
Number of bonded IOBs	13	232	5%
Number of BUFGMUXs	1	16	6%

The area in terms of the cells used is shown in figure 10.

```

rc:/> report area
=====
Generated by:      Encounter(R) RTL Compiler RC14.10 - v14.10-p008_1
Generated on:     May 04 2016 10:52:40 am
Module:           tappp
Technology library: tsmc18 1.0
Operating conditions: slow (balanced_tree)
Wireload mode:   enclosed
Area mode:       timing library
=====

Instance  Cells  Cell Area  Net Area  Total Area  Wireload
-----
tappp     35     652        0         652        <none> (D)

(D) = wireload is default in technology library

```

Fig 10 Area consumption of TAPC

Alternatively the area may also be represented using number of gate equivalent (GE). The GE is an estimation of hardware design complexity independent from circuit realisation and fabrication technology. One approach to calculate the GE is by dividing design area over area of one GE. Using the values from fig.5 the approximate GE of TAPC is obtained to be 66

```

rc:/> report gates

Module:           tappp
Technology library: tsmc18 1.0

  Gate      Instances  Area
-----
AND2X1           4   53.222
AOI21XL          1   13.306
AOI22X1          1   16.632
AOI2BB1X1        1   16.632
DFFSXL           3  189.605
DFFSXL           1   63.202
INVX1            1    6.653
INVXL            1    6.653
NAND2X1          4   39.917
NAND3BXL         1   16.632
NAND4BXL         1   19.958
NOR2BXL          1   13.306
NOR2X1           6   59.875
NOR2XL           2   19.958
OAI21XL          4   53.222
OAI222XL         1   26.611
OAI22X1          1   19.958
OAI2BB1X1        1   16.632
-----
total            35  651.974

  Type      Instances  Area  Area %
-----
sequential      4  252.806  38.8
inverter         2   13.306   2.0
logic            29  385.862  59.2
-----
total            35  651.974 100.0

```

Fig 11. Gate count of TAP

Power usage of TAPC in terms of leakage and dynamic power dissipation is shown in figure 12. The leakage power gives the static power dissipation during quiescent condition and dynamic power is the power usage during its normal operation.

```

rc:/> report power
=====
Generated by:      Encounter(R) RTL Compiler RC14.10 - v14.10-p008_1
Generated on:     May 04 2016 10:52:57 am
Module:          tappp
Technology library: tsmc18 1.0
Operating conditions: slow (balanced_tree)
Wireload mode:   enclosed
Area mode:       timing library
=====

          Leakage Dynamic Total
Instance Cells Power(nW) Power(nW) Power(nW)
-----
tappp      35    17.512 54645.195 54662.707

```

Fig 12. Power usage of TAPC

Power usage of IP core using XPower analysis tool in terms of leakage and dynamic power dissipation is shown in figure 13. The leakage power gives the static power dissipation during quiescent condition and dynamic power is the power usage during its normal operation.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	
Device		On-Chip	Power (W)	Used	Available	Utilization (%)			Supply	Summary	Total	Dynamic	Quiescent	
Family	Spartan6	Clocks	0.000	1	--	--			Source	Voltage	Current (A)	Current (A)	Current (A)	
Part	xc6sxl6	Logic	0.000	19	9112	0			Vccint	1.200	0.006	0.000	0.006	
Package	csj324	Signals	0.000	31	--	--			Vccaux	2.500	0.003	0.000	0.003	
Grade	C-Grade	IOs	0.000	13	232	6			Vcco25	2.500	0.002	0.000	0.002	
Process	Typical	Leakage	0.020											
Speed Grade	-2	Total	0.020											
											Supply Power (W)	Total	Dynamic	Quiescent
												0.020	0.000	0.020
Environment			Thermal Properties			Effective TJA	Max Ambient	Junction Temp						
Ambient Temp (C)	25.0		(C/W)	(C)	(C)	27.8	84.4	25.6						
Use custom TJA?	No													
Custom TJA (C/W)	NA													
Airflow (LFM)	0													
Heat Sink	None													
Custom TSA (C/W)	NA													
Characterization														
Production	v1.3.2011-05-04													
The Power Analysis is up to date.														

Fig. 13. Power usage of IP core

6. CONCLUSIONS

The thesis proposes design and implementation of the IEEE 1149.1-2013 TAP controller IP core. The design is synthesised using Xilinx® ISE and implemented in Xilinx nexys 3 evaluation board. The proposed work is fully compatible with the standard and provides a reusable module with a robust and easily testable design. This work is expected to serve as a ready to use module that can be directly inserted in to a new digital IC designs with little modifications.

REFERENCES

- [1] IEEE Standard 1149.1-2013, "Standard Test Access Port and Boundary-Scan Architecture"
- [2] IEEE Standard 1149.1-2001, "Standard Test Access Port and Boundary-Scan Architecture".
- [3] Kenneth P. Parker, "Surviving State Disruptions Caused by Test: the "Lobotomy Problem"," IEEE International Test Conference 2010.

- [4] Kenneth P. Parker, David Dubberke, Shuichi Kameyama, “Surviving State Disruptions Caused by Test: A Case Study,” Keysight Technologies, August, 2014.
- [5] David B. Lavo, “A Good Excuse for Reuse: “Open” TAP Controller Design,” ITC International Test Conference, 2000, p. 1090-1099.
- [6] Dave Stang, and R. Dandapani, “An implementation of IEEE 1149.1-To avoid violation and other practical In Compliance”, IEEE-2002.
- [7] Lima, M., F. Santos, J. Bione, T. Lins, and E. Barros, “ ipPROCESS: A Development Process for Soft IP-core with Prototyping in FPGA “, Forum on Design Languages (FDL 2005), Swiss, Sept. 2005

BIG GRAPH: TOOLS, TECHNIQUES, ISSUES, CHALLENGES AND FUTURE DIRECTIONS

Dhananjay Kumar Singh¹ and Ripon Patgiri²

Department of Computer Science & Engineering
National Institute of Technology Silchar
Assam, India-788010

¹eng.dksingh@gmail.com

²ripon@cse.nits.ac.in

ABSTRACT

Analyzing interconnection structures among the data through the use of graph algorithms and graph analytics has been shown to provide tremendous value in many application domains (like social networks, protein networks, transportation networks, bibliographical networks, knowledge bases and many more). Nowadays, graphs with billions of nodes and trillions of edges have become very common. In principle, graph analytics is an important big data discovery technique. Therefore, with the increasing abundance of large scale graphs, designing scalable systems for processing and analyzing large scale graphs has become one of the timeliest problems facing the big data research community. In general, distributed processing of big graphs is a challenging task due to their size and the inherent irregular structure of graph computations. In this paper, we present a comprehensive overview of the state-of-the-art to better understand the challenges of developing very high-scalable graph processing systems. In addition, we identify a set of the current open research challenges and discuss some promising directions for future research.

KEYWORDS

Big Data, Big Graph, Graph Processing, Graph Analytics, Graph Parallel Computing, Distributed Processing, Graph Algorithms

1. INTRODUCTION

The Big Data delineates huge data set to store, process and analyze. These data are generated by the daily aggrandizement of data from various sources, call for Big Data to knob these perplex data. Therefore, the conventional means of handling data are now obsolete, emerging Big Data with full-fledged, and qui vive for research in these data. Therefore, NoSQL is a prominent field in Big Data. The Big Graph is part of NoSQL which is gigantic in size, perplex to handle, and arduous to visualize.

Recently people, devices, processes and other entities have been more connected than at any other point in history. In general, the complex relationships, interactions and interdependencies Natarajan Meghanathan et al. (Eds) : ACITY, VLSI, AIAA, CNDC - 2016 pp. 119–128, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60911

between objects are naturally modelled as graphs. Therefore, graphs have been used to represent data sets in a wide range of application domains, such as social science, astronomy, computational biology, telecommunications, semantic web, protein networks, and many more. In practice, graph analytics is an important and effective big data discovery tool. For example, it enables identifying influential persons in a social network, inspecting fraud operations in a complex interaction network and recognizing product affinities by analyzing community buying patterns.

Nowadays, graphs with millions and billions of nodes and edges have become very common. For example, in 2012, Facebook has reported that its social network graph contains more than a billion users (nodes) and more than 140 billion friendship relationships (edges). The enormous growth in graph sizes requires huge amounts of computational power to analyze. In practice, distributed processing of large scale graphs is a challenging task due to their size in addition to their inherent irregular structure and the iterative nature of graph processing and computation algorithms. Graph algorithms are becoming increasingly important for analyzing large datasets in many fields. Real-world graph data follows a pattern of sparsity that is ¹not uniform, but highly skewed towards a few items. Implementing graph traversal, statistics and machine learning algorithms on such data in a scalable manner is quite challenging. As a result, several graph analytics frameworks such as Giraph, FlashGraph, GraphChi, X-Stream and many more, have been developed, each offering a solution with different programming models and targeted at different users.

The rest of this paper is organized as follows: Section 2 provides basic information about Big Graph. In Section 3, we discussed about Big Graph processing systems like Apache Giraph, GPS, and many more. We present Big Graph Analytics frameworks like Ringo, PowerGraph and so on in Section 4. Graph Algorithms for solving many problems in scientific computing, data mining and other domains, are discussed in Section 5. The future directions of Big Graph are discussed in Section 6. And finally, Section 7 concludes the paper.

2. BIG GRAPH

We can simply define Big Graph as,

“Big Data + Structure = Big Graph”

Big graphs are ubiquitous, ranging from social networks and mobile call networks to biological networks and the World Wide Web. The sources of real-world large-scale graphs include:

- Social graphs (Facebook, Twitter, Google+, LinkedIn, etc.)
- Endorsement graphs (web link graph, paper citation graph, etc.)
- Location graphs (map, power grid, telephone network, etc.)

The size of large scale graphs used in the recent literature is given in table 1.

¹<https://www.ibm.com/developerworks/library/os-giraph>

Table 1: Large Scale Graphs in current literature¹

Name	Nodes	Edges
Web Graph	More than 20 billion nodes (pages)	More than 160 billion edges (hyperlinks)
Facebook	More than a billion nodes (users)	More than 140 billion edges (friendship relationships)
LinkedIn	Almost 8 million nodes	Almost 60 million edges
SemanticWeb	3.7 million nodes (objects)	400 million edges (facts)

3. BIG GRAPH PROCESSING

The growth of graph-structured data in modern applications such as social networks and knowledge bases creates a crucial need for scalable platforms and parallel architectures that can process it in bulk.

3.1. Pregel

Pregel[3] is a scalable, general-purpose system for implementing graph algorithms in a distributed environment. It is known as the first Bulk Synchronous Parallel (BSP), an implementations that provides a native API specifically for graph algorithms using a “think like a vertex” computing paradigm. The basic computation model of Pregel is shown in figure 1.

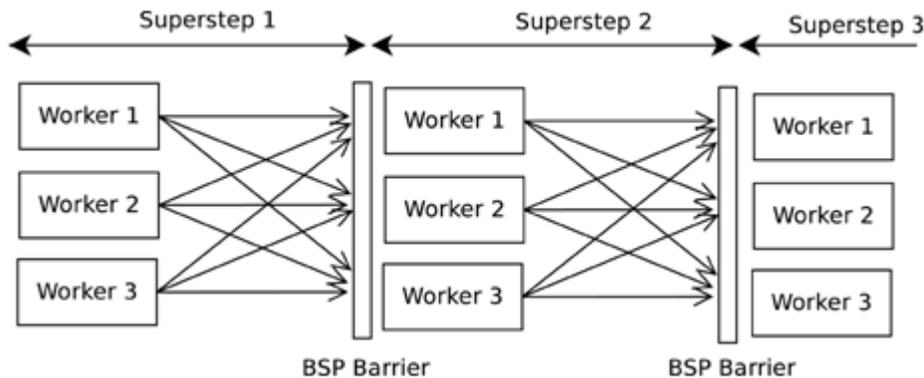


Figure 1. Pregel Computation Model with three supersteps and three workers[4].

In Pregel, programs are expressed as a sequence of iterations (supersteps), in each of which a vertex can receive messages sent in the previous iteration, send messages to other vertices, and modify its own state and that of its outgoing edges. The input graph is loaded once at the start of a program and all computations are executed in-memory.

3.2. Giraph

Apache Giraph[1] is an iterative graph processing system built for high scalability. It runs workers as map-only jobs on Hadoop and uses HDFS for data input and output. Giraph adds several features including master computation, sharded aggregators, edge-oriented input, out-of-

core computation, and more. It also uses Apache ZooKeeper for coordination, checkpointing, and failure recovery schemes. With a steady development cycle and a growing community of users worldwide, Giraph is a natural choice for unleashing the potential of structured datasets at a massive scale. For example, it is currently used at Facebook to analyze the social graph formed by users and their connections.

3.3. Mizan

Mizan[8] is a scalable framework for supporting graph mining algorithms in large parallel computing infrastructures. It is a layer between the users' code and a variety of computing infrastructures, such as Linux clusters, cloud environments and supercomputers. Mizan examines the graph structure and decides the placement of the data and the low level message passing mechanism transparently to the users' code. It uses message passing and implements an instance of the Bulk Synchronous Parallel model.

3.4. GPS

GPS[9] is an open-source system for scalable, fault-tolerant, and easy-to-program execution of algorithms on extremely large graphs. It is a distributed system, designed to run on a cluster of machines, such as Amazon's EC2. GPS offers Large Adjacency List Partitioning (LALP), an optional performance optimization of algorithms that send to all of its neighbors the same message. GPS also features an optional dynamic migration scheme. Dynamic migration repartitions the graph during the computation by migrating vertices between workers, to improve workload balance and network usage.

3.5. GraphLab

GraphLab[11] is a framework for asynchronous parallel graph computations in machine learning. It differs from Pregel in that it does not work in bulk synchronous steps, but rather allows the vertices to be processed asynchronously based on a scheduler. The vertex functions can run at any time as long as specified consistency rules are obeyed. It is therefore well-suited for the machine learning types of applications for which it is defined, where each vertex accumulates information from its neighbors' states and updates its state, possibly asynchronously. It extends the shared memory GraphLab abstraction to the distributed setting by refining the execution model, relaxing the scheduling requirements, and introducing a new distributed data-graph, execution engines, and fault-tolerant systems.

3.6. Graph Sample and Hold

Graph Sample and Hold (gSH)[7] is a stream sampling framework that supports analytics over big graphs. The nice property of gSH consists in building theoretically sound unbiased estimators derived from the sample graph, but still robust for the estimation of different properties of the target (big) graph, yet conveying in high accuracy and tolerable approximation errors.

3.7. Pregelix

Pregelix[6] is a dataflow-based Pregel-like system built on top of the Hyracks parallel dataflow engine. It combines the Pregel API from the systems world with data-parallel query evaluation

techniques from the database world in support of large scale graph analytics. This combination leads to effective and transparent out-of-core support, scalability, and throughput, as well as increased software simplicity and physical flexibility. To the best of our knowledge, Pregelix is the only open source Pregel-like system that scales to out-of-core workloads efficiently, can sustain multi-user workloads, and allows runtime flexibility.

4. BIG GRAPH ANALYTICS

Analytics is the ability to discover meaningful patterns and interesting insights into data. Graph analytics is a special piece of analytics where the underlying data can be modeled as a set of graphs. Graph analytics is a rapidly developing area where a combination of graph-theoretic, statistical and database techniques are applied to model, store, retrieve, and perform analyses on graph-structured data.

4.1. In-Memory Big Graph Analytics

PowerGraph: A scalable, distributed graph computation framework written in C++. PowerGraph[12] supports both the highly-parallel bulk-synchronous Pregel model of computation as well as the computationally efficient asynchronous GraphLab model of computation. PowerGraph exploits the Gather-Apply-Scatter (GAS) model of computation to factor vertex-programs over edges, splitting high-degree vertices and exposing greater parallelism in natural graphs. It allows vertex-partitioning to effectively place its large scale graph in a distributed environment.

GraphX: GraphX[14] enables distributed dataflow frameworks such as Spark to naturally express and efficiently execute iterative graph algorithms. To achieve performance parity with specialized graph systems, GraphX recasts graph-specific optimizations as distributed join optimizations and materialized view maintenance. By leveraging advances in distributed dataflow frameworks, GraphX brings low-cost fault tolerance to graph processing. GraphX API enables the composition of graphs with unstructured and tabular data and permits the same physical data to be viewed both as a graph and as collections without data movement or duplication.

Ringo: A system for construction and analysis of large scale graphs on a single large memory multicore machine that combines high productivity analysis with fast and scalable execution times. Ringo [10] table operations, transformations between tables and graphs, and several graph algorithms are fully parallelized to take full advantage of the multi-core environment, and the set of graph algorithms available for parallel execution is under constant expansion.

4.2. SSD-Based Big Graph Analytics

FlashGraph: It stores vertex state in memory and edge lists on SSDs. FlashGraph[19] runs on top of the set-associative file system (SAFS), a user-space filesystem designed to realize both high IOPS, and lightweight caching for SSD arrays on non-uniform memory and I/O systems. It uses an asynchronous user-task I/O interface to reduce overhead associated with accessing data in the filesystem and overlap computation with I/O. FlashGraph selectively accesses edge lists required by a graph algorithm from SSDs to reduce data access; it conservatively merges I/O requests to increase I/O throughput and reduce CPU consumption; it further schedules the order of processing vertices to help merge I/O requests and maximize the page cache hit rate.

4.3. Disk-Based Big Graph Analytics

GraphChi: A disk-based system for computing efficiently on graphs with billions of edges. GraphChi[13] is able to execute several advanced data mining, graph mining, and machine learning algorithms on very large graphs, using just a single consumer-level computer. It partitions the vertices into disjoint intervals and breaks large edge list into smaller shards containing edges with destinations in corresponding intervals. GraphChi uses a vertex-centric processing model, which gathers data from neighbors by reading edge values, computes and applies new values to the vertices, and scatters new data to neighbors by writing values on the edges.

X-Stream: An edge-centric graph processing system, uses streaming partitions to utilize the sequential streaming bandwidth of the storage medium for graph processing. X-Stream[17] introduces an edge-centric scatter-gather processing model. In the scatter phase, it streams every edge, and generates updates to propagate vertex states. In the gather phase, it streams every update, and applies it to the corresponding vertex state.

TurboGraph: A disk-based graph engine that process billion-scale graphs very efficiently by using modern hardware on a single PC. TurboGraph[15] is the first truly parallel graph engine that exploits full parallelism, including multi-core parallelism and FlashSSD IO parallelism, and full overlap of CPU processing and I/O processing as much as possible.

Chaos: A graph processing system designed for analytics on big graphs using small clusters. Chaos [16] builds on the X-Stream single-machine graph processing system, but scales out to multiple machines. It treats the aggregate storage of all machines as a single at disk and uses work stealing to balance the load across nodes in the cluster. With very limited pre-processing, Chaos achieves sequential storage access, computational load balance and I/O load balance.

GridGraph: A system for processing large-scale graphs on a single machine using 2-level hierarchical partitioning. GridGraph[18] breaks graphs into 1D-partitioned vertex chunks, and 2D-partitioned edge blocks using a first fine-grained level partitioning in preprocessing. It uses a new streaming-apply the model that streams edges sequentially and applies updates onto vertices instantly.

4.4. Issues and Challenges of Big Graph Analytics

High-degree vertex: Graphs with high-degree vertices are computationally challenging and contribute heavily communication and storage overhead. In addition, these graphs are difficult to partition.

Sparseness: Splitting sparse graph requires more communication, more computation and more synchronization.

Data-driven computations: Graph computations are often completely data-driven. The computations performed by a graph algorithm are dictated by the vertex and edge structure of the graph on which it is operating rather than being directly expressed in code. As a result, parallelism based on partitioning of computation can be difficult to express because the structure of computations of the algorithm is not known apriori.

Unstructured problems: Similar to difficulties encountered in parallelizing a graph problem based on its computational structure, irregular structure of graph data makes it difficult to extract parallelism by partitioning the problem data.

In-memory challenge: The large-scale graph does not fit in a single memory location, because of its immense in size. Instead of SSD or HDD, the graph data should reside in the RAM, such that the response time would become minimal.

Poor locality: Because graphs represent the relationships between entities and because these relationships may be irregular and unstructured, the computations and data access patterns tend not to have very much locality. Performance in contemporary processors is predicated upon exploiting locality. Thus, high performance can be hard to obtain for graph algorithms, even on serial machines.

Communication overhead: The high-degree vertices incur communication overheads. Today, the high-degree vertices are in millions, but tomorrow will be in the billions, and beyond.

Load balancing: When we analyze graphs such as power-law graphs, then we have to pay extra attention to load balancing.

4.5. Applications of Big Graph Analytics

Graph analytics has wide ranging applications in many diverse domains such as Internet and overlay management, road networks, online social networks, etc.

4.5.1. Machine Learning

One of the most popular application of machine learning is recommendation systems. One approach to the design of recommendation systems that has wide use is collaborative filtering. The Netflix movie recommendation task uses collaborative filtering to predict the movie ratings for each user, based on the ratings of similar users.

4.5.2. Social Network Analysis

Graphs are employed heavily in online social networks. The reason for this popularity is that graphs offer a natural way of representing various kinds of relationships that are important for these applications.

4.5.3. Semantic Networks

A semantic network is a graph structure for representing knowledge in patterns of interconnected nodes and arcs. Declarative graphic representation is common to all semantic networks that can be used to represent knowledge and support automated systems for reasoning about the knowledge.

4.5.4. Biological Networks

Interactions arise naturally in biology and it can be assembled into networks of graphs where the nodes are biological entities and edges represent molecular interactions, associations between diseases.

4.5.5. Friend Recommendations

Existing social networking services recommend friends to users based on their social graphs, which may not be the most appropriate to reject a user's preferences on friend selection in real life.

5. GRAPH ALGORITHMS

Graph algorithms are becoming increasingly important for solving many problems in scientific computing, data mining and other domains.

5.1. PageRank

PageRank[20] is a method for computing a ranking for every web page based on the graph of the web. It is used by Google to rank webpages based on the idea that more important websites likely receive more links from other websites. PageRank has applications in search, browsing, and traffic estimations.

5.2. Connected Components

A connected component, in an undirected graph, is a connected subgraph of the graph. In a directed graph, connected component can either be weakly connected or strongly connected. Weakly and strongly connected component are respectively weakly and strongly connected subgraphs of a graph.

5.3. Single Source Shortest Path (SSSP)

The single-source shortest path problem is a classical problem in the research field of graph algorithm. In SSSP problem, we have to find the shortest paths from a source vertex 'v' to all other vertices in a graph.

5.4. Triangle Counting

In this algorithm, each vertex shares its neighborhood list with each of its neighbors. Each vertex, then checks if any of their neighbors overlap with the neighborhood list(s) they received. With directed edges and no cycles, the total number of such overlaps gives the number of triangles in the graph. With undirected edges, the total number of overlaps gives 3-times the number of triangles.

5.5. Collaborative Filtering

It is used, for example, to recommend products based on purchases of other users with similar interests. This is a machine learning algorithm that estimates how a given user would rate an item given an incomplete set of (user, item) ratings.

6. FUTURE DIRECTION

In the era of big data, interest in analysis and extraction of information from large data graphs is increasing rapidly. Graphs are now widely used for data modeling in application domains for which identifying relationship patterns, rules, and anomalies are useful. These domains include web graph, social networks, semantic web, protein-protein interaction networks, bibliographical networks, etc. The ever-increasing size of graph-structured data for these applications creates a critical need for scalable systems that can process large amounts of it efficiently.

7. CONCLUSION

The usage of large scale graph processing platforms is rapidly expanding in both academia and industry. In principle, large scale graph processing platforms are increasingly important as more and more problems require dealing with graphs. To this end, we presented a thorough survey of the state-of-the-art of the emerging platforms in this domain. In addition, we have provided an overview of the recent studies for benchmarking and evaluating some of the existing platforms. Finally, we identified and presented a set of the current open research challenges and also presented some of the promising directions for future research. In general, we believe that there are still many opportunities for new innovations and optimizations in the domain of large scale graph processing. Hence, we consider this article as an important step in helping researchers to understand the domain and guiding them towards the right direction to improve the state-of-the-art.

REFERENCES

- [1] Apache Giraph. <http://giraph.apache.org/>
- [2] Twitter: Most Followers. <http://friendorfollow.com/twitter/most-followers/>
- [3] Grzegorz Malewicz, Matthew H. Austern, Aart J.C Bik, James C. Dehnert, Ian Horn, Naty Leiser, and Grzegorz Czajkowski. Pregel: A System for Large-Scale Graph Processing. In *Proceeding of SIGMOD'10*, Pages 135-146, June 6-11, 2010.
- [4] Minyang Han, Khuzaima Daudjee, Khaled Ammar, M. Tamer zsu, Xingfang Wang, and Tianqi Jin. An Experimental Comparison of Pregel-like Graph Processing Systems. In *VLDB'14*, Volume 7 Issue 12, Pages 1047-1058, August 2014.
- [5] Avery Ching, Sergey Edunov, Maja Kabiljo, Dionysios Logothetis, and Sambavi Muthukrishnan. One Trillion Edges: Graph Processing at Facebook-Scale. In *VLDB'15 Endowment*, Volume 8 Issue 12, Pages 1804-1815, August 2015.

- [6] Yingyi Bu, Vinayak Borkar, Jianfeng Jia, Michael J. Carey, and Tyson Condie. Pregelix: Big(ger) Graph Analytics on A Dataow Engine. In Proceeding of VLDB Endowment, Volume 8 Issue 2, Pages 161-172, October 2014.
- [7] Ahmed, N.K., Du_eld, N.G., Neville, J., and Kompella, R.R. Graph Sample and Hold: A Framework for Big-Graph Analytics. In KDD'14, Pages 1446-1455, 2014
- [8] Zuhair Khayyat, Karim Awara, Amani Alonazi, Hani Jamjoom, Dan Williams, and Panos Kalnis. Mizan: A System for Dynamic Load Balancing in Large-scale Graph Processing. In EuroSys'13, Pages 169-182, April, 2013.
- [9] Semih Salihoglu and Jennifer Widom. GPS: A Graph Processing System. In SSDBM'13, Article 22, 2013. DOI=<http://dx.doi.org/10.1145/2484838.2484843>.
- [10] Yonathan Perez, Rok Sosi, Arijit Banerjee, Rohan Puttagunta, Martin Raison, Pararth Shah, and Jure Leskovec. Ringo: Interactive Graph Analytics on Big-Memory Machines. In SIGMOD'15, Pages 1105-1110, May 31-June 4, 2015.
- [11] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J. M. Hellerstein. Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud. In VLDB'12, Volume 5 Issue 8, Pages 716-727, April 2012.
- [12] Joseph E. Gonzalez, Yucheng Low, Haijie Gu, Danny Bickson, and Carlos Guestrin. PowerGraph: Distributed Graph-Parallel Computation on Natural Graphs. In OSDI'12, Pages 17-30, 08 October 2012.
- [13] Aapo Kyrola, Guy Blelloch, and Carlos Guestrin. GraphChi: Large-Scale Graph Computation on Just a PC. In OSDI'12, Pages 31-46, 08 October 2012.
- [14] Joseph E. Gonzalez, Reynold S. Xin, Ankur Dave, Daniel Crankshaw, Michael J. Franklin, and Ion Stoica. GraphX: Graph Processing in a Distributed Dataow Framework. In OSDI'14, Pages 599-613, 06 October 2014.
- [15] Wook-Shin Han, Sangyeon Lee, Kyungyeol Park, Jeong-Hoon Lee, Min-Soo Kim, Jinha Kim, and Hwanjo Yu. TurboGraph: A Fast Parallel Graph Engine Handling Billion-scale Graphs in a Single PC. In KDD'13, Pages 77-85, 11 August 2013.
- [16] A. Roy, L. Bindschaedler, J. Malicevic, and W. Zwaenepoel. Chaos: Scale-out Graph Processing from Secondary Storage. In SOSP'15, Pages 410-424, 2015.
- [17] Amitabha Roy, Ivo Mihailovic, and Willy Zwaenepoel. X-Stream: Edge-centric Graph Processing using Streaming Partitions. In SOSP'13, Pages 472-488, 2013.
- [18] Xiaowei Zhu, Wentao Han, and Wenguang Chen. 2015. GridGraph: Large-Scale Graph Processing on a Single Machine Using 2-Level Hierarchical Partitioning. In USENIX ATC'15, Pages 375-386, 2015.
- [19] Da Zheng, Disa Mhembere, Randal Burns, Joshua Vogelstein, Carey E. Priebe, and Alexander S. Szalay. FlashGraph: Processing Billion-Node Graphs on an Arrayof Commodity SSDs. In FAST'15, Pages 45-58, 2015.
- [20] L Page, S Brin, R Motwani, and T Winograd. The PageRank Citation Ranking: Bringing Order to the Web. In Stanford InfoLab, 1999.

AUTHOR INDEX

Bharat Bhushan 81

Deeptimanta Ojha 23

Dhananjay Kumar Singh 119

Durga Prasad Mohapatra 23

Keshav Kaushik 81

Liu Jianqiang 69

Manas Ranjan Patra 23

Marakonda Patnaikuni Vasanthi 41

Mohammad Alnuem 01

Mukesh Sukla 41

Muruganantham C 107

Nandakumar R 97, 107

Nandini V 53

Panda A K 41

Rajesh Kumar Sahoo 23

Rajkumar P 97

Rana Alosaimi 01

Ripon Patgiri 119

Ronak Shah 33

Sahoo G 81

Setua S.K 13

Shelja A S 107

Soumajit Adhya 13

Sowpati Santhi 41

Sriranjitha R 53

Sruthi N 97

Sun Shuting 69

Suresh M 41

Yazhini T. P 53

Zou Bin 69