Advances in Intelligent Systems and Computing 1089

Bibudhendu Pati
Chhabi Rani Panigrahi
Rajkumar Buyya
Kuan-Ching Li *Editors*

Advanced Computing and Intelligent Engineering

Proceedings of ICACIE 2018, Volume 2

Springer

Advances in Intelligent Systems and Computing

Volume 1089

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing such as: computational intelligence, soft computing including neural networks, fuzzy systems, evolutionary computing and the fusion of these paradigms, social intelligence, ambient intelligence, computational neuroscience, artificial life, virtual worlds and society, cognitive science and systems, Perception and Vision, DNA and immune based systems, self-organizing and adaptive systems, e-Learning and teaching, human-centered and human-centric computing, recommender systems, intelligent control, robotics and mechatronics including human-machine teaming, knowledge-based paradigms, learning paradigms, machine ethics, intelligent data analysis, knowledge management, intelligent agents, intelligent decision making and support, intelligent network security, trust management, interactive entertainment, Web intelligence and multimedia.

The publications within "Advances in Intelligent Systems and Computing" are primarily proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

**** Indexing: The books of this series are submitted to ISI Proceedings, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink ****

More information about this series at http://www.springer.com/series/11156

Bibudhendu Pati · Chhabi Rani Panigrahi ·
Rajkumar Buyya · Kuan-Ching Li
Editors

Advanced Computing and Intelligent Engineering

Proceedings of ICACIE 2018, Volume 2

Springer

Editors
Bibudhendu Pati
Department of Computer Science
Rama Devi Women's University
Bhubaneswar, Odisha, India

Chhabi Rani Panigrahi
Department of Computer Science
Rama Devi Women's University
Bhubaneswar, Odisha, India

Rajkumar Buyya
Cloud Computing
The University of Melbourne
Melbourne, VIC, Australia

Kuan-Ching Li
Department of Computer Science
and Information Engineering
Providence University
Taichung, Taiwan

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-981-15-1482-1 ISBN 978-981-15-1483-8 (eBook)
https://doi.org/10.1007/978-981-15-1483-8

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

This volume contains the papers presented at the 3rd International Conference on Advanced Computing and Intelligent Engineering (ICACIE) 2018: The 3rd International Conference on Advanced Computing and Intelligent Engineering (www.icacie.com) was held during 22–24 December 2018 at the Institute of Technical Education and Research (ITER), Siksha ‘O’ Anusandhan (Deemed to be University), Bhubaneswar, India, in collaboration with Rama Devi Women’s University, Bhubaneswar, India, during 22–24 December 2018. There were 528 submissions, and each qualified submission was reviewed by a minimum of two Technical Program Committee members using the criteria of relevance, originality, technical quality and presentation. The committee accepted 103 full papers for oral presentation at the conference, and the overall acceptance rate is 20%.

ICACIE 2018 was an initiative taken by the organizers which focuses on research and applications on topics of advanced computing and intelligent engineering. The focus was also to present the state-of-the-art scientific results, to disseminate modern technologies and to promote collaborative research in the field of advanced computing and intelligent engineering.

Researchers presented their work in the conference and had an excellent opportunity to interact with eminent professors, scientists and scholars in their area of research. All participants were benefitted from discussions that facilitated the emergence of innovative ideas and approaches. Many distinguished professors, well-known scholars, industry leaders and young researchers participated in making ICACIE 2018 an immense success.

We also had a panel discussion on the emerging topic entitled “*Intellectual Property and Standardisation in Smart City*” consisting of panellists from software industries like TCS, Infosys, educationalist and entrepreneurs.

We thank all the Technical Program Committee members and all reviewers/ sub-reviewers for their timely and thorough participation during the review process.

We express our sincere gratitude to Prof. Manojranjan Nayak, President, S ‘O’ A Deemed to be University and Chairman, Advisor Committee Prof. Damodar Acharya for their endless support towards organizing the conference. We extend our sincere thanks to Honourable Vice Chancellor Dr. Amit Banerjee for allowing us to organize

ICACIE 2018 on the campus and also for his timely support. We also express our sincere gratitude to Honourable Patron Vice Chancellor Prof. Padmaja Mishra, Rama Devi Women's University for her moral supports and ceaseless encouragement towards successful completion of the conference. We thank Prof. Pradeep Kumar Sahoo, Dean Faculty in Engineering and Technology, for his guidance in the local organization of ICACIE 2018. We also thank Prof. P. K. Nanda, Prof. Bibudhendu Pati and Prof. A. K. Nayak, General Chairs for their valuable guidance during review of papers as well as other aspects of the conference. We appreciate the time and efforts put in by the members of the local organizing team at Institute of Technical Education and Research (ITER), Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, India, especially the faculty members of the Department of Computer Applications, Computer Science and Information Technology, Computer Science and Engineering, student volunteers, administrative and hostel management staff, who dedicated their time and efforts to make ICACIE 2018 successful. We thank Er. Subhashis Das Mohapatra for designing and maintaining ICACIE 2018 web site.

Bhubaneswar, India	Bibudhendu Pati
Bhubaneswar, India	Chhabi Rani Panigrahi
Melbourne, Australia	Rajkumar Buyya
Taichung, Taiwan	Kuan-Ching Li

About This Book

The book focuses on theory, practice and applications in the broad areas of advanced computing techniques and intelligent engineering. This two volumes book includes 109 scholarly articles, which have been accepted for presentation from 528 submissions in the 3rd International Conference on Advanced Computing and Intelligent Engineering held at Institute of Technical Education and Research (ITER), Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, India, in collaboration with Rama Devi Women's University, Bhubaneswar, India, during 22–24 December 2018. The first volume of this book consists of 54 papers, and second volume contains 49 papers with a total of 103 papers. This book brings together academic scientists, professors, research scholars and students to share and disseminate their knowledge and scientific research works related to advanced computing and intelligent engineering. It helps to provide a platform for the young researchers to find the practical challenges encountered in these areas of research and the solutions adopted. The book helps to disseminate the knowledge about some innovative and active research directions in the field of advanced computing techniques and intelligent engineering, along with some current issues and applications of related topics.

Contents

About the Editors

Dr. Bibudhendu Pati is an Associate Professor and the Head of the Department of Computer Science at Rama Devi Women's University, Bhubaneswar, India. He received his Ph.D. degree from IIT Kharagpur. Dr. Pati has 21 years of experience in teaching, research, and industry. His areas of research include wireless sensor networks, cloud computing, big data, Internet of Things, and networks virtualization. He is a life member of the Indian Society of Technical Education (ISTE), a senior member of IEEE, member of ACM, and life member of the Computer Society of India (CSI). He has published several papers in leading international journals, conference proceedings and books. He is involved in organizing various international conferences.

Dr. Chhabi Rani Panigrahi is an Assistant Professor at the Department of Computer Science at the Rama Devi Women's University, Bhubaneswar, India. She holds a doctoral degree from the Indian Institute of Technology Kharagpur. Her research interests include software testing and mobile cloud computing. She has more than 18 years of teaching and research experience, and has published numerous articles in leading international journals and conferences. Dr. Panigrahi is the author of a number of books, including a textbook. She is a life member of the Indian Society of Technical Education (ISTE) and a member of IEEE and the Computer Society of India (CSI).

Dr. Rajkumar Buyya is a Redmond Barry Distinguished Professor of Computer Science and Software Engineering and Director of the Cloud Computing and Distributed Systems (CLOUDS) Laboratory at the University of Melbourne, Australia. He is also serving as the founding CEO of Manjrasoft, a spin-off company of the university, commercializing its innovations in cloud computing. He served as a future fellow of the Australian Research Council during 2012–2016. He has authored over 525 publications and seven textbooks including "Mastering Cloud Computing" published by McGraw Hill, China Machine Press, and Morgan Kaufmann for Indian, Chinese, and international markets, respectively. He has also edited several books including "Cloud Computing: Principles and Paradigms"

(Wiley Press, USA, Feb 2011). He is one of the highly cited authors in computer science and software engineering worldwide (h-index=112, g-index=245, 63,900+ citations). Microsoft Academic Search Index ranked Dr. Buyya as #1 author in the world (2005–2016) for both field rating and citation evaluations in the area of Distributed and Parallel Computing. Recently, Dr. Buyya is recognized as "2016 Web of Science Highly Cited Researcher" by Thomson Reuters.

Prof. Kuan-Ching Li is a Professor in the Department of Computer Science and Information Engineering at the Providence University, Taiwan. He has been the vice-dean for the Office of International and Cross-Strait Affairs (OIA) at the same university since 2014. Prof. Li is the recipient of awards from Nvidia, the Ministry of Education (MOE)/Taiwan and the Ministry of Science and Technology (MOST)/Taiwan, and also guest professorship from various universities in China. He received his Ph.D. from the University of Sao Paulo, Brazil, in 2001. His areas of research include networked and graphics processing unit (GPU) computing, parallel software design, and performance evaluation and benchmarking. He has edited two books: Cloud Computing and Digital Media and Big Data published by CRC Press. He is a fellow of the IET, senior member of the IEEE, and a member of TACC.

Advanced Electronics Applications-I

Design and Analysis of Slotted Microstrip Patch Antenna Using DGS for S Band

Nikita Saxena and Asmita Rajawat

Abstract In this paper, a rectangular microstrip antenna has been designed. The design has been implemented with a defected ground surface (DGS) to obtain the desired gain and reflection coefficient at both the frequencies 2.45 and 3.75 GHz. In the proposed design, a microstrip patch antenna having an Inset Fed giving the dual-band characteristics is implemented using FR-4 as the substrate for the designed simple microstrip patch antenna, which has got a dielectric constant of 4.3. The designed microstrip patch antenna was simulated on computer simulation technology (CST), and the desired results were obtained. Four "L"-shaped slots have been made on the surface of the patch leading to an increase in the density of the current on the surface of the simple microstrip patch antenna which also led to multiband characteristics, thus improving the results. The S11 parameters were found to be below −10 dB mark for both the frequencies. The gain obtained at 2.45 and 3.75 GHz was found to be 4.1 dB and 3.57 dB, respectively. The design proposed in this paper has been found to be useful for power transmission applications, such as wireless headphones, cordless phones, keyless vehicle locks, and Bluetooth.

Keywords DGS · Microstrip patch antenna · Return loss · Gain · Dual band

1 Introduction

A. Related Work

Antennas are a part of every small and big thing around us ranging from insects to space satellites; they find their application in every field. One of the most efficient and easy-to-fabricate antenna is the microstrip patch antenna. In a microstrip patch

N. Saxena (✉) · A. Rajawat
Amity University, Noida, Uttar Pradesh, India
e-mail: nikita01saxena01@gmail.com

A. Rajawat
e-mail: arajawat@amity.edu

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_1

antenna, the ground and the patch are made up of the same material, whereas the substrate is made from the different material. Substrate layer is an insulating dielectric layer between the patch of the antenna and the ground. The signals are generated in antenna when the radio frequency current is applied between the patch and the ground, which is connected by a microstrip feed line.

This antenna which has been designed in this paper operates at two frequencies 2.45 and 3.75 GHz which are lying in S band; a defect in the ground was introduced in the implemented design to increase the gain and get better S11 parameter.

A multiband antenna gives more than one band of frequencies, whereas a wideband antenna produces one band consisting of many frequencies together. The major drawback of the wideband antenna is that it provides many undesirable frequencies as well. Therefore, a multiband antenna is considered to be a better option for many practical purposes.

In [1], a microstrip patch antenna was designed whose working range is 2–2.8 GHz. It uses FR-4 as the substrate, and a return loss of −23 dB is obtained at the central frequency of 2.25 GHz. In [2], the design consists of circular slots and it works at the Wi-Fi frequency. The antenna obtained the gain of 3.93 and 3.73 dB for lower band and the upper band, respectively.

Defected ground structure is the technique in which slots or defects are made on the ground plane of microwave planar circuits; this helps in improving various parameters of RMPA, namely low gain, cross-polarization, narrow bandwidth, etc. In [3], four DGS antennas were implemented, namely U shape, E shape, double E shape, and psi shape. On comparison, the psi-shaped DGS antenna had the maximum bandwidth and the least return loss at 2.45 GHz frequency. An L-shaped microstrip patch antenna along with DGS was designed for Wi-Fi security and WLAN applications. This design can also be used in RF applications [4]. For S band application, a UI slot microstrip patch antenna was developed. It operates at 2.95 GHz frequency giving a bandwidth of 3.26 dB and a return loss of −19 dB [5]. Enhancement of gain has been catered to by applying different techniques such as array configuration [6], EBG [7], and techniques which use parasitic elements [8] etc.

In this paper, the antenna has been designed by implementing a DGS structure which led to the enhancement of the parameters of the microstrip patch antenna like the gain of the antenna and the return loss.

Introduction of the DGS technique in the simple microstrip patch antenna brought some significant changes in the results of the designed antenna. The results obtained were then analyzed and were found to be compatible with the S band applications. In our research, the DGS technique has been found useful for enhancing the gain of the antenna designed and thereby giving the desired results at 2.45 and 3.75 GHz frequency.

B. **Contribution**

This paper will contribute toward the future studies and simulation of slotted microstrip patch antenna implemented using DGS technique. Inset feeding technique is used for the design because it is one of the simplest feeding techniques, and therefore, it makes the fabrication of the antenna also very simple and easy. Another reason for using the Inset Fed is its simple impedance matching.

C. **Organization of Paper**

The paper consists of four divisions, namely Section 1, Section 2, Section 3, and Section 4, which discuss the introduction of the designed antenna, proposed antenna design along with its specifications, results and discussion, and conclusion of the design, respectively.

2 Proposed Antenna Design

Nowadays, simple microstrip patch antennas are gaining popularity in the communication domain. Some major reasons behind this success are its ease of fabrication, they are compact and lightweight, therefore compatible with many existing systems, and they have got a high efficiency rate too.

The designed antenna in this paper was optimized to obtain the desired results by cutting the four slots on the patch of the antenna. The slots were cut according to the current density patterns obtained by simulating the simple microstrip patch antenna at the same frequencies. Cutting the slots on the patch of the antenna also led to the increase in the virtual length of the antenna due to which the single antenna began to work as three different antennas at two frequencies lying in the S band.

This paper throws light on the slotted microstrip patch antenna designs which have used defected ground structure technique. The design has been simulated on computer simulation technology (CST) software [9]. The design has been implemented on 2.45 GHz. The antenna has been designed with FR-4 (lossy) as the substrate which has a dielectric constant of 4.3. Since the dielectric constant of FR-4 is quite high, it proved to be beneficial in reducing the size of the antenna and therefore making the design more compact and more compatible with the already existing systems. The thickness of the substrate is 1.6 mm. The designed antenna uses the insert feeding technique because of its simple implementation and ease of fabrication. The impedance matching in the inset feed is also very convenient.

The implemented design has got four slots on the surface of the patch, and the ground of the antenna has been defected to increase the gain of the antenna.

The standard antenna equations [10] have been used to determine the dimensions of the patch which were further optimized for more desirable results. The final optimized parameters of the design of the microstrip patch antenna model are shown in Table 1. The parameters were optimized according to the density of the current on the surface of the patch of the antenna. Cutting of slots and the dimensions of the slots were positioned in such a way on the patch of the antenna that they gave the best possible results for the gain and the bandwidth at the chosen frequencies. Similarly, Fig. 1 shows the design of the patch antenna along with the slots that were cut on the surface of the patch of the antenna.

Figure 2 shows the DGS slots cut on the ground surface, whereas Fig. 3 shows

Table 1 Optimized parameters

S. No.	Parameters	Values (mm)
1	Length of the patch	28
2	Width of the patch	37
3	Width of the feed	2.9
4	Length of the feed	30.6
5	Depth of the feed	10.1
6	Height of the substrate	1.6

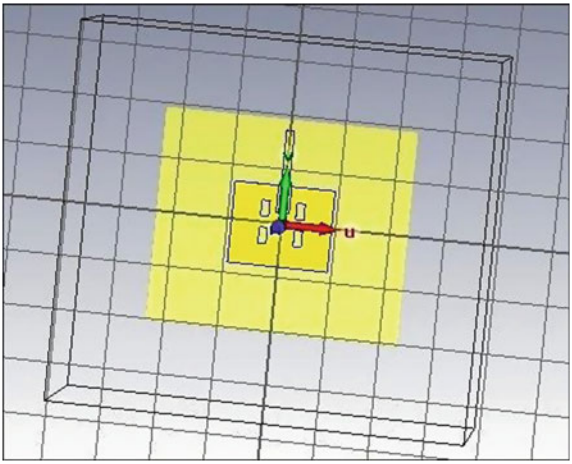


Fig. 1 Microstrip patch antenna with slots on patch

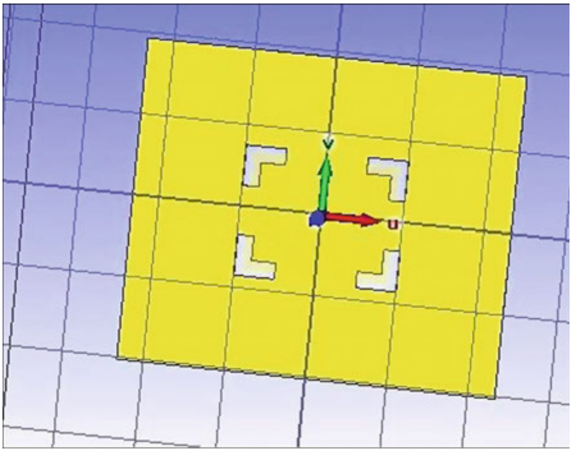


Fig. 2 Ground plane with DGS

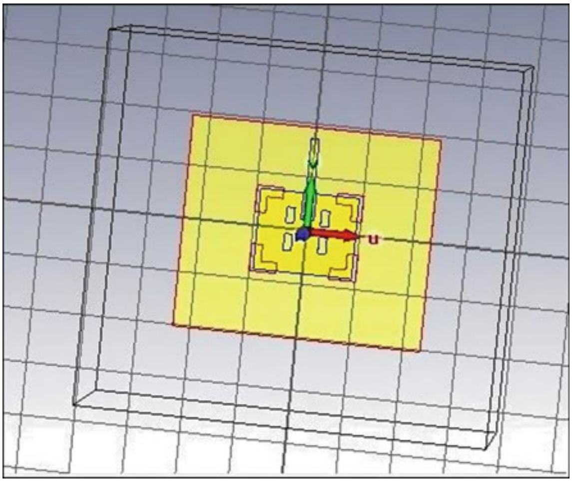


Fig. 3 Ground plane and patch together

the full design having the patch along with the DGS, to obtain the dual-band characteristics of patch antenna.

3 Results and Discussion

Slotted microstrip patch antenna with an inset feed technique has been designed. The gain and reflection coefficient obtained at 2.45 GHz are 4.1 dB and −30.4 dB, respectively. The gain and reflection coefficient obtained at 3.75 GHz are 3.19 and −34.8 dB. The gain obtained after the simulation and the S11 graph indicates that the designed antenna is applicable for various multiband applications.

A. **Return Loss**

The graph shown in Fig. 4 has been plotted between S11 (dB) and frequency (GHz). It signifies the return loss (in dB) of the dual-band antenna at the two prominent frequencies that is 2.45 and 3.75 GHz. The proposed microstrip antenna has been simulated at 2.45 and 3.75 GHz in which the return loss obtained is −31.4 dB and −34.8 dB, respectively. Thus, the designed antenna fulfills the requirements for S band applications.

B. **Gain of the antenna**

It is the measure of the transmission power of the antenna designed.

(a) **Gain of the antenna at 2.45 GHz**

Figure 5 elaborates the gain of the antenna which is 4.098 dB at 2.45 GHz frequency. This gain has been enhanced due to the DGS on the ground of the antenna.

(b) **Gain of the antenna at 3.75 GHz**

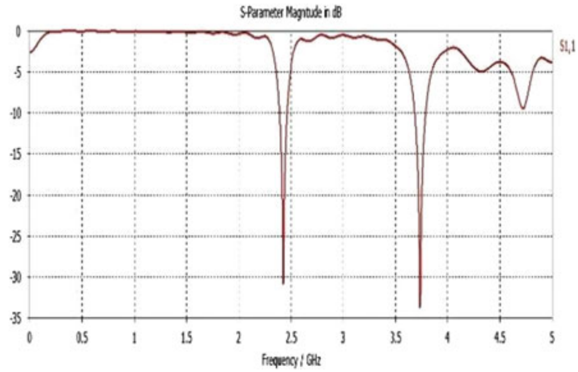


Fig. 4 S11 of the antennas operating on 2.45 and 3.75 GHz

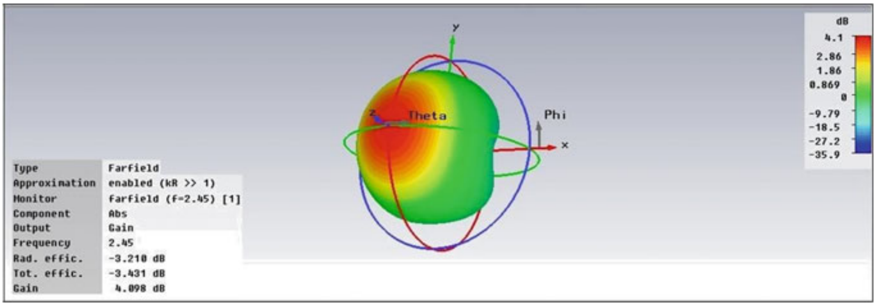


Fig. 5 Gain of the antenna at 2.45 GHz

Figure 6 shows the gain of the antenna which is 3.574 dB at 3.75 GHz frequency. However, the slots have been cut on the patch a dual-band antenna has been obtained.

C. Surface current

The surface currents define the current distribution around the slots that have been cut on the ground and the patch as shown in Figs. 7 and 8.

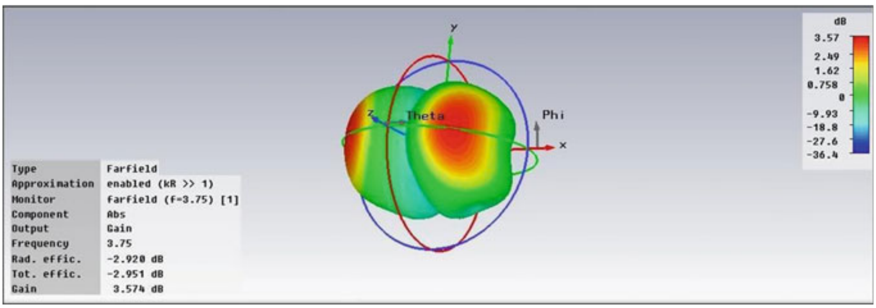


Fig. 6 Gain of the antenna at 3.75 GHz

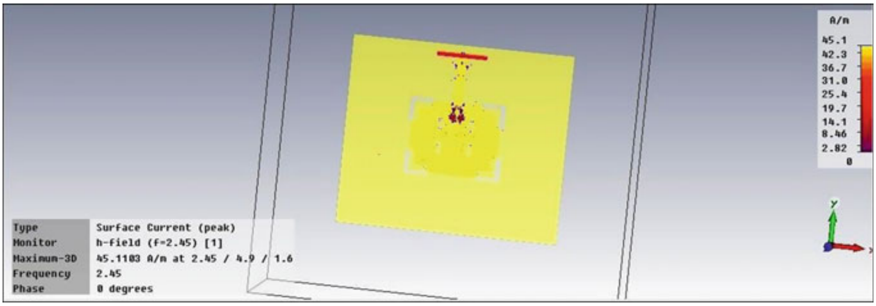


Fig. 7 Surface current at 2.45 GHz

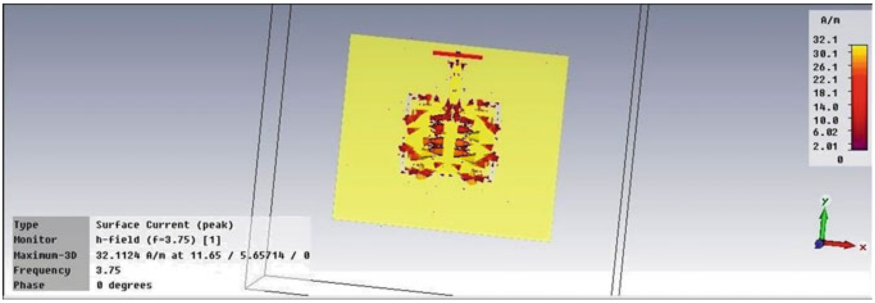


Fig. 8 Surface current at 3.75 GHz

(a) **Surface current at 2.45 GHz**

Figure 7 shows the surface current density on the patch of the antenna designed at the frequency of 2.45 GHz.

(b) **Surface current at 3.75 GHz**

Figure 8 throws light upon the surface current flowing on the patch of the designed antenna at the frequency of 3.75 GHz.

Table 2 shows the results of the antenna design at 2.45 and 3.75 GHz.

Table 2 Parameters of antenna for results and discussion

S. No.	Frequency (GHz)	Gain (dB)	S11 parameter (dB)
1	2.45	4.1	−30.4
2	3.75	3.19	−34.8

4 Conclusion

The designing and simulation of a dual-band Inset Fed microstrip antenna has been done. The design operates at 2.45 and 3.75 GHz frequency. Due to the implementation of DGS, the density of current on the surface of the patch has been improved due to which improved multiband characteristics have been obtained. Since the S11 parameter for both the operating frequencies is below −10 dB mark, we can say that this is a very suitable design for many S band applications such as cordless phones, Bluetooth, keyless unlocking of car doors, and wireless headphones.

References

1. Ramkumar Prabhu, M., Rajalingam, A.: Rectangular patch antenna to operate in flame retardant 4 using coaxial feeding technique. Int. J. Electron. Eng. Res. **9**(3) (2017)
2. Mishra, B., Singh, V., Dwivedi, A.K., Pandey, A.K., Sarwar, A., Singh, R.: Slots loaded multilayered circular patch antenna for WiFi/WLAN applications. Comput. Netw. Sustain. (2017)
3. Marotkar, D.S., Zade, P.: Bandwidth enhancement of microstrip patch antenna using defected ground structure. In: International conference on electronics, electrical and optimization techniques (2016)
4. Afzal, W., Ahmad, M.M.: L-Shaped Microstrip Patch Antenna for DGS and WLAN Applications (2014 Mar)
5. Kaur, A., Bharti, G.: U-I slot microstrip patch antenna for S band applications. Int. J. Recent Sci. Res. **7**(4), 10410–10412 (2016 Apr). Yadavindra College of Engineering Punjabi University Gurukashi Campus Talwandi Sabo, Punjab, India
6. Khayat, M., Williams, J.T., Jakson, D.R., Long, S.A.: Mutual coupling between reduced surface-wave microstrip antennas. IEEE Trans. Antenna Propag **48**, 1581–1593 (2000 Oct)
7. Park, J., Herchlein, A., Wiesbeck, W.: A Photonic Bandgap (PBG) structure for guiding and suppressing surface waves in millimeter-wave antennas. IEEE Trans. Antennas Propag. **49**, 1854–1857 (2001 Oct)
8. Rojas, R.G., Lee, K.W.: Surface wave control using nonperiodic parasitic strips in printed antennas. IEE Pros. Microwave Antennas Propag. **148**, 25–28 (2001 Feb)
9. EM simulator, CST Microwave studio TM, V.9
10. Balanis, C.A.: Antenna Theory Analysis and Design, 2nd edn. Wiley, Singapore (2002)

Current Control of a CHB Multilevel Inverter Using PR and Adaptive Fuzzy PI Controller: A Comparison

Gayatri Mohapatra and Asim Kumar Dey

Abstract With the rapidly increasing power demand with renewable energy, the control is also equally important to be considered. The VSI, when interfaced with grid, is to be controlled for different disturbances. The current-controlled technique is discussed with two controllers. The PR controller is helping to remove the steady-state error, and the fuzzy has the advantage of updated PI parameter. At the same time, THD is compared in terms of harmonic percentages with different controllers. The work is simulated in MATLAB/Simulink, and it is found that the fuzzy is giving better response as compared to PR regulator.

Keywords CHB-MLI · Current control · PR controller · Adaptive fuzzy PI controller

1 Introduction

Development of the renewable energy needs a different inverter topology and control for a huge power requirement specially dealing with the medium to high voltage, resulting in interest of multilevel inverters. Handling of a large amount of power with appreciable control technique can be a constraint in the recent era. This has a requirement of optimal choice of inverter and adequate control. Cascaded H-bridge (CHB) with multistring PV cell is used to remove per phase unbalance powers, proper control method [1–4].

Here, THD of different multilevel inverters with their optimized values as per the IEEE standard is discussed up to a specific level. The load current is controlled

G. Mohapatra (✉) · A. K. Dey
Department of Electrical Engineering, S'O'A (Deemed to be University), Bhubaneswar, Orissa, India
e-mail: gayatrim79@gmail.com

A. K. Dey
e-mail: asimiter@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_2

using current control technique for different disturbance ranges using proportional resonant (PR) controller and adaptive fuzzy PI controller. A comparison is established between the two controllers.

The paper is aligned as: Sect. 2 briefs the idea of cascaded multilevel inverter, the control technique is explained in Sect. 3, PR controller is derived in Sect. 4, and fuzzy PI controller in adaptive mode is given in Sect. 5. The simulation and result are in Sect. 6 and the discussion with conclusion in Sects. 7 and 8.

2 Multilevel Inverter

The multilevel inverter (MLI) is an inverter with different levels containing an array of power semiconductor and capacitor voltage sources. This has a set of semiconductor fully controllable IGBTs and capacitors acting like a DC source. The switching sequence adds the voltages of each bridge and finally amplified at the end point, thereby resisting only reduced voltage by each switch. The CHB inverter, as in Fig. 1a, has an arrangement of single H-bridge units in series, and Fig. 1b gives the output voltage with different levels. It has the advantages of accruing less number of components as compared to other topology, hence cost effective [1, 5].

2.1 *Modeling of Inverter*

Having n as the number of steps in output voltage as in Fig. 1b, the levels in the load voltage can be as Eq. (1). The delay in switching of IGBTs is predicted to remove

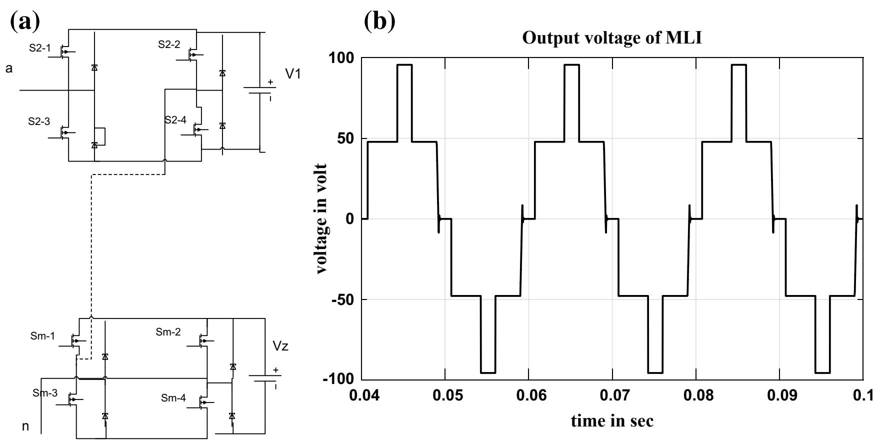


Fig. 1 **a** explains the cascaded multilevel inverter and **b** shows the output voltage per phase

the highest low-order harmonics, at the output point. The expansion of respective equation made in Fourier can be used on the stepped waveform in output voltage (V_{an}) expressed as in Eq. (2). [6, 4]

$$k = 2s - 1 \tag{1}$$

$$V_{an}(wt) = \sum_{x=1,3,5}^{\infty} \frac{4E_{dc}}{x\pi}[\cos(h\beta_1) + \cos(h\beta_2) + \cdots + \cos(h\beta_s)] \sin(hwt) \tag{2}$$

where

S gives the number of cells of the inverter.
E_{dc} is the DC capacitor voltage.
h explains the no. of harmonic order staring from $5 - \infty$.

Here, k can be 1 if S is even and 2 if odd.

$\beta_1 - \beta_S$ are the angles of delay depending upon the no. of level.

The desired voltage of output (rms) (V_{an}) is the sum of the bridge potential, as in Eq. (3) such that

$$V_{an} = \sum_{z=1}^{S} V_{az} \tag{3}$$

where

V_{an} is the voltage per phase of the inverter.
V_{az} explains the individual source voltage for S number of sources.

3 Control Technique

The multilevel inverter must satisfy the following objective in order to have better control.

- Low THD will be a consideration as that the output voltage will be close to sinusoidal [7] and each photovoltaic generator must operate at its MPP.
- The current injected from the source must be in the same phase with that of the grid terminal voltage [8].
- Voltage and current stability for any generated disturbance must be achieved.

Considering the practical implementations of the proposed model, there must be operational constraints as follows:

- DC component and harmonics of even order are crossed out with the constant DC voltage across the capacitor.

- Switching angle of IGBT, β, takes the values between $0°$ *and* $90°$ with an operating switching frequency of 10 kHz.

The system is developed in MATLAB/Simulink explaining a control technique with renewable energy as well as the inverter as given in Fig. 2a, b. Load current is considered to be the factor to be sensed for the optimal control with reference to the current reference. The current reference is being calculated considering the grid code into account with PLL [9, 10]. The data obtained are converted to DC component, and the error signal is generated.

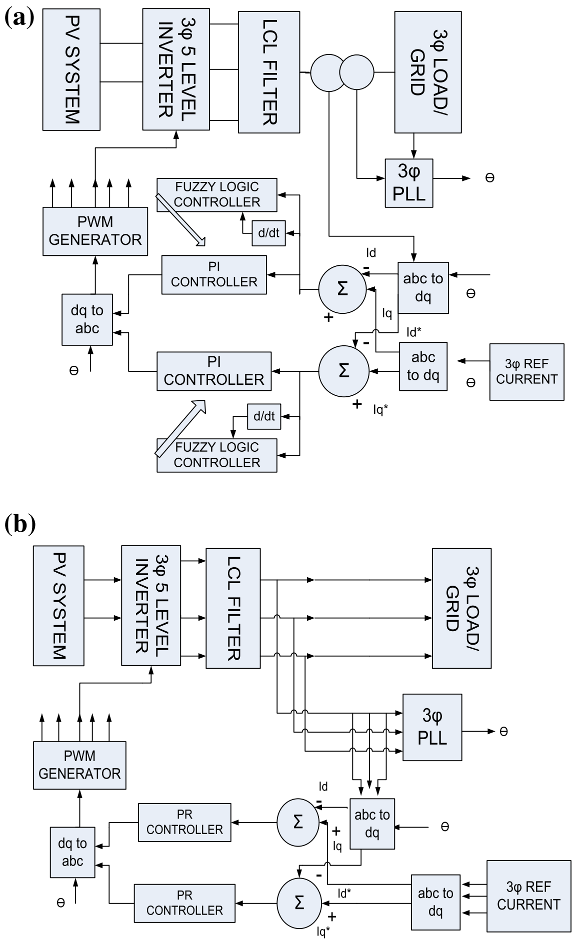


Fig. 2 **a** explains the overall block diagram of a PV integrated multilevel inverter with adaptive fuzzy PI controller [11], **b** explains the overall block diagram of a PV integrated multilevel inverter with PR controller

4 Proportional Resonant (PR) Controller

A PR controller as in Fig. 4 is defined as a current controller [12, 13] $H_c(s)$ which is defined as in Eq. (4) which converts the error signal into a sinusoid resulting in less difference. The harmonic to a specific order can be eliminated as in Eq. (5).

$$H_c(s) = K_{P1} + K_{I1}\frac{s}{s^2+\omega^2} \tag{4}$$

$$G_h = \sum_{h=3,5,7} K_{\text{th}}\frac{s}{s^2+(\omega h)^2} \tag{5}$$

K_{P1} proportional gain
K_{I1} integrator gain
G_{h} gain required for the working compensator.

A PR controller as explained in Eqs. (4) and (5) is same as that of a PI controller in configuration with an additional factor of harmonic compensator [11, 14]. Equation (4) as shown transforms the signal into a sine or cosine waveform, thereby yielding less error. It can also be interpretable as the resultant of an LC filter. As from Figs. 3, 4, the sharp Bode plot gives the roots of the characteristic equation in a better manner. Addition of gain can enhance the regulator frequency to make the system fixed to resonate at the required value. A small K_{P1} yields a narrow band, but a large K_{P1} leads to an expanded band.

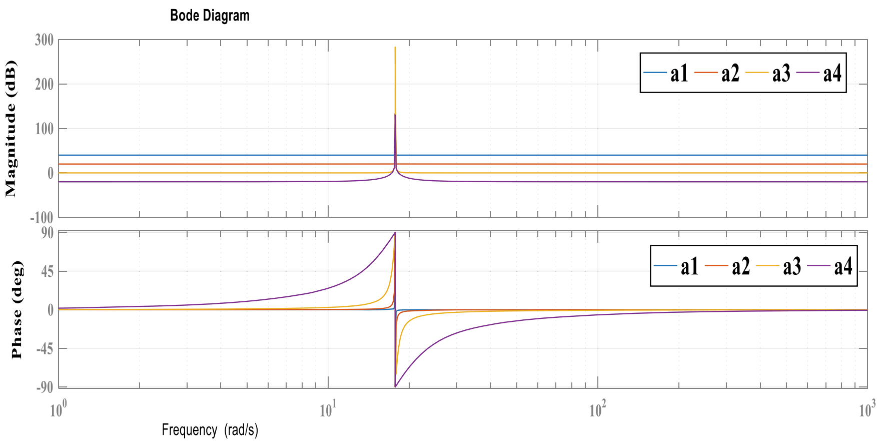
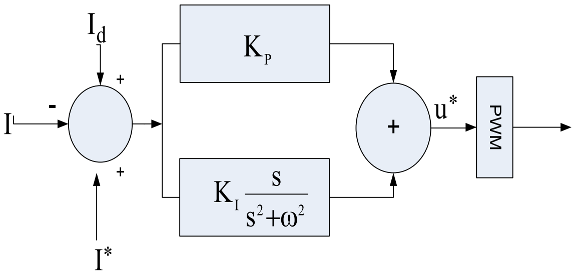


Fig. 3 Bode plot diagram of PR controller

Fig. 4 PR controller



5 Adaptive Fuzzy PI Controller

The structure of the fuzzy controller shown in Fig. 5 [11, 14] has taken error and the rate of change of error as input, and K_{P} and K_{I} are taken as output. The parameters are discretized as defined below in Eqs. (6) and (7). The fuzzy controller is set to minimize the error and get optimized values.

$$E_i(j) = I_a(j) - I_b(j) \tag{6}$$

$$\partial E_i(j) = E_i(j) - E_i(j-1) \tag{7}$$

where

$E_i(j)$	error
$\partial E_i(j)$	change of error
$I_a(j)$	reference current
$I_b(j)$	actual current
j	parameter of discretization.

A rule matrix table is developed as in Table 1 where the conditions defined for the rule are explained in Table 2. Five different conditions are set for consideration, and the defuzzification is done accordingly to evaluate the error. The output of the fuzzy is allowed to pass through the zero-order hold and the discrete-time integrator to get the required response.

Fig. 5 Fuzzy controller

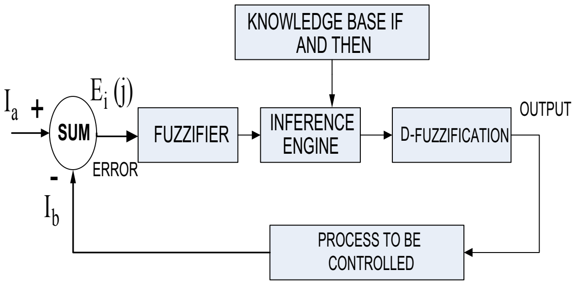


Table 1 Rule matrix table for K_P and K_I

Output	Error	R1	R2	R3	R4	R5	Output	R1	R2	R3	R4	R5
Change in error K_P	R1	L2	L2	M1	M1	S0	Change in error K_I	S0	S0	M1	L2	L2
	R2	L2	M1	M1	S0	S0		S0	S0	M1	L2	L2
	R3	M1	L2	M1	M1	M1		M1	M1	M1	M1	M1
	R4	S1	M1	M1	L2	L2		L2	L2	M1	S0	S0
	R5	S1	S1	M1	M1	L2		L2	L2	M1	S0	S0

Table 2 Fuzzy parameters used in the controller

R1	Negative large	R4	Positive small	L2	Large
R2	Negative small	R5	Positive large	M1	Medium
R3	Zero			S0	Small

The fuzzy parameters are explained in Table 2.

6 Result

A THD analysis of a multilevel inverter is done, a graph is plotted as per Fig. 6, and it can be observed that THD is inversely proportional to the number of levels.

The filter parameter set and different cases are explained in Table 3.

The inverter is operated for 20 V, 10 A constant parameters, the current control technique is applied by both the controllers, and the result is observed as in Figs. 7 and 8. Two different case studies are being done for both the controllers so that the current control output can be checked as in Figs. 9, 10.

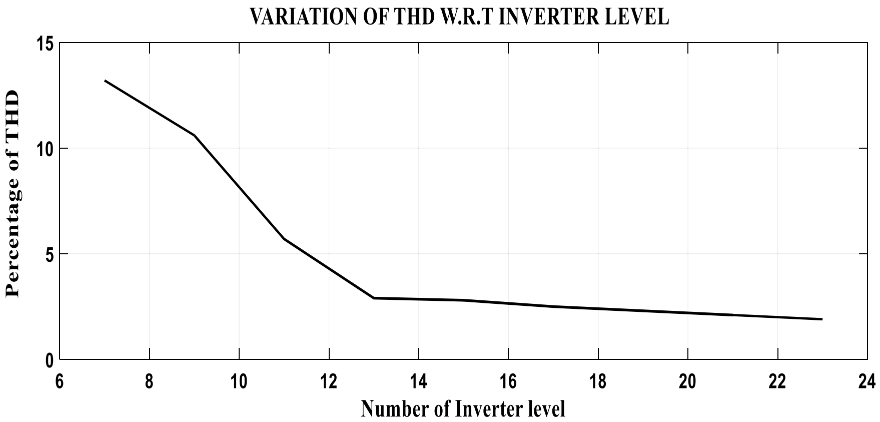


Fig. 6 THD variation w.r.t. inverter level

Table 3 Cases used in the controller

Filter parameter	1 mH,100 μF
Load parameter	5−10 Ω, 2 mH

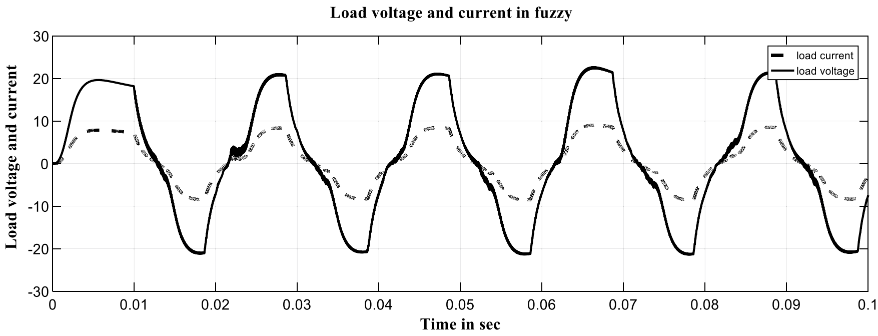


Fig. 7 Comparison of the load voltage and current with fuzzy controller

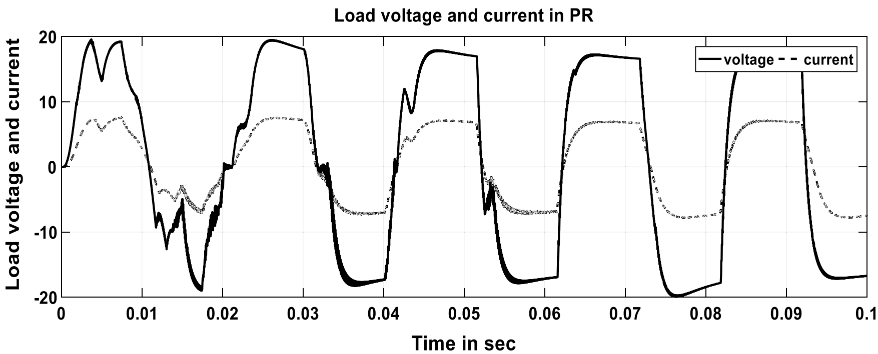


Fig. 8 Comparison of the load voltage and current with PR controller

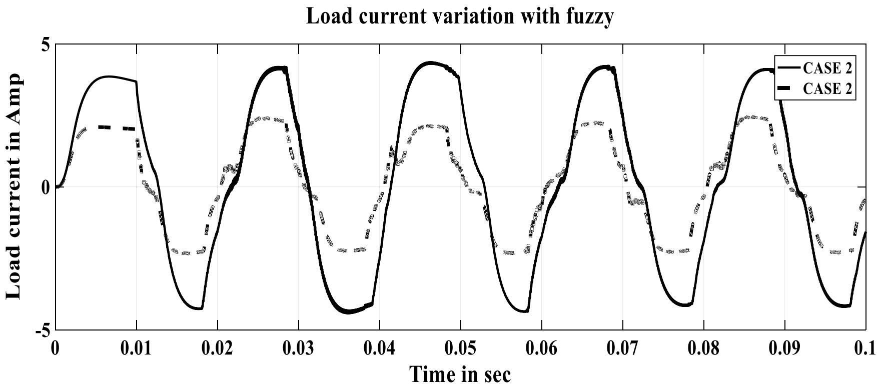


Fig. 9 Load current for different cases as explained

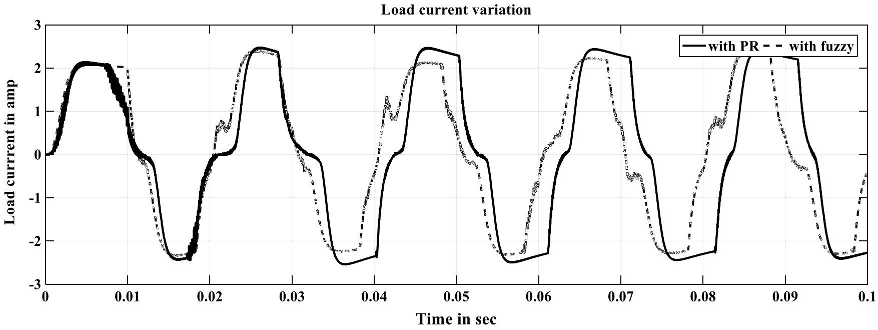


Fig. 10 Comparison of the load current with PR and fuzzy controller

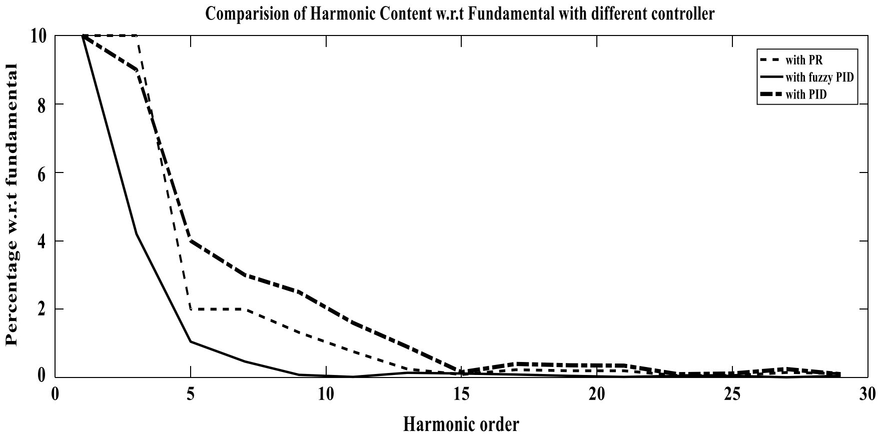


Fig. 11 Comparison of the percentage of harmonic order w.r.t. fundamental for different types of controller

7 Result Analysis

The simulated waveforms are analyzed which leads to the following points.

- The current waveform is smoother with the fuzzy PI controller.
- With the insertion of disturbance, the current is controlled within the reference values.
- The waveforms are analyzed with some harmonic content as a constraint in designing proper filter values to prepare a prototype model.
- The load voltage and current explain their phase shift.
- The comparison of THD explains that the fuzzy PID is giving better result as per Fig. 11.

8 Conclusion

In this article, PR controller and adaptive fuzzy PI controller are implemented in a three-phase five-level PV integrated multilevel inverter for a specific load current. Load is varied in the form of R or RL load with grid-connected system, and the response is observed. It can be concluded that the fuzzy finds better result in the form of harmonic reduction and waveform response.

References

1. Rodriguez, J., Lai, J.S., Peng, F.Z.: Multilevel inverters: a survey of topologies, controls, and applications. IEEE Trans. Ind. Electron. (2002 Aug)
2. Forzani, A., Nazarzadeh, J.: Precise loss calculation in cascaded multilevel inverters. In: Second international conference on computer and electrical engineering, vol. 2, pp. 563–568 (2009 Dec 28–30)
3. Lahlou, T., Malik, S.N., Herzog, H.G.: Simulation and analysis of control and modulation scheme for a multilevel cascaded H bridge inverter for use in a battery energy storage system. IEEE Trans. Ind. Electron. (2017 Aug)
4. Alesina, A., Venturini, M.: Solid-state power conversion: a Fourier analysis approach to generalized transformer synthesis. IEEE Trans. Circuits Syst. **28**(4), 319–330 (1981)
5. Mohapatra, G., Nayak, M.R.: Switching angle and power loss calculation for THD minimization in CHB multilevel inverter using DEA. In: Satapathy, S., Tavares, J., Bhateja, V., Mohanty, J. (eds.) Information and Decision Sciences. Advances in Intelligent Systems and Computing, vol. 701. Springer, Berlin (2018)
6. Mohapatra, G.: Current control of a PV integrated CHB-multilevel inverter using PR controller. In: Recent Technologies for Smart-City Energy Security and Power (ICSESP_2018) (2018 Mar)
7. Escobar, G., Martinez-Rodriguez, P.R, Ho, C.N.M., Sosa, J.M.: Design of inverter side current reference and controller for a single phase LCL based grid current inverter. Int. Trans. Electr. Energy Syst. **2**(3), 573–581 (2018 May)
8. Gholizade_Narm, H.: A novel control strategy for a single phase grid connected power injection system. IJE Trans. C Aspect **27**(12), 1841–1849 (2014 Dec)
9. Teodorescu, R., Blaabjerg, F., Liserre, M., Dell, A.: A stable three phase LCl filter based active rectifier without damping. In: IAS, pp. 1552–1557 (2003)
10. Iyer, N.P.R.: A dual programmable AC to DC rectifier using three phase matrix converter topology analysis aspects. Springer Electr. Engg. **100**(2), 1183–1194 (2018)
11. Sefa, I., Altin, N., Ozdemir, S., Kaplan, O.: Fuzzy PI controlled inverter for grid interactive renewable energy system. IET Renew. Power Gener. **9**(7), 729–738 (2015)
12. Escobar, G., Martinez-Rodriguez, P.R., Ho, C.N.M., Sosa, J.M.: Design of inverter side current reference and controller for a single phase LCL based grid current inverter. Int. Trans. Electr. Energy Syst. (2018 May)
13. Zammit, D., Staines, C.S., Apap, M., Licari, J.: Design of PR control with selective harmonic compensators using Matlab. J. Electr. Syst. Inf. Technol. (2017 Aug)
14. Teodorescu, R., Blaabjerg, F., Liserre, M.: Proportional-resonant controllers. A new breed of controllers suitable for grid-connected voltage-source converters. IEEE Proc. Electr. Power Appl. **153**(5) (2006 May)

Pyramid Entropy Source for True Random Number Generator on FPGA

Sivaraman Rethinam, Sundararaman Rajagopalan, Sridevi Arumugham, Siva Janakiraman, C. Lakshmi and Amirtharajan Rengarajan

Abstract True random number generator (TRNG) is a crucial part of any modern cryptographic primitives. Designing high entropy TRNG is still a challenging task to compete. Usually, TRNGs use physical mean as random source to generate random numbers which are unpredictable. TRNG has different application but not limited to symmetric key cryptography, random simulation, nonce, user authentication protocols, and gaming. In this work, ring oscillator (RO)-based pyramid entropy source is proposed for generating device-independent true random numbers. The de-synchronization technique has been utilized for extracting the randomness from entropy source wherein linear-feedback shift register (LFSR) is used as post-processor to improve the statistical characteristics of proposed TRNG. The design of TRNG has been implemented on Altera Cyclone II EP2C20F484C6 FPGA which consumed 892 logic elements (5%), 67.75 mW power to generate 131,072 true random bits. It yields maximum entropy of 0.999987 through entropy analysis. Statistical properties of the proposed TRNG have been evaluated using NIST tests. Restart experiment is also conducted to evidence the true randomness of the proposed TRNG.

Keywords Cryptography · De-synchronization · Randomness · Entropy and NIST tests

1 Introduction

Increasing demand for information security requires strong cryptographic primitives. Strength of any cryptographic algorithm is mainly based on its key which is to be random. Many key generation schemes have been proposed such as linear-congruential generator (LCG), linear-feedback shift register (LFSR), and chaotic maps and cellular automata (CA) which are pseudorandom number generators (PRNGs). Though PRNG produces good random numbers, it is deterministic and weak to resist critical attacks. The eventual solution to the key generation problem is true random

S. Rethinam · S. Rajagopalan (✉) · S. Arumugham · S. Janakiraman · C. Lakshmi · A. Rengarajan
Department of ECE, School of EEE, SASTRA Deemed to be University, Thanjavur, India
e-mail: raman@ece.sastra.edu

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_3

number generators (TRNGs). TRNG achieves immense equidistribution of zeros and ones in random number generation which can be used for key generation, initialization vectors, random simulation, etc. True random numbers are so random as well as unpredictable in nature since it uses physical mean as a random source. Post-processing is used for enhancing the statistical properties of TRNG [1].

Plenty of TRNG works have been proposed so far based on software and hardware approaches. Specifically, FPGA-based TRNGs have attained extreme importance due to its easy prototyping, portability, upgradability, and reconfigurability properties. In FPGA, logic gates [2], open-loop structures [3], phase-locked loops (PLLs) [4], BRAM write collisions [5], digital clock manager (DCMs) [6], and chaotic oscillator [7] were used for generating true random numbers. Among all available methods of TRNG, ring oscillator (RO) is an auspicious entropy source which produces high jittery clocks to generate true random bits with low power and less hardware. In addition, it offers device-independent fashion for TRNG to work in any type of FPGA. Such a model of TRNG was proposed by Sunar et al., in which 114 identical rings have been used with three stages RO were XORed to generate true random bits. The design consists of a random source, harvesting mechanism, sampler, and resilient function as post-processor. It was a pure digital logic-based design which obtained entropy of 0.97. This work also suggests using non-identical rings of RO rather than identical rings RO which causes coupon collector problem [8].

Schellekens et al. continued the work presented in [8] using Xilinx Virtex XC2VP30 FPGA. TRNG was designed using 110 rings with three stages RO which occupied 1664 slices. Sampling frequency was taken as 40 MHz. Statistical properties of this TRNG were evaluated through NIST and DIEHARD tests [9]. Wold et al. presented a post-processor less TRNG implemented on Altera Cyclone II FPGA. This design uses identical rings of RO but the higher-order transition problem was eliminated through dual sampling using two D flip flops. It consumed 83 logic elements for 25 rings with 75 inverters and 167 logic elements for 50 rings with 150 inverters, respectively. It achieved 100 Mbps as throughput for the sampling frequency of 100 MHz [10]. Marghescu et al. also represented a RO-based TRNG reported in [8] on Xilinx ZYNQ 700 SoC platform. This design of TRNG was divided into two parts in which FPGA generates true random bits using RO as noise generator and ARM verifies the statistical properties of the generated bits. Von Neumann Corrector (VNC) was utilized as post-processing technique which was implemented on ARM Cortex A9 [11]. A new type of RO called Transitions Effect RO (TERO) based TRNG was proposed in [12]. It is so sensitive to intrinsic noise present in FPGA. Highly correlated true random bits generated from this TRNG have satisfactory statistical properties and passed in NIST tests. The design was implemented on Xilinx Spartan 3E wherein the throughput was not reported.

Jessa and Matuszewkski illustrated a RO-based TRNG design where independent jitter generators were combined to produce true random bits. Further, the auxiliary source of randomness (ASR) which was constructed using Galois field RO with feedback polynomial was added to improve randomness. The design was implemented on Xilinx Virtex 5 FPGA with different sampling frequencies such as 100, 150, 200, and 250 MHz. To prove the presence of true randomness, a restart experiment has been conducted [13].

2 Proposed Methodology

Commonly, TRNG comprising of three basic building blocks namely entropy source, harvesting mechanism, and post-processing as shown in Fig. 1a. In this work, RO is chosen as noise source which generates perturbed jittery clock signals induces the randomness. De-synchronization technique is adopted for extracting the randomness from the RO entropy source. Further, harvested true bits are post-processed to achieve adequate statistical characteristics.

Totally, 64 inverters are used to construct this entropy source presented in Fig. 1b. Non-identical rings of RO have been considered in this work to avoid coupon collector problem [8]. Theoretically, RO which has less inverter will generate high frequency of oscillation. Pyramid structure has a collection of inverters to generate multiple frequencies of oscillation listed in Table 1. Each frequency of RO constitutes jitters in clock which is further combined to generate true random bits. Frequency of oscillation in RO can be calculated as,

$$f = \frac{1}{\kappa t \mathrm{pd}} \tag{1}$$

where

f—Frequency
k—Number of inverters in RO
t_{pd}—Propagation delay.

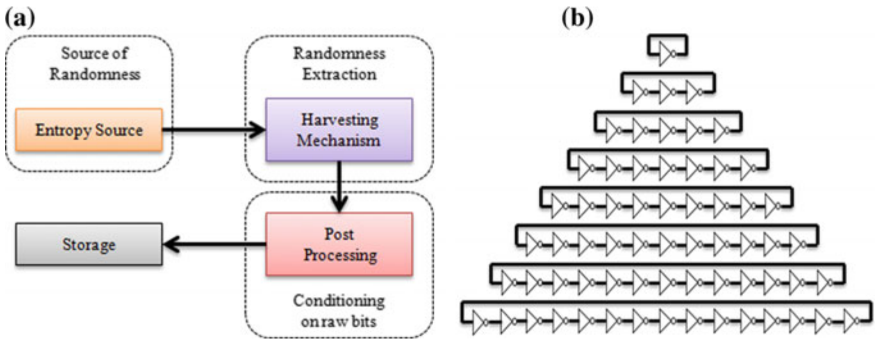


Fig. 1 **a** Architecture of TRNG, **b** Structure of pyramid RO

Table 1 Pyramid entropy source frequencies

Number of rings	Number of inverters	Propagation delay (nS)	Frequency (MHz)
Ring 1	1	8.050	125.0000
Ring 2	3	16.008	20.1822
Ring 3	5	24.046	8.3173
Ring 4	7	32.015	4.4621
Ring 5	9	39.928	2.7827
Ring 6	11	48.118	1.8892
Ring 7	13	55.987	1.3739
Ring 8	15	64.032	1.0411

2.1 Harvesting Mechanism

It is a process of extracting randomness from entropy source and samples them to convert into bits. Two methods of randomness extraction have been used namely jitter counting and de-synchronization. In this work, de-synchronization is adopted because of individual entropy sources have been combined together to generate multiple jittery. De-synchronization is a XOR tree, in which the individual results of all eight rings of RO would be XORed to produce final true random bit sequence as shown in Fig. 2a.

Further, it is sampled using a D flip flop with a sampling frequency of 27 MHz. Further, sampled true random bits are post-processed through post-processing mechanism explained in Sect. 2.2. Finally, post-processed bits are stored in a Block Random Access Memory (BRAM) which is constructed from an inbuilt IP core of Altera Cyclone II FPGA. BRAM of size 1024 × 128 bits have been considered to store 131,072 true random bits. Figure 2b expresses the snapshot of generated true random numbers on Block RAM through In System Memory Content Editor of Quartus II 8.0 Electronic Design Automation (EDA) tool.

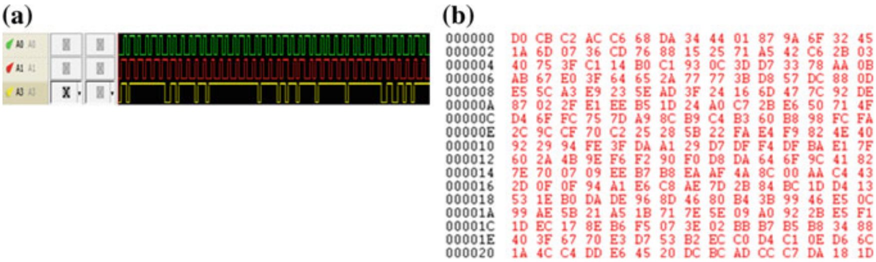


Fig. 2 **a** Raw true random bits, **b** BRAM snapshot of TRNG

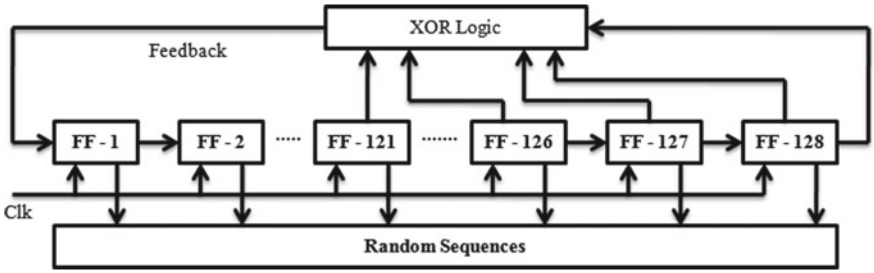


Fig. 3 Generic architecture of LFSR

2.2 Post-Processing

Post-processing is an additional unit which is useful to enhance the results of raw true random bits. It has been used to improve the statistical properties of TRNG to alleviate the bias problems when ring oscillator has been used as an entropy source. Usually, Von Neumann Corrector (VNC) has been suggested as a post-processor. Though it improves the equidistribution of true random numbers, it affects the throughput of the design. So, in this work, LFSR has been adopted as post-processing technique with a polynomial of $X^{128} + X^{126} + X^{101} + X^{99} + 1$.

Generic architecture of 128 bit LFSR is shown in Fig. 3. In general, LFSR is a shift register which shifts the binary data from one register to another register with respect to the clock. In addition, the pseudorandomness of the LFSR has been decided by the XOR operation in feedback loop of LFSR. Since LFSR is a pseudorandom sequence generator, it requires an initial seed to start the random perturbations where F3EA9FF87807FFFFF3EA9FF87807FFFF is taken as seed in this work.

3 Results and Discussion

To evaluate the properties of proposed TRNG and to evidence the true randomness, standard analyses have been performed namely entropy, NIST test, and restart experiment. Hardware analyses such as resource utilization, power consumption, and throughput have also been presented to show the hardware capability and performance of the proposed design.

3.1 Entropy Analysis

It is an important metric to analyze the equidistribution property of TRNG. For a good TRNG, entropy value must be close to one which exhibits that the number of occurrences of zeros and ones has been distributed equally. Table 2 depicts the entropy

Table 2 Entropy analysis

No. of experiments conducted	Raw true random bits	Post-processed true random	No. of experiments conducted	Raw true random bits	Post-processed true random
1	0.978577	0.999997	6	0.981178	0.999999
2	0.978282	0.999994	7	0.975934	1
3	0.975166	1	8	0.984573	0.999996
4	0.977103	0.999997	9	0.978116	0.999996
5	0.977759	1	10	0.978212	1

values obtained for raw and post-processed true random bits in which post-processed true random bits yield maximum entropy.

3.2 NIST SP 800-22 Battery of Test

National Institute of Standards and Technology (NIST) has specified standard tests for evaluating random number generators [14]. It is used to evaluate the statistical properties of any random number generator to verify its randomness. Tests concentrate on basic properties of TRNG such as equidistribution of zeros and ones, unpredictability, and irreproducibility. In this work, NIST SP 800-22 tests have been performed for ten different sets of TRNG and the results of three experiments of 131,072 post-processed true random bits have been tabulated in Table 3.

From the table, it is inferred that the proposed TRNG has passed the NIST tests such that the statistical characteristics of TRNG have been evaluated and evidenced.

3.3 Restart Experiment

The major difference between a PRNG and TRNG is its initial sequence. PRNG produces the same sequence for all the time wherein TRNG generates different sequences at each time. This can be determined through restart experiments [15]. Capturing the initial bitstream of TRNG for multiple times will depict the true random nature of TRNG. For this proposed TRNG, restart experiment has been conducted for several times and 10 sets of results have been presented in Fig. 4.

Table 3 NIST test results

Tests	p value for 3 experiments		
	E1	E2	E3
Frequency	0.91141	0.35048	0.73991
Block frequency	0.53414	0.73991	0.53414
Cumulative sums—I	0.12232	0.73991	0.35048
Cumulative sums—II	0.53414	0.91141	0.35048
Runs	0.73991	0.53414	0.35048
Longest run	0.35048	0.35048	0.73991
Rank	0.91141	0.73991	0.35048
FFT	0.12232	0.53414	0.35048
Non-overlapping Template	0.73991	0.99146	0.91141
Overlapping template	0.53414	0.91141	0.12232
Approximate entropy	0.35048	0.91141	0.53414
Serial—I	0.53414	0.73991	0.35048
Serial—II	0.35048	0.91141	0.35048
Linear complexity	0.53414	0.35048	0.53414
$P \geq 0.001$ to pass the test			

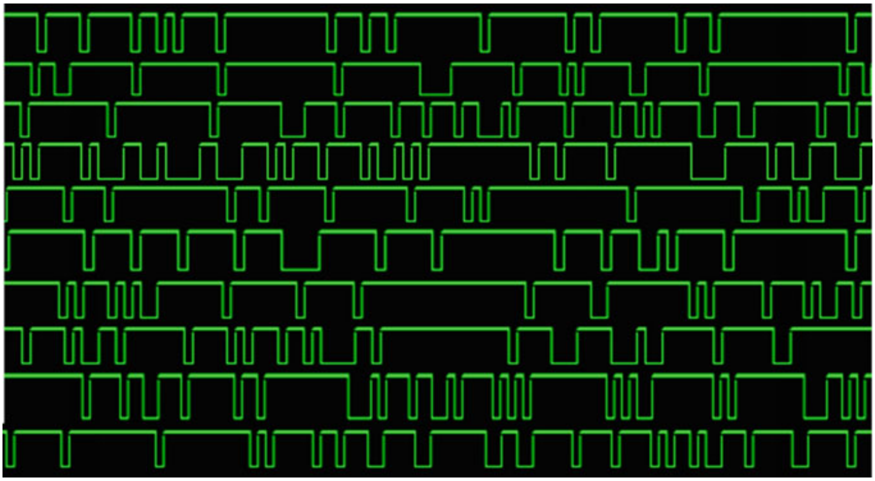


Fig. 4 Restart experiment results

3.4 Resource Utilization

FPGA implementation of digital design includes logic elements, logic registers, PLLs, BRAMs, and their power consumption. Table 4 lists out the resource utilization of proposed TRNG implanted on Altera Cyclone II FPGA.

Table 4 Resource utilization

Parameters	RO TRNG	RO TRNG with post-processing
Target device	Cyclone II EP2C20F484C6 FPGA	Cyclone II EP2C20F484C6 FPGA
Logic elements	620	892
Combinational functions	465	730
Logic registers	520	791
Memory bits	131,072	131,072
Total power dissipation (mW)	67.75	67.75

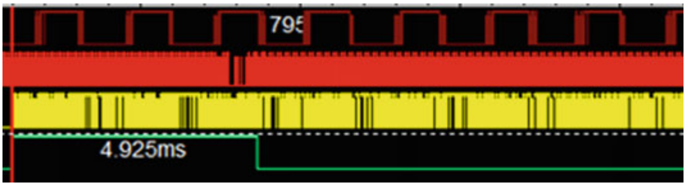


Fig. 5 Timing analysis

3.5 Throughput Analysis

Speed of the design can be determined through throughput analysis in which number of bits per second is calculated. Figure 5 expresses the time taken for generating 210 128 bit true random bits such as 4.925 mS, and further, throughput is calculated as 26.613660 Mbps. Table 5 shows the throughput comparison of proposed work with other earlier TRNGs. From the comparison, it is witnessed that the proposed TRNG yields high throughput.

4 Conclusion

Pyramid ROs-based TRNG has been proposed in this work. De-synchronization and D flip flop were used for randomness extraction and sampling. LFSR is used as post-processor to enhance the TRNG. Statistical properties of proposed TRNG have been validated using entropy and NIST tests. Restart experiment was also conducted to prove the true randomness of TRNG. Future work will be on improving the harvesting mechanism through chaotic maps.

Table 5 Performance comparisons—throughput

Design Category	Reference	Target device	Throughput in Mbps
PLL/DLL	Fischer and Drutarovsky [16]	Altera EP20K200E	0.07
	Fischer et al. [17]	Altera EPIS25	1
	Kwok and Lam [18]	Xilinx XC2VP20	6
	Deak et al. [19]	Xilinx Kintex – 7 XC7K410T	6.25
Free running oscillator	Kohlbrenner, Gaj [20]	Xilinx XCV1000	0.5
	Schellekens et al. [9]	Xilinx XC2VP30	2.5
	Sunar et al. [8]	Xilinx XC2VP30	2.5
Transition effect ring oscillator	Varchola and Drutarovsky [21]	Xilinx XC3S500E	0.25
Dual metastability	Wieczorek [22]	Digilent's Nexys 2 board	1–5
Coherent sampling	Valchanov et al. [23]	Actel Fusion M7AFS600 and Xilinx XC3S2005ft256	2
Block memory Von Neumann method Central limit theorem	Guneysu and Paar [24]	Xilinx XC3S200	7
	Guneysu and Paar [25]	Xilinx XC3S200	13.1
RS latch metastability (LUT latch 64)	Hisashi and Ichikawa [26]	Xilinx XC4VFX20	3.85
(LUT latch 128)		Xilinx XC4VFX20	8.33
(LUT latch 256)		Xilinx XC4VFX20	12.5
Logic gates	Dichtl and Golić [2]	Xilinx XC3S200–4FT256C	12.5
Open loop	Danger et al. [3]	Altera Stratix EP1S25	20
Pyramid RO TRNG + LFSR	**Proposed work**	Cyclone II EP2C20F484C6 FPGA	**26.613660**

Acknowledgements The authors wish to thank SASTRA University for providing infrastructure through the Research & Modernization Fund (Ref. No: R&M/0026/SEEE–010/2012–13) to carry out the research work.

References

1. Bagini, V., Bucci, M.: A design of reliable true random number generator for cryptographic applications. Cryptogr. Hardw. Embed. Syst., 204–218 (1999)
2. Dichtl, M., Golić, J.D.: High-speed true random number generation with logic gates only. Cryptogr. Hardw. Embed. Syst., 45–62 (2007). https://doi.org/10.1007/978-3-540-74735-2_4
3. Danger, J.L., Guilley, S., Hoogvorst, P.: High speed true random number generator based on open loop structures in FPGAs. Microelectron. J., **40**, 1650–1656 (2009). https://doi.org/10.1016/j.mejo.2009.02.004
4. Deak, N., Gyorfi, T., Marton, K., Vacariu, L., Cret, O.: Highly efficient true random number generator in FPGA devices using phase-locked loops. In: 20th International Conference on Control Systems and Computer Science (2015), 453–458, (2015)
5. Güneysu, T., Paar, C.: Transforming write collisions in block RAMs into security applications. In: Proceedings of the 2009 International Conference on Field-Programmable Technology, FPT'09. (2009) 128–134. https://doi.org/10.1109/fpt.2009.5377631
6. Johnson, A.P., Chakraborty, R.S., Mukhopadhyay, D.: An improved DCM-based tunable true random number generator for xilinx FPGA. IEEE Trans. Circuits Syst. II Express Briefs **99** (2016)
7. Park, M., Rodgers, J.C., Lathrop, D.P.: True random number generation using CMOS Boolean chaotic oscillator. Microelectronics J. **46**(12), 1364–1370 (2015)
8. Sunar, B., Martin, W.J., Stinson, D.R.: A provably secure true random number generator with built-in tolerance to active attacks. IEEE Trans. Comput. **56**(1), 109–119 (2007)
9. Schellekens, D., Preneel, B., Verbauwhede, I.: FPGA vendor agnostic true random number generator. In: 2006 International Conference on Field Programmable Logic and Applications FPL, 139–144 (2006). https://doi.org/10.1109/fpl.2006.311206
10. Wold, K., Tan, C.H.: Analysis and enhancement of random number generator in FPGA based on oscillator rings. In: Proceedings of International Conference on Reconfigurable Computing and FPGAs, ReConFig, vol. 2009, pp. 385–390 (2008)
11. Marghescu, A., Teseleanu, G., Maimut, D.S., Neacsa, T., Svasta, P.: Adapting a ring oscillator-based true random number generator for Zynq system on chip embedded platform. In: 2014 IEEE 20th International Symposium for Design and Technology in Electronic Packaging, SIITME, 197–202 (2014)
12. Varchola, M., Drutarovsky, M.: New high entropy element for FPGA based true random number generators in cryptographic hardware and embedded systems. In: Mangard, S., Standaert, F.-X. (eds.) CHES 2010: 12th International Workshop, Santa Barbara, USA, August 17–20, 2010. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 351–365 (2010)
13. Jessa, M., Matuszewski, L.: The use of delay lines in a ring-oscillator-based combined true random number generator. 2012 International Conference on Signals and Electronic Systems ICSES (2012)
14. Bassham, L.E., et al.: A statistical test suite for random and pseudorandom number generators for cryptographic applications. Natl. Inst. Stand. Technol., Gaithersburg, MD, USA, Tech. Rep., no. April, 2010
15. Martin, H., Peris-Lopez, P., Tapiador, J.E., San Millan, E.: A new TRNG based on coherent sampling with self-timed rings. IEEE Trans. Ind. Inf. **12**(1), 91–100 (2016)
16. Fischer, V., Bernard, F.: True random number generators in FPGAs. Security trends for FPGAS, 73–100 (2011). https://doi.org/10.1007/978-94-007-1338-3.2011
17. Fischer, V., Drutarovský, M., Šimka, M., Bochard, N.: High performance true random number generator in Altera Stratix FPLDs. In: Becker, J., Platzner, M., Vernalde, S. (eds.) Field Programmable Logic and Application, 14th International Conference FPL 2004, Leuven, Belgium, Aug 30–Sept 1, 2004. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 555–564 (2004). https://doi.org/10.1007/978-3-540-30117-2_57
18. Kwok, S.H.M., Lam, E.Y.: FPGA-based high-speed true random number generator for cryptographic applications. In: TENCON 2006—2006 IEEE Region 10 Conference, pp. 1–4 (2006). https://doi.org/10.1109/tencon.2006.344013

19. Deak, N., Gyorfi, T., Marton, K., Vacariu, L., Cret, O.: Highly efficient true random number generator in FPGA devices using phase-locked loops. In: 2015 20th International Conference on Control Systems and Computer Science, pp. 453–458 (2015)
20. Kohlbrenner, P., Gaj, K.: An embedded true random number generator for FPGAs. In: Proceedings of the 2004 ACM/SIGDA 12th International Symposium on Field Programmable Gate Arrays, ACM, New York, NY, USA, pp. 71–78 (2004). https://doi.org/10.1145/968280.968292
21. Varchola, M., Drutarovsky, M.: New high entropy element for FPGA based true random number generators. In: Mangard, S., Standaert, F.-X. (eds.) Cryptographic Hardware and Embedded Systems CHES 2010 12th International Workshop. St. Barbar. USA, Aug 17–20, 2010. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 351–365 (2010). https://doi.org/10.1007/978-3-642-15031-9_24
22. Wieczorek, P.Z.: Dual-metastability FPGA-based true random number generator. Electron. Lett. **49**, 744–745 (2013). https://doi.org/10.1049/el.2012.4126
23. Valtchanov, B., Fischer, V., Aubert, A.: Enhanced TRNG based on the coherent sampling. In: 3rd International Conference on Circuits, Signals, and Systems SCS 2009, 1–6 (2009). https://doi.org/10.1109/icscs.2009.5412601
24. Güneysu, T., Paar, C..: Transforming write collisions in block RAMs into security applications. In: Proceedings of the 2009 International Conference on Field-Programmable Technology, FPT'09, 128–134 (2009). https://doi.org/10.1109/fpt.2009.5377631
25. Güneysu, T.: True random number generation in block memories of reconfigurable devices. In: 2010 International Conference on Field-Programmable Technol. FPT'10, 200–207 (2010). https://doi.org/10.1109/fpt.2010.5681499
26. Hisashi, H., Ichikawa, S.: FPGA implementation of metastability-based true random number generator. IEICE Trans. Inf. E95–D, 426–436 (2012). https://doi.org/10.1587/transinf.e95.d.426

A Multi-objective Approach to Study the Effects of Ball Race Conformity on Optimum Design of Rolling Element Bearing Using Metaheuristics

S. N. Panda, S. Panda and D. S. Khamari

Abstract One of most critical performance parameters of deep groove ball bearing is the achievable highest fatigue life. The life of the bearing may be affected by more than one factor as types of lubrication, thermal attributes, etc. This present work emphasizes on optimization of life factors, life of bearing along with dynamic load capacity using metaheuristic algorithm based upon the particle swarm methodology (PSO). The two objectives as life factor and dynamic load capacity have been simultaneously optimized using a multi-objective approach under realistic constraint conditions, and also for efficient constraint handling, penalty function method has been used. Right here for formulation of life factors, critical assumptions of operating condition, reliability and materials and processing are being considered. From the above assumptions, it is also considered that the total bearing system is following a series combination of subsystems. A convergence study has been applied to the stated optimization problem. The outcome of this multi-objective approach shows the better efficacy and generosity of the particle swarm algorithm, which can also be implemented for future engineering problems.

Keywords Particle swarm optimization · Life factors · Life of bearing · Penalty function · Constraint violation · Ball race conformity

S. N. Panda (✉)
Department of Production Engineering, B.I.T., Sindri, Jharkhand 828123, India
e-mail: suryanarayan.uce@gmail.com

S. Panda
Department of Mechanical Engineering, V.S.S. University of Technology, Burla 768018, Odisha, India
e-mail: sumanta.panda@gmail.com

D. S. Khamari
Department of Mechanical Engineering, Govt. College of Engineering, K alahandi 766002, Odisha, India
e-mail: debanshushekhar@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_4

1 Introduction

The design imperfection of rolling element bearing has a noticeable impact on the performances like endurance life, wear life and working under sophisticated operating conditions. Thus, an efficient design of bearing can result in better quality low-cost operation. Optimum design of rolling element bearing is an increasingly important area in the development of precision bearings, but it is a tedious task for bearing engineer. Since the beginning of development of rolling element bearings, two major issues have been analyzed in different ways for efficient design of rolling element bearings as, life of the bearing considering fatigue endurance of rolling element bearings and load carrying capacity of rolling element bearings. In recent days, with the ease of developing computational methods the optimal design of rolling element bearing can be predicted for effective bearing life.

Goodman [1] is the first investigator to identify the phenomenon of rolling contact fatigue and put forward the concept of endurance rating of the bearings, and later on, the author also provided the concept of spalling, which is an important aspect for determining the endurance of rolling element bearings. Palmgren [2] predicted bearing life or bearing life rating using probability of failure and formulates about bearing life measures developing an analytical formula as L_{10} life. Lundberg and Palmgren [3, 4] to develop a basic equation for fatigue life in terms of probability of survival relating to numbers of stress cycles and volume of stress materials and finally the relation between bearing fatigue life and applied load were determined by considering the point contact and line contact with rolling surfaces, and also prediction of bearing life is performed by the authors considering the inner and outer race lives. Though this attempt of empirical representation was not in good agreement with the theoretical explanations, continuous attempts were made for improvement of the formulation, and finally, Palmgren [5] proposed the most approximate relation of life rating of bearing specific to ball bearing with compliance to the test data. The author concluded that the impact of the lubrication on endurance life can also be empirically determined under adequate lubricating condition. Zaretsky [6] concluded in his study that the bearing life is affected by race geometry and also depend on the Hertzian stresses at respective races. In another work, Zaretsky et al. [7] incorporate the life factors using the Lundberg–Palmgren theory and concluded that the bearing life is affected by inner and outer race conformity. In this work, the authors also proposed algebraic relations for calculating life factors (LF_c) and determine L_{10} life for ball bearing.

A constraint optimization technique based on GA has been first adopted for finding the optimal design of rolling element bearing by Chakraverty et al. [8], considering five design variables which are affecting the bearing life. Rao and Tiwari [9], in their study, considered the assembly angle and some constant of constraints as different input design variables in the scope of optimum design of rolling element bearing. The authors have used some of realistic constraints that contain constants whose values are decided by parametric studies through the initial optimization run. Tiwari et al. [10] also proposed optimization methodology for optimum fatigue life of bearing considering tapper rolling element bearing. The authors used evolutionary-based optimization algorithm for deriving the optimum solution for best achievable fatigue life. Much of the work has been proposed considering single-objective approach of optimization, and in contrast, Gupta et al. [11] described optimum design of rolling element bearing considering multi-objective method based on NSGA-II with dynamic load and static load carrying capacity along with elastohydrodynamic minimum film thickness as various objectives. Most of the investigators in the field of optimum design of bearing used the analytical model as proposed by Changsen [12] for modeling of performance parameters in scope to rolling element bearing. A close look on the above review states that the application of soft computing strategies in the optimum design of rolling element bearing is the recent trend, but research scope is open as most of the authors have not considered the Lundberg and Palmgren theory and Zaretsky formulation for analyzing the performance aspects of rolling element bearing.

The aim of this proposed work is to incorporate Lundberg–Palmgren theory along with Zaretsky approach and to predict the optimum design of rolling element bearing using multi-objective constraint optimization approach, considering simultaneous optimization of two objectives as dynamic load capacity and life factors of bearing, which is affected by inner and outer race conformities. The analytical model used for aforesaid problem based upon the prediction of Zaretsky et al. [7], which includes the algebraic relation among the bearing endurance life with the Hertzian stress as 9th power. The author also considered the life of the total bearing system is a series combination of that of subcomponents of bearing that is the life of inner race and outer races, respectively.

2 Bearing Macro-geometry

The macro-geometry of deep groove ball bearing can be described with geometric parables like internal bore diameter (d), external diameter (D), width of bearing (w), diameter of ball (D_b), pitch diameter (D_m), inner and outer raceway curvature coefficients (f_i and f_o) and numbers of rolling element (Z), (Fig. 1).

Here, optimum design deep groove ball bearing is predicted, considering optimum dynamic load capacity and life factors, which are influenced by inner and outer conformity ratios for bearing geometry of $D = 30$ mm, $d = 10$ mm and $w = 9$ mm, respectively.

2.1 Objective Function

In case of rolling element bearing specifically for deep groove ball bearing, the most prominent performance characteristics are dynamic load capacity (C_d), and thus, this performance parameter can be optimized for maximization of value as to achieve best

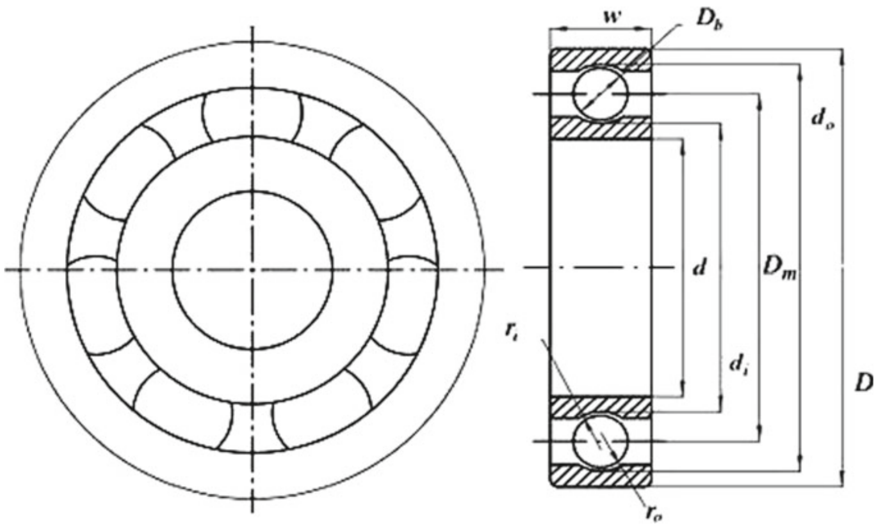


Fig. 1 Radial deep groove ball bearing internal geometries

performing rolling element bearing. The dynamic load carrying capacity is directly influenced by the geometrical design variables [12] and can be formulated as Eq. (1).

Fatigue life in terms of 10^6 revolutions,

$$L = \left(\frac{C_\mathrm{d}}{F}\right)^a \tag{1}$$

where applied load is termed by F and for ball bearing, $a = 3$.

$$C_\mathrm{d} = \begin{cases} \max\left[-f_\mathrm{c} Z^{2/3} (i\ \cos\alpha)^{0.7} D_\mathrm{b}^{1.8}\right] & D_\mathrm{b} \le 25.4\ \mathrm{mm} \\ \max\left[-3.647 f_\mathrm{c} Z^{2/3} (i\ \cos\alpha)^{0.7} D_\mathrm{b}^{1.4}\right] & D_\mathrm{b} > 25.4\ \mathrm{mm} \end{cases} \tag{2}$$

$$f_\mathrm{c} = 37.91 \left\{ 1 + \left[1.04 \left(\frac{1-\gamma}{1+\gamma}\right)^{1.72} \left(\frac{f_\mathrm{i}(2f_\mathrm{o}-1)}{f_\mathrm{o}(2f_\mathrm{i}-1)}\right)^{0.41} \right]^{10/3} \right\}^{-0.3}$$
$$\left[\frac{\gamma^{0.3}(1-\gamma)^{1.39}}{(1+\gamma)^{1/3}}\right] \left[\frac{2f_\mathrm{i}}{2f_\mathrm{i}-1}\right]^{0.41} \tag{3}$$

where $\gamma = D_\mathrm{b}\cos\alpha/Dm$ and α is the contact angle, which depend upon the variety of bearing under consideration. In this study, the deep groove ball bearing is analyzed, for which $\alpha = 0$. The term i is the number of rows in deep groove ball bearings, and in this study, $i = 1$ (single row deep groove ball bearing).

2.2 Bearing Life Prediction

Some early researchers [3–5] concluded that the survival s depends upon the orthogonal shear stress, life η, depth of orthogonal shear stress maximum Z_o and stressed volume V. This relation can be written as Eqs. (4), (5).

$$\ln\frac{1}{s} \sim \tau_o \frac{\eta^e}{Z_o^h} V \tag{4}$$

$$V = aZ_o l \tag{5}$$

where l is the circumference of races which indicates the length of the rolling path for rolling element, a, Z_o are functions of the maximum Hertzian stress.

Formulation for the bearing life [4],

$$L_{10} = \left(\frac{C_d}{p_{eq}}\right)^p \tag{6}$$

The value of exponent $p = 3$ for deep groove ball bearing.

It is a fair possibility that some aspect like reliability, materials and processing and operating conditions may impact the bearing life, and thus, the above formulation can be expressed as Eq. (7) considering a_1, a_2 and a_3 coefficients of stated aspects.

$$L = a_1 a_2 a_3 Z_o L_{10} \tag{7}$$

The analytical relation between the fatigue life of the bearing and maximum Hertzian stress S also with the equivalent bearing load P_{eq} can be expressed in Eq. (8).

$$L \sim \left(\frac{1}{p_{eq}}\right)^p \sim \left(\frac{1}{S_{max}}\right)^n \tag{8}$$

Accordance with Hertzian theory $p = n/3$ and as per Lundberg–Palmgren Hertzian stress life exponent $n = 9$.

2.3 Life Ratio of Bearing

The endurance life L of the individual races as inner and outer can be formulated as follows,

$$L \sim \left(\frac{1}{S_{max}}\right)^n \left(\frac{1}{l}\right)^{\frac{1}{e}} \left(\frac{1}{N}\right) \tag{9}$$

where number of load cycle per revolution is represented by *N,* and Weibull's slope is represented by e whose value is 1.11 for standard cases.

The life ratio indicates the ratio between life of outer race and inner race and expressed as X, Eq. (10).

$$X = \frac{L_{\mathrm{o}}}{L_{\mathrm{i}}} \approx \left(\frac{S_{\mathrm{maxi}}}{S_{\mathrm{maxo}}}\right)^{n} \left(\frac{1}{k}\right)^{\frac{1}{e}} \tag{10}$$

2.4 Life Factor of Bearing

The formulation for life factors can be determined considering the life factors of inner and outer races, $\mathrm{LF_i}$ and $\mathrm{LF_o}$, respectively, and considering the race conformities after normalizing the value of Hertzian stress for inner and outer races to a standard conformity of 0.52 [6].

$$\mathrm{LF} = \left(\frac{S_{\mathrm{max0.52}}}{S_{\mathrm{max}}}\right)^{n} \tag{11}$$

The race conformity values depend upon transcendental functions, whose values are different for respective races, and these functions are represented as μ, ν, respectively.

2.5 Series System Reliability

Lundberg–Palmgren [3] first considered the total bearing life can be expressed as combination of individual life of subcomponents, i.e., primarily inner and outer races. This concept has been implemented by Zaretsky [6] so to express total bearing life and the author formulated a series combination of life of subcomponent gives the total bearing life as in Eq. (12).

$$\frac{1}{L_{10}^{e}} = \frac{1}{L_{10_i}^{e}} + \frac{1}{L_{10_o}^{e}} \tag{12}$$

2.6 Life Factors and L_{10m} Life of Bearing

The life factors for respected race life as inner and outer are expressed as in Eqs. (13) and (14) [7],

$$\mathrm{LF_i} = \left[\frac{\left(\frac{2}{D_\mathrm{m}-D_\mathrm{b}} + \frac{4}{D_\mathrm{b}} - \frac{1}{0.52D_\mathrm{b}}\right)^{2/3} (\mu\nu)_i}{\left(\frac{2}{D_\mathrm{m}-D_\mathrm{b}} + \frac{4}{D_\mathrm{b}} - \frac{1}{f_\mathrm{i}D_\mathrm{b}}\right)^{2/3} (\mu\nu)_{0.52}} \right]^n \tag{13}$$

$$\mathrm{LF_o} = \left[\frac{\left(-\frac{2}{D_\mathrm{m}-D_\mathrm{b}} + \frac{4}{D_\mathrm{b}} - \frac{1}{0.52D_\mathrm{b}}\right)^{2/3} (\mu\nu)_o}{\left(-\frac{2}{D_\mathrm{m}-D_b} + \frac{4}{D_\mathrm{b}} - \frac{1}{f_\mathrm{o}D_\mathrm{b}}\right)^{2/3} (\mu\nu)_{0.52}} \right]^n \tag{14}$$

Life factor and $L_{10\mathrm{m}}$ life of rolling element bearing can be expressed as per Lundberg–Palmgren approach, Eqs. (15) and (16).

$$\mathrm{LF_c} = \left[\frac{(\mathrm{LF_i})^e (\mathrm{LF_o})^e (X^e + 1)}{(\mathrm{LF_o})^e X^e + (\mathrm{LF_i})^e} \right]^{1/e} \tag{15}$$

$$L_{10_m} = \frac{(\mathrm{LF_i})(\mathrm{LF_o}) X L_i}{\left[(\mathrm{LF_o})^e X^e + (\mathrm{LF_i})^e\right]^{1/e}} \tag{16}$$

Life factor and $L_{10\mathrm{m}}$ life of rolling element bearing can be expressed as per Zaretsky approach, Eq. (17) and Eq. (18).

$$\mathrm{LF_c} = \left[\frac{(\mathrm{LF_i})^e (\mathrm{LF_o})^e (X + 2)}{(\mathrm{LF_o})^e X^e + 2(\mathrm{LF_i})^e} \right]^{1/e} \tag{17}$$

$$L_{10_m} = \frac{(\mathrm{LF_i})(\mathrm{LF_o}) X L_i}{\left[(\mathrm{LF_o})^e X^e + 2(\mathrm{LF_i})^e\right]^{1/e}} \tag{18}$$

2.7 Synthesis of Multi-objective Function

The objective of this proposed approach of design optimization for rolling element bearing is to maximize dynamic load carrying capacity and minimize life factors simultaneously. Thus, in this proposed work using goal attainment method the multi-objective optimization problem can be formulated as Eq. (19).

$$\text{Max } F(x) = w_1((C(x))/(C_k)) + w_2((-\mathrm{LF}(x))/(\mathrm{LF}_k)) \tag{19}$$

Equation (19) can be formulated for both Lundberg–Palmgren and Zaretsky approaches, respectively, where C_k and LF_k are the optimized value of $C(x)$ and $LF(x)$ when they are optimized as a single-objective case. w_1 and w_2 are the weighing coefficients, and their values are 0.5 and 0.5, respectively (such that $w_1 + w_2 = 1$). The weighing coefficients are selected depending upon the importance of objective function in the multi-objective optimization problem. Many constraint conditions [11] are being imposed upon the optimization problem, so to get the global optimum value with the ease of design space.

2.8 Synthesis of Design Variables

The various design variables involved in this proposed optimization problem are mostly geometric features, along with some constant parameters. The various geometric features are D_m, D_b, Z, f_i and f_o. The constant parameters in different constraints are K_Dmin, K_Dmax, ε, e and β. All the geometric features and constant terms are considered as the input design variables in the proposed design optimization of rolling element bearing. The ten input parameters are presented in Eq. (20).

$$X = [D_\mathrm{b}, Z, D_\mathrm{m}, f_\mathrm{o}, f_\mathrm{i}, K_{\mathrm{D\,min}}, K_{\mathrm{D\,max}}, \varepsilon, e, \beta] \tag{20}$$

3 Optimization Algorithm (PSO)

A population-centered evolutionary algorithm proposed, particle swarm optimization (PSO), has been developed by Kennedy and Eberhart [13]. Bound to the search space, each and every particle maintains track of its positions, which is associated with the most effective solution (fitness) it has observed up to now, *pBest*. Another best esteem followed by the global best version of the particle swarm optimizer is the global best value, *gBest* and its position, obtained thus far by any particle in the population. The strategy for implementing PSO is to randomly initialize a population of particles, evaluate the fitness value of each and every particle and compare the evaluated value with global fitness value if not equal then update velocity and position of particle to find the global optimum fitness value. Updation of the velocity and position of the particle can be performed as per Eqs. (21), (22). The above steps can be looped for further evaluating fitness value of each particle till the terminating criterion is agreed. The stopping criterion generally is taken as the maximum number of iterations.

$$\begin{aligned} v[] = v[] &+ c1 * \mathrm{rand}() * (\mathrm{pBest}[] - \mathrm{present}[]) \\ &+ c2 * \mathrm{rand}() * (\mathrm{gBest}[] - \mathrm{present}[]) \end{aligned} \tag{21}$$

$$\text{present}[] = \text{present}[] + v\,[] \tag{22}$$

4 Constraint Handling

In this proposed constrained nonlinear optimization problem, the “penalty function” approach is a widely accepted technique for constraint handling. In this proposed method, a penalty function $P(x)$ is added to the original objective function, finally given the pseudo-objective function, hence to govern the violations of constraint. In the first step after finding individual constraint violation, all the violations are summed up to get an overall constraint violation. The total constraint violation, then after multiplied with a penalty parameter added to the main objective function, so to formulating the pseudo-objective function. The objective function is penalized by a constant factor, r_p, when it meets an active constraint. A higher value of r_p is chosen to ensure the feasibility of the solution and improvement of efficiency of algorithm. The pseudo-objective function is represented in Eq. (23).

$$\phi(X) = F_\mathrm{c}(X) + r_\mathrm{p} * P(X) \tag{23}$$

$$P(X) = \left[\sum_{i=1}^{m}\left\{\max[0, g_i(X)]^2\right\} + \sum_{i=1}^{p}[h_i(X)]^2\right] \tag{24}$$

In this case, the actual objective function is denoted as $F_\mathrm{c}(x)$, and the added penalty function is $P(x)$. The equality and inequality constraints are presented as g_i and h_i, respectively. Using the penalty function approach, the stated constrained optimization problem is converted into an unconstrained optimization problem.

5 Outcomes and Analysis

Tables 1, 2 give the outcomes for optimum values of design variables and dynamic load capacity and life factors using PSO under realistic constraint conditions [9] in Lundberg–Palmgren and Zaretsky approach, respectively. In this multi-objective approach, the dynamic load capacity in Lundberg–Palmgren approach and Zaretsky approach is found to be 5327 and 5810.3, respectively. These values are quite better on average of 8.51 and 2% than the catalog values and values from early researches [11], respectively. The life factor values in multi-objective approach in the above two approaches found to be 1.0924 and 1.0637, respectively, which are found to be 3% better than the values obtain by Zaretsky et al. [7]. Lesser values of life factor indicated that the inner race is under higher Hertzian stress as compared to outer

Table 1 Optimization outcomes for Lundberg–Palmgren approach

D	d	W	D_b	D_m	Z	f_i	f_o	K_{Dmin}	K_{Dmax}	ε	e	β	C	LF_c
30	10	9	5.8	20.94	7	0.515	0.515	0.447	0.667	0.3	0.0468	0.7242	5327	1.0924

Table 2 Optimization outcomes for Zaretsky approach

D	d	W	D_b	D_m	Z	f_i	f_o	K_{Dmin}	K_{Dmax}	ε	E	β	C	LF_{cz}
30	10	9	5.7	20.76	7	0.514	0.515	0.492	0.702	0.3	0.0475	0.7135	5810.3	1.0637

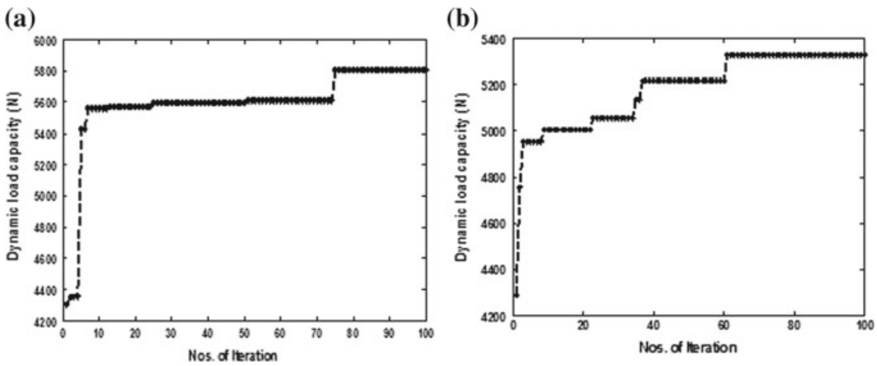


Fig. 2 Convergence traits of dynamic load capacity using PSO, **a** Lundberg–Palmgren approach, **b** Zaretsky approach

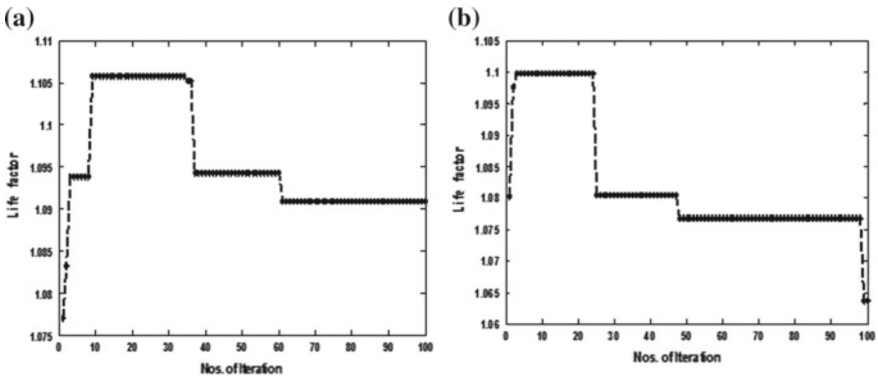


Fig. 3 Convergence traits of life factors using PSO, **a** Lundberg–Palmgren approach, **b** Zaretsky approach

race, which suggests the efficiency of design approach agreement with Lundberg–Palmgren approach. Further, Figs. 2, 3 depict the convergence characteristic dynamic load capacity and life factor in both the stated approaches. The general trend found for dynamic load capacity which suggests the upside variation of optimum value in early iterations before settling in downward side. Same type of variation is observed for life factor, but opposite in trend.

6 Conclusions

The proposed population-based algorithm as PSO algorithm has been successfully applied to the stated multi-objective optimization problem under realistic constraint conditions involving the dynamic load capacity and life factor as two objectives for rolling element bearing specifically for deep groove ball bearing. The algorithm efficiently handled the above nonlinear mixed integer problem. The results reported indicate that the proposed population-based algorithm approach can be efficiently used in design-oriented problems in bearings. The analytical and convergence characteristics agree with the effect of race conformity on performance along with the contact stress which is quite noticeable, and thus, this can impact the life of the bearing. Tiwari et al. [10] also reported the same prospects as conformity ratios can affect the smooth operation and life of bearing, in the field of optimum design of bearings. So it can be concluded that this study is acceptable compliance with the investigation of Zaretsky et al. [7] for formulating endurance life of rolling element bearing. The above study can enable the bearing designer to analyze the changes in life of bearing which includes the effect of race conformity for real-time problems in rolling element bearing.

References

1. Goodman, J.: Roller and ball bearings and (2) the testing of antifriction bearing materials. Proc. Inst. Civ. Eng. **189**(4) (1911)
2. Palmgren, A.: Die Lebensdauer von Kugellagern (The service life of ball bearings). Z. Ver. Devt. Ingr. **68**(14), 339–341 (1924)
3. Lundberg, G., Palmgren, A.: Dynamic capacity of rolling bearings. Acta Polytech. Mech. Eng. Ser. **1**, 3 (1947)
4. Lunberg, G., Palmgren, A.: Dynamic capacity of roller bearings. Acta Polytech. Mech. Eng. Ser **2**, 96 (1952)
5. Palmgren, A.: Ball and Roller Bearing Engineering. SKF Industries Inc., Philadelphia (1959)
6. Zaretsky, E.V.: STLE Life Factors for Rolling Bearings, STLE SP–34. Society of Tribologists and Lubrication Engineers, Park Ridge, IL (1992)
7. Zaretsky, E.V., Poplawski, J.V., Root, L.E.: Reexamination of ball-race conformity effects on ball bearing life. Tribol. Trans. **50**(3), 336–349 (2007)
8. Chakraborthy, I., Vinay, K., Nair, S.B., Tiwari, R.: Rolling element bearing design through genetic algorithms. Eng. Optim. **35**(6), 649–659 (2003)
9. Rao, B.R., Tiwari, R.: Optimum design of rolling element bearings using genetic algorithms. Mech. Mach. Theory **42**(2): 233–250 (2007)
10. Tiwari, R., Kumar, S.K., Prasad, P.V.V.N.: An optimum design methodology of tapered roller bearing using genetic algorithm. Int. J. Comput. Methods Eng. Sci. Mech. **13**(2), 108–127 (2012)
11. Gupta, S., Tiwari, R., Nair, S.B.: Multi-objective design optimisation of rolling bearings using genetic algorithms. Mech. Mach. Theory **42**(10), 1418–1443 (2007)
12. Changsen, W.: Analysis of rolling element bearings. Mechanical Engineering Publications Ltd., London (1991)
13. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks (ICNN), 1942–1948, (1995)

Design and Analysis of Higher-Order Sigma Delta Modulator

Deepti Malhotra and Alpana Aggarwal

Abstract The paper focuses on basics of analog-to-digital conversion. Various types of analog-to-digital converters are available. Here, the basics of one of its type, sigma delta, is discussed, starting from quantization and then moving to first-, second-, and third-order structures. The SNR obtained is 62 dB and 61.4 B, 91.9, respectively. Moving toward higher order with same approach causes instability. Therefore, an architecture using feedback or feedforward may be used. Some of the popular higher-order topologies are based on feedback and feedforward arrangements: cascade of resonator-feedback form (CRFB), cascade of resonator-feedforward form (CRFF), cascade of integrator-feedback form (CIFB), and cascade of resonator-feedforward form (CIFF) with SNR achieved as 140 dB, 153.6 dB, 141.8 dB, and 164.4 dB, respectively.

Keywords Sigma delta · Quantization · Higher order · Nyquist · Oversampling · ADC · NTF · STF

1 Introduction

There have been tremendous advancements in digital circuits with its speed density increasing constantly. However the signals available, majority times are analog in nature. Their vulnerability to distortion limits their use and encourages their conversion to digital domain. This leads to the extensive use of various types of ADC's classified according to their sampling techniques. One category of Nyquist rate ADC operates at sampling frequency twice the baseband frequency, and the other one operates at frequency much greater than Nyquist frequency popularly called oversampling converters. Oversampling helps in distributing the noise spectrum over wide frequency range. Sigma delta ADC which utilizes oversampling technique has found wide range of applications in instrumentation, communication, biomedical, etc. Here, difference of input and output waveform is digitized instead of one-to-one mapping

D. Malhotra (✉) · A. Aggarwal
Electronics and Communication Engineering, Thapar University, Patiala, Punjab, India
e-mail: deeptimalhotra1981@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_5

as done in Nyquist converters. Apart from oversampling another striking feature of $\sum\Delta$ ADC which stands it out from the other ADC is the feature of noise shaping. This feature helps in pushing most of the noise spectrum to higher frequencies to reduce the in-band noise. The resolution obtained from Nyquist reaches maximum till 12, whereas for oversampling it is greater than 20. All these features have made oversampling converters to take over many Nyquist-based applications.

2 Quantization

The main element that contributes to noise in analog-to-digital conversion is quantization. However, being uniformly distributed it can be easily evaluated. SNR can be calculated as

$$\mathrm{SNR} = 6.02\mathrm{N} + 1.76\mathrm{dB} \tag{1}$$

With half a decrease in LSB noise power decreases by 4 and SNR rises by 6 dB and with every 1-bit increase, there is 6-bit increase in SNR. Figure 1 shows the input quantized into 10 levels, and the mean error uniformly distributed over entire frequency range with average value of 0.305. To avoid overloading of quantizer, the input should be restricted to no-overloading range.

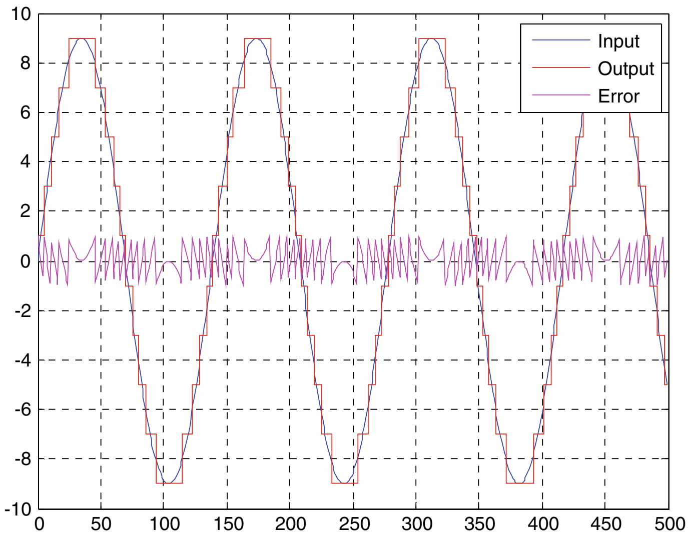


Fig. 1 Quantization of a sine signal

3 Order of the Modulator

The first order of $\sum\Delta$ described in Fig. 2a can be defined by Eq. (1)

$$V(z) = E(z)\left(1 - z^{-1}\right) + U(z) \tag{2}$$

Here, signal transfer function (STF) is 1, and noise transfer function (NTF) is $(1 - z^{-1})$. The transfer function emphasizes that NTF makes modulator that behaves like HPF for noise and LPF for input signal. The in-band noise is given by Eq. (3)

$$\text{IBN} = \frac{\Delta^2 \pi^2 \text{OSR}^{-3}}{36} \tag{3}$$

For every double in OSR, the noise reduces by 9 dB and SQNR rises by 9B dB. This results in improvement of 1.5 bit of resolution [1]. Figure 2a shows the block diagram for first-order $\sum\Delta$ and b, c shows the simulated output and the output spectrum for OSR of 80, respectively. Figure 2b shows simulation of sine wave input which results in 1-bit serial output signal. Higher amplitude level in sine wave leads to more number of ones at output and low signal level means more number of zeros

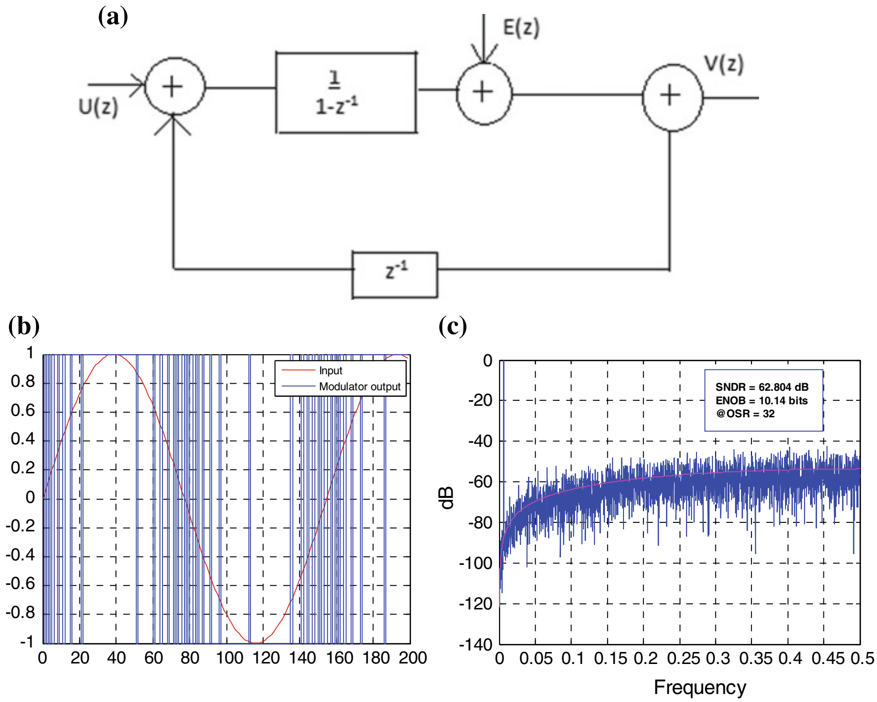


Fig. 2 **a** First-order modulator, **b** Time-domain simulations of first order, **c** Output spectrum

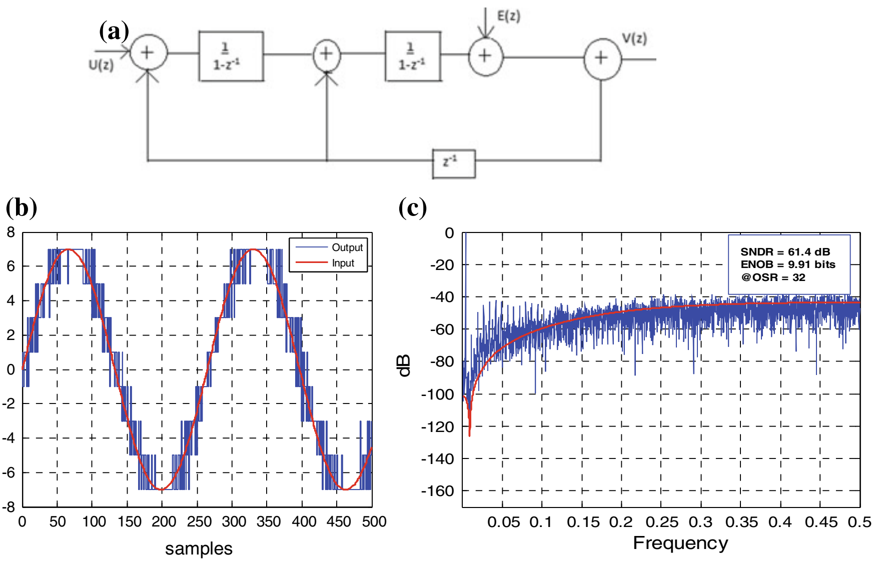


Fig. 3 **a** Second-order modulator, **b** Time-domain simulations of second order, **c** Output spectrum

at output. At ‘0’ amplitude equal number of ones and zeros are observed at output. Figure 2c shows the power spectrum of output bitstream. Clearly, it shows the shaping of quantization noise away from desired frequency band.

Figure 3a shows the block diagram for second-order $\sum\Delta$ and b, c shows the simulated output and the output spectrum for OSR of 80, respectively. The ideal second-order modulator (replicating first order) will be

$$V(z) = E(z)\left(1 - z^{-1}\right)^2 + U(z) \tag{4}$$

$$\text{Here the IBN} = \frac{\Delta^2 \pi^4 \mathrm{OSR}^{-5}}{60} \tag{5}$$

According to Eq. (5) for every double in OSR, the SNR rises by 15 dB and resolution is improved by 2.5 dB [2]. Use of multiple feedbacks introduces more flexibility to enhance stability and dynamic range and stability. Each structure results in different NTF and STF. The one which offers minimum IBN is preferred. Figure 3b, c shows the simulated output in time domain and frequency domain, respectively, corresponding to second-order modulator at OSR of 80. The feedback loop processes the average error instead of the individual samples. The out of band gain (OBG) of first order is 2, and for second order, it is 4 making larger LSB jumps for second order [3].

The nonlinearity of quantizer cannot be neglected at higher orders as the overloading causes unstability. As a result of this, the quantizer’s output starts wiggling. One way of avoiding overloading and improving stability of the quantizer is by increasing the number of levels by using multibit quantizer. This will reduce LSB (LSB = FS/2^n) and increase maximum stable input (MSA) [4]. Apart from lowering the IBN quantization noise, it also lowers the slew rate requirements in loop filter. Both input and output of DAC consist of two points $\pm V_{\mathrm{ref}}$ making its operation highly linear. But in cases where multibit quantizer is used, dynamic matching techniques may be used [5]. Figure 4a, b shows the time-based simulations and spectral density of its contents. Table 1 describes the comparisons of all three orders. Figure 5 plots SNR of third order as a function of input levels for smaller amplitudes. Table 2 correctly shows the variation of vital performance parameters of the ADC by choosing different architectures.

As the complexity increases and other parameters like finite op-amp slew rate, quantizer delay varies. Various other structures for higher orders have been suggested, viz. Booser Wooley and Silva-Steensgaard. Both vary in terms of placement of feedback and feedforward coefficient. Method of describing function analysis is also used in which approximation is done using curve fitting [6]. Here, noise is eliminated during clipping. Also at higher orders, MSA reduces as OBG increases. One way by which OBG can be reduced is by introducing poles in signal band while not altering the IBN. Butterworth, inverse Chebyshev can be used to alter the responses [7].

$$\mathrm{SNR} = \frac{9M^2(\mathrm{OSR})^3}{2\pi^2} \tag{6}$$

$$\mathrm{SNR} = \frac{15M^2(\mathrm{OSR})^5}{2\pi^4} \tag{7}$$

Figure 6 shows the theoretical tradeoff between SNR and OSR for first order and second order. For second order, SNR varies with fifth power of OSR unlikely to first order where it varies for third power of OSR. Hence, doubling of OSR causes improvement in SNR by 9 dB in order 1, and same change improves SNR by 15 dB in order 2.

4 Higher-Order NTF

The architectures can be extended to higher orders also. These depict better performances but at the cost of reduced signal range and increased hardware. Higher-order noise shaping has reduced IBN and high SNR. For NTF = $(1 - z^{-1})^N$, the IBN is given as (8)

$$\mathrm{IBN} = \frac{\Delta^2}{12}\frac{\pi^{2N}\mathrm{OSR}^{-(2N+1)}}{(2N+1)} \tag{8}$$

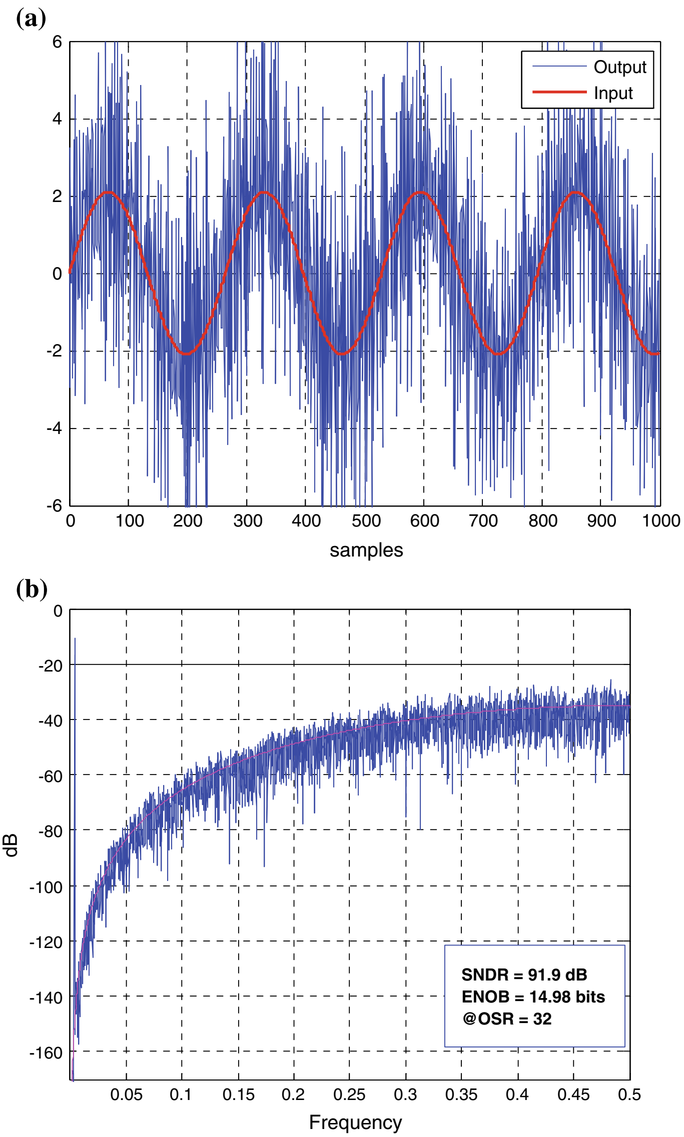


Fig. 4 **a** Time-domain simulations, **b** Output spectrum for third order

Table 1 Parameter variation w.r.t order

Order	OSR	SNR (dB)	Resolution
First	32	62	10.14
Second	32	61.4	9.91
Third	32	91.9	14

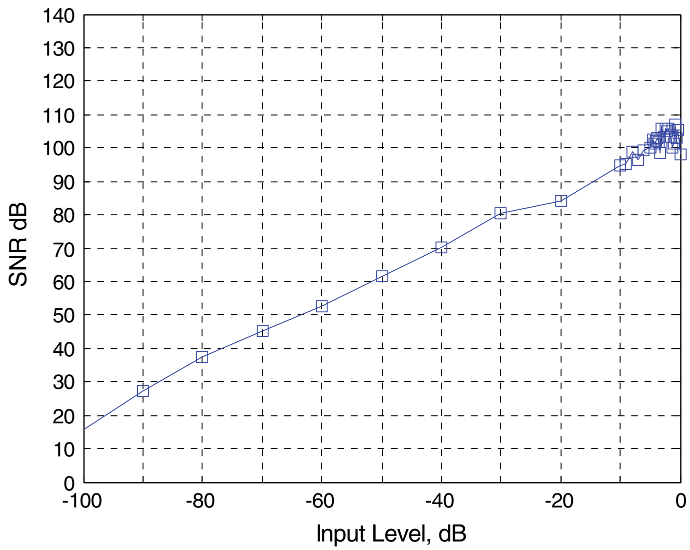
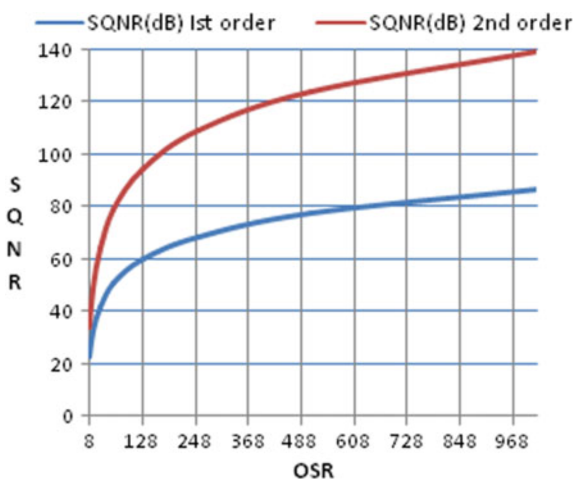


Fig. 5 SNR versus input level

Table 2 Parameter variation w.r.t architectures

Architecture	OSR	SNR (dB)	Resolution
CIFB	80	140	23.09
CRFB	80	153.6	25.23
CIFF	80	141.8	23.23
CRFF	80	164.4	26.36

Fig. 6 SNR versus OSR



The modulator basically constitutes two parts: loop filter and quantizer as shown in Fig. 7.

$$V(z) = \mathrm{STF}(z)U(z) + \mathrm{NTF}(z)\,E(z)$$

where

$$\mathrm{NTF}(z) = \frac{1}{1 - L_1(z)} \text{ and } \mathrm{STF}(z) = \frac{L_0}{1 - L_1(z)}$$

Or $L_0 = \frac{\mathrm{STF}(z)}{\mathrm{NTF}(z)} \quad L_1 = 1 - \frac{1}{\mathrm{NTF}(z)}$

The NTF and STF are the quantization noise and input signal, respectively, represented as transfer function. These relations stand irrespective of the choice of the architectures. It can also be concluded that the stability of the modulator can be determined by loop gain $L_1(z)$ and NTF(z). All higher-order NTFs have stability issues. Also, the OBG (2^N) starts increasing. This result in large oscillations at quantizer's input. Hence, to reduce OBG and high IBN shaping, poles can be introduced into the NTF. Higher-order NTFs can be obtained by adding more integrator and feedback/feedforward branches to the loop. Some of the popular higher-order topologies are based on feedback and feedforward arrangements: cascade of resonator-feedback form (CRFB), cascade of resonator-feedforward form (CRFF), cascade of integrator-feedback form (CIFB), and cascade of resonator-feedforward form (CIFF) [8]. Figures 8, 9, 10, and 11 depict the poles zeros distribution of NTF, STF, $L_{0,}$ and L_1 along with the simulated output in frequency domain.

A number of alternative structures exist which perform second order/third order and give unity gain STF and same NTF. Here, all delay-free integrators are converted into delaying integrators and removing feedback delays varies NTF. Some of these are as follow: Booser Wooley, Silva-Steensgaard, Error feedback, multiple feed-in/feedback coefficients. While choosing any architecture, the choice should be on the basis of optimal NTF. STF may be ignored as its role is to filter the signal only as peak SNR and IBN are calculated on basis of NTF only [9, 10]. By default, LPF structure has been considered; however, BPF and HPF can also be considered using suitable transforms. This shall extend the role of sigma delta in wireless also.

Another most widely used structure is multistage noise shaping (MASH). These are the cascade structures where one or more modulators are aligned together. In

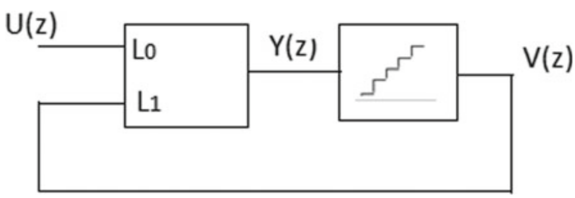


Fig. 7 Block diagram for higher orders

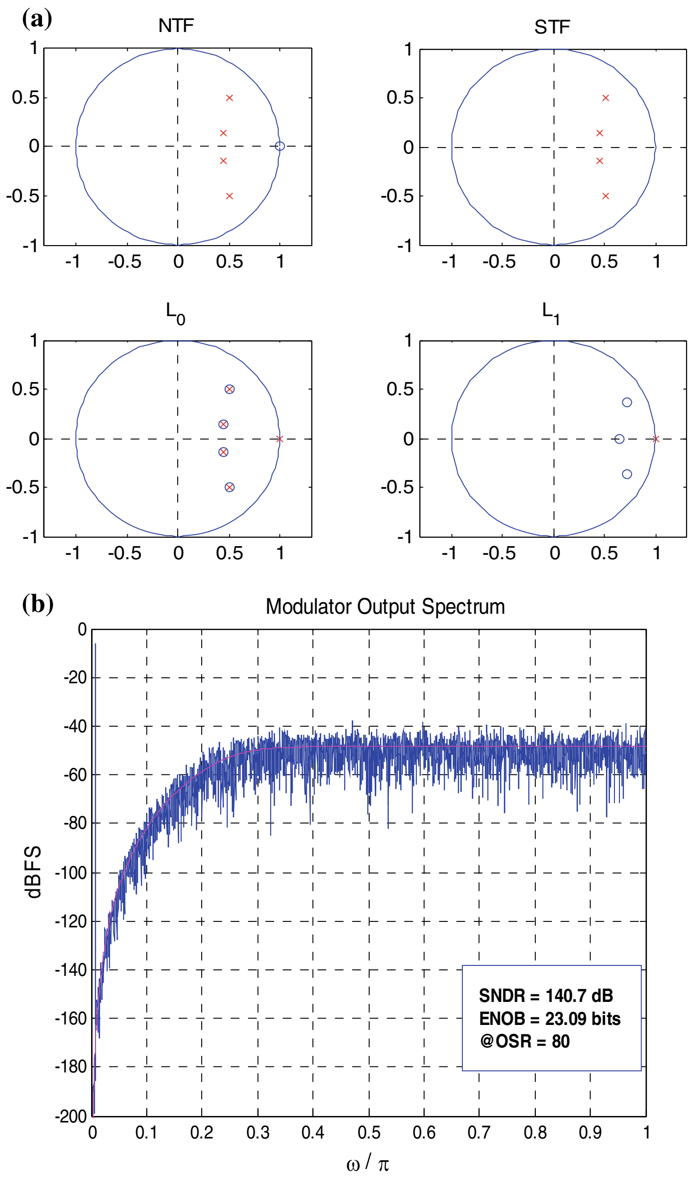


Fig. 8 CIFB. **a** Pole zero plot, **b** Output spectrum

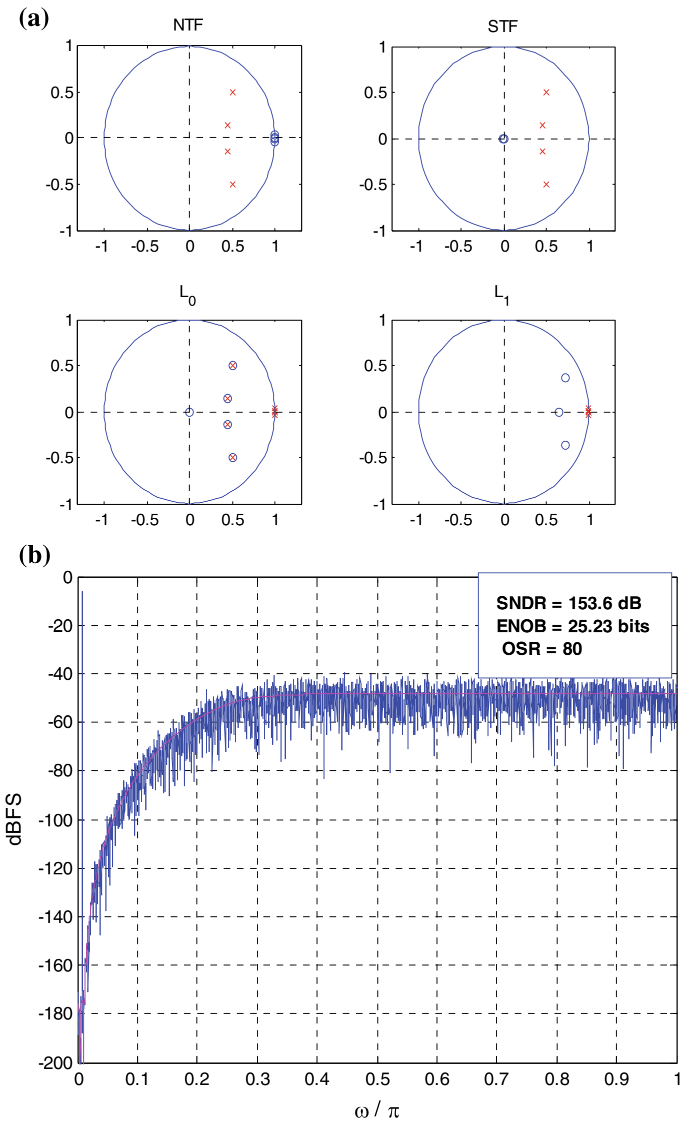


Fig. 9 CRFB. **a** Pole zero plot, **b** Output spectrum

this, the difference of the loop filter output and first-stage output is fed as input to second stage. In order to keep the error in maximum stable amplitude, an interstage coupling coefficient is used at the input of second stage. Since perfect noise cancelation between two stages is not possible, not very high-order converters can be designed [11]. These noise leakages can change STF and NTF but do not alter the SNR to large extent as long as the loop filter is larger than the signal band.

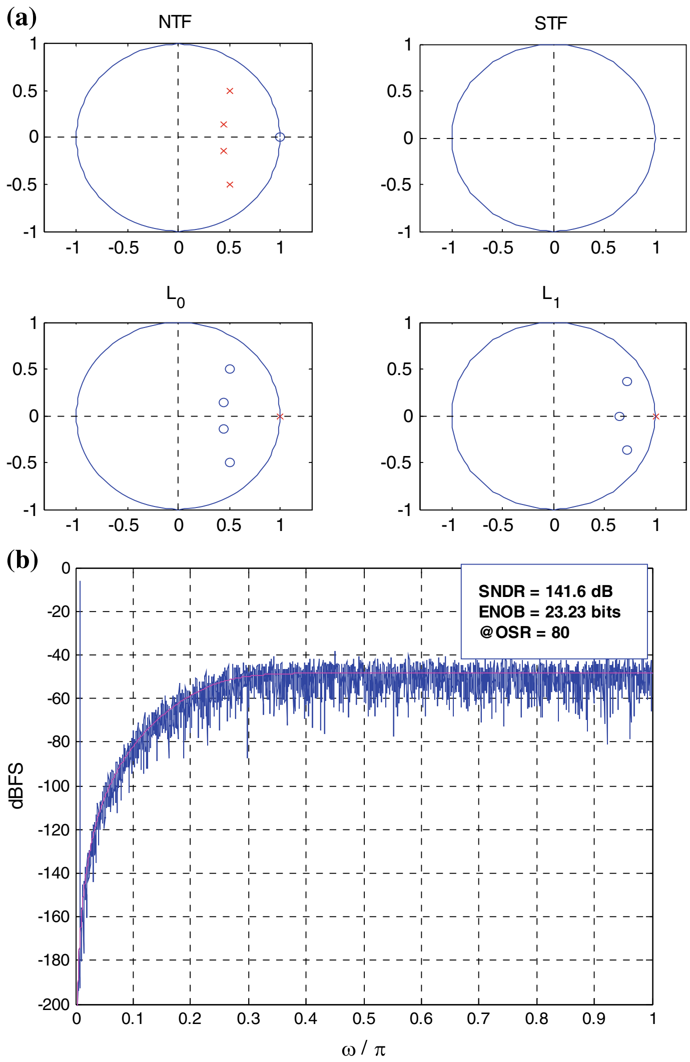


Fig. 10 CIFF. **a** Pole zero plot, **b** Output spectrum

5 Conclusion

In this paper, an architecture of first-, second-, and third-order sigma delta modulator is presented and implemented in the popular MATLAB/SIMULINK environment. The SNR obtained is 62 dB and 61.4 B, 91.9, respectively. Some of the popular higher-order topologies, feedback, and feedforward arrangements: CRFB, CRFF, CIFB, and CIFF were simulated and achieved SNR of 140 dB, 153.6 dB, 141.8 dB, and 164.4 dB, respectively.

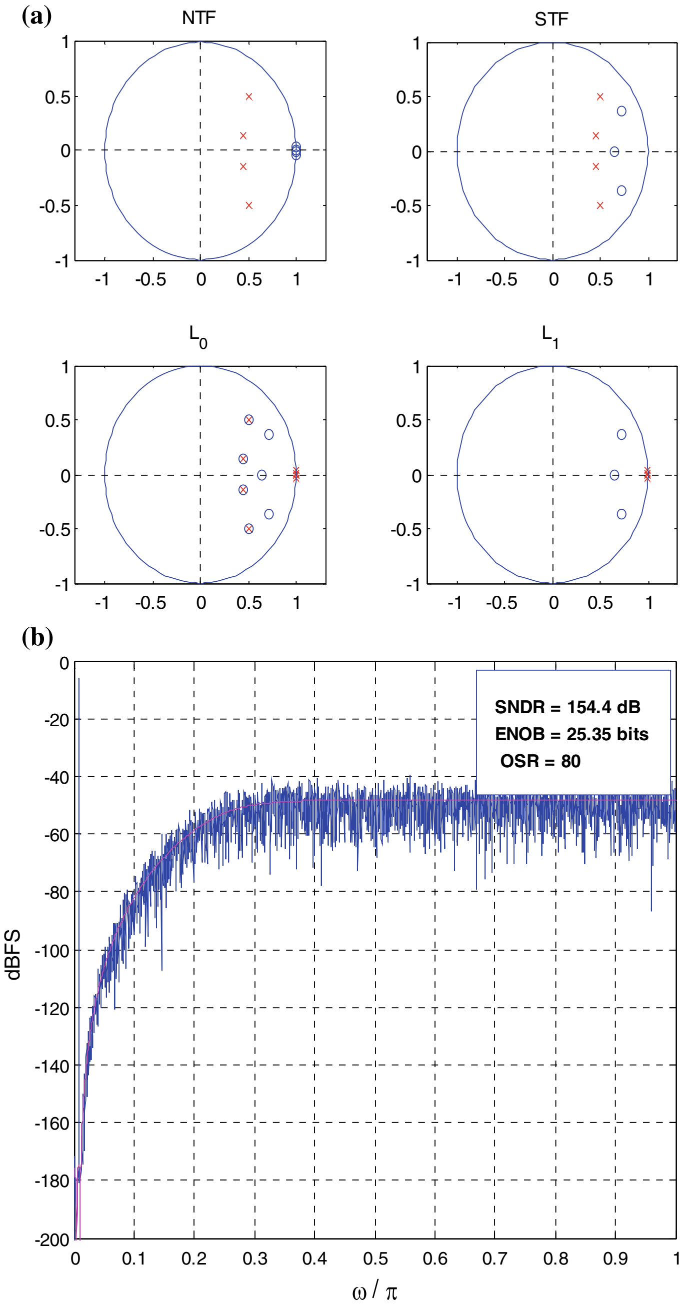


Fig. 11 CRFF. **a** Pole zero plot, **b** Output spectrum

References

1. Schreier, R., Temes, G.C.: Understanding Delta Sigma Converters. IEEE press
2. Cherry, J., Snelgrove, W.: Continuos Time Delta Sigma Modulators for High Speed A/D Conversion. Kluwer Academic Publishers
3. Park, M., Perrott, M.: A 78 dB SNDR 87 mW, 20 MHz bandwidth continuous–time $\sum\Delta$ ADC with VCO-based integrator and quantizer implemented in 0.13 μm at CMOS. IEEE J. Solid State Circuits **44**(12), 33 (2009)
4. Dendouga, A., Bouguachal, N., Kouda, S., Barra, S., Lakehal, B.: Contribution to modeling of a non ideal $\sum\Delta$ modulator. J. Comput. Electron. (2012)
5. Silva, J., Moon, U., Steensgaard, J., Temes, G.C.: Wideband low-distortion delta sigma ADC topology. Electron. Lett. **37**(12), 737–738 (2001)
6. Temes, G.: Finite amplifier gain and bandwidth effects in switched capacitor filters. IEEE J. Solid States **15**, 358–361 (1980)
7. Harrison, R.R., Charles, C.: A low power low noise CMOS amplifier for neural recording applications. IEEE J. Solid State Circuit **38**(6) (2003)
8. Matsukawa, K., Nitani, Y., Takayama, M., Obata, K., Dosho, S., Matsukawa, A.: A fifth order continuos time $\sum\Delta$ modulator with single op-amp resonator. IEEE J. Solid State Circuit **45**(4) (2010)
9. Mitteregger, G., Ebner, C., Mechnig, S., Blon, T., Holuigue, C., Romani, E.: A 20-mW, 640-MHz, CMOS continuous-time ADC with 20 MHz signal bandwidth, 80-dB dynamic range and 12-bit ENOB. IEEE J. Solid State Circuits **41**(12), 2641–2649 (2006)
10. Steensgaard, J.: High performances data converters. Ph.D. thesis, Technical University of Denmar, Department of Information Technology (1999)
11. Donida, A., Cellius, R., Nagari, A., Baschirotto, A.: A 40 nm CMOS 1.1 V, 101 dB dynamic range, 1.7 mW continuous time. IEEE Regul. Pap. **62**(3) (2015)

A New Variant of MTF-TRANS List Accessing Hybrid Algorithm

Richismita Rout, Shiba Prasad Dash and Pratyashi Satapathy

Abstract The whole problem of effectively reorganizing the list and also accessing its elements in order to obtain the optimal cost is list accessing problem. Here, the input is an unsorted linear list of distinct elements along with a sequence of requests; each element present in the request sequence is operated on an element in the list. The algorithm which is used to reorganize the list in order to obtain a minimum cost is list accessing algorithm. MTF and TRANS are two basic list accessing algorithms. In this thesis, we have analyzed both MTF and TRANS algorithms, and on the basis of their strengths, we have developed a new hybrid algorithm. The newly developed hybrid algorithm utilizes both the MTF and TRANS algorithms. We have performed experimental analysis on the newly developed hybrid algorithm, and it is found that it performs better in comparison with MTF and TRANS algorithms.

Keywords List accessing problem · Request sequences · MTF · TRANS · Hybrid algorithm

1 Introduction

Linear search is a fundamental searching method for an unsorted list, whose efficiency can be improved, if the list self-organizes itself. List update problem (LUP) is a basic common model for self-organizing linear search. In a list accessing problem, the input is a list and a request sequence, where the list contains a set of distinct elements, and in a request sequence, there may or may not be repetition of elements. The requested element is searched sequentially from the starting position of the list till it is found, and then, it is accessed by incurring some cost. So the problem of accessing

R. Rout (✉) · S. P. Dash · P. Satapathy
Biju Patnaik University of Technology, Rourkela, Odisha, India
e-mail: rout.richismita@gmail.com

S. P. Dash
e-mail: titun.tiki@gmail.com

P. Satapathy
e-mail: pratyashi1234@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_6

a requested element in a list and thereby reorganizing the list efficiently in order to obtain an optimum cost is a list accessing problem.

1.1 List Accessing Cost Model

In a list accessing problem, when the requested element from the sequence of request is accessed in a linear list, some cost is assigned to that element. The method of assigning cost is known as cost model. Cost model can be classified into two types, viz. full cost model and partial cost model. In case of full cost model, the cost for accessing a requested element is determined by the position of that element in the list from the beginning of the list. Full cost model is further divided into two types: free transposition and paid exchange. In free transposition, no charge is incurred for the exchange of elements, whereas in paid exchange, cost of 1 is incurred for the same. The access cost in partial cost model is calculated by the number of comparisons made between the accessed element and the elements present before the accessed element in the list.

1.2 List Accessing Algorithm

An algorithm which is used to reorganize the list in order to minimize the access cost is known as list accessing algorithm. LAA can be either offline algorithm or online algorithm. In the former case, the whole request sequence is completely known before the algorithm starts, whereas in the latter one it is partially known. Both online algorithm and offline algorithm can be either deterministic or randomized. In deterministic after accessing the element, the rule for movement of an element is fixed, whereas in randomized, there is no fixed rule for the same.

1.3 Related Work

In 1965, McCabe initialized the methods to solve list accessing problem by developing MTF algorithm and TRANS algorithm. Efforts have been done by many researchers to minimize the total cost. Some researchers have developed hybrid algorithms by combining the existing algorithms [1, 2]. A survey of online algorithm which gives an introduction to the competitive analysis of online algorithms is proposed by Albers [3] in 1991. Again, Albers [4] has proposed competitive analysis of the LUP with look ahead. Mohanty and Tripathy [5] have developed an improved move-to-front (IMTF) offline algorithm, and their experimental study showed that IMTF's performance was better than that of MTF for two new types of dataset. Mohanty et al. [6] developed some special types of request sequences where TRANS performed better than that of MTF algorithm. In the year 2011, Mohanty et al. [7] characterized the request sequences for list accessing. By using TRANS algorithm, Rout [1] has introduced a hybridized MTF-TRANS-FC (H-M-T-F-C) algorithm. Samanta and Dash [2] have developed a hybridized algorithm by combining MTF and TRANS algorithms using BIT algorithm.

1.4 Our Contribution

In this paper, MTF and TRANS algorithms are studied and its strengths and limitations are explored. On the basis of its strength, a new hybrid algorithm is developed. Our experimental analysis shows that the new hybrid algorithm performs better than that of MTF and TRANS algorithm.

2 Analysis of Existing Algorithms

Till now several list accessing algorithms have been developed among which MTF and TRANS algorithms are the fundamental list accessing algorithms.

MTF. In this algorithm, once the requested element is accessed from the list, it moves to the first position of the list without altering the respective order of other elements present in the list.

TRANS. In this algorithm, if the accessing element is not present in front of the list, the requested element will swap with the previous element of the list.

2.1 Analysis of MTF Algorithm

Assume the configuration of the request sequence and the list to be ⟨1, 1, and 1⟩ and ⟨1, 2, and 3⟩ respectively. So as per MTF algorithm, the requested element will move to the first position of the list after that element is being accessed. After accessing all the elements, the total access cost will be calculated using free exchange of full cost model. So the total cost of MTF algorithm for the above request sequences after accessing all the element from the list is: $1 + 1 + 1 = 3$. From our study, we have analyzed the following interesting result about MTF algorithm (taking list as ⟨1, 2 and 3⟩).

1. It gives a better result, whenever there is a repetition of elements in the request sequence.

   Illustration: Considering a request sequence as ⟨3, 3 and 3⟩, for the above-mentioned list, the total access cost obtained is 5.

2. It gives a worst result, when the order of the request sequence and the lists are in reverse to each other [5].

   Illustration: Here, for the above-mentioned list and the request sequences ⟨3, 2, and 1⟩, the total access cost obtained is 9.
3. Also MTF algorithm performs poorly for distinct element set in request sequences [5].

2.2 Analysis of TRANS Algorithm

Let the configuration of the request sequence and the list be ⟨3, 1, and 2⟩ and ⟨1, 2, and 3⟩, respectively. So, as per TRANS algorithm, the requested element will swap itself with the previous element in the list, after being accessed. If the accessing element is present in first position of the list, then there is no need of any further movement of element. After accessing all the element of the list, the total access cost will be calculated using free exchange of full cost model. So the total cost of TRANS algorithm for the above request sequences after accessing all the element from the list is: $3 + 1 + 3 = 7$. From our study, we have analyzed the following interesting result about TRANS algorithm.

1. TRANS algorithm gives best result, when the elements are in reverse order of request sequences.

   Illustration: Let the configuration of list be ⟨1, 2, and 3⟩ and the request sequences be ⟨3, 2, and 1⟩. The total access cost obtained by TRANS algorithm is 7.
2. Also TRANS algorithm performs good for distinct element set in request sequences.
3. TRANS algorithm gives worst result, when the elements of request sequences are repeated.

   Illustration: Let the configuration of list be ⟨1, 2, and 3⟩ and the request sequences be ⟨3, 3, and 3⟩. The total access cost obtained by TRANS algorithm is 6.

After analyzing the MTF and TRANS algorithms, we observed that MTF algorithm gives best results for the request sequences where the elements are repeated whereas TRANS performs good in case of distinct element set in a request sequence. So based on this advisability of the two algorithms, we have developed a new HYBRID algorithm which combines both MTF algorithm and TRANS algorithm.

3 Our Approach

We have introduced a new algorithm based on the look-ahead concept and online algorithm. We have developed a new variant of MTF-TRANS list accessing HYBRID algorithm, which gives better result as compared to MTF and TRANS algorithms for specific request sequence.

3.1 Hybrid Algorithm

In this algorithm before accessing any element of the list, it will check the next element of the request sequences. If the access element is same as the next element present in the request sequences, then it will perform MTF, otherwise TRANS. After performing MTF and/or TRANS operation, the list will reorganize accordingly.

3.2 Illustration of Hybrid Algorithm

Considering the configuration of the list and the request sequence to be ⟨1, 2, 3, 4, and 5⟩ and ⟨5, 4, 4, 3, and 2⟩ respectively, the first element of the request sequence is 5 which is in the 5th position of the list. Before accessing the element (5), it will check the next element of the request sequences which is 4. Since the access element (5) and the next element (4) are not same, according to the HYBRID algorithm it will perform TRANS operation. So the access cost for accessing the element (5) in the list is 5. After accessing the element, the list will reorganize. The new list is ⟨1, 2, 3, 5, and 4⟩. Next element of in the request sequence is 4 which is in 5th position of the reorganized list. Before accessing the element (4), it will check the next element of the request sequences which is 4. Since the access element (4) and the next element (4) are same, according to the HYBRID algorithm it will perform MTF operation. So the access cost for accessing the element (4) in the list is 5. After accessing the element, the list will reorganize. The new list is ⟨4, 1, 2, 3, and 5⟩. The third element of the request sequence is 4 which is in the first position of the list. As it is present in the first position of the list, no movement takes place and the access cost for accessing the element is 1, and the list remains the same. The fourth element of the request sequence is 3, and the element next to the fourth element is 2. As the fourth and fifth elements are not same, TRANS will be performed. Since the element is present in the fourth position, the access cost is 4. The new list is ⟨4, 1, 3, 2, 5⟩. The last element of the request sequence is 2, as it is the last element so TRANS will be performed. Since 2 is present in the fourth position of the list, the access cost is 4. The new final list is ⟨4, 1, 2, 3, 5⟩. Here, the total access cost is $5 + 5 + 1 + 4 + 4 = 19$, whereas in MTF and TRANS the total access cost is 21 and 20, respectively (as shown in Tables 1 and 2). By applying hybrid algorithm, the unnecessary movement of element can be reduced and the cost can be minimized by reorganizing the element of the list (Table 3).

Similarly, the total access cost obtained for different request sequences, for the HYBRID algorithm, MTF algorithm, and TRANS algorithm is shown in Table 4.

Table 1 Illustration of MTF algorithm

Request sequence	List					Cost
5	1	2	3	4	5	5
4	5	1	2	3	4	5
4	4	5	1	2	3	1
3	4	5	1	2	3	5
2	3	4	5	1	2	5
	2	3	4	5	1	
Total access cost						21

Table 2 Illustration of TRANS algorithm

Request sequence	List					Cost
5	1	2	3	4	5	5
4	1	2	3	5	4	5
4	1	2	3	4	5	4
3	1	2	4	3	5	4
2	1	2	3	4	5	2
	2	1	3	4	5	
Total access cost						20

Table 3 Illustration of hybrid algorithm

Request sequence	List					Cost
5	1	2	3	4	5	5
4	1	2	3	5	4	5
4	4	1	2	3	5	1
3	4	1	2	3	5	4
2	4	1	3	2	5	4
	4	1	2	3	5	
Total access cost						19

Table 4 Total access cost by MTF, TRANS, and hybrid algorithms

List size	Request sequences	MTF	TRANS	HYBRID
12,345	52,311	17	15	14
12,345	51,233	15	13	12
12,345	53,211	18	13	13
12,345	55,425	17	19	15
12,345	55,423	20	21	19
12,345	54,432	21	20	19
12,345	54,331	17	20	16
12,345	54,113	15	19	15

4 Experimental Analysis

An experiment was performed by implementing MTF, TRANS, and HYBRID algorithms for various request sequence and lists. The total access cost is obtained by either of the algorithms.

4.1 Experimental Result

The experiment was carried out by implementing the two basic algorithms (MTF, TRANS) and the newly developed HYBRID algorithm in C++ language. Here, two singly linked lists are used, one for the list and another for the request sequence. A number of user-defined functions are used, viz. List (), MTF (), TRANS (), HYBRID (). The function list generates a list of distinct elements as given input by user. MTF () function is used to perform MTF algorithm and to calculate access cost based on the same, whereas TRANS () is used to perform the TRANS algorithm and calculate its cost. HYBRID () is used to perform our newly developed HYBRID algorithm. In this, before accessing the element, that element is checked with the next element; if they are same, MTF () is performed; else, TRANS () is performed. For the last element of the request sequence, TRANS () is performed. Let C_{MTF} be the total access cost attained by using move-to-front algorithm, C_{TRANS} be the total access cost obtained by TRANS algorithm, and C_{HYBRID} be the total access cost obtained by HYBRID algorithm. Let the gain of HYBRID with respect to MTF algorithm be denoted as G_{TRANS} and that with respect to TRANS algorithm be denoted as G_{MTF}. G_{MTF} and G_{TRANS} are defined mathematically as follows (Table 5):

$$G_{\mathrm{TRANS}} = \left(\frac{C_{\mathrm{HYBRID}} - C_{\mathrm{TRANS}}}{C_{\mathrm{TRANS}}} \right) * 100 \tag{1}$$

Table 5 Experimental results of MTF, TRANS, and HYBRID algorithm

List size	Request sequences	C_{MTF}	C_{TRANS}	C_{HYBRID}	% gain (G_{MTF})	% gain (G_{TRANS})
3	5	12	10	10	16.66	0
4	6	20	24	19	5	20.83
5	6	27	27	23	14.81	14.81
6	8	35	33	32	8.57	3.03
5	10	35	39	32	8.57	17.94
5	15	48	54	42	12.5	22.22
10	15	82	86	79	3.6	8.13
10	20	111	113	99	10.81	12.38
10	30	171	173	151	11.7	12.71
10	40	229	246	198	13.53	19.51
10	50	277	293	249	10.10	15.01
20	30	307	318	251	18.24	21.06
30	50	743	921	668	10.09	27.47
40	80	1322	1929	1232	6.80	36.13
50	100	2068	2514	1775	14.16	29.39

$$G_{\text{MTF}} = \left(\frac{C_{\text{HYBRID}} - C_{\text{MTF}}}{C_{\text{MTF}}} \right) * 100 \tag{2}$$

Figures 1 and 2 show the performance of MTF, TRANS, and new hybrid algorithm, in the form of a line chart and column chart, respectively, where x-axis represents total access cost and y-axis represents size of request sequences and list size. From

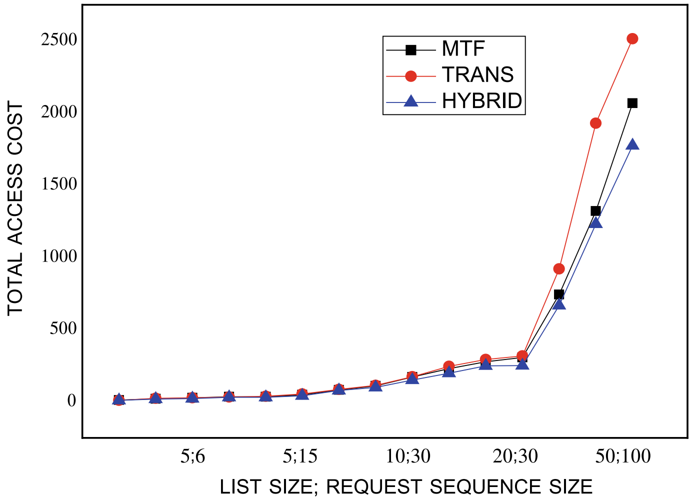


Fig. 1 Line chart of experimental result

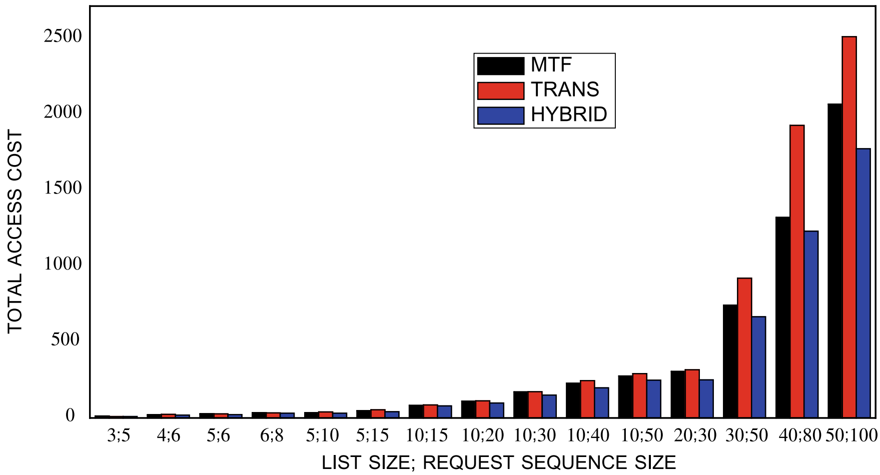


Fig. 2 Column chart for experimental result

the graph, it is observed that for a particular request sequence and list the total access cost obtained by the newly developed hybrid model is less than that of MTF and TRANS algorithms.

5 Conclusion

In this paper, a newly developed HYBRID algorithm is presented, and it is observed that its performance is better than that of MTF and TRANS algorithms. The gain percentage is also calculated, which shows that HYBRID algorithm is more efficient as compared to MTF and TRANS algorithms.

References

1. Rout, D.: Experimental analysis of hybrid MTF-TRANS-FC (HMTFC) algorithm. Int. J. Res. Advent Technol. E. **1**(5) (2013). ISSN: 2321-9637
2. Samanta, B., Dash, S.P.: A novel hybrid approach to list accessing problem using BIT algorithm. In: International Conference on Man and Machine Interfacing (MAMI) (2015)
3. Albers, S.: Online algorithms: a survey. Math. Program. **97** (2003). Print: ISSN 0025-5610/Online ISSN 1436-4646
4. Albers, S.: A competitive analysis of list update problem with lookahead. Theor. Comput. Sci. **197**(1–2), 95–109 (1998)
5. Mohanty, R., Tripathy, S.: An improved move to front offline algorithm for list accessing problem. IJACC **3**(1) (2011)
6. Mohanty, R., et al.: TRANS outperforms MTF for two special types of request sequence without locality of reference. Procedia Technol. **6**, 556–563 (2012). https://doi.org/10.1016/j.protcy.2012.10.067
7. Mohanty, R., Sharma, B., Tripathy, S.: Characterization of request sequences for list accessing algorithm and new theoretical results for MTF algorithm. IJCA **22**(8) (2011). ISSN: 0975-8887
8. McCabe, J.: On serial files with relocatable records. Oper. Res. **12**, 609–618 (1965)

Design and Implementation of a Factorial Circuit for Binary Numbers: An AVM-Based VLSI Computing Approach

Siba Kumar Panda, Ankita Sahoo and Dhruba Charan Panda

Abstract The arithmetic operation is the most crucial part in VLSI technology. Multiplication plays a significant role in arithmetic operations of various VLSI systems. The digital multipliers are used in realizing many DSP computations like FFT, MAC, and DFT. By calculating the silicon area, power, speed, and delay, the performance of the circuit is evaluated. In this paper, we used the ancient Vedic mathematics for implementation of the factorial circuit in the physical design level. The ancient Vedic mathematics (AVM)-based calculations are faster than the conventional-based calculations. An innovative idea to design an AVM-based factorial circuit is suggested in this work. In this paper, the proposed multiplier is designed by using AVM-based algorithm called Urdhva-Tiryagbhyam (UT), which is used to calculate the factorial of binary numbers. The proposed factorial circuit is designed by using the UT-based Vedic multiplier, incrementer, decrementer, zero detector, and a comparator with the help of Verilog HDL. It is synthesized and simulated using the Cadnce EDA (NC Sim,RTL Compiler, Encounter) tool in digital level. The result obtained in this work is found to be correct as well as meticulous.

Keywords Factorial circuit · UT sutra · Ancient Vedic mathematics (AVM) · Verilog HDL · VLSI

1 Introduction

Design of digital multipliers has always been an active area of research due to its importance in digital signal processing applications. So, many interesting multiplication algorithms have been formulated. Multipliers are the key constituents for various

S. K. Panda (✉) · A. Sahoo
Department of ECE, Centurion University of Technology & Management, Jatni, Odisha 752050, India
e-mail: panda.sibakumar08vssut@gmail.com

D. C. Panda
PG Department of Electronic Science, Berhampur University, Berhampur Odisha, 760007, India
e-mail: dcpanda@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_7

signal processing applications. For the implementation of a factorial circuit, we need a high-speed multiplier. In this context, many arithmetic circuits are designed and implemented using ancient Vedic mathematics. Many researchers had already excellently proposed their works. In 2018, Panda and Panda [1] published a review report in explaining the development of VLSI systems using Vedic mathematics. This innovative work helped a lot in the design and implementation of various higher order arithmetic circuits. Singh and Gupta [2] in 2017 explained the design of multipliers and divisor by using different formulas of Vedic mathematics which are useful in various applications like DSP, image processing, cryptography, convolution, etc. In 2017, Jie and Ruslan [3] proposed a 2×2 multiplier design by using UT algorithm. Similarly in 2016 Gadakh and Khade [4], Ram and Rani [5], and Mistri and Somani [6] explained about the design of 16-bit, 32-bit multipliers by using different algorithms like Urdhva, Nikhilam, and Anurupyena to study the various parameters like execution time, area, power consumption, and speed. In 2015, Dafe and Bhagat [7] explained the blueprint of a speedy 4×4 multiplier using Vedic techniques. The design of 4×4 Vedic multiplier is done by the help of 4 number of 2×2 multipliers which are the basic blocks for any digit of multiplier. Panda and Das [8] in 2015 and Patil and Manjunatha [9] in 2014 explained the design of efficient multipliers using UT sutra and concluded that by the use of UT sutra, carry propagation delay can be reduced from LSB to MSB and the use of Nikhilam sutra reduces the complexity in multiplication process. In 2011, Saha and Banerjee presented a design of factorial circuit [10]. In 2001, Maharaja had explained the apprehension of ancient Vedic mathematics [11], which is a unique technique consisting of 16 sutras and 13 subsutras. Vedic-implemented systems or circuits in VLSI are more efficient than conventional technique-based systems. These sutras are used to reduce the partial products, area, power consumption, and delay and increase the speed. The entire system of ancient mathematics is intelligently interrelated and also integrated well with the VLSI theory. The most important feature of AVM is coherence, which helps to reduce the complexity of the circuit design.

Section 2 presents Urdhva-Tiryagbhyam Sutra and its application to calculate factorial of binary numbers. Section 3 presents the factorial circuit design. The execution environment is presented in Sect. 4. Section 5 describes the result analysis. The conclusion is drawn in Sect. 6.

2 The Urdhva-Tiryagbhyam Sutra: Factorial of Binary Numbers

For determining the factorial of a positive binary number, many algorithms are designed and developed such as Braun array, Baugh–Wooley method, Booth's algorithm, and Wallace tree. However, these algorithms have many drawbacks like a large area, less speed, and more power consumption. The implementation of the ancient Vedic mathematics (AVM)-based circuit design can overcome these drawbacks. The objective of this work is to investigate the digital VLSI design style from RTL design to the generation of GDS-II file of the concerned circuit. Generally, the factorial of a number is determined using a multiplier unit. It is one of the most time-consuming operations and requires greater hardware environment. A keen factorial circuit may radically perk up the computation process and reduce the delay for a large coverage. Figure 1 shows the flowchart representation for factorial calculation.

The UT algorithm:

In order to find factorial of a number, we used 'Urdhva-Tiryagbhyam' sutra from ancient Vedic mathematics. The meaning of this sutra is 'vertically and crosswise' multiplication.

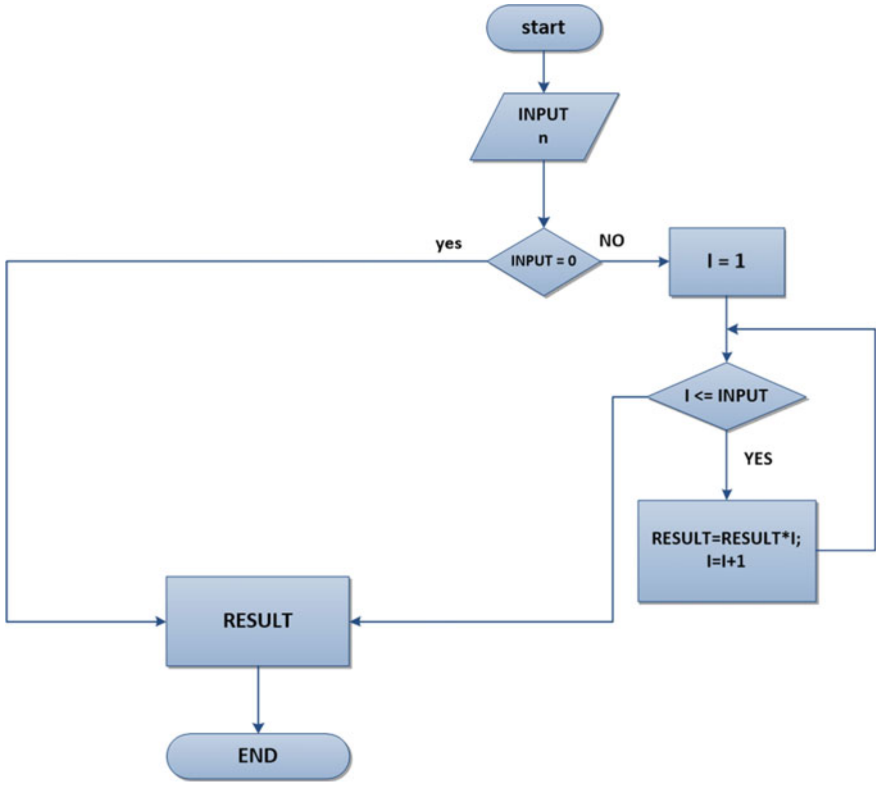


Fig. 1 Schema chart of factorial calculation

Multiplicand – C3 C2 C1 C0

Multiplier – D3 D2 D1 D0

S0 = C0*D0 C3 C2 C1 C0

D3 D2 D1 D0

S1= C1*C0 + C0*D1 C3 C2 C1 C0

D3 D2 D1 D0

S2= C2*D0+C0*D2+C1*D1 C3 C2 C1 C0

D3 D2 D1 D0

S3=C3*D0+C0*D3+C2*D1+C1*D2 C3 C2 C1 C0

D3 D2 D1 D0

S4=C3*D1+C1*D3+C2*D2 C3 C2 C1 C0

D3 D2 D1 D0

S5=C3*D2+C2*D3 C3 C2 C1 C0

D3 D2 D1 D0

S6=C3*D3 C3 C2 C1 C0

D3 D2 D1 D0

Stepwise factorial calculation using UT sutra: An AVM-based VLSI Computing approach

Calculating the factorial of a binary number 0101(5)

i. Calculate the factorial of the binary number by iterative multiplication from 1(0001) to 5(0101).

ii. 0001 (1) × 0010 (2) = 0000010(2)

```
  0  0  0  1
  0  0  1  0
-------------
0 / 0/0 /0 /0 /1 /0
     = 0000010
```

iii. 0010 (2) × 0011(3) = 0000110(6)

```
  0  0  1  0
  0  0  1  1
-------------
0/ 0 /0 /0 / 1/1/0
    = 0000110
```

iv. 0110 (6) × 0100(4) = 11000(24)

```
  0  1  1  0
  0  1  0  0
-------------
0/ 1/ 1/ 0/ 0/ 0
  = 011000
```

v. 11000 (24) ×0101 (5) =1111000(120)

```
  1  1  0  0  0
  0  0  1  0  1
----------------
0 / 0 / 1 / 1 / 1 / 1 / 0 / 0 / 0
          = 001111000
```

3 The Factorial Circuit

The factorial circuit is used to calculate the factorial of a positive binary number '*n*'. It is obvious that product of all the positive integers leads to calculation of factorial of a number. It is premeditated by iterative multiplications, accurately, by either decrementing the number up to 1 or incrementing from 1 up to the given number. Figure 2 explains the diagrammatic outline of factorial circuit with its submodules. For iterative multiplication, we need a high-speed multiplier. The speed of the multiplication operation is a vital factor in design aspects. With the use of Vedic sutra, we can design a high-speed multiplier as well as high-performance-based factorial circuit. The architecture is based on 'Urdhva-Tiryagbhyam' sutra which is used for the multiplication process facilitating factorial circuitry design as well as implementation. This sutra is more efficient and easy to execute for calculating the factorial of a number in case of decimal as well as binary [11] as compared to other sutras used for multiplication in AVM.

Factorial circuit modules and design
These circuit modules are designed with Verilog HDL which is based on 'C'. The circuit module to calculate factorial of any binary number is described in terms of various submodules.

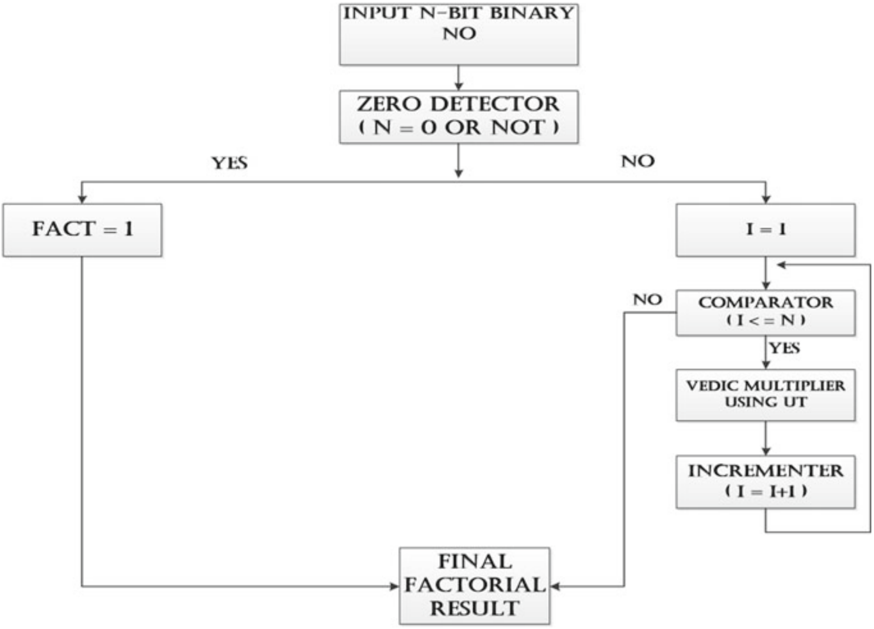


Fig. 2 Complete block diagram for implementation of the factorial circuit

Zero detector Submodule

The zero detector submodule plays an important role to check the input number is zero or not at the entry level. If it is zero, then it settles the factorial to 1 directly. Otherwise, it will be passed through the incrementer submodule.

Comparator Submodule

The basic function of this submodule is to compare the incrementer value (I) with the input value (N) by applying three conditions.

- If $I \leq N$, then the incremented value (I) goes to the Vedic multiplier for multiplication process and calculates the factorial of the number. The loop is continued till the incremented value is greater than the input value.
- If $I > N$, then the multiplication process is stopped and the previous result of Vedic multiplier is stored as the final result of factorial calculation.

If we use a decrementer circuit to calculate the factorial of a number, comparator checks three conditions for input value (N) and decremented value (D).

- If $D \leq N$, then the decremented value goes to the multiplier and the process is continued same as incrementer.
- If $D = 0$, then the multiplication process is stopped and the previous result of Vedic multiplier is stored as the final result of factorial calculation.

Incrementer/Decrementer Submodule
This circuit is required to increment the value by 1 or decrement the value by 1 during the calculation process, i.e., $Y = X + 1$ or $Y = X - 1$.

Scheming Vedic Multiplier
An AVM-based algorithm, Urdhva-Tiryagbhyam is used for scheming of the multiplier. This sutra shows the multiplication of large-sized binary number ($N \times N$ bits) by breaking it into small extent numbers. The area required for this multiplier is smaller as compared to the other multiplier architectures. For multiplier design, the basic blocks, i.e., 2 × 2 multipliers are made and the implementation is done for 4 × 4, 8 × 8, 16 × 16, and 32 × 32 bits. The 2 × 2 bit Vedic multiplier design requires four AND gates, two full adders, and two half adders. The equations of 2 × 2 Vedic multiplier modules are:

- S0 = b0a0
- S1 = b0a1 + b1a0
- S2 = b1a1 + c1
- Product = S2 & S1 & S1.

4 Execution Environment

In this part, the factorial circuit for binary numbers is coded in Verilog hardware description language (HDL) and then synthesized and simulated in Cadence EDA tool. After that, the physical design core of the factorial circuit is also obtained with the use of gpdk90 technology.

5 Result Analysis

The design of ancient Vedic mathematics (AVM)-based factorial circuit [1] is entirely different from the conventional method. AVM-based factorial circuit is designed in RT level and also synthesized successfully. The individual block is implemented using Verilog HDL. Figure 3 shows the RTL schematic of the factorial circuit.

Figure 4 shows the physical design layout of the factorial circuit with grids using gpdk-90.

During simulation, the output waveform is tested for calculation of factorial of 0, where the input is of 4-bit, i.e., 0000(0) and output is of 16-bit, i.e., factorial of 0 is 0000000000000001. Figure 5 shows the output waveform of factorial 5(0101), i.e., 0000000001111000(120). It shows the result is correct.

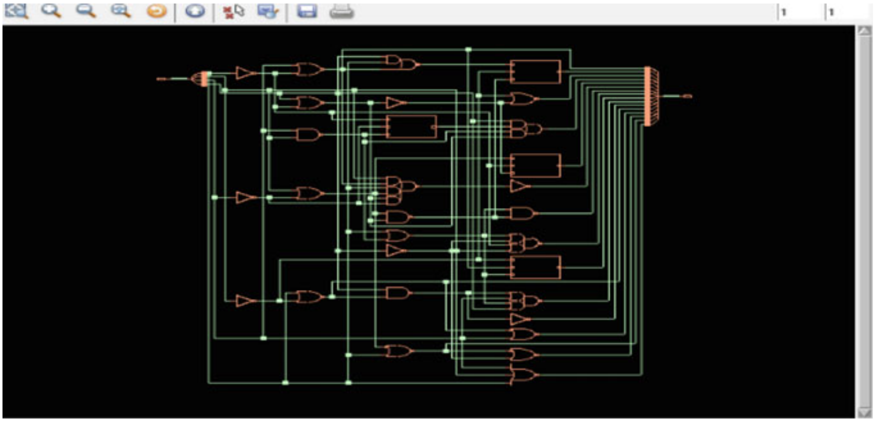


Fig. 3 RTL schematic of the factorial circuit

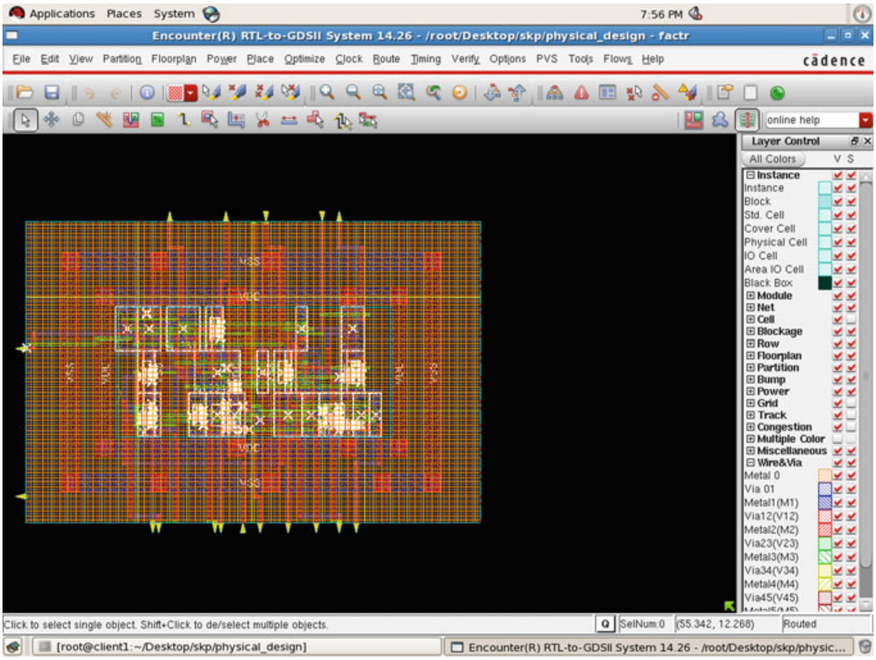


Fig. 4 PD layout of factorial circuit with grids using gpdk-90

6 Conclusion

The performance of the projected factorial architecture using Urdhva-Tiryagbhyam sutra of ancient Vedic mathematics (AVM) was proved to be efficient for quick

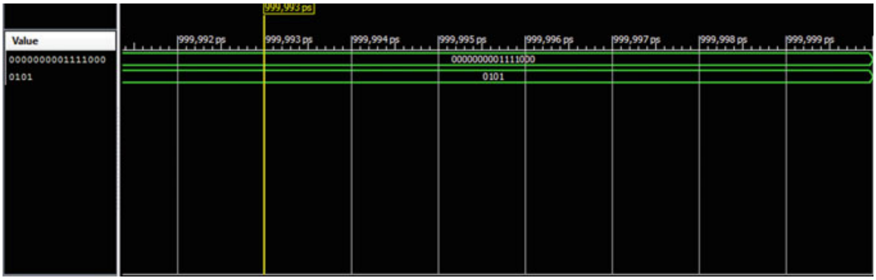


Fig. 5 Output waveform of factorial 5

calculations. The various performance reports have been studied by synthesizing using gpdk-90 technology in Cadence tool. The physical layout of the designed factorial circuit is obtained successfully with the generation of GDS-II file. The performances of the method in provisions of simulation results have been offered which manifest that the proposed method is superior in comparison with conventional methods. So, we may suggest this method of design to be used for many VLSI computing systems as well as competent VLSI signal processing perspectives.

Acknowledgements All the simulations, synthesis, and implementation are done in Laboratory of Centurion University of Technology & Management, Odisha, India. Authors gratefully acknowledge to the University for providing research scope.

References

1. Panda, S.K., Panda, D.C.: Developing high-performance AVM based VLSI computing system: a study. In: Computing Analytics and Networking. Advances in the Intelligent System in Computing, vol. 710, pp. 315–325. Springer, Singapore (2018)
2. Singh, A., Gupta, N..: VEDIC mathematics for VLSI design. Int. J. Eng. Sci. Res. Technol. (2017)
3. Jie, L.S., Ruslan, S.H.: A 2 × 2 bit Vedic multiplier with different adders in 90 nm CMOS technology. AIP Conf. Proc. **1883**, 020017-1–020017-7 (2017)
4. Gadakh, S.N., Khade, A.K.: Design and optimization of 16 × 16 bit multiplier using Vedic mathematics. In: IEEE International Conference on ACDOT, pp. 460–464 (2016)
5. Ram, G.C., Rani, D.S.: VLSI architecture for delay efficient 32-bit multiplier using Vedic mathematics sutras. In: IEEE International Conformance on TRETCT, pp. 1873–1877 (2016)
6. Mistri, N.R., Somani, S.B.: the study of Vedic multiplier algorithm using Nikhilam method. IJAREEJE **5**, 3618–3623 (2016)
7. Dafe, S.A., Bhagat, S.A.: Design of speed multiplier using Vedic mathematics technique. Int. J. Emerg. Res. Manag. Technol. **4**, 43–46 (2015)
8. Panda, S.K., Das, R.: VLSI implementation of vedic multiplier using Urdhva-Tiryagbhyam sutra in VHDL environment: a novelty. IOSR J. VLSI Signal Process. **5**, 17–24 (2015)
9. Patil, S., Manjunatha, D.V.: Design of speed and power efficient multipliers using Vedic mathematics with VLSI implementation. In: IEEE International Conference on Advances in Electronics, Computer & Communications (2014)

10. Saha, P., Banerjee, A.V.: ASIC design of a high speed low power circuit for factorial calculation using ancient vedic mathematics. J. Microelectron. (2011)
11. Maharaja, B.K.T.: Vedic Mathematics. Motilal Banarsidass Publishers Private Ltd., Delhi, India (2001)

Current Perspectives and Advancements of Perovskite Photovoltaic Cells

Chandni Devi and Rajesh Mehra

Abstract Perovskite photovoltaic cells have pulled in much consideration on account of their fast ascent to 23% PCE. Here, in this paper the quick development of PSCs has been reviewed, as they entered into the stage that could upgrade the industry of photovoltaics. Specifically, in this paper, the recent advancements in the architectures of perovskite photovoltaic cells, various blends used to develop perovskite photoactive layers have been portrayed. The remarkable advances of long-haul strength are talked about, and this paper gives an attitude towards what the eventual fate of PSCs may soon bring the photovoltaic group.

Keywords ETM · HTM · Perovskite layer · Photovoltaics · Solar cells

1 Introduction

The immoderate usage of fossil fuel which leads to emission of greenhouse gases annexes its contribution to the changes in the climatic. The finite amount of resources on earth based on fossil fuels and nuclear energy is in strong contrast to the regularly increasing requirements for electricity. In 2012, the global electricity production was about 22,600 TWh as compared to the production of electricity in year 2000 (15,500 TWh) [1]. Therefore, it is obvious that there is a need to search for other alternatives using renewable sources of energy. The sunlight is the primary energy input to life on the surface of the earth. It is versatile, and the solar technology is making significant advances over the years. However, human has begun utilizing huge amount of energy of fossil fuels, altering the anatomy of the earth's atmosphere and impelling global climate change. Photovoltaic cells suggest a way to capture sunlight and transform that energy into electric energy for human use. Figure 1

C. Devi (✉) · R. Mehra
Department of Electronics and Communication Engineering, National Institute for Technical Teachers Training and Research, Chandigarh, India
e-mail: chandni.ece@nitttrchd.ac.in

R. Mehra
e-mail: rajeshmehra@yahoo.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_8

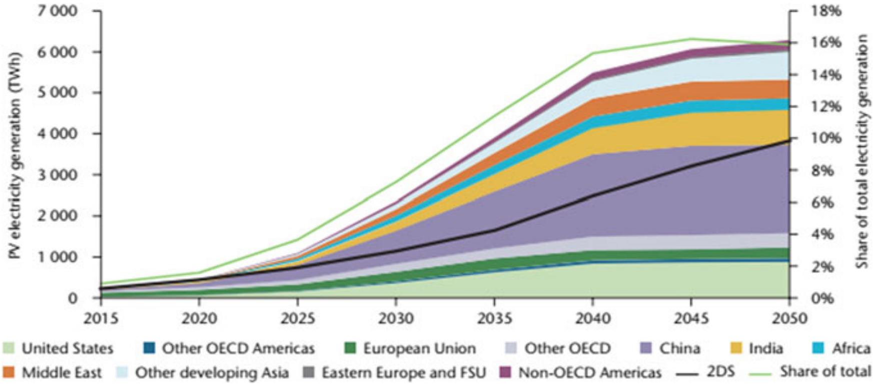


Fig. 1 Regional production of PV electricity envisioned in the road map [1]

shows the road map for PV's part of global electricity approaching 16% by 2050. A basal challenge in the development of photovoltaic technology is to lower the analogous energy production cost [2]. The production cost can be reduced if the efficiency of the device is optimized because the increased energy output per unit area lowers the overall system costs, from raw materials and manufacturing of module to installation and maintenance [3]. To perpetuate progress towards higher conversion efficiencies, photovoltaic cells are being fabricated with complex designs including heterojunctions, nanostructures, novel contact and different passivation schemes. These complex designs demand a comprehensive and amalgamate understanding of the electrical and optical device physics at the minuscule scale.

2 Perovskite Solar Cells

Among the foremost hot topics of materials science within the last few years has been perovskites, which have climbed to acclaim due to their distinctive properties. Particularly, they have improved the domain of photovoltaics, with fantastic achievements. Initially, the values of conversion efficiencies are up to the mark as required for purpose of commercialization. Joined with amazing quality and fast progress in upscaling, this overview will give an audit of the today's headways in this growing area. Solar cells composed of these materials have increased incredible growth in just two or three years with most outrageous conversion efficiency progressing from 3.8% [4] in year 2009 to an asserted 23% in 2018 [5]. This paper is a survey of the fast advancements in the architectures of perovskite photovoltaic cells; various mixtures are used to develop perovskite photoactive layers of the PSC device.

2.1 Architecture

The solar cells have been customarily planned with the utilization of a mesoporous framework and a stack design as found in Fig. 2a and b. These are made out of FTO, an ETL layer of various thickness values, the perovskite material, which may be penetrated in the mesoporous framework or as a "topping" layer on the electron transport layer and a HTL with a metal contact. The scanning electron microscopy images of some best PSC (efficiency over 20%) for SnO_2 and mesoporous TiO_2 planar designs are shown in Fig. 2b and c. Another planar arrangement has been produced where the stack depicted is completely inverted (TCO/HTM/Active layer/ETM/Metal) [16–19].

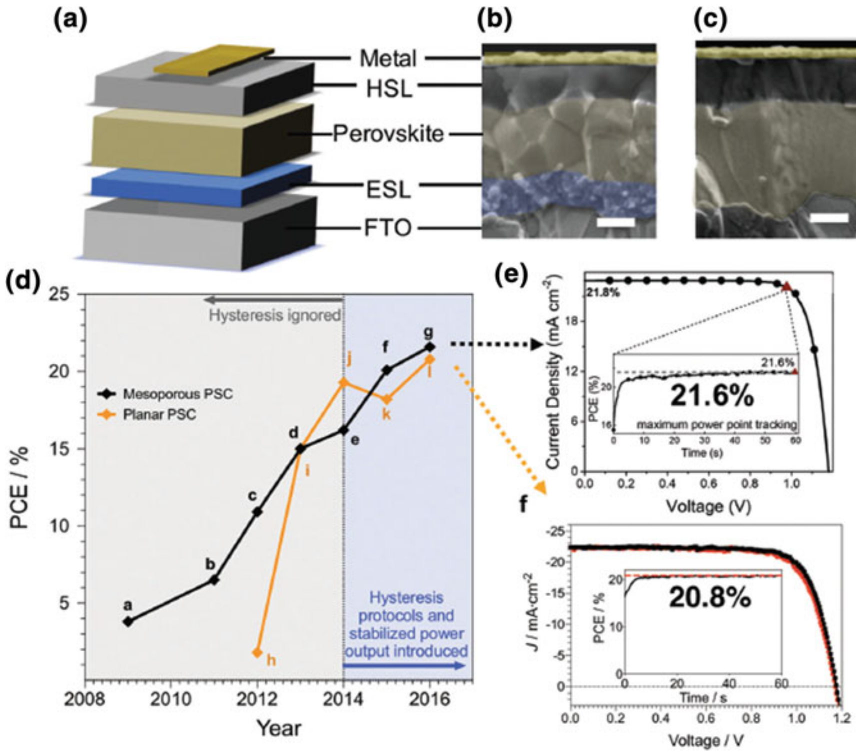


Fig. 2 High-efficient PSC. **a** Schematic stack of a PSC; **b** SEM image of a mesoporous TiO_2 device [6]; **c** planar (ALD) atomic layer deposited tin oxide junction [6]; **d** PCE over the years a = Ref. [4]; b = Ref. [7]; c = Ref. [8], d = Ref. [9]; e = Ref. [10]; f = Ref. [11]; g = Ref. [12]; h = Ref. [8] i = Ref. [13]; j = Ref. [14]; k = Ref. [6]; l = Ref. [15]; **e** V-I and MPP tracking curve [12]; **f** planar tin oxide heterojunction [15]

2.1.1 Infiltrated Mesoporous and Perovskite-Topped Devices

PSCs were spearheaded in a DSSC set-up in year 2009, where Kojima utilized a mesoporous TiO_2 and invasion of the perovskite materials in view of iodine or bromine in blend with a fluid electrolyte. Consequent work on mesoporous platform set-ups, generally made out of titanium oxide, produced record efficiencies approaching 22.1%. The most noteworthy reported efficiency each year is given in Fig. 2d, where the increment in efficiency value is from 3.8 to 23% in few years. The next research work on PSCs came into picture in 2011 and furthermore utilized a DSC arrangement with a fluid electrolyte producing 6.5% efficiency. An achievement in solar cells landed with the presentation of Spiro-OMeTAD as HTM by researchers Kim et al. [20] and Lee et al. [8] which, not at all like the fluid analogues, break up the layer of perovskite, and consequently opened the door in the field for investigation. From that point forward, a progression of record breaking efficiencies has been dynamically published. For the previous 2 years, a number of endeavours have been made to enhance measuring conventions, and these days maximum power point tracking and slow scan rates of the current–voltage hysteresis bends are characterizing the record efficiencies. The primary affirmed power conversion efficiency was detailed by Burschka et al. in the year 2013 [9]. Other consequent PSC efficiency records have been accomplished utilizing a mesoporous layer of TiO_2. In the 2014, confirmed 16.2% efficiency was accomplished by researcher Seok, and revealed in the research of Jeon et al. [10] by depositing methyl ammonium lead iodide TiO_2 layer, with the help of a method called "antisolvent". Another set of leaps forward by a similar group were accounted for in research works by Yang et al. [11], giving 18.5 and 20.1% guaranteed efficiencies, in the year 2015. The enhanced value of V_{OC} esteems by the FA/MA lead-based iodine/bromine perovskites gave the path to accomplishing this elite. A guaranteed 21.0% proficiency was accounted for in 2015 by NREL's chart of record high efficiencies. Introduction of inorganic cations [21, 22] turned into a conspicuous road to enhance the properties of the MA/FA perovskite material blends. Two extra PCE have been revealed by various cation details including the triple cation introducing Cs into perovskites of MA/FA having a stable efficiency of 21.1%, and recently a perovskite introducing Rb into the MA/Cs/FA perovskite blend with an efficiency value of 21.6% value which is stabilized (Fig. 2e). Along with this the V_{OC}, i.e., open-circuit voltage values up to 1.24 V were revealed [18]. A guaranteed efficiency of 22.1% by the KRICT has been accounted for by National Renewable Energy Laboratory. The pursuit for PCE has been a somewhat rushed which is clear from the chart in Fig. 2d, for the group; however, it has been an exceptionally compensating couple of years.

2.1.2 Configuration of the Perovskite Layer

A "perovskite" for the most part alludes to a material which possess the similar precious crystal structure of $CaTiO_3$ [23] (ABY_3 [13]), as appeared in Fig. 3. Regularly, the perovskites utilized as a part of PSCs photoactive layers are made up of

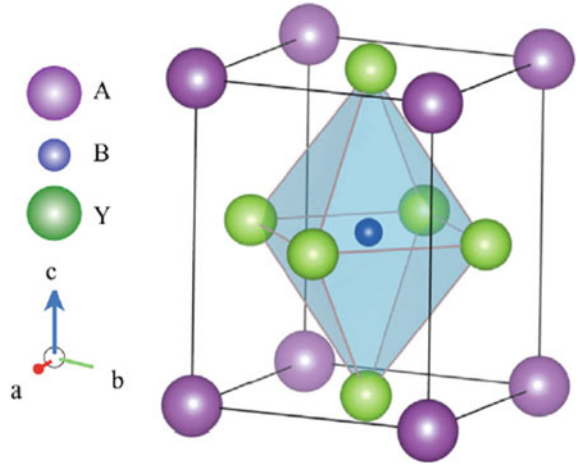


Fig. 3 Crystal structure of a perovskite [13]

a mixture of organic material, a halide material and a metal. Here, average organic cations incorporate $CH_3NH_3^+$ and $HC(NH_2)^{2+}$ having a scope for metal cations being used like Pb^{2+}, Sn^{2+}, Cu^{2+} and many more [24–26]. These organic and metal cations have been consolidated with the anions of halide like bromine, iodine and chlorine. Via consolidating these different components in various proportions, the scope of these various perovskite materials can be figured out, like $MAPbI_{(3-x)}Cl_x$ [27], $MAPbBr_{(3-x)}Cl_x$ [28] and $MAPbI_{(3-x)}Br_x$ [29]. This effective technique opened the path for the research in parameter space for material advancement and basically permits the tuning of the bandgap of the photoactive layer in the range of 1.3–2.3 eV [30]. In the accompanying sections, the three primary perovskite intensifies that have gotten most serious investigations ($MAPbI_3$, $MAPbI_{(3-x)}Cl_x$, $FAPbI_3$) will be examined in more detail.

$MAPbI_3$

The most broadly examined perovskite material framework is methylammonium lead iodide $MAPbI_3$, having the state-of-the-art power conversion efficiency as 18.1% [31]. The $MAPbI_3$ was at first consolidated in the fluid electrolyte of dye-sensitized solar cell. In that work, it was analysed that it was extremely unstable, steadily dissolved in the fluid electrolyte, and that device got deteriorated after a short span of time approximately 10 min when it was put in light. In the year 2011, $MAPbI_3$ quantum dot PV cells achieved the most astounding efficiency (6.5%) [7]. In the year 2012, a PSC device ($FTO/TiO_2/MAPbI_3$ + mp-TiO_2 + HTM/gold) was then exhibited which produced better results in terms of stability and efficiency and provided stimulus for the advancement of new sorts of perovskite devices [32]. In the year 2013, a sequential deposition process utilizing $MAPbI_3$ forerunners was shown for the very first time, producing the films with less morphological imperfections, furthermore improved the efficiency of the device to 15%. This work cleared the route for advancement of sequential deposition techniques utilizing some other types of perovskite.

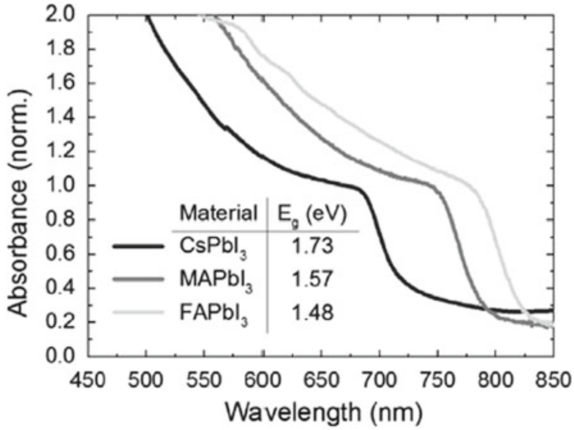
$MAPbI_{(3-x)}Cl_x$

The blended perovskite $MAPbI_{(3-x)}Cl_x$ could be made via utilizing both iodine and chlorine. From the past research work, it was concluded that $MAPbI_{(3-x)}Cl_x$ perovskite has a better stability and long charge diffusion length contrasted with $MAPbI_3$. Grätzel et al. [32] proposed that the state doped with chlorine could not survive in $MAPbI_{(3-x)}Cl_x$ in the light of the fact that Cl^- ions are a part of phase which is amorphous in nature and contains lead which would melt at an approx. temperature of 103 °C. Some other researchers found that Cl^- influences the nucleation act of perovskites yet most certainly not recognizable among the last film of perovskite layer. It was recommended that an intermediary $MAPbCl_3$ phase resulted through solution casting, yet later on vanished at thermal annealing. The presence of Cl^- in $MAPbI_{(3-x)}Cl_x$ was confirmed by a researcher [33] via the location of hydrogen chloride gas. Utilizing x-beam fluorescence estimations, Unger et al. [34] proposed that Cl^- exists inside the lattice of the crystal and at the grain boundaries or, at the interface with TiO_2 upon warm toughening. Later on Cl^- was found to go about as a doping material and also as a passivation for the surface in $MAPbI_{(3-x)}Cl_x$ perovskite. It was also demonstrated that $MAPbI_{(3-x)}Cl_x$ and $MAPbI_3$ have comparative photoluminescence decay lifetimes and absorption, fermi levels and bandgaps. However, devices having a photoactive layer of $MAPbI_{(3-x)}Cl_x$ normally produced higher-efficient PSCs.

$FAPbI_3$

Formamidinium lead halide perovskites $CH(NH_2)_2PbI_3$ ($FAPbI_3$) were utilized to make PSCs with improved efficiency and more prominent stability if contrasted with their generally utilized MA partners [35]. Compared with $MAPbI_3$ and $CsPbI_3$, the expanded ionic radius of FA brings about $FAPbI_3$ having a more extensive absorption and decreased bandgap (as appeared in Fig. 4) [36]. $FAPbI_3$ is moreover stable as it does not experience a phase transition inside the temperature scope of the PV device. Lee et al. [37] investigated a $FAPbI_3$ film that was topped by a $MAPbI_3$ layer over TiO_2. These solar cells were appeared to possess an efficiency of 16.01%, with a J_{sc} of 20.9 mA cm^{-2} and fill factor (FF) of 74%. The high fill factor watched was credited to a diminished resistance (Rs) and an expanded shunt resistance (Rsh). $FAPbI_3$ is normally prepared using a blend of FAI and PbI_2, although Wang et al. [38] used a reaction among the antecedent $HPbI_3$ and FAI, with $HPbI_3$ being integrated by blending PbI_2 with HI in proportion of 1:1.5. The device using $FAPbI_3$ manufactured utilizing this new technique accomplished an efficiency of 17.5%. Also,

Fig. 4 Ultraviolet–visible spectroscopy of various films of perovskite



the $FAPbI_3$ films manufactured from the FAI/$HPbI_3$ forerunner were observed to be uniform, thick and higher crystallinity when compared with their partners manufactured from PbI_2/FAI, thus in this way resulted in improved carrier transport and enhanced efficiency of device. These upgraded properties were credited to the way that it was not important to include HI to the forerunner solution keeping in mind the end goal to deliver a uniform, smooth perovskite film, not at all like past work utilizing formamidinium.

2.1.3 Planar Devices and ETL

The short history of these devices begins with the work of Lee et al. [8] in which a working planar device was demonstrated. An expansive change was appeared by a similar group and then by Liu in 2013, in which the films of evaporated perovskite were deposited in a planar set-up producing a conversion efficiency of about 15%. As clarified before, no hysteresis was appeared in that work. Another researcher Zhou and his colleagues demonstrated a conversion efficiency of 19.3%. In that work, the hysteresis was not discussed, using TiO_2 as electron transport layer. Correa demonstrated in the year 2015 that utilizing tin oxide by ALD and a blended ion perovskite layer produced a conversion efficiency of 18.2% which is stable one. In the year 2016, the utilization of tin oxide in the inverted PV arrangement likewise resulted a stable efficiency value of 18.8% [39]. Utilization of ionic fluid to control the morphology of the active layer produced an efficiency of 19.5% with SnO_2 as electron transport layer [40].

2.1.4 Hole Transport Layers

Despite the device design, the hole transport layer is one of the important segments to plan exceedingly efficient and stable PV cells. Presently, there are many p-type semiconductors, which are presented with their corresponding device performances, in some review research papers [41]. Small molecules support the processing in terms of flexibility. Small molecule optical properties, specifically bandgap and redox potential, are generally simple to adjust all together to alter the molecular backbone to the specific perovskite material [42, 43]. Doping of the small molecule hole selection layer is a fundamental stride to plan exceedingly efficiency perovskite solar cells. The combination and the part of each doping segment are progressively a consequence of exact advancements instead of the utilization of a careful outline. Expanding the hole transporting capacities while keeping up a low rate of recombination at the interface with the absorption/active layer is the most apparent impact so far saw by doping the hole transport layer [44, 45].

3 Conclusion

The perovskite materials which are solution-processed have accomplished amazing turning points in only a couple of years. These materials find their usage as a part of a bunch of utilizations from LEDs and photodetectors to photovoltaic cells. In this paper, authors have laid out and talked about the advancements in the perovskite solar cells and various blends used as photoactive layers. This has proceeded with the eagerness in this field, yet some extra efforts need to be done to comprehend and further enhance the parameters of PV and long-haul stability under operating perquisites. Going ahead, PSCs should diminish the recombination (non-radiative) and enhance charge transport with a specific end goal to accomplish the most astounding conceivable V_{OC} and fill factor values. Regardless, the noteworthy progression in such a little measure of time is unrivalled and might be in future achieving efficiency record of 29%, i.e., of crystalline silicon and GaAs. When calculating the efficiency values which are published, it is vital to utilize MPP tracking or, on the other hand stable output power, as these qualities are more illustrative than current voltage curves that are prone to hysteresis. In addition to this, latest work has demonstrated that both UV illumination and humidity have no significant impact on the long-haul stability, in so far as proper UV absorbers/encapsulations are used. There are unquestionably challenges staying in endeavouring to comprehend the part of movement of ions in long-haul stability, and how to further enhance the already superb PV parameters of solar cells. It is believed nevertheless that this field has extra years of thrilling new discoveries which will keep up the enthusiasm of both the industry and academia.

References

1. International Energy Agency (IEA): Solar Photovoltaic Energy. Technology Roadmap (2014)
2. OECD, IEA: Renewable Energy. Medium-Term Market Report 2014. Market Analysis and Forecasts to 2020. Executive Summary (2014)
3. Todorov, T., Gunawan, O., Guha, S.: A road towards 25% efficiency and beyond: perovskite tandem solar cells. Mol. Syst. Des. Eng. **1**, 370–376 (2016)
4. Kojima, A., Teshima, K., Shirai, Y., Miyasaka, T.: Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. J. Am. Chem. Soc. **131**, 6050–6051 (2009)
5. https://www.nrel.gov/pv/assets/pdfs/pv-efficiencies-07-17-2018
6. Domanski, K., et al.: Not all that glitters is gold: metal-migration-induced degradation in perovskite solar cells. ACS Nano **10**, 6306–6314 (2016)
7. Im, J.-H., Lee, C.-R., Lee, J.-W., Park, S.-W., Park, N.-G.: 6.5% efficient perovskite quantum-dot-sensitized solar cell. Nanoscale **3**, 4088–4093 (2011)
8. Lee, M.M., Teuscher, J., Miyasaka, T., Murakami, T.N., Snaith, H.J.: Efficient hybrid solar cells based on meso-superstructured organometal halide perovskites. Science **338** (2012)
9. Burschka, J., et al.: Sequential deposition as a route to high-performance perovskite-sensitized solar cells. Nature **499**, 316–319 (2013)
10. Jeon, N.J., Noh, J.H., Kim, Y.C., Yang, W.S., Ryu, S., Il Seok, S.: Solvent engineering for high-performance inorganic–organic hybrid perovskite solar cells. Nat. Mater. **13**, 897–903 (2014)
11. Yang, W.S., et al.: High-performance photovoltaic perovskite layers fabricated through intramolecular exchange. Science **348**, 1234–1237 (2015)
12. Saliba, M., et al.: Incorporation of rubidium cations into perovskite solar cells improves photovoltaic performance. Science **354**, 206–209 (2016)
13. Liu, M.: Efficient planar heterojunction perovskite solar cells by vapour deposition. Nature **501**, 395–398 (2013)
14. Zhou, H.P., et al.: Interface engineering of highly efficient perovskite solar cells. Science **345**, 542–546 (2014)
15. Anaraki, E.H., et al.: Highly efficient and stable planar perovskite solar cells by solution-processed tin oxide. Energy Environ. Sci. Energy Environ. Sci **9**, 3128–3134 (2016)
16. Malinkiewicz, O., et al.: Perovskite solar cells employing organic charge-transport layers. Nat. Photonics **8**, 128–132 (2014)
17. Jeng, J.Y., et al.: $CH_3NH_3PbI_3$ perovskite/fullerene planar-heterojunction hybrid solar cells. Adv. Mater. **25**(27), 3727–3732 (2013)
18. Sun, S., et al.: The origin of high efficiency in low-temperature solution-processable bilayer organometal halide hybrid solar cells. Energy Environ. Sci. **7**, 399–407 (2014)
19. Nie, W., et al.: High-efficiency solution-processed perovskite solar cells with millimeter-scale grains. Science **347**, 522–525 (2015)
20. H.-S. Kim, et al.: Lead iodide perovskite sensitized all-solid-state submicron thin film mesoscopic solar cell with efficiency exceeding 9%. Sci. Rep. **2** (2012)
21. Yi, C., Luo, J.: Entropic stabilization of mixed A-cation ABX3 metal halide perovskites for high performance perovskite solar cells. Energy Environ. Sci. **9**, 656–662 (2016)
22. David, P., et al.: A mixed-cation lead mixed-halide perovskite absorber for tandem solar cells. Science **351**, 151–155 (2017)
23. Snaith, H.J., et al.: Perovskites: the emergence of a new era for low-cost, high-efficiency solar cells. J. Phys. Chem. Lett. **4**, 3623–3630 (2013)
24. Sum, T.C., Mathews, N.: Advancements in perovskite solar cells: photophysics behind the photovoltaics. Energy Environ. Sci. **7**, 2518–2534 (2014)
25. Grätzel, M., Gao, P.: Organohalide lead perovskites for photovoltaic applications. Energy Environ. Sci. **7**, 2448–2463 (2014)
26. Boix, P.P., Agarwala, S., Koh, T.M., Mathews, N., Mhaisalkar, S.G.: Perovskite solar cells: beyond methylammonium lead iodide. J. Phys. Chem. Lett. **6**, 898–907 (2015)

27. Bai, S., Wu, Z., et al.: Efficient planar heterojunction perovskite solar cells employing graphene oxide as hole conductor. Nanoscale **6**, 10505–10510 (2014)
28. Zhang, M., Yu, H., Lyu, M., Wang, Q., Yun, J.-H., Wang, L.: Composition-dependent photoluminescence intensity and prolonged recombination lifetime of perovskite $CH_3NH_3PbBr_{3-x}Cl_x$ films. Chem. Commun. **50**, 11727–11730 (2014)
29. Qiu, Y., Qiu, J., et al.: All-solid-state hybrid solar cells based on a new organometal halide perovskite sensitizer and 1D TiO_2 nanowire arrays. Nanoscale **5**, 3245–3248 (2013)
30. Jung, H.S., Park, N.G.: Perovskite solar cells: from materials to devices. Small **11**, 10–25 (2015)
31. Heo, J.H.: Hysteresis-less inverted $CH_3NH_3PbI_3$ planar perovskite hybrid solar cells with 18.1% power conversion efficiency. Energy Environ. Sci. **8**, 1602–1608 (2015)
32. Grätzel, M., et al.: The light and shade of perovskite solar cells. Nat. Mater. **13**, 838–842 (2014)
33. Holliman, P.J., Williams, A.E.: Perovskite processing for photovoltaics: a spectro-thermal evaluation. J. Mater. Chem. A **2**, 19338–19346 (2014)
34. Unger, E.L., et al.: Chloride in lead chloride-derived organo-metal halides for perovskite-absorber solar cells. Chem. Mater. **26**, 7158–7165 (2014)
35. Minemoto, T., Murata, M.: Impact of work function of back contact of perovskite solar cells without hole transport material analyzed by device simulation. Curr. Appl. Phys. **14**, 1428–1433 (2014)
36. Stranks, S.D., Eperon, G.E.: Formamidinium lead trihalide: a broadly tunable perovskite for efficient planar heterojunction solar cells. Energy Environ. Sci. **7**, 982–988 (2014)
37. Lee, J.W., Seol, D.J., Cho, A.N., Park, N.G.: High-efficiency perovskite solar cells based on the black polymorph of $HC(NH_2)_2PbI_3$. Adv. Mater. **26**, 4991–4998 (2014)
38. Wang, F., Yu, H., Xu, H., Zhao, N.: $HPbI_3$: a new precursor compound for highly efficient solution-processed perovskite solar cells. Adv. Funct. Mater. **25**, 1120–1126 (2015)
39. Bai, Y., Zhu, Z., et al.: Enhanced efficiency and stability of inverted perovskite solar cells using highly crystalline SnO_2 nanocrystals as the robust electron-transporting layer. Adv. Mater. **28**, 6478–6484 (2012)
40. Matsui, T., Seo, J.-Y., et al.: Ionic liquid control crystal growth to enhance planar perovskite solar cells. Adv. Energy Mater. **6** (2016)
41. Rub, M.A., Ameen, S., et al.: Perovskite solar cells: influence of hole transporting materials on power conversion efficiency. CAMSUSCHEM **8**, 10–27 (2016)
42. Shirahata, Y., Hamatani, T.: Arsenic and chlorine co-doping to $CH_3NH_3PbI_3$ perovskite solar cells. Adv. Mater. Phys. Chem. **7** (2013)
43. Abate, A., et al.: Lithium salts as 'redox active' p-type dopants for organic semiconductors and their impact in solid-state dye-sensitized solar cells. Phys. Chem. Chem. Phys. **15**, 2572–2579 (2013)
44. Staff, D.R., Abate, A., et al.: Influence of ionizing dopants on charge transport in organic semiconductors. Phys. Chem. Chem. Phys. **16**, 1132–1138 (2014)
45. Park, S., et al.: Effect of multi-armed triphenylamine-based hole transporting materials for high performance perovskite solar cells. Chem. Sci. **7**, 5517–5522 (2016)

Advanced Electronics Applications-II

J1939 Functionality Development in Diagnostic Event Manager Module Based on AUTOSAR 4.2.1

Ankit Sharma, R. K. Sharma and Narayan Kamath

Abstract Diagnostic system is an important part of vehicle's software architecture. At present, every high-end vehicle has software to control various operations and also assist driver. AUTOSAR is open system architecture for a vehicle's ECU software, and diagnostic is part of its service layer [1]. In AUTOSAR, diagnostic system is group of three modules—DEM, DCM, and FIM modules. Diagnostic system in a vehicle is responsible for detecting and logging errors or faults and also alerting driver about it through various lamps like malfunction indicator lamp (MIL), etc [2]. It plays an important role in safety of passengers. AUTOSAR allows adding different protocols in diagnostic system to support a range of diagnostic services those are well defined in every protocol. Main parts of vehicle those are covered by diagnostic system are engine and emission system. Adding J1939 functionalities in DEM module boosts up diagnostic system usability. After implementation, this diagnostic system can be used in cars as well as heavy-duty vehicle. SAE_J1939 is developed by Society of Automotive Engineers (SAE) and recommended for heavy-duty vehicle on and off road. J1939 protocol defines diagnostic messages (DMs) to carry different fault-related data. It helps mechanic to identify and fix the faults in vehicle. It also defines different lamps to indicate condition of various subsystem of vehicle to the driver. Through lamps' behavior, users (driver and mechanic) can easily observe fault condition in vehicle. Lamps' feature is an efficient way of human–machine interaction and makes vehicle more user-friendly. J1939 is developed for heavy-duty vehicles used in industry also for vehicles those are used in forestry and agriculture. This paper describes development process of J1939 functionalities in DEM module based on AUTOSAR 4.2.1 step by step.

Keywords AUTOSAR · Diagnostic system · Diagnostic event manager (DEM) · SAE_J1939 · Diagnostic communication manager (DCM)

A. Sharma (✉) · R. K. Sharma
NIT Kurukshetra, Kurukshetra, Haryana, India
e-mail: nktsharma2807@gmail.com

N. Kamath
KPIT Technologies Ltd., Bangalore, Karnataka, India

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_9

1 Introduction

The features of vehicle according to their use in different field are increasing day by day. Utilization of vehicles is not simply restricted to transportation, yet it is in numerous fields like development, ranger service, farming, safeguard activities, and military. Heavy-duty vehicles are in use for industrial purpose. In construction field, vehicle with trailer to carry material is used. In rescue operations, vehicle with supported system is added to them used in fire rescue operation. These vehicles have an added system that supports different operations other than transportation system. As number of features in vehicles expanded, the framework necessities related with vehicle end up complex consequently [3]. To control and perform this system, more ECUs are required in vehicle that made framework more mind-boggling [4]. An institutionalized engineering is an answer for decrease in the unpredictability of the framework [5]. Henceforth, AUTOSAR (AUTomotive Open System Architecture) showed up and got much popularity in the car space. AUTOSAR is an open and institutionalized stage for car programming [6]. By utilizing AUTOSAR institutionalization, adaptability, expanded quality, and well-being of vehicle framework can be accomplished [7]. The AUTOSAR is an open and institutionalized layered auto-rationale programming design mutually created via car makers, providers, and apparatus engineers [8]. The difference between AUTOSAR and the other software architectures used in vehicle domain is the concept that allows making the functionality of standardized layers independent from hardware ECU [9]. At present, the engineering of the inserted frameworks is not totally secluded and a change in hardware modules influences the hardware abstraction modules but the other software modules, which might be changed for this situation [9, 10].

AUTOSAR is a layered architecture that defines various leveled structure of programming with connection and mapping of programming layers with fundamental programming modules [11]. This building gives strong relationship between application programming and shrouded gear sections like sensors, actuators, and microcontroller hardware. Figure 1 demonstrates the AUTOSAR layered architecture that recognizes three programming layers. Application layer, runtime environment (RTE), and basic Software are three software layers which run on a microcontroller. The top layer is called application layer which comprises application software components which are connected through abstract component, named the virtual function bus. The application programming parts are the smallest that have singular usefulness. RTE is the middle layer of AUTOSAR layered architecture. RTE provides communication services to the application software. The basic software layer is lower layer that consists of ECU abstraction layer, service layer, microcontroller abstraction layer, and complex drivers; these functional groups are corresponding to system, memory, and communication services. Microcontroller abstraction layer is the bottom layer of basic software (BSW). Microcontroller abstraction layer makes higher software layers independent on microcontroller. The service layer is the top layer of the basic

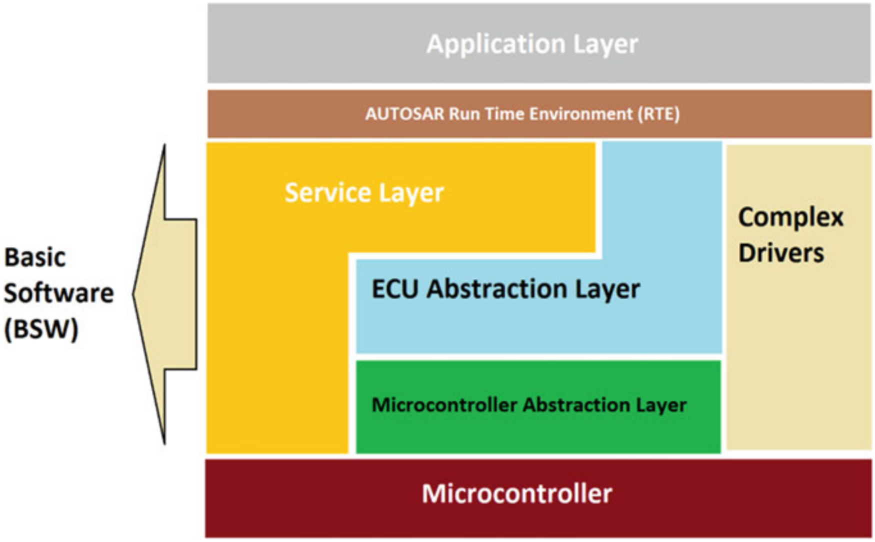


Fig. 1 AUTOSAR layered architecture with BSW description

software. Service layer provides basic software modules and basic services for applications. RTE provides communication services to the software components in application layer. RTE makes AUTOSAR software components independent from any specific ECU. RTE made the AUTOSAR software components able to communicate with other components (inter and/or intra ECU) and/or services.

Rest of the paper comprises development steps as diagnostic event manager and J1939 functionalities described in Sect. 2. Design as sequence diagram is described in Sect. 3. Configuration explained in Sect. 4. Section 5 contains results, and conclusion is given in Sect. 6.

2 J1939 Functionalities in Diagnostic Event Manager Module

Diagnostic event manager (DEM) is one of the modules in service layer of basic software (BSW). DEM handles and stores the events detected by diagnostic monitors in both software components (SW-Cs) and basic software (BSW) modules. The stored event information is available through an interface to other BSW modules or SW-Cs. Diagnostic event is the status (pass/fail) of a system under test for a set of test conditions. If an event gets qualified as failed, it becomes active. If the event gets qualified as passed, it becomes passive. The system/function that monitors these conditions is said to be diagnostic monitor or a diagnostic function. Relevant errors are reported either from application layer (SWC) or basic software modules (BSWM).

- BSWs report the new status of the event with the “Dem_ReportErrorStatus” API.
- SWCs report the new status of the event with the “Dem_SetEventStatus” API through the RTE layer.

A standardized interface and various debounce algorithms for diagnostic monitors empower uniform and cross-venture advancement of the embedded programming. At least one fault or event can be mapped to a diagnostic trouble code (DTC). The DEM is additionally configured from the ECU Configuration Description. It contains information related to faults in vehicle as freeze frames and DTC numbers. J1939 is a set of standards defined by SAE (Society of Automotive Engineers). They are used in heavy-duty vehicles such as trucks and buses, mobile hydraulics, etc. The application of J1939 focus is on the power trains and chassis of commercial vehicles and heavy-duty vehicles like bus and truck industry. The protocol is used in heavy vehicles for on-street and off-road operations (construction machines), forestry, and agriculture also for vehicle used in armed forces.

SAE J1939-73 defines the message structures and behavior of “diagnostic messages” (DMs) which are used for diagnostic communication in J1939 networks. J1939 is also used during vehicle operation to report immediate diagnostic information into the vehicle like periodically broadcasting active DTCs to the instrument cluster to communicate to the driver status of the vehicle using different lamp status. J1939 defines six functionalities in diagnostic event manager those are covered with the help of fourteen APIs. These functionalities are as follows: read DTC, clear diagnostic data, broadcasting lamps’ status, provide freeze frame on request, provide IUMPR ratio, and readiness information. Reading DTC is supported by three APIs:

- Dem_J1939DcmSetDTCFilter
- Dem_J1939DcmGetNumberOfFilteredDTC
- Dem_J1939DcmGetNextFilteredDTC.

All three APIs work together in order to provide DTC information to tester tool when requested by tester tool. Likewise other functionalities are supported by respective APIs.

3 Design

The DEM module describes a set of application programming interfaces (APIs) for J1939 and provides requirements those are to be realized through embedded code for implementing J1939 functionality in DEM module. Design of DEM for J1939 protocol consists six functionalities. There are two level of design one is high-level design (HLD) and second is low-level design (LLD). High-level design gives the general framework design as far as useful architecture and the outline of framework improvement.

Low-level design describes the flow of the program code which is outlined in light of high-level outline. It explains the working of the internal logic of the corresponding APIs. Designs are made using the tool named Enterprise Architect version 12.0.1. Enterprise Architect supports various methods of modeling business processes using UML as the foundation modeling language [12]. The design of some of J1939 functionalities in DEM module is shown as sequence diagram in Figs. 2, 3 and 4.

In Fig. 2, tester tool sends diagnostic message DM11 to clear all J1939-related DTCs. Inside vehicle system J1939_DCM will trigger Dem_J1939DcmClearDTC API to perform the operation. When diagnostic message DM11 is requested, all DTCs as well as additional information like freeze frame will be cleared. If operation executed successfully, DEM returns E_OK to J1939_DCM.

In Fig. 3, tester tool sends a request message DM1 to read active DTCs. According to request J1939_DCM module sequentially call three APIs those supports read DTC functionality. First API Dem_J1939DcmSetDTCFilter will set input parameter on that basic DTC will be filtered out in further processing. Second API Dem_J1939DcmGetNumberOfFilteredDTC will provide count of DTCs those met the filter criteria set by first API to J1939_Dcm module. Now J1939_DCM module will call third API Dem_J1939DcmGetNextFilteredDTC to fetch DTCs.

In Fig. 4, one ECU sends request for DM31 that is reading lamps' status and second ECU receives the request message DM31 and call-related APIs sequentially in order to provide lamp status.

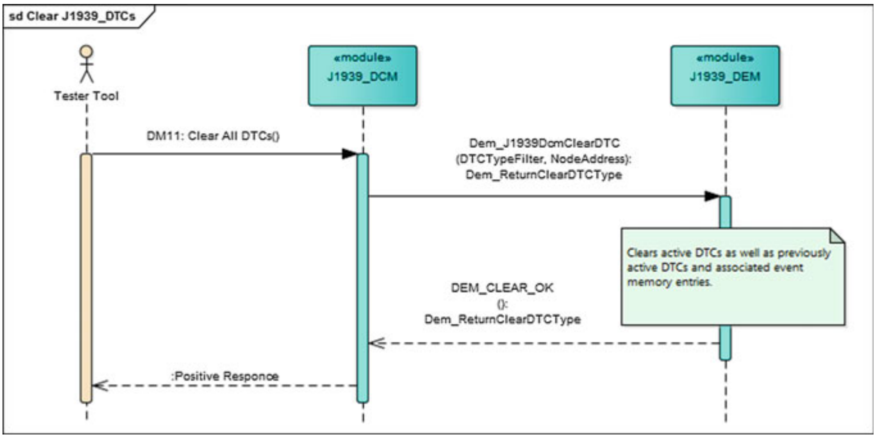


Fig. 2 Sequence diagram for clear J1939_DTC functionality

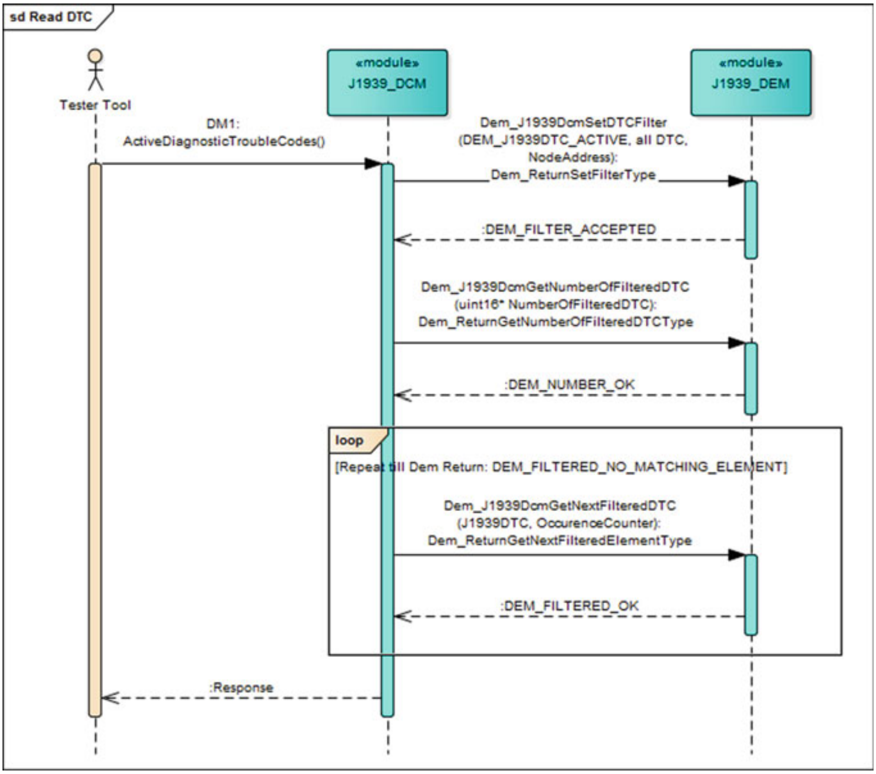


Fig. 3 Sequence diagram for read J1939_DTC functionality

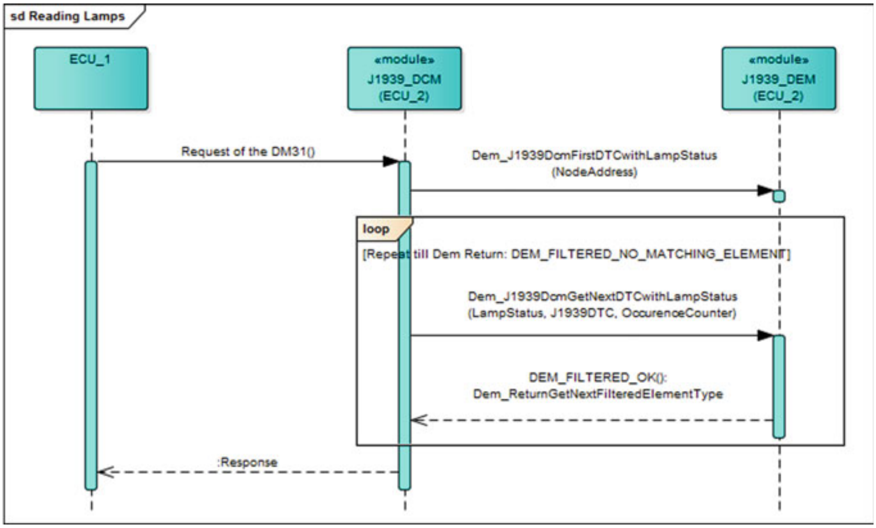


Fig. 4 Sequence diagram for reading lamp status functionality

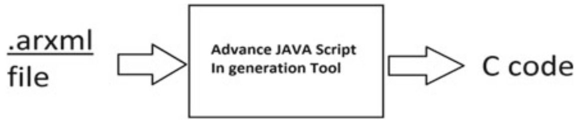


Fig. 5 Procedure for C code generation in generation tool

4 Configuration

Configuration tool that is used to configure J1939 part of DEM module is Compose4K-SAR. These kinds of configuration software are developed by AUTOSAR stack supplier. Compose4K-SAR is developed by KPIT Technologies. To develop AUTOSAR-compliant software for an ECU, software engineer depends on configuration tools, in order to develop software in lesser time since it is time-consuming doing configuration manually [12, 13]. AUTOSAR also allows plug-in from car manufacturers that can be configured using configuration tool in very less time it can be done through configuration tool easily in lesser time. For configuration to develop J1939 functionalities in DEM Compose4K-SAR tool has been used. This tool generates C code file according to arxml and advance java script code files (Fig. 5).

Code of arxml file is generated according to configuration done by programmer in generator tool that is specific to configuration chapter in AUTOSAR specification sheet of DEM module. An advance JAVA script file works on Eclipse platform; it takes arxml as an input in configuration tool and generates the C code [14, 15]. Generated C code by generation tool (Compose4K-SAR) compiled together with static C code to form executable file (Fig. 6).

5 Results

The DEM module has a well-defined code file. It includes the header file and generated files those are the output of generation tool according to configuration file required for DEM. The DEM module implementation provides a file named DiagEvtM_J1939DiagCom.h which contains extern declaration of every API of DiagEvtM_J1939DiagCom.c file. DiagEvtM_J1939DiagCom.c file contains code that defines all APIs related to J1939 functionalities. For J1939 functionalities DEM module comprises DiagEvtM_J1939Cfg.h and DiagEvtM_J1939Cfg.c files; these are generated files. DiagEvtM_J1939Cfg.h consists of macro values according to configuration parameters, and DiagEvtM_J1939Cfg.c consists of data in structure form that also according to configuration parameters. SchMgr_DEM.h and MemoryMap.h are included in DEM module implementation. MemoryMap.h uses to map the code and the data of the DEM module into specific memory sections. The APIs of J1939 functionality in DEM module are developed. Design is done with the help of the tool Enterprise Architect version 12.0.1. The embedded C code is written in Notepad++, and compilation is done by using gcc compiler. The testing is done by

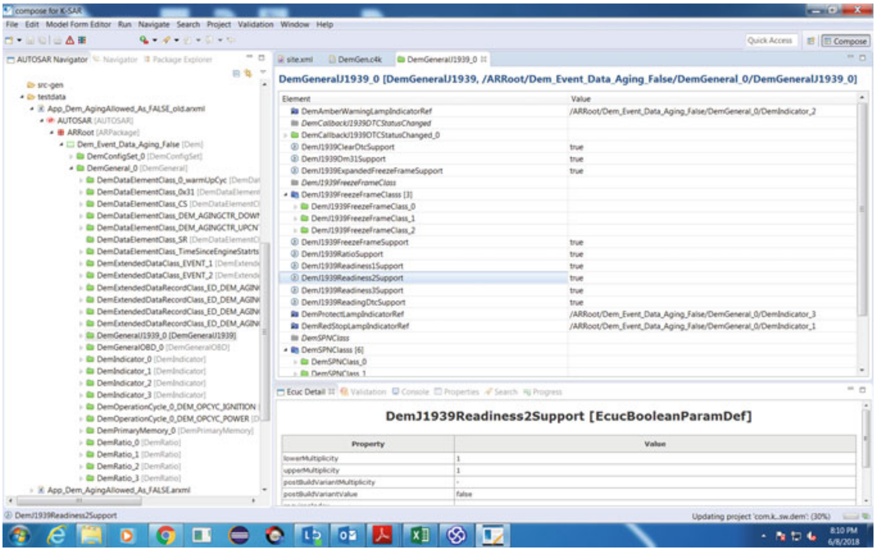


Fig. 6 Snap shot of configuration in generation tool

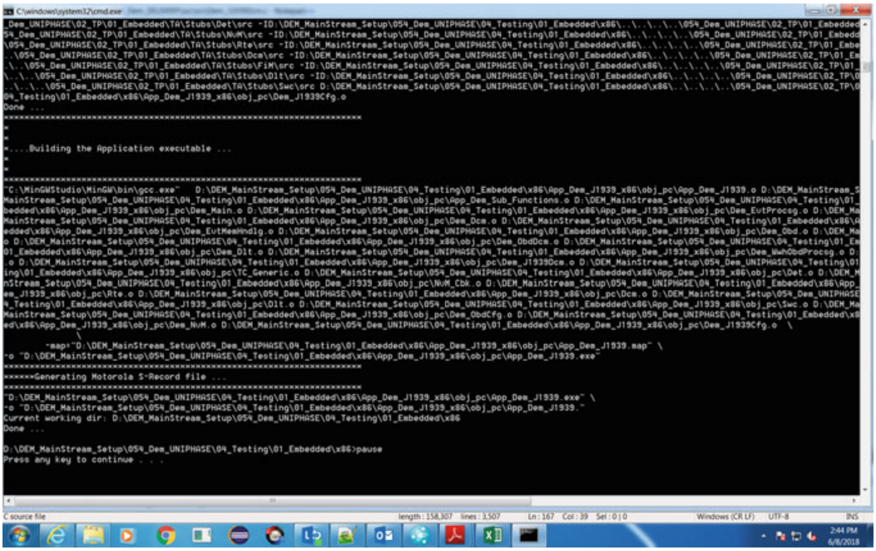


Fig. 7 Snap shot of compilation using gcc compiler with no compilation error

writing different test cases code that covers main functionalities and working scenarios. In testing, each of the application programming interfaces (API) is covered. By successfully running the test application, API can be validated [12]. Figure 7 shows successful compilation using the gcc compiler without any compilation error.

6 Conclusion

As the improvement in car industry is in a quick stage, standard software architecture is important for vehicle's ECUs to decrease intricacy and to build well-being. Consequently, AUTOSAR engineering is presently broadly utilized as a standard design in car industry due to its particular highlights. The DEM module present inside the services layer of AUTOSAR architecture and J1939 functionalities of DEM will make diagnostic module able enough to diagnose not only cars but also heavy-duty vehicles. It covers bus and truck industry, and vehicle those are used in forestry and agriculture. All the application programming interfaces (APIs) related to J1939 in DEM module are implemented based on the requirements and specifications given by DEM module version 4.2.1.

References

1. Wang, Q., Xin, B., Li, C., Chen, H.: The Realization of Reusability in Vehicular Software Development under AUTOSAR
2. AUTOSAR: Specification Diagnostic Event Manager V4.2.1 (online). Available from: http://www.autosar.org
3. Durisic, D., Nilsson, M., Staron, M., Hansson, J.: Measuring the impact of changes to the complexity and coupling properties of automotive software systems. J. Syst. Softw. **86**(5), 275–1293 (2013)
4. Durisic, D., Staron, M., Hansson, J.: Evolution of Long-Term Industrial Meta-Models—An Automotive case Study of AUTOSAR
5. Kitchenham, B., Pfleeger, S., Pickard, L., Jones, L., Hoaglin, P., Emam, K., Rosenberg, J.: Preliminary guidelines for empirical research in software engineering. J. IEEE Trans. Softw. Eng. **28**, 721–734 (2002)
6. Mjeda, A., Leen, G., Waish, E.: The AUTOSAR Standard—The Experience of Applying Simulink According to Its Requirements. SAE International (2007)
7. Hardung, B., Kolzow, T., Kruger, A.: Reuse of Software in Distributed Embedded Automotive Systems (2004)
8. Kim, J.-W., Lee, K.J., Ahn, H.: Development of software component architecture for motor-driven power steering control system using AUTOSAR methodology. In: International Conference on Control, Automation and Systems (ICCAS 2015)
9. Briciu, C.V., Filip, I., Heininger, F.: A new trend in automotive software: AUTOSAR concept (2013)
10. Broy, M., Kruger, I., Pretschner, A., Salzmann, C.: Engineering automotive software. Proc. IEEE **95**(2) (2007)
11. Sung, K., Han, T.: Development process for AUTOSAR-based embedded system. Int. J. Control Autom. (2013)
12. Dhanamjayan, P.R., Jose, K., Manjusree, S.: ECU state manager module development and design for automotive platform software based on AUTOSAR 4.0. Int. J. Tech. Res. Appl. (2015)
13. Broy, M.: Challenges in automotive software engineering. In: International Conference on Software Engineering (2006)

14. Liu, Y., Li, Y., Zhuang, R.: The Application of Automatic Code Generation Technology in the Development of the Automotive Electronics Software. Trans Tech Publications, Switzerland (2013)
15. Pagel, M., Brorkens, M.: Definition and generation of data exchange formats in AUTOSAR. Springer, Berlin, Heidelberg (2006)

Design and Investigation of Compact Microstrip Patch Array Antennas for Narrowband Applications

Swapnil S. Thorat and Sangeeta R. Chougule

Abstract This paper investigates possibility of designing microstrip patch array antenna without feed network. A simple rectangular patch antenna, fed with coaxial feeding, on readily available FR4 substrate with larger ground of size 60 mm × 60 mm is designed to operate in C band. Improvement in gain is achieved by designing multiple-level array of 2 × 2, separated by air gap on second layer of FR4 on top of first antenna, and its results are optimized by doing parametric analysis. Gain of the proposed antenna system is observed to be improved in three-layer antenna up to 9.63 dB than simple rectangular patch antenna. VSWR bandwidth of proposed antenna is 500 MHz.

Keywords Multilayer · Array · Parasitic feeding · Space feed · Electromagnetic coupling · Microstrip patch antenna

1 Introduction

Microstrip patch array antennas, despite having numerous advantages, suffer application limitations due to complex feed network design which occupies precious space and causes unwanted radiations at higher frequency [1–3]. Advancement in array antenna design was investigated in substrate integrated waveguide arrays (SIW) with Yagi-like parasitic element, which accumulates losses in the form of metallic and dielectric loss resulting in lower efficiency [4, 5]. Microstrip grid array antennas give better performance based on properly synchronized surface currents. Gain of array antenna system is in direct relation with number of radiating elements and substrate parameters like permittivity, thickness, etc. [6]. Leaky wave antennas (LWAs)

S. S. Thorat (✉)
Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, Maharashtra 411046, India
e-mail: ssthorat1@bvucoep.edu.in

S. R. Chougule
Kolhapur Institute of Technology, Kolhapur, Maharashtra 416006, India
e-mail: kit.hodetc@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_10

use dielectric or metallic superlayers to increase the directivity by placing a partially reflecting surface in front of an antenna with a ground plane. Larger dielectric constant of upper layer substrate helps in reducing side lobe levels of antenna ultimately improving radiation pattern [7]. One of early attempts of spatially fed structures appears in local oscillator feeding where local oscillators fed mixers via array of slot-coupled patch antennas. This approach proved very helpful as it completely eliminates need of feed network design. Many other unorthodox combinations of array antenna can be created due to elimination of physical feed network [8]. Arrangement of the antenna elements in a longitudinal direction becomes problematic due to the element spacing constraint for alias-free sampling; hence, a 16-port shared-arm dipole array (SADA) was proposed. This reduces size and feed network, but for better efficiency, all the 16 ports must work in mathematical synchronization and have to energize separately which is quite complex [9, 10]. A very unconventional approach of using null directions or null points as a reference in setting beam direction of electronically steered antennas is applied for five-element linear array; extension of idea to planar arrays is still to be investigated [11]. Use of electromagnetic band gap (EBG) structures along with array antenna reduces size of radiating patch and further improves gain of antenna, but choice of EBG structure, their spacing, and size or shape and increased fabrication complexity, cost, etc., proved to be disadvantage of the system [12]. Some multilayered array antennas suggested feeding of array elements in top layer parasitically by middle layer of properly arranged slot antennas. Due to multilayered nature, unwanted radiations from feed network are minimized but still there is need of designing the feed network in base plate [13–16]. A circular array composed of 25 wire monopole elements center fed and positioned in two concentric rings over a ground plane was investigated where one driven element and 24 parasitic elements, 12 of which are loaded with an electronically controllable capacitive load are present. Though there is limitation of need of electronically controlled switches or capacitive loads, this work shows the potential in parasitic feeding technique [17]. Feed network elimination can also be achieved by parasitic coupling at edges of central active feed antenna to achieve compact low-cost phased array of dielectric resonator antenna, but very less improvements in terms of gain are observed [18]. Recently observed use of parasitic strips and sleeve feed for mobile phone applications for bandwidth improvement shows promising application field for such nontraditionally feed antennas [19].

2 Structure of Antenna

Three different antenna structures are proposed in this paper. First antenna—'Antenna A'—is single-layer simple rectangular microstrip patch antenna (RMSA). Second antenna—'Antenna B'—is two layer antenna with upper layer containing 2 × 2 array of rectangular patches without any feed network. Third antenna—'Antenna C'—is three-layer antenna with first layer of single rectangular microstrip patch antenna and upper two layers containing array of 2 × 2 array of rectangular patches. Detailed design and results are discussed in sections to follow.

2.1 Antenna A

2.1.1 Structure

A square microstrip patch antenna designed using commonly available FR4 substrate with dielectric constant of 4.4 and loss tangent of 0.0002. Centrally situated on substrate of 60 mm × 60 mm and thickness of 1.53 mm, square radiating patch was optimized for its dimensions to work on resonating frequency of 5.5 GHz. Entire opposite plane of substrate is acting as ground plane. Coaxial feeding is used to avoid spurious radiations from feedline structure. Position of coaxial feeding was optimized to yield better impedance matching. Impedance of coaxial feed is assumed to be 50 Ω (Fig. 1).

Design equations for dimensions of patch are as follows [20]

$$\text{Width (W)} = \frac{c}{2f_0\sqrt{\frac{\varepsilon_R+1}{2}}} \tag{1}$$

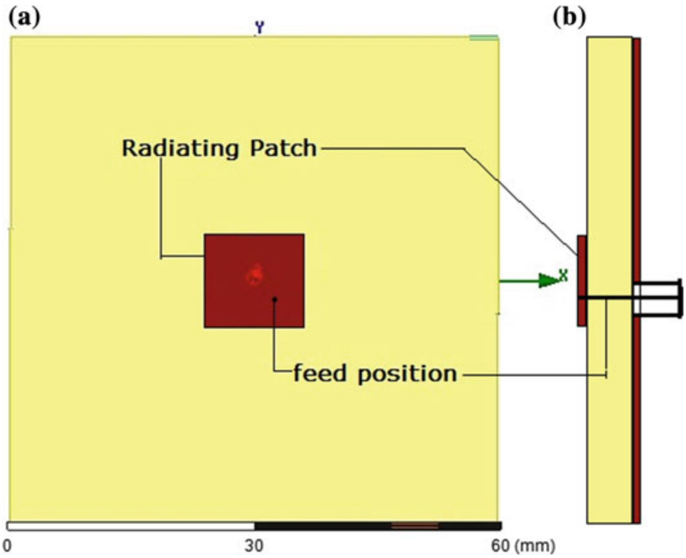


Fig. 1 **a** Top view; **b** side view of simple microstrip patch Antenna A

$$\varepsilon_{\text{eff}} = \frac{\varepsilon_{\text{R}} + 1}{2} + \frac{\varepsilon_{\text{R}} - 1}{2}\left[\frac{1}{\sqrt{1 + 12\left(\frac{h}{W}\right)}}\right] \tag{2}$$

$$\text{Length} = \frac{c}{2 f_0 \sqrt{\varepsilon_{\text{eff}}}} - 0.824h\left[\frac{(\varepsilon_{\text{eff}} + 0.3)\left(\frac{W}{h} + 0.264\right)}{(\varepsilon_{\text{eff}} - 0.258)\left(\frac{W}{h} + 0.8\right)}\right] \tag{3}$$

where ε_{R} is relative dielectric constant of substrate material and ε_{eff} is effective dielectric constant. 'h' is height of substrate material or thickness. 'c' is velocity of light in free space.

2.1.2 Results

a. **Reflection Coefficient**

The above structure is simulated in HFSS software, and Fig. 2 shows return loss of −23.59 dB at resonating frequency of 5.5 GHz. Return loss of less than −10 dB signifies frequencies at which antenna is radiating effectively. Figure 2 shows total operating band of 248.1 MHz with lower cutoff frequency 5.4 GHz and upper cutoff frequency of 5.6481 GHz.

b. **Radiation Pattern**

Isotropic radiation pattern is observed in horizontal plane. Vertical plane radiation pattern shows directive radiation pattern with 60° half power beamwidth as shown in Fig. 3. No side lobes or back lobes are observed as a result of sufficiently large ground plane. Maximum gain observed in broadside direction is 5.13 dB. Slight shift of radiating lobe from broadside axis is observed due to optimized position of coaxial feed.

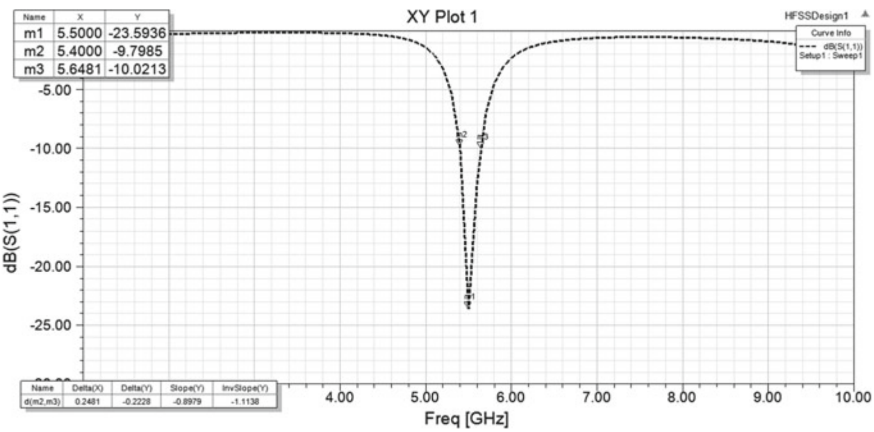


Fig. 2 Graph of reflection coefficient S11 versus frequency of Antenna A

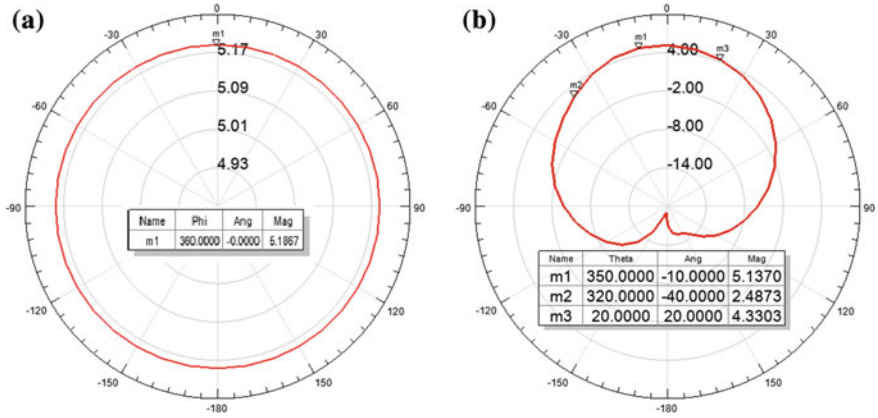


Fig. 3 Radiation pattern **a** horizontal plane; **b** vertical plane of Antenna A

2.2 Antenna B

2.2.1 Structure

A 2 × 2 array of RMSA was designed over previously optimized square 'Antenna A.' Upper array of antennas was supported by same FR4 substrate as shown in green color in Fig. 4. Upper 2 × 2 array does not have any direct feed network; rather, it is fed by radiations from first layer (bottom) rectangular antenna. This shows improvement over performance of antenna.

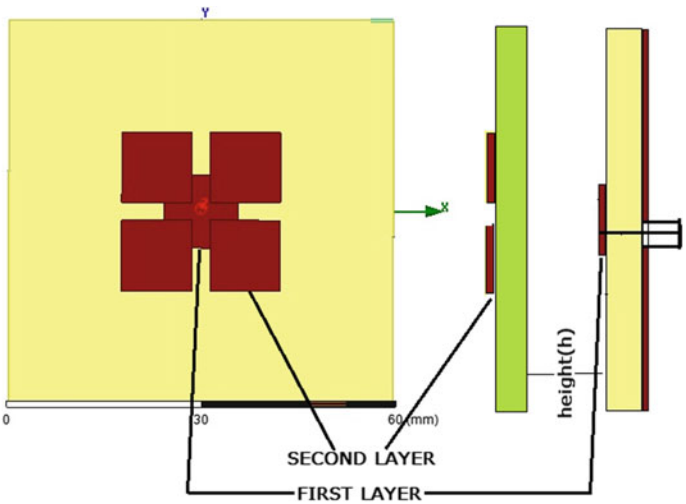


Fig. 4 Top view and side view of simple microstrip patch Antenna B

2.2.2 Results

a. **Reflection Coefficient**

Return loss −33.58 dB obtained at 6.1 GHz shows very good impedance matching at resonating frequency. Figure 5 shows total operating band of 230 MHz with lower cutoff frequency 5.99 GHz and upper cutoff frequency of 6.22 GHz. Resonating frequency is shifted toward upper side of band due to increased effective height.

b. **Radiation Pattern**

Isotropic radiation pattern in horizontal and directive in vertical plane is observed similar to 'Antenna A,' but maximum gain observed in broadside direction is 8.99 dB which is almost 4 dB more than gain with single patch, i.e., 5.13 dB. Slight shift of radiating lobe toward broadside axis is observed due to optimized position of upper array antenna (Fig. 6).

c. **Parametric analysis of Air Gap between the first and second layers**

Resonating frequency of antenna shifts from 6.2 to 5.7 GHz, i.e., toward lower side of band with significant change in impedance matching but without any change in bandwidth of antenna as distance between upper and lower patch is increased from 5 to 10 mm. Variations in air gap can be used as potential parameter in tuning antenna to different frequencies (Fig. 7).

d. **Parametric analysis of Upper Patch dimension**

As observed from graph in Fig. 8; size of radiating element in upper patch increases from 10.8 to 14 mm resonating frequency decreases with best impedance matching at 6.1 GHz; relatively very small change is observed in frequency.

e. **Parametric analysis of Gap width between array elements**

As gap between array elements is widened from 1 to 5 mm, no change in resonating frequency is observed but greater impedance matching is happening when array elements are very close to each other as observed in Fig. 9.

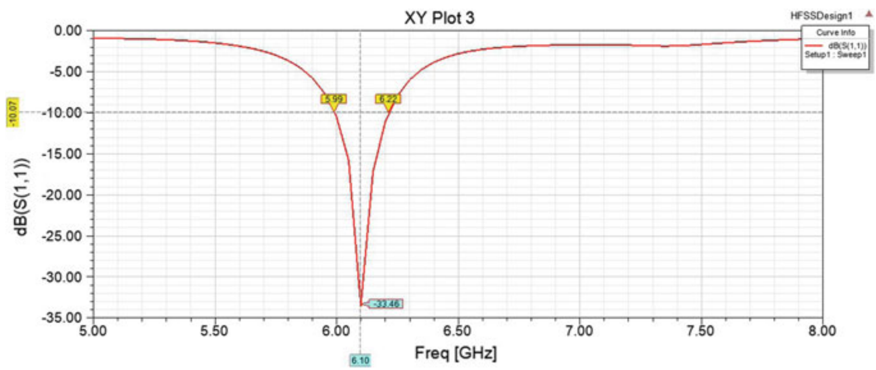


Fig. 5 Graph of reflection coefficient S11 versus frequency of Antenna B

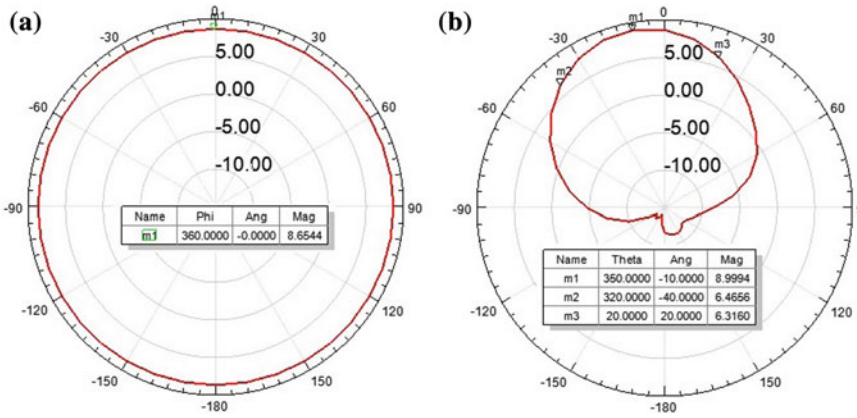


Fig. 6 Radiation pattern **a** horizontal plane; **b** vertical plane of Antenna B

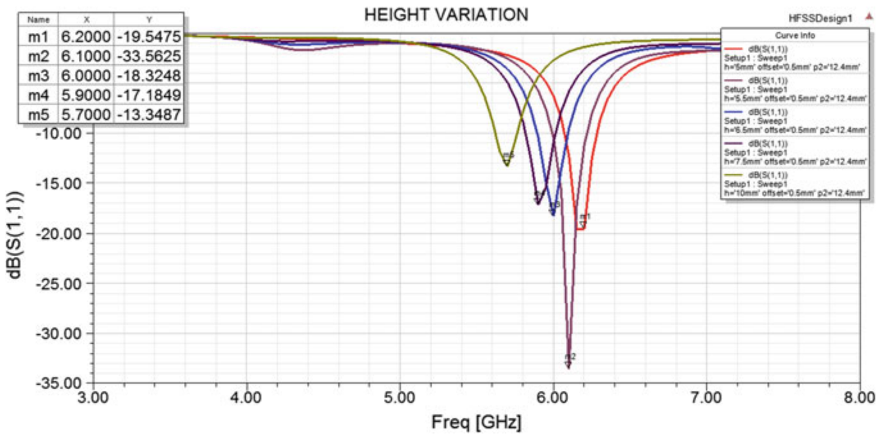


Fig. 7 Parametric analysis of air gap between the first and second layers of Antenna B

2.3 Antenna C

2.3.1 Structure

A three-layered 2 × 2 array of RMSA was designed over previously optimized 'Antenna B.' Gap between array elements is increased in the third layer as compared to the second layer of antenna (Fig. 10).

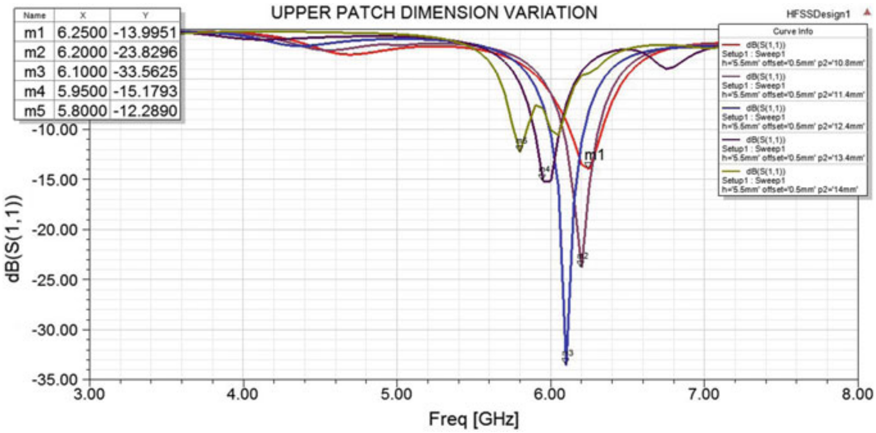


Fig. 8 Parametric analysis of upper patch dimension of Antenna B

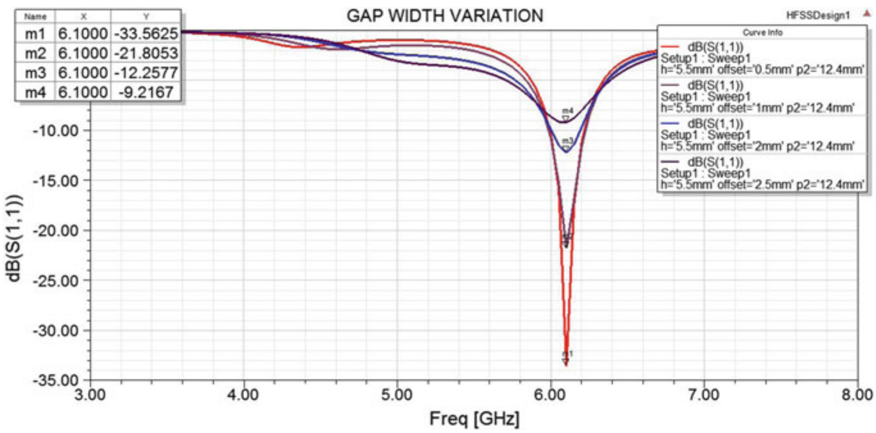


Fig. 9 Parametric analysis of gap width between array elements

2.3.2 Results

a. **Reflection Coefficient**

Figure 11 shows total operating band of 500 MHz with lower cutoff frequency 5.90 GHz and upper cutoff frequency of 6.40 GHz. Resonating frequency is shifted slightly toward upper side of band due to increased effective height.

b. **Radiation Pattern**

Shape of radiation pattern remains unchanged with maximum gain observed in broadside direction as 9.63 dB which is almost 5.5 dB more than gain with single patch,

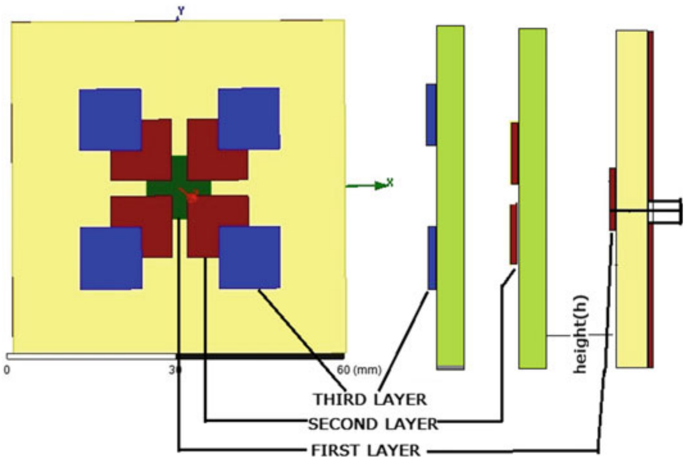


Fig. 10 Top view and side view of Antenna C

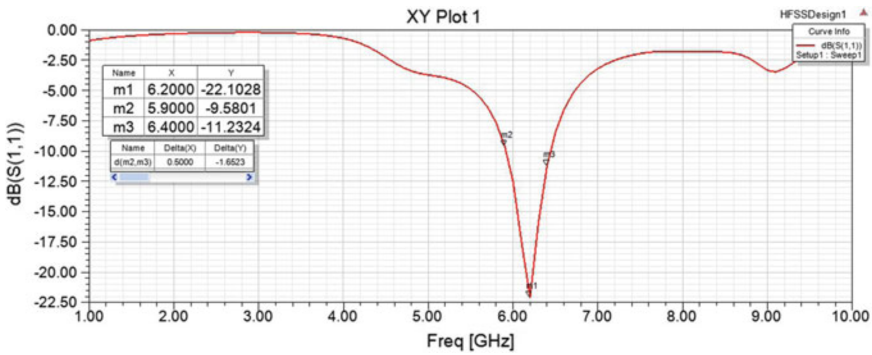


Fig. 11 Graph of reflection coefficient S11 versus frequency of Antenna C

i.e., 5.13 dB. Radiating lobe is aligned along broadside axis is due to optimized position of elements of upper array antenna (Fig. 12).

3 Comparative Result Analysis

The proposed Antenna C is observed to be radiating exactly in bore side direction, eliminating effect of position of feed and also has improved gain of 9.50 dB, while traditional Antenna A shows slightly shifted major lobe with less gain of 5.13 dB as demonstrated in Table 1.

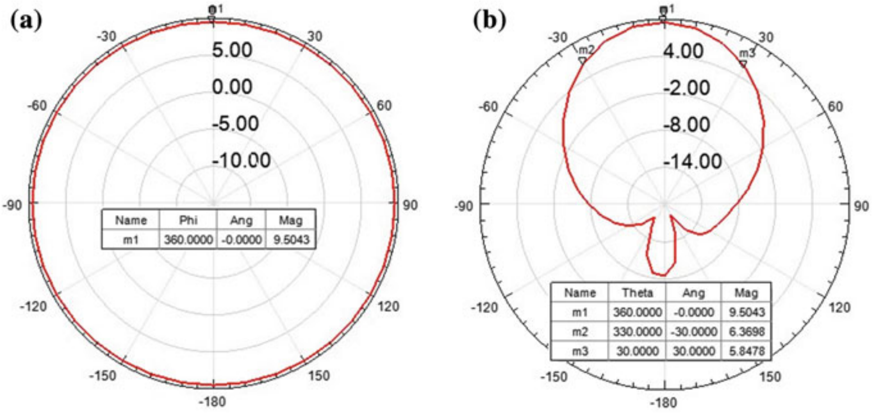


Fig. 12 Radiation pattern **a** horizontal plane; **b** vertical plane of Antenna B

Table 1 Comparative analysis of antennas

Antenna type	Number of layers	Array elements	Max. gain in dB
Antenna A	Single layer	Not present	5.13
Antenna B	Double layer	2 × 2 (second layer)	8.65
Antenna C	Triple layer	2 × 2 (second and third layers)	9.50

4 Conclusion

The proposed antenna types B and C demonstrate feasibility of designing array antenna without incorporation of physical feed network which eliminates complexity of design, saves precious space in antenna, avoids spurious radiations from feed network, and saves power consumed. Significant improvement of 3.52 dB in gain is observed in space-fed double-layered array antenna, while addition of third layer of array slightly improves gain by 0.85 dB. Hence, it can be concluded that adding further layer of array over top will be less helpful as far as the gain is concerned. Bandwidth of operation of antenna increases with increase in number of layers due to coupling between elements. Antenna operating frequency can be tuned by varying parameters like air gap between layers, size and spacing of array elements, etc., which provides more control as far as design of antenna is concerned. Air gap between consecutive layers is very useful in tuning antenna to desired frequency and hence can be used in adaptive radio applications. Gap between array elements is deciding factor for proper coupling of elements. Smaller gap gives better coupling and good impedance matching. This paper serves as a proof of concept for space-fed array antennas. Further investigations can be made by varying dielectric constant of substrate material, using more number of elements in array, using different configurations of array, or using wideband antenna as base antenna for array design, etc.

References

1. Falahati, A., NaghshvarianJahromi, M.: Wideband fan-beam low-sidelobe array antenna using grounded reflector for DECT, 3G, and ultra-wideband wireless applications. IEEE Trans. Antennas Propag. **61**(2) (2013)
2. Zhang, B., Zhang, Y.P.: Analysis and synthesis of millimeter-wave microstrip grid-array antennas. IEEE Antennas Propag. Mag. **53**(6) (2011)
3. Li, J., Wang, P., Jiang, T., Bao, L., He, Z.: UHF stacked hilbert antenna array for partial discharge detection. IEEE Trans. Antennas Propag. **61**(11) (2013)
4. Ghassemi, N., Wu, K., Claude, S., Zhang, X., Bornemann, J.: Low-cost and high-efficient W-band substrate integrated waveguide antenna array made of printed circuit board process. IEEE Trans. Antennas Propag. **60**(3) (2012)
5. Stoneback, M., Stoneback, M., Kuga, Y.: Feasibility study of a wirelessly controlled and powered space-fed phased array antenna. IEEE Trans. Antennas Propag. **61**(12) (2013)
6. Angeletti, P.: Multiple beams from planar arrays. IEEE Trans. Antennas Propag. **62** (2014)
7. Blanco, D., Llombart, N., Rajo-Iglesias, E.: On the use of leaky wave phased arrays for the reduction of the grating lobe level. IEEE Trans. Antennas Propag. **62** (2014)
8. Forsyth, W.E., Shiroma, W.A.: A retrodirective antenna array using a spatially fed local oscillator. IEEE Trans. Antennas Propag. **50**(5) (2002)
9. Kang, W., Kim, K., Kim, W.: Design of a 16-port shared-arm dipole array for monostatic imaging radar. IEEE Trans. Antennas Propag. **62**(1), 454–459 (2014)
10. Abbak, M., Tekin, I.: RFID coverage extension using microstrip-patch antenna array. IEEE Antennas Propag. Mag. **51**(1) (2009)
11. Peng, Z., Hu, T., Cui, W., Huangfu, J., Li, C., Ran, L.: Unconventional beamforming for quasi-hemispheric coverage of a phased array antenna. IEEE Antennas Wireless Propag. Lett. **12**, 1654–1657 (2013). AWPL-11-13-1588
12. Nashaat, D., Elsadek, H.A., Abdallah, E.A.: Ultrawide bandwidth 2×2 microstrip patch array antenna using electromagnetic band-gap structure (EBG). IEEE Trans. Antennas Propag. **59**(5) (2011)
13. Karmakar, N.C., Bialkowski, M.E.: Microstrip circular phased array design and development using microwave antenna CAD tools. IEEE Trans. Antennas Propag. **50**(7) (2002)
14. Seki, T., Honma, N.: A 60-GHz multilayer parasitic microstrip array antenna on LTCC substrate for system-on-package. IEEE Microwave Wireless Compon. Lett. **15**(5) (2005)
15. Rezk, M., Kim, W., Yun, Z., Iskander, M.F.: Performance comparison of a novel hybrid smart antenna system versus the fully adaptive and switched beam antenna arrays. IEEE Antennas Wireless Propag. Lett. **4** (2005)
16. Álvarez-Folgueiras, M., Rodríguez-González, J.A., Ares-Pena, F.: Low-sidelobe patterns from small, low-loss uniformly fed linear arrays illuminating parasitic dipoles. IEEE Trans. Antennas Propag. **57**(5) (2009)
17. Scott, H., Fusco, V.F.: 360° electronically controlled beam scan array. IEEE Trans. Antennas Propag. **52**(1) (2004)
18. Nikkhah, M.R., Rashed-Mohassel, J., Kishk, A.A.: Compact low-cost phased array of dielectric resonator antenna using parasitic elements and capacitor loading. IEEE Trans. Antennas Propag. **61**(4) (2013)

19. Zuo, S.-L., Zhang, Z.-Y., Yang, J.-W.: Planar meander monopole antenna with parasitic strips and sleeve feed for DVB-H/LTE/GSM850/900 operation in the mobile phone. IEEE Antennas Wireless Propag. Lett. **12** (2013)
20. Balanis, C.A.: Antenna Theory—Analysis and Design, 3rd edn. Wiley

Multi-verse Optimizer for Dynamic Stability Analysis Using STATCOM and Power System Stabilizer

P. K. Dhal

Abstract This paper investigates a novel nature-inspired algorithm called as multi-verse optimizer (MVO) used for solving optimization problem in power system 39-bus system. The main objective in this paper is the dynamic stability analysis using static synchronous (STATCOM) and power system stabilizer (PSS). The power system disturbance occurs mainly due to three-phase fault. The voltage profile and system stability are affected. It needs PSS controller to compensate and to improve the stability after the disturbances. The STATCOM location is vital and should be optimized to place certain point in the network for maximum benefit of the system. The different operating conditions are considered: without STATCOM, PSS, with STATCOM, PSS and STATCOM, and PSS tuned by MVO algorithm. It is analyzed by PSAT software. It shows that MVO gives novel optimization values in the systems.

Keywords Dynamic stability · STATCOM · Power system stabilizer · Multi-verse optimizer

1 Introduction

In recent days, the optimal power flow problem is very significance. The optimal power flow is the essential tool which permits that consumers are to identify the economic operation, i.e., security of the system. The OPF problem is one of must operations of electrical power system network [1, 2]. The OPF problem found out the optimal operational state in the electrical 39-bus system. It is minimizing the specific objective function with the limits of constraints, i.e., equality and inequality. Generally, the OPF problem may be determined as a nonlinear network condition of the system. So many researchers are used for the solution of the optimal power flow problems [3, 4]. The conventional methods are not able to solve in nonlinear problems, because of some limitation. This paper introduces multi-verse optimizer.

P. K. Dhal (✉)
Department of Electrical and Electronics Engineering, Vel Tech Rangarajan
Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India
e-mail: pradyumna.dhal@rediffmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_11

Table 1 Selection of parameter boundaries value

Boundaries	K_r	T_r	$K_{w,pss}$	T_{pss}
Lower boundaries	1	0.01	5	0.01
Upper boundaries	50	20	10	10

K_r Gain regulation of FACTS controller; T_r Time constant; $K_{w,pss}$ Gain value of PSS; T_{pss} Time constant

Multi-verse optimizer is applied for the solution of OPF. So the power flow can be made flexible using flexible AC transmission systems (FACTS). The FACTS device such as STATCOM helps to increase loadability of the system, reduction of overloaded condition, and losses reduced. The STATCOM device effectively tackled the voltage collapse problem and security of the system [5, 6]. The optimal location of STATCOM controller plays a vital role for system performance and economic benefits. The several approaches are proposed by many researchers, but dynamic stability is highlighted in this paper.

2 Significance Role of STATCOM and PSS

The role of STATCOM is controlled the voltage in the system. It is shunt-connected device. It contains of a synchronous voltage source. It supplies or receives the reactive power. Similarly, the role of PSS is that the damping oscillation is the fundamental character in achieving stability. The rotor angle oscillation of the generator is achieved through PSS, but the dynamic power stability of some system could not be reached in certain point due to high gain of fast exciters, inertia constant, etc.; in this case to neutralize the oscillation, the supporting controller, i.e., PSS, is used [7]. The gain of the PSS has been adjusted by multi-verse optimizer with STATCOM device (Table 1).

3 Basic Concept of Multi-verse Optimizer

Multi-verse optimizer (MVO) was proposed by Seyedali Mirjalili in 2015. The MVO consists of two theory concepts. There are (a) multi-verse theory and (b) big bang theory. The MVO concepts were developed and are performed like the exploration, exploitation, and local search. Similarly, big bang theory will have white hole, black hole, and worm hole. The many authors explained about MVO [8, 9]. The objects in the all universes may have random movement toward the best universe. It is always high possibility to move objects from a universe.

4 Proposed Flowchart-Based MVO Solving Optimal Power Flow

See Fig. 1.

The voltage magnitude indicates the 39-bus system as shown in Fig. 3. It is mentioned that the rotor angles of two generators, i.e., generator 1 and generator 2, are shown in Figs. 4 and 5, but the rotor angle stability without STATCOM is not improved due to severe fault, even the system contains power system stabilizer because it is not tuned properly. So the rotor angle goes to unstable conditions and voltage conditions of buses are 7, 8, 12, 15, 16, 20, 24, 32, and 33 affected in the systems compared to other buses. At this point, a STATCOM controller and PSS controller are required to improve the voltage values as well as rotor angle. It is shown in Fig. 3.

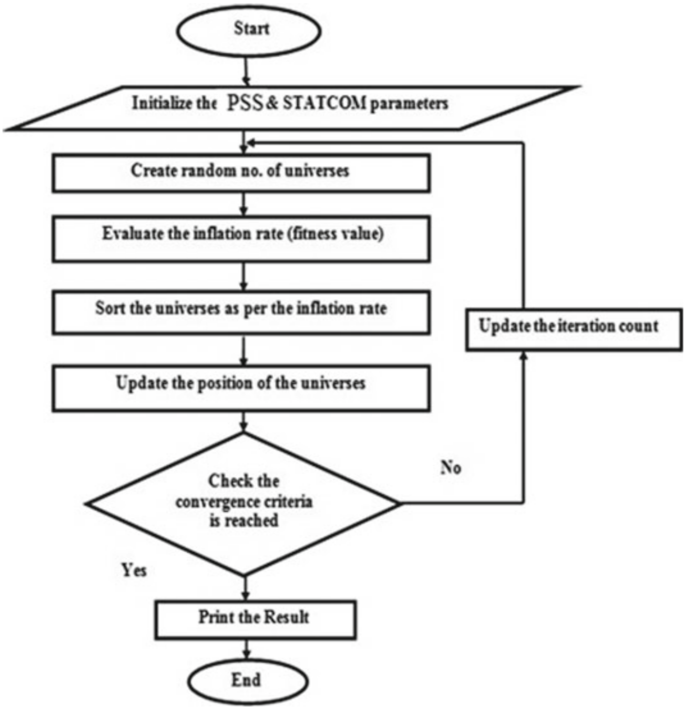


Fig. 1 Proposed flowchart using MVO

5 Results and Discussions

The multi-verse optimizer technique is studied to find the optimal power flow problem for standard IEEE 39-bus test system [10–12]. The problem solutions are done by PSAT software. IEEE 39-bus test systems have 10 numbers of generators and 10 loads considered to analyze the system. The author is assuming the fault duration is at 1.0 s and the clearing duration 1.10 s in 39-bus systems. This time is producing better results. The IEEE 39-bus system is shown in Fig. 2 (Table 2).

The remaining generators are also unstable conditions except generator 4, 5. It goes to infinity (Figs. 3, 4, 5, 6, 7, 8, 9, 10, and 11; Tables 3, 4, 5, and 6).

To exhibit the proposed MVO algorithm technique, its performance produces better result. Here, MVO algorithm is applied at bus 7, 8, 12, 15, 16, 20, 24, 32, and 33, because of low voltage due to three-phase fault without STATCOM, with STATCOM and PSS. When it tuned to STATCOM and PSS, it has improved to 0.9528, 0.9664, 1.0067, 1.0476, 1.0268, 0.9891, 1.0233, 0.9862, and 0.9994, respectively.

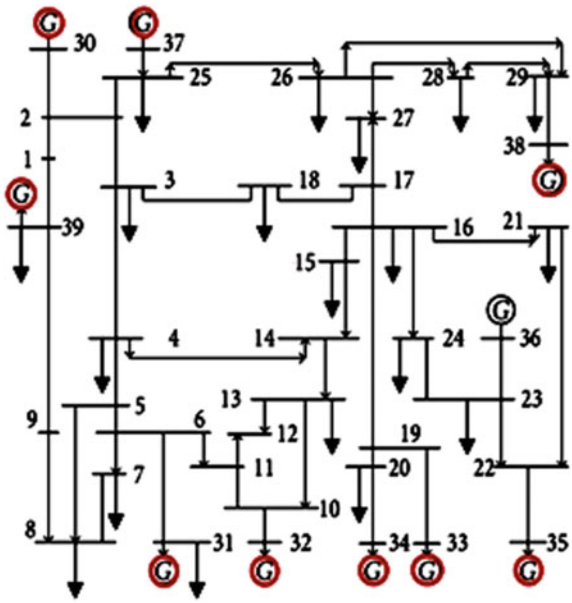


Fig. 2 Structure of IEEE 39 bus

Table 2 Different loaded condition of 39 buses

Different load at buses	Power (P) in per unit	Reactive power (Q) in per unit
03	3.220	0.024
04	5.000	0.845
07	2.338	8.400
08	5.220	1.760
12	0.085	0.880
15	3.294	1.530
16	3.294	3.230
18	1.580	0.300
20	6.800	1.030
21	2.740	1.150
23	2.475	0.846
24	3.086	0.922
25	2.240	0.472
26	1.390	0.170
27	2.810	0.755
28	2.060	0.276
29	2.835	1.269
31	0.092	0.046
39	11.04	2.500

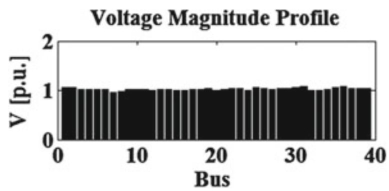


Fig. 3 Voltage per unit without controller in 39 systems

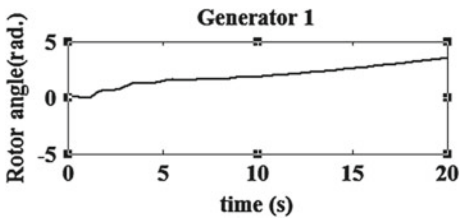


Fig. 4 Rotor angle in per unit value of generator 1 without controller in 39 system

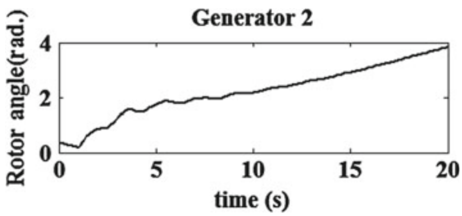


Fig. 5 Rotor angle in per unit of generator 2 without controller in 39 system

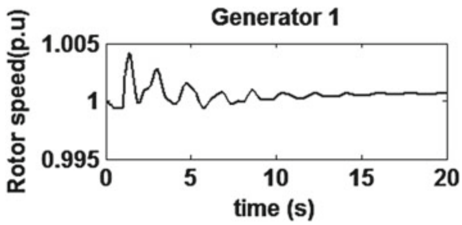


Fig. 6 Rotor speed per unit of generator 1 without controller in 39 system

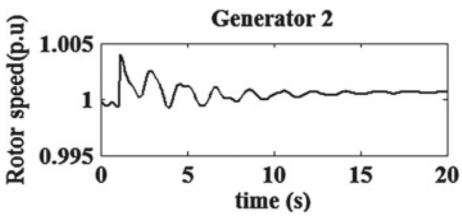


Fig. 7 Rotor speed per unit of generator 2 without controller in 39 system

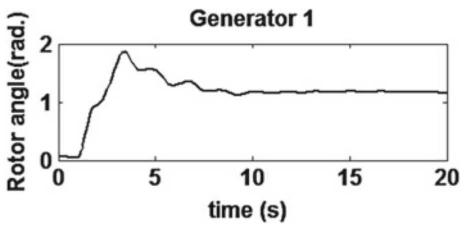


Fig. 8 Rotor angle per unit of generator 1 with controller tuned by MVO in 39 system

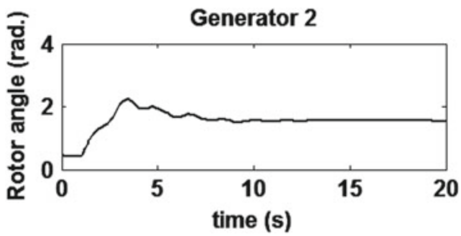


Fig. 9 Rotor angle per unit of generator 2 with controller tuned by MVO in 39 system

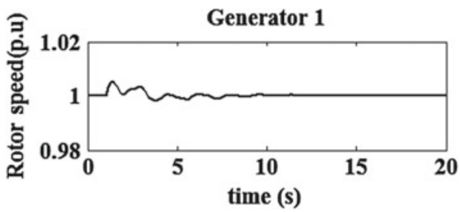


Fig. 10 Rotor speed per unit of generator 1 with controller tuned by MVO in 39 system

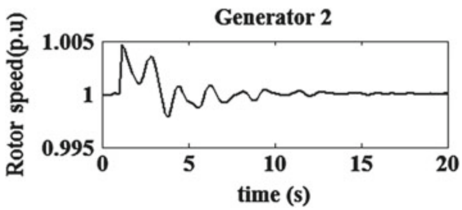


Fig. 11 Rotor speed per unit of generator 2 with controller tuned by MVO in 39-bus system

Table 3 System bus parameters without STATCOM

No. of buses	V (per unit)	Phase (rad.)	P-gen (per unit)	Q-gen (per unit)
01	1.0460	3.2729	0.0000	0.0000
02	1.0430	3.3208	0.0000	0.0000
03	1.0188	3.2712	0.0000	0.0000
04	1.0017	3.2579	0.0000	0.0000
05	1.0033	3.2799	0.0000	0.0000
06	1.0075	3.2927	−0.2460	−2.0320
07	0.9506	3.2550	0.0000	0.0000
08	0.9639	3.2441	0.0000	0.0000
09	1.0157	3.2434	0.0000	0.0000
10	1.0140	3.3341	0.0000	0.0000
11	1.0105	3.3200	0.0000	0.0000
12	0.9963	3.3196	0.0000	0.0000
13	1.0094	3.3215	0.0000	0.0000
14	1.0025	3.2918	0.0000	0.0000
15	0.9853	3.2851	0.0000	0.0000
16	0.9933	3.3119	0.0000	0.0000
17	1.0048	3.2923	0.0000	0.0000
18	1.0088	3.2763	0.0000	0.0000
19	1.0356	3.3951	0.0000	0.0000
20	0.9829	3.3696	0.0000	0.0000
21	1.0038	3.3564	0.0000	0.0000
22	1.0338	3.4372	0.0000	0.0000
23	1.0270	3.4338	0.0000	0.0000

(continued)

Table 3 (continued)

No. of buses	V (per unit)	Phase (rad.)	P-gen (per unit)	Q-gen (per unit)
24	0.9924	3.3144	0.0000	0.0000
25	1.0510	3.3455	0.0000	0.0000
26	1.0339	3.3247	0.0000	0.0000
27	1.0144	3.2889	0.0000	0.0000
28	1.0325	3.3883	0.0000	0.0000
29	1.0328	3.4381	0.0000	0.0000
30	1.0484	3.3634	2.5000	1.8848
31	1.0741	3.4182	5.9750	11.341
32	0.9845	3.4742	6.5000	2.4265
33	0.9972	3.4869	6.3200	2.0681
34	1.0120	3.4607	5.0800	2.1135
35	1.0491	3.5253	6.5000	3.2888
36	1.0629	3.5733	5.6000	1.7108
37	1.0282	3.4645	5.4000	0.3409
38	1.0266	3.5627	8.3000	1.3639
39	1.0314	3.2445	10.068	1.6359

Table 4 System parameters with STATCOM using MVO

No. of buses	V (per unit)	Phase (rad.)	P-gen (per unit)	Q-gen (per unit)
01	1.0485	0.9278	0.0000	0.0000
02	1.0494	0.9728	0.0000	0.0000
03	1.0331	0.9231	0.0000	0.0000
04	1.0158	0.9087	0.0000	0.0000
05	1.0069	0.9296	0.0000	0.0000
06	1.0092	0.9422	−0.1242	−0.9890
07	0.9528	0.9050	0.0000	0.0000
08	0.9664	0.8944	0.0000	0.0000
09	1.0168	0.8974	0.0000	0.0000
10	1.0233	0.9829	0.0000	0.0000
11	1.0174	0.9691	0.0000	0.0000
12	1.0067	0.9686	0.0000	0.0000
13	1.0229	0.9704	0.0000	0.0000
14	1.0264	0.9412	0.0000	0.0000
15	1.0476	0.9334	0.0000	5.3407
16	1.0268	0.9608	0.0000	0.0000
17	1.0303	0.9430	0.0000	0.0000
18	1.0301	0.9279	0.0000	0.0000

(continued)

Table 4 (continued)

No. of buses	V (per unit)	Phase (rad.)	P-gen (per unit)	Q-gen (per unit)
19	1.0474	1.0422	0.0000	0.0000
20	0.9891	1.0176	0.0000	0.0000
21	1.0273	1.0035	0.0000	0.0000
22	1.0460	1.0819	0.0000	0.0000
23	1.0398	1.0787	0.0000	0.0000
24	1.0233	0.9634	0.0000	0.0000
25	1.0567	0.9976	0.0000	0.0000
26	1.0460	0.9763	0.0000	0.0000
27	1.0329	0.9406	0.0000	0.0000
28	1.0386	1.0392	0.0000	0.0000
29	1.0370	1.0887	0.0000	0.0000
30	1.0476	1.0151	2.5000	1.4792
31	1.0436	1.0661	5.8532	7.5966
32	0.9842	1.1217	6.5000	2.0101
33	0.9964	1.1334	6.3200	1.1960
34	1.0112	1.1084	5.0800	1.7095
35	1.0483	1.1689	6.5000	2.3064
36	1.0624	1.2164	5.6000	1.1523
37	1.0276	1.1161	5.4000	0.0562
38	1.0263	1.2130	8.3000	1.0571
39	1.0313	0.9008	10.0534	1.5038

Table 5 Comparison of voltage profile without STATCOM, tuned by MVO with STATCOM and PSS

No. of buses	Without STATCOM (per unit)	With STATCOM & PSS-MVO (per unit)
01	1.0460	1.0485
02	1.0430	1.0494
03	1.0188	1.0331
04	1.0017	1.0158
05	1.0033	1.0069
06	1.0075	1.0092
07	0.9506	0.9528
08	0.9639	0.9664
09	1.0157	1.0168
10	1.0140	1.0233
11	1.0105	1.0174
12	0.9963	1.0067
13	1.0094	1.0229

(continued)

Table 5 (continued)

No. of buses	Without STATCOM (per unit)	With STATCOM & PSS-MVO (per unit)
14	1.0025	1.0264
15	0.9853	1.0476
16	0.9933	1.0268
17	1.0048	1.0303
18	1.0088	1.0301
19	1.0356	1.0474
20	0.9829	0.9891
21	1.0038	1.0273
22	1.0338	1.0460
23	1.0270	1.0398
24	0.9924	1.0233
25	1.0510	1.0567
26	1.0339	1.0460
27	1.0144	1.0329
28	1.0325	1.0386
29	1.0328	1.0370
30	1.0484	1.0476
31	1.0741	1.0436
32	0.9845	0.9862
33	0.9972	0.9994
34	1.0120	1.0112
35	1.0491	1.0483
36	1.0629	1.0624
37	1.0282	1.0276
38	1.0266	1.0263
39	1.0314	1.0313

Table 6 Rotor angle per unit value of generators in the 39-bus system

Generators	Without STATCOM (rad.)	With both STATCOM & PSS tuned by MVO (per unit)
1	3.50	1.15
2	3.83	1.54
3	4.26	1.95
4	4.30	2.02
5	4.39	2.09
6	4.20	1.90
7	4.36	2.05

(continued)

Table 6 (continued)

Generators	Without STATCOM (rad.)	With both STATCOM & PSS tuned by MVO (per unit)
8	4.38	2.07
9	4.47	2.14
10	3.41	1.07

6 Conclusion

In the novel work, the problem of optimal power flow is optimized for a IEEE 39-bus system using MVO algorithm. The MVO gives better results because the system is subjected to a three-phase fault. The proposed method has recovered voltage profile and rotor angle which confirm the effectiveness for recommended this algorithm. The future work can be carried out with hybrid algorithm.

References

1. Trived, I.N., et al.: Voltage stability enhancement and voltage minimization using multi-verse optimizer algorithm. In: 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)
2. Bhavani, T., Lavanya, K., Narasimharao, N.: Enhancement of dynamic stability and voltage control of a grid-connected OWF and MCF using a STATCOM. Int. J. Adv. Eng. Global Technol. **04**(01) (2016)
3. Mirjalili, S., Mirjalili, S.M., Hatamlou, A.: Multi-verse optimizer: a nature-inspired algorithm for global optimization. Nat. Comput. Appl. (2015)
4. Siddiqui, A.S., Deb, T.: Voltage stability improvement using STATCOM and SVC. Int. J. Comput. Appl. 88(14) (2014)
5. El-Sinnary, A.-R.I.H., Ibrahim El-Sayed, M., Mostafa, A.-E.-R.E.A.: Dynamic voltage stability of power systems. IJEDR **2**(2) (2014)
6. Khalkho, P., Singh, A.K.: Transient stability improvement by using PSS and increasing inertia of synchronous machine. Am. J. Electr. Power Energy Syst. **3**(2), 45–49 (2014)
7. Min, R., Xu, F., Yuan, F., Gao, Z.: Coordinated control of multi-FACTS to enhance the dynamic stability of the power system. Energy Power Eng. **5**, 1187–1191 (2013)
8. Falehi, A.D., Rostami, M., Mehrjardi, H.: Transient stability analysis of power system by coordinated PSS-AVR design based on PSO technique. Energy Power Eng. **3**, 478–484 (2011)
9. Farahani, S.S.S., Nikzad, M., Tabar, M.B., Yousefpour, B., Tourang, H.: Dynamic stability enhancement of power system using STATCOM. Indian J. Sci. Technol. **4**(11) (2011)
10. Dhal, P.K.: Improvement of stability by optimal location with tuning STATCOM using particle swarm optimization algorithm. Adv. Intell. Syst. Comput. **434**, 583–593 (2016)
11. Karthikeyan, K., Dhal, P.K.: Transient stability enhancement by optimal location and tuning of STATCOM using PSO'. Procedia Technol. **21**, 345–351 (2015)
12. Karthikeyan, K., Dhal, P.K.: Multiverse optimization (MVO) technique based voltage stability analysis through continuation power flow in IEEE 57 bus. Energy Procedia **117**, 583–591 (2017)

Performance Analysis of a Coordinated PID- and UPFC-Based Stabilizer for Stability Improvement

Ravi Prakash Mahobia, Shimpy Ralhan, Mahesh Singh and Ritu Sharma

Abstract In this paper, proportional–integral–derivative controller variables are tuned by the two algorithms, namely genetic algorithm and ant colony optimization, which are used to design the proposed power system having UPFC. These optimization algorithms are well-known and mostly used in finding the variables of PID controller. The computational methods improve the parameters of the proposed system, and attributes that are obtained are also stable in nature. System which has been scheduled with the GA-PID and ACO-PID controller has been modeled by application of MATLAB. The performance of ACO-PID is better than GA-PID, establishes the fact that ACO-PID system achieves faster response.

Keywords Genetic algorithm · Ant colony optimization · Proportional–integral–derivative · Unified power flow controller

1 Introduction

The current installed capacity of electricity generation in India is 334.40 GW as of the end January 2017. Nowadays, the continuous interconnection of regional electric grid is the recent trend of modern power system all over the world, such as interconnection of national grids of India, Europe network, the Japan power grids, the national grids of China, and North American power grids. Moreover, during the operation, in the event of fault or disturbances, an additional supporting power of each area of interconnected grids results in increased reliability of generation, transmission, and distribution

R. P. Mahobia (✉) · S. Ralhan · M. Singh · R. Sharma
Department of EEE, SSGI-SSTC, Bhilai 490020, India
e-mail: raviprakash93.eee@gmail.com

S. Ralhan
e-mail: shimpys@gmail.com

M. Singh
e-mail: singhs004@gmail.com

R. Sharma
e-mail: ritusharma10893@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_12

of power system. To make the power system intact from the disturbance, many schemes are used, which includes active and reactive power compensation, power factor correction, and phase angle correction. All these terms are subjected to the stability. The term stability can be defined as the ability to maintain the synchronism after subjected to the disturbance. It is also important, on the market point of view because in grid penalty is charged if there is no proper power supply, from one state to another state. So stability plays very important role in the power system, and many researches are going on the stability improvement. FACTS device is the result of one of them. Stability can be explained as the ability to manage its synchronism, when subjected to any disturbance and fault happens [1]. PID controllers are generally utilized in improving the system damping and also for small signal stability. In 1922, first time Minorsky introduced the classical PID controller. Most of the properties had already been represented by Elmer Sperry in 1911, because of its simple structure and robustness as well as operation; PID controllers have been using in the plants [2]. The gain of PID controller of UPFC will be tuned by process of optimization algorithm. Here, two algorithms have been utilized in tuning process: One is genetic algorithm, and another is ant colony optimization.

2 Problem Identification

2.1 PID Controller

In different industries, PID controllers are used in process control systems, which is a generic feedback controller [3, 4].

Tuning of parameters of PID is an important factor in its application as shown in Fig. 1 with gain parameters [5]. P represents the plant which has PID controller, whereas $e(t)$ can be defined as difference between desired input ($u(t)$ and output $y(t)$). The transfer function of the PID can be written as shown in (1).

$$u(t) = K_{\mathrm{p}}e(t) + K_i \int e(t) + K_{\mathrm{d}} \frac{\mathrm{d}}{\mathrm{d}x} e(t) \tag{1}$$

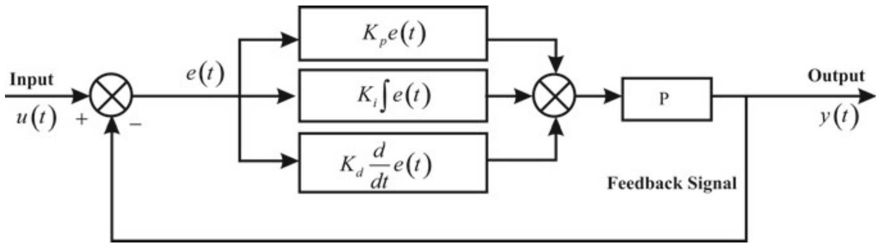


Fig. 1 Schematic of PID controller

Rise time can only be reduced by the proportional controller (K_p), but it does not cancel steady-state error. Steady-state error can be reduced by integral controller (K_i), but it generates adverse effect on transient response. Steady-state stability can be increased by derivative controller (K_d), and also maximum overshoot (M_p) is reduced.

The PID controller has three control parameters, K_p, K_i, K_d. When we increase the value of K_p, the peak overshoot of the system increases and the steady-state error decreases. That means as we increase the value of K_p more and more the system becomes unstable due to the increase in overshoot. On the other hand, when K_i increases, the overshoot becomes smaller, and the speed of response becomes slower, and when we increase the value of K_d, the overshoot is very small with a slow rise time, but the settling time is similar. In practical applications, the pure derivative action is never used as it produces a derivative kick due to which an undesirable noise is also produced.

2.2 Error Minimization Technique

Certain performance index has to be minimized by adjusting the PID gains of the system for optimizing the performance of the system, i.e., PID controlled as shown in (2). Here, we have implemented integral of time multiplied by absolute error (ITAE) criteria for minimizing the error as shown in (2).

$$J = \int_0^{\infty} t|e(t)|\mathrm{d}t \quad (2)$$

In the controller of UPFC, PID is used to damp out the oscillations, for which change in speed is taken as the input and added to UPFC controller input V_{ref}.

2.3 Unified Power Flow Controller

The unified power flow controller (UPFC) is the most adaptable device of the FACTS controllers [6]. The UPFC comprises of two branches, i.e., series and shunt branches, and both have its own voltage- or current-source converter. Series branch injects voltage to line, through the series arrangement of transformer. The other one is shunt branch which infuses current to the line through the shunt arrangement of transformer. The voltage and current can be fed to the line at any phase angle. Series branch injects active power to framework. Figure 2 shows the schematic diagram of UPFC. UPFC is not able to ingest or supply real power, except when UPFC is having power source in its DC terminals [7]. Active and reactive powers can be independently managed by the series and shunt transformers of UPFC, respectively [8]. UPFC is better than

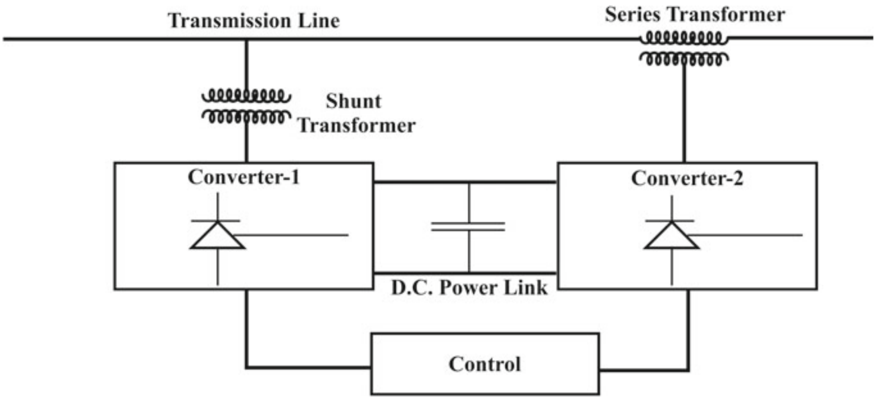


Fig. 2 Schematic diagram of UPFC

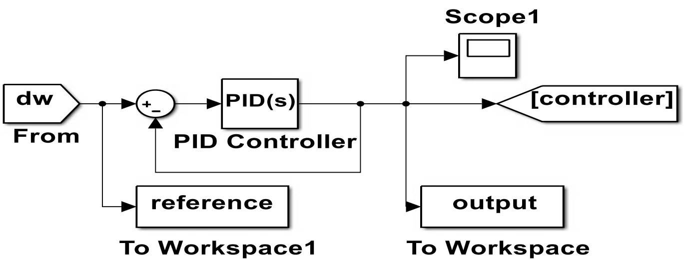


Fig. 3 PID controller with AC voltage regulator

STATCOM and SSSC [9]. It is conceivable to manage active and reactive power circulation freely at the input of UPFC control circuit with the regulation of the DC capacitor link voltage and varying both modulation index and phase angle of the input inverter. The main parameters of a power system that can change according to the requirement are line impedance (X_L), terminal voltage (Vt), and rotor angle (δ). Here, the effectiveness of UPFC is analyzed by analyzing damping of the oscillation or change in angular speed (dw) of the two-machine six-bus system. Figure 3 shows block diagram of the UPFC. The change in speed is fed to PID controller in UPFC which improves the stability of the system.

2.4 Single Generator Connected with Infinite Bus System

The power framework or system structure is of higher order and complex system with nonlinear component. Multi-machine system can be simplified by reducing it to the single generator connected to infinite bus system as shown in Fig. 4 [10].

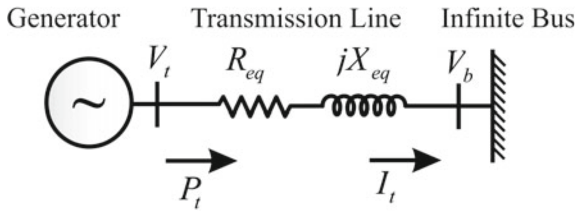


Fig. 4 Single generator connected with infinite bus system

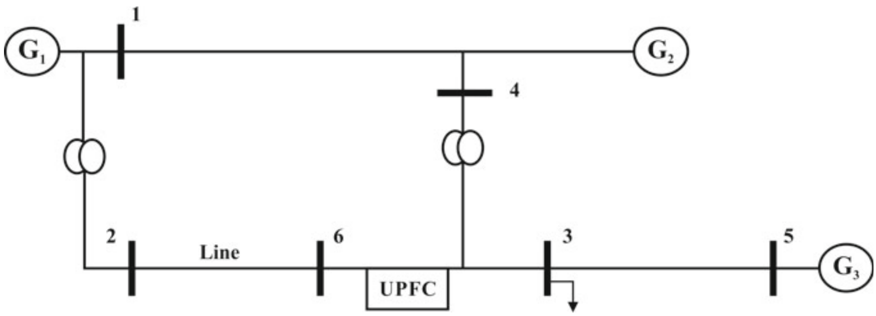


Fig. 5 Two-machine six-bus system

2.5 Two-Generator Six-Bus System

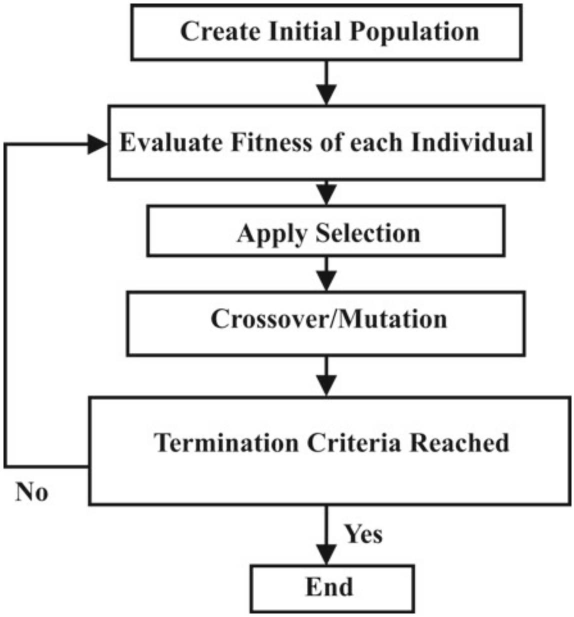
The power system under the test is shown in Fig. 5 which consists of two generators: G_1 and G_2 connected with six-bus system and G_3 is taken as grid.

3 Methodology

3.1 Genetic Algorithm

Genetic algorithm (GA) is a robust method utilized for solving optimization problems. According to idea of fittest function in year 1989, a probabilistic technique was proposed by Goldberg. It is the technique by which optimization and universal search are done and which is prepared by natural and evolution selection. Figure 6 shows flowchart of GA. GA is the technique in which optimization and universal search have been done by evolution and natural selection [11].

Fig. 6 Flowchart of GA



3.2 *Ant Colony Optimization*

Ant colony optimization is the optimization procedure for development of system that was presented inside the early 1990s.

It takes the inspiration from the real ant colonies, that is, foraging behavior. This phenomenon of ants has been taken in artificial ant colonies to search approximate solutions to distinct optimization issues (problem), and to continuous improvement issues (problem), and to special issues in the telecommunications, load equalization, and routing [12, 13]. Figure 7 shows flowchart of ant colony optimization.

4 Results

The results under fault condition with coordinated UPFC and PID controller are shown below. Here, the comparison is done on the basis of speed deviation graph plotted for conventional PID, GA-PID, and ACO-PID. Comparison is done on the basis of maximum overshoot and settling time [14].

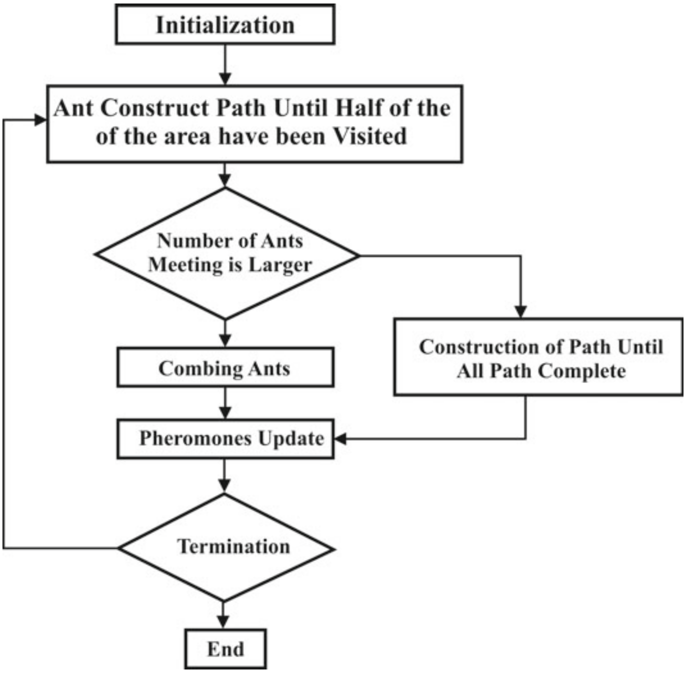


Fig. 7 Flowchart of ant colony optimization

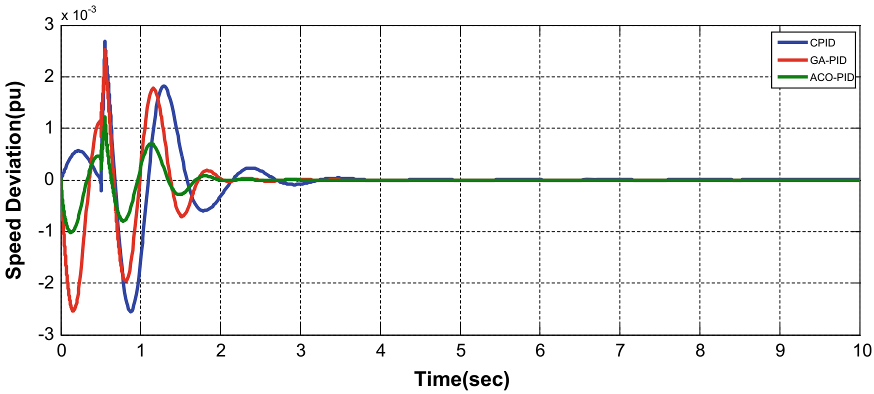


Fig. 8 Examination of deviation in speed SMIB system

4.1 Results for SMIB

Figure 8 shows the speed deviation in p. u. of SMIB under faulty with variation in time. According to the waveforms obtained, it is evident that system gets stable in

Table 1 Settling time and maximum overshoot of the speed deviation of SMIB

Parameter	CPID	GA-PID	ACO-PID
Settling time (s)	4	3	2
Maximum overshoot (p.u.)	2.8×10^{-3}	2.5×10^{-3}	1.3×10^{-3}

ACO-PID faster among the three optimization techniques. Speed change or deviation in ACO-PID attains to negligible point in scale and then zero in lesser. From Table 1, it is clear that ACO-PID performance is best from the CPID and GA-PID for SMIB.

4.2 *Results for Two-Machine System*

Speed deviation of machine 1 for the system considered under faulty condition is shown in Fig. 9. Here, also the system gets stable in ACO-PID faster as compared to other PIDs. Speed change in ACO-PID approaches to zero in lesser time.

From Table 2, it is clear that ACO-PID performance is best as compared to the CPID and GA-PID for machine 1 in the system considered.

Figure 10 shows the result of speed deviation of machine 2 under faulty condition. From Table 3, it is clear that ACO-PID performance is best from the CPID and GA-PID for machine 2.

Table 4 shows the tuned PID parameters after optimization using GA-PID, and ACO-PID and compared with CPID. In case of CPID, its value is default value, and below given PID parameters are for all conditions.

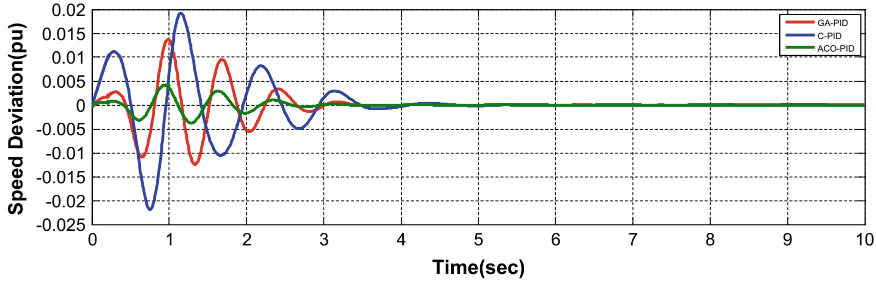


Fig. 9 Examination of deviation in speed of two-machine system

Table 2 Settling time and maximum overshoot of speed deviation of machine 1

Parameter	CPID	GA-PID	ACO-PID
Settling time (s)	5	4	3
Maximum overshoot (pu)	0.019	0.013	0.004

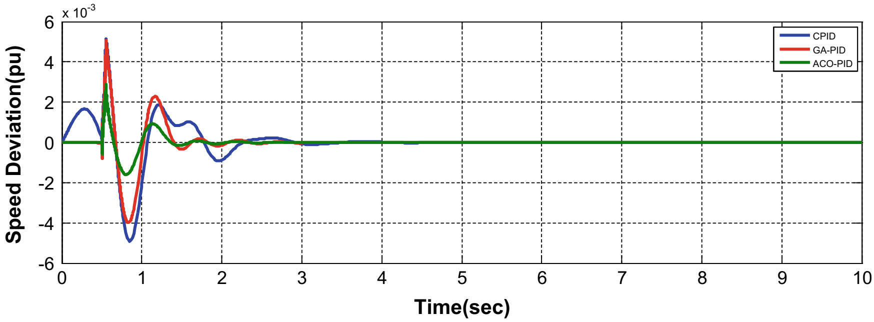


Fig. 10 Examination of deviation in speed of two-machine system

Table 3 Settling time and maximum overshoot of speed deviation of machine 2

Parameter	CPID	GA-PID	ACO-PID
Settling time (s)	4.5	3	2.5
Maximum overshoot (pu)	5.4×10^{-3}	5.1×10^{-3}	2.8×10^{-3}

Table 4 PID parameters coordinated with UPFC

Parameter	CPID	GA-PID	ACO-PID
K_{p}	1	0.8241	0.57749
K_{i}	1	1.8941	0.012997
K_{d}	0	0.0059	0.025986

5 Conclusion

The attribute of coordinated PID and UPFC is investigated for two-machine six-bus system as well as SMIB by two algorithms, and conventional PID (only UPFC without PID) is carried out. By the investigation of these three, it is clear that attributes of ACO-PID are best among CPID, GA-PID, and ACO-PID. From ACO-PID graph, and values of settling time and maximum overshoot, it is clear that system stabilizes faster and maximum overshoot is also minimized. The zero value of speed deviation reaches in lesser time in case of ACO-PID.

References

1. Kamdar, R., Kumar, M., Agnihotri, G.: Transient stability analysis and enhancement of IEEE-9 bus system. Electr. Comput. Eng. Int. J. (ECIJ) **3** (2014)
2. Mo, H., Yin, Y.: Research on PID tuning of servo system based on bacterial foraging algorithm. In: Seventh International Conference on Natural Computation, vol. 3, pp. 1758–1762 (2011)

3. Fan, X., Cao, J., Yang, H., Dong, X., Liu, C., Gong, Z., Wu, Q.: Optimization of PID parameters based on PSO. In: International Conference on Information Science and Cloud Computing Companion, pp. 393–397 (2013)
4. Jalilvand, A., Vahedi, H., Bayat, A.: Optimal tuning of PID controller for buck converter using BFO. In: International Conference on Intelligent and Advanced Systems, pp. 1–5 (2010)
5. Lakshmi, K., Sriniwas, V.: Optimal tuning of PID controller using PSO. In: International Conference on Electrical, Electronics, Signal, Communication and Optimization (EESCO), pp. 1–5 (2015)
6. Gyqyi, L., Rietman, T.R., Edris, A., Schaudar, C.D., Williams, S.L.: The unified power flow controller: a new approach to power transmission control. IEEE Trans. Power Deliv. **10**(2), 1085–1097 (1995)
7. Makombe, T., Jenkins, N.: Investigation of a unified power flow controller. In: IEE Proceedings-Generation Transmission and Distribution (1999)
8. Padiyar, K.R., Kulkarni, A.M.: Control design and simulation of unified power flow controller. IEEE Trans. Power Deliv. **13**, 1348–1354 (1998)
9. Amara, S., Hsan, H.A.: Power system stability improvement by FACTS devices: a comparison between STATCOM, SSSC and UPFC. In: International Conference on Renewable Energies and Vehicular Technology. IEEE (2012)
10. Singh, M., Patel, R.N., Jhapte, R.: Performance comparison of optimized controller tuning techniques for voltage stability. In: IEEE First International Conference on Control, Measurement and Instrumentation (CMI), pp. 11–15. IEEE (2016)
11. Booker, L.B., Goldberg, D.E., Holland, J.H.: Classifier systems and genetic algorithms. Computer Science and Engineering, 3116 EECS Building, The University of Michigan, Ann Arbor, MI 48109, U.S.A.
12. Blum, C.: Ant colony optimization: introduction and recent trends. Phys. Life Rev. **2**, 353–373 (2005)
13. Jaiswal, U., Aggarwal, S.: Ant colony optimization. Int. J. Sci. Eng. Res. **2**(7) (2011)
14. Singh, M., Patel, R.N., Neema, D.D.: Robust tuning of excitation controller for stability enhancement using multi-objective metaheuristic Firefly algorithm. Swarm Evol. Comput. **44**, 136–147 (2018). (Elsevier)

Design of Low Power Reduced Complexity Wallace Tree Multiplier Using Positive Feedback Adiabatic Logic

M. G. Ganavi and B. S. Premananda

Abstract The demand for low power devices is exponentially increasing with the increase in the use of portable equipment such as laptops and mobile phones. Adiabatic logic is an emerging technique which proves to be efficient in reducing the power dissipation of the system. Positive Feedback Adiabatic Logic (PFAL) is an efficient adiabatic logic. Multipliers are the fundamental arithmetic operators in the digital circuits. Addition is the integral part of multiplication, to add the partial products. A 16-bit Wallace Tree Multiplier (WTM) is implemented using carry-save addition. The Conventional WTM (CWTM) is modified to minimize complexity, termed as Reduced Complexity WTM (RCWTM) which has minimum number of Half Adders compared to the CWTM, which results in the reduction of the WTM area. The RCWTM is designed using both static CMOS logic and PFAL. The design is analyzed in Cadence Virtuoso 180 nm technology and simulated in Cadence Spectre. The PFAL based RCWTM dissipates 81.8% less power compared to static CMOS design.

Keywords Adiabatic · Carry-save adder · CMOS · PFAL · Wallace Tree Multiplier

1 Introduction

Power is the critical parameter in the design of computational and arithmetic circuits. Low power devices have importance in the field of VLSI, multimedia, and digital systems. Multiplication is the basic arithmetic operation in the design of systems such as FIR filter, ALU, etc.. The multiplication process involves the addition of partial products. The partial products can be added by using different adders. There are different types of multipliers such as Array multiplier, Booth multiplier, Wallace Tree Multiplier, Dadda multiplier, Sequential multiplier, and Vedic multiplier [1]. The Conventional WTM (CWTM) is different from that of regular multipliers such

M. G. Ganavi (✉) · B. S. Premananda
Department of Telecommunication Engineering, R.V. College of Engineering, Bengaluru, India
e-mail: ganavimg6@gmail.com

B. S. Premananda
e-mail: premanandabs@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_13

as array multiplier. The grouping mechanism used in the CWTM reduces the area occupied, delay, and also power dissipation. In CWTM, the three rows of partial products are grouped and Full Adder (FA) is used where the number of inputs is three and Half Adders (HAs) are used where the number of inputs is two. The FAs and HAs add the input bits using the carry-save addition process, which in turn implies that Carry-Save Adder (CSA) adds the partial products obtained by multiplication of the input bits. The CSA adds three bits at a time, outputs the sum in the present stage and the carry is considered as addend with the sum of next stage.

Bennet et al. [2] have designed an asynchronous CSA which is energy efficient. The CSA is designed using a double pass transistor with asynchronous adiabatic logic, which dissipates less power as compared to that of conventional CSA. The schematic of different adder architectures such as Ripple Carry Adder (RCA), Carry Lookahead Adder (CLA), CSA, Carry Select Adder (CSLA), and some of the parallel prefix adders such as Kogge Stone adder, Han Carlson adder, and Ling adder are compared with respect to the performance parameters such as area, power dissipation, and delay [3]. The results infer that CSLA has maximum power dissipation of 1.109 mW followed by CSA having 1.802 mW power dissipation and least power dissipation is 0.206 mW in RCA. The CSLA occupies more area with a gate count of 600 whereas RCA occupies less area with a gate count of 288. RCA is the slowest among all the other adders. CSA proves to be an efficient adder that can be used in the process of multiplication to reduce the delay and area of the circuit which in turn reduces the power dissipation in the circuit.

The low power techniques account for reducing the power dissipation in the circuits. To minimize the power dissipation in the circuits, different approaches such as sleepy stack, reversible logic, and Adiabatic Logic (AL) techniques exist, among which adiabatic logic proves to be an efficient technique [4]. Adiabatic in thermodynamics means that the heat is not dissipated to the environment. In static CMOS circuits, the energy is dissipated to the ground, unlike adiabatic systems in which the energy is recycled back to the power clock [5]. AL family is partitioned into two categories, namely, full and quasi/partial adiabatic. The charge in capacitance is completely recovered back to the power clock in case of full adiabatic circuits and in case of partial adiabatic circuits, only a part of energy is recovered back to the power clock [4]. The circuit complexity is more in case of full adiabatic circuit, which makes its implementation difficult though the power dissipation is less than that of the quasi adiabatic circuit. Quasi adiabatic techniques are used more as it is difficult to implement the full adiabatic circuits due to the increased complexity. PFAL is an efficient technique in minimizing the power dissipation than the other quasi adiabatic techniques.

A modified FA is designed using the multiplexers to minimize power dissipation in the WTM [6]. The WTM designed using modified FA has reduced the power dissipation by 37.45%, area by 45.75% and delay by 17.65% compared to the WTM designed using conventional FA. WTM is implemented using compressors to reduce the number of adders [7]. The 5:3 and 7:3 compressors are implemented using the property of both binary counters and compressor. The use of compressors has reduced the number of stages and the delay in the circuit. 7:3 compressor increases the speed by reducing the delay. The delay of WTM is 49.101 ns and that of the conventional multiplier is 51.090 ns, which indicates a reduction in the delay. WTM saves the area by employing the carry-save addition process. RCWTM incorporates less number of adders to add the partial products thereby reducing the area.

Alam et al. [4] have implemented a 4-bit comparator using ECRL and PFAL. The adiabatic circuits are compared with the static CMOS circuits with respect to power and area. Compared to the static CMOS circuits, ECRL and PFAL circuits dissipate 13.64% and 54.54% less power, respectively. The area occupied by ECRL comparator is less as compared to the PFAL comparator. ECRL technique is used to design conventional CSLA and CSLA using Binary to Excess-1 Converter (BEC), which proves to be efficient in minimizing the power dissipation [5]. CSLA using BEC dissipates less power than the conventional CSLA in CMOS design as BEC has less number of gates than RCA. The use of ECRL to design the adder further reduces the power dissipation in the adder. AL promises to be an efficient technique in minimizing the power dissipation in the CMOS circuits though it increases the complexity of the circuits.

Power is an important consideration in the design of digital and arithmetic circuits. PFAL proves to be an efficient adiabatic technique in reducing power dissipation. WTM occupies less area, thereby minimizing the power dissipation in the circuit compared to the other multipliers such as array multiplier. The delay in WTM is also less as it uses the concept of carry-save addition to add the partial products. CSA does not propagate the carry to the next stage instead it retains the carry and adds it as addend in the next stage thereby minimizing the delay due to carry propagation. Reduced complexity WTM (RCWTM) reduces the area and also the complexity of the multiplier. The proposed RCWTM in the paper provides better efficiency with respect to power in WTM. The designed RCWTM using PFAL is a promising approach in minimizing power dissipation which can be implemented in low power applications.

Organization of the paper is as follows: Sect. 2 provides the design and implementation of CSA, CWTM and RCWTM using static CMOS logic and focuses on the detailed description of AL technique, the working of adiabatic circuits and the design of circuits using AL in Sect. 3. Section 4 includes results and comparative analysis of CWTM and RCWTM with respect to area and power dissipation at different frequencies. Section 5 provides the conclusions and the future scope of the work.

2 Designs and Implementation of Wallace Tree Multiplier

2.1 Carry-Save Adder

Adders are the crucial elements in the digital circuits and are an integral part of multiplication process. There are various types of adders such as RCA, CLA, CSA, and CSLA [2]. The CSA is same as the FA except that the carry input in the FA is considered as one of the three inputs in the CSA. CSA provides the partial sum bits in one sequence and the corresponding carries bits in another sequence shifted by one bit to the left. The CSA considers three bits as input and provides two-bit output, i.e., sum and carry. CSA is used to add large number of operands which becomes more difficult in conventional adders. The CSA does not propagate carry from one stage to another stage thereby reducing the delay in the circuit; instead the carry is retained in present stage and added in the next stage [3]. In the example of carry-save addition the three inputs are added and the partial sum is considered as one sequence and carry out is retained as another sequence, instead of propagating the carry [6].

The final sum is formed by adding the partial sum and the carryout. A 4-bit CSA has six FAs and two HAs and an 8-bit CSA has fourteen FAs and two HAs, which makes it clear that whatever may be the number of FAs, the number of HAs remain two. The FAs used here follow the procedure of carry-save addition, which makes them appear as CSAs.

2.2 Wallace Tree Multiplier

A multiplier is an indispensable part of any arithmetic operation in systems such as Digital Signal Processors, microprocessors and FIR filters [1]. WTM is efficient with respect to area and speed. WTM is also power-efficient and has a high switching speed than the other multipliers. The process of multiplication involves three steps namely, generation of partial products, grouping of partial products and reduction and finally addition [8]. The partial products are formed by multiplying the inputs, three rows of partial products are considered as a single group as shown in Fig. 1. FA is used where there are three inputs and HA is used where there are two inputs. The FA and HA add the inputs by following the procedure of carry-save addition, i.e., retaining the partial sum in the present stage and passing the carry to the next stage as addend, by which the FA can also be called as a CSA in the addition process involved in WTM. The procedure of carry-save addition is repeated till the N $\times$ N partial products are summed. At the final stage a carry propagates adder is used to sum the bits.

2.3 Reduced Complexity Wallace Tree Multiplier

WTM uses both HAs and FAs for the addition of partial products using the concept of carry-save addition. A FA considers three inputs and provides two outputs thereby reducing the number of stages, in case of HAs the stage is not reduced, it takes two inputs and provides two outputs. From this it is evident that using the HAs is a waste of area, so minimizing the number of HAs in WTM makes it more efficient with respect to area [9].

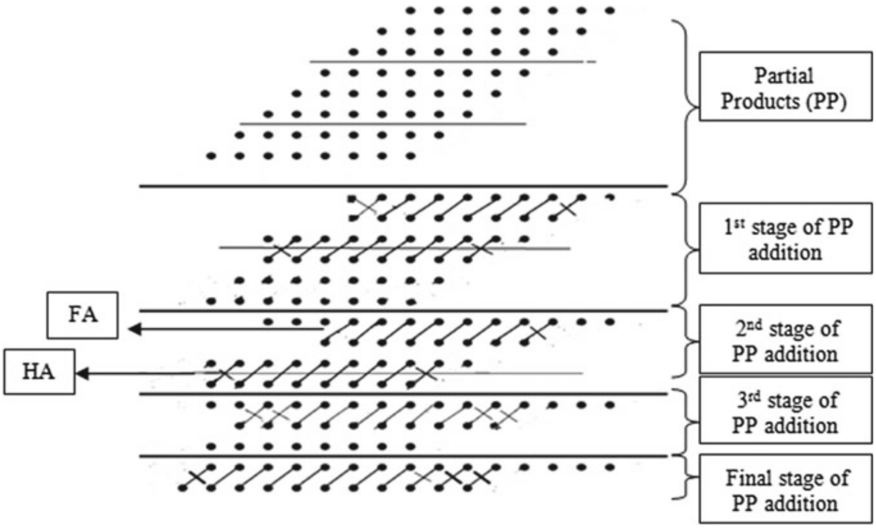


Fig. 1 8 × 8-bit CWTM [8]

The partial products form a pyramidal shape in the RCWTM, three bits in a single column are added using CSA and single bits are passed to the next row for addition [10]. The procedure is repeated till the last step is reached and carry propagate adder is used in the last step. The dot representation of RCWTM is represented in Fig. 2, which uses minimum number of HAs with a small increase in the number of FAs.

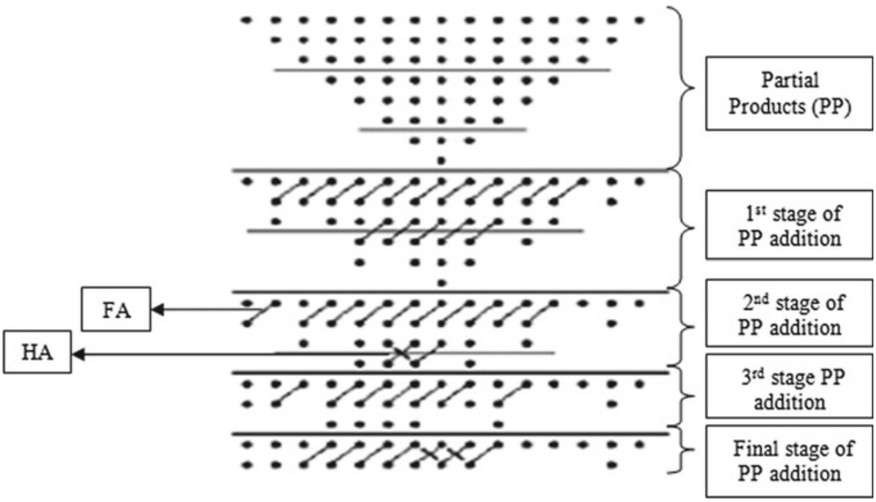


Fig. 2 8 × 8-bit reduced complexity WTM [6]

3 Adiabatic Logic

Adiabatic logic is an emerging and promising low power technique that efficiently reduces the power dissipation in the circuits but increases the complexity. PFAL reduces the power dissipation by a large amount as compared to other quasi adiabatic techniques. The switching power dissipation in static CMOS circuits with capacitive load has a lower limit of $CV_{\mathrm{dd}}^2/2$ in which V_{dd} and C represent supply voltage and capacitive load, respectively. The adiabatic charging/discharging is as depicted in [4]. The power supply used in the AL is a pulsed power supply to which the stored charge in the capacitor is recycled back and minimizes the power dissipation. There are four phases of power clocks, and each succeeding phase of power clock leads the preceding phase by 90° which is as shown in Fig. 3, designed and simulated in Cadence Virtuoso and Spectre simulator, respectively. Each phase of power clock has four intervals namely Evaluate (*E*), Hold (*H*), Recovery (*R*) and Wait (*W*). The output is evaluated from the stable input signal in the *E* interval. The output is provided to the next stage as input by keeping it stable in the *H* interval. The energy is recovered in the *R* interval and symmetry is provided by *W* interval [11]. The energy dissipated is supplied through a constant current source in an AL when considering the charge is provided by,

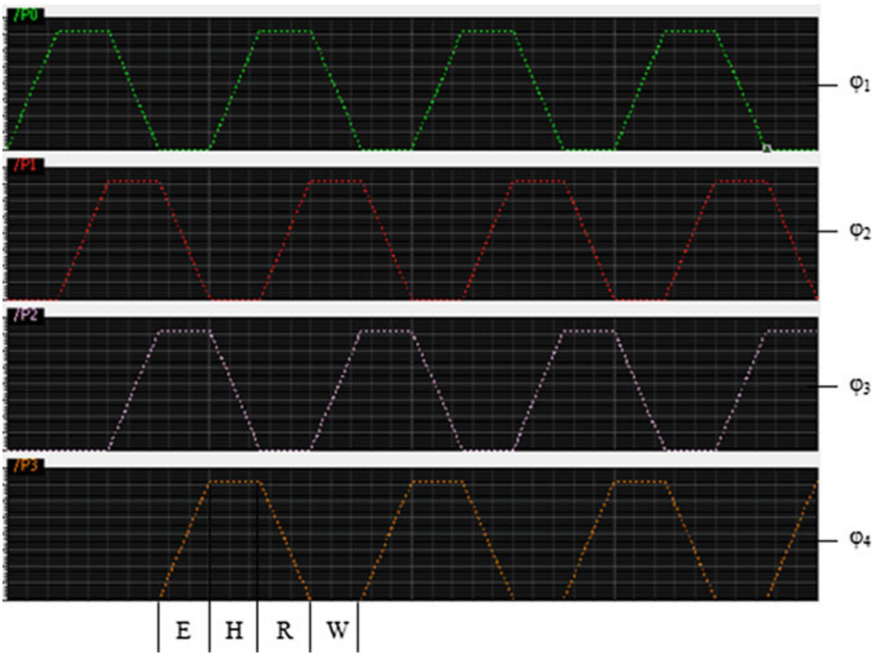


Fig. 3 Power clock for adiabatic logic

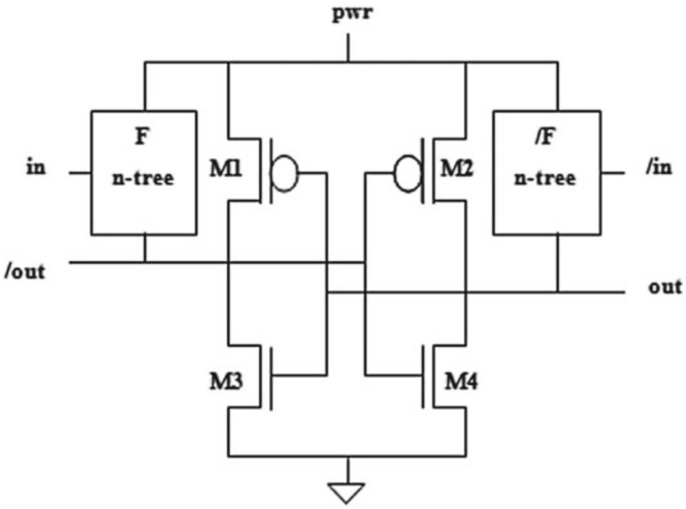


Fig. 4 Basic PFAL structure

$$E = \frac{RC}{T}V^2 \tag{1}$$

where C represents the load capacitance, V_{dd} is the power clock or power supply and T is charging or discharging time.

The adiabatic circuits have less energy dissipation than the static CMOS circuits if $T \gg 2RC$ [5]. There are different adiabatic techniques, among which PFAL proves to be efficient w.r.t. power dissipation. The basic structure of PFAL depicted in Fig. 4, consists of two PMOS and two NMOS circuits and feedback is provided to the functional blocks represented as F-tree and /F-tree [4]. A direct input and the inverted input are provided to circuit and both direct and inverted outputs are obtained. The power clocks are applied to the circuit in such a way that, each stage will have different phases of power clocks [12]. If φ_1 is provided to the first stage then φ_2 is provided to the next stage, φ_2 leads φ_1 by 90°. After applying all the four phases of power clocks to the circuit, the phases are repeated to the succeeding stages. In adiabatic circuits, the output of one stage is provided as input to the next stage and not the same stage and each stage has different phases of power clocks. To propagate the output from one stage to another, buffers are used wherever necessary.

4 Results and Discussions

The WTM and RCWTM are designed and simulated in Cadence Virtuoso 180 nm technology and spectre simulator, respectively. The CWTM is designed in Cadence Virtuoso using the concept of carry-save addition as shown in Fig. 5. The partial

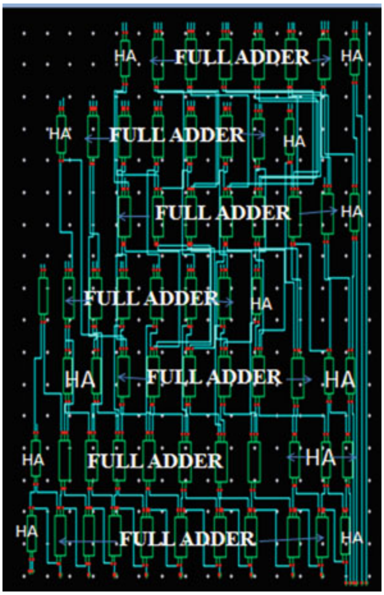


Fig. 5 Snapshot of CMOS based 8 × 8-bit CWTM

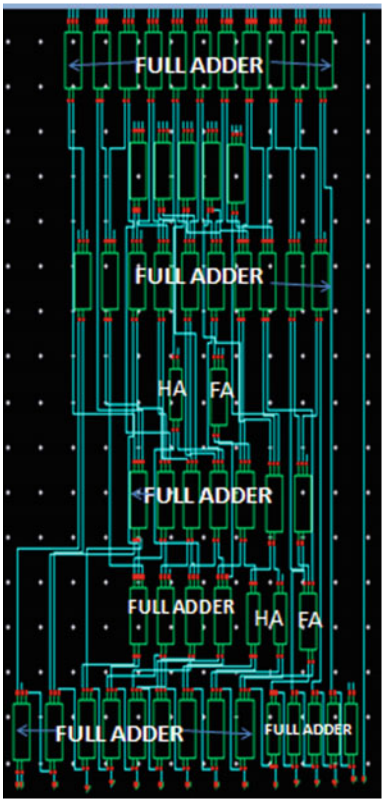
products are formed by multiplying the input bits which is performed by using the AND gates. The partial products are then added by using the FA and HA, which is nothing but the CSA and in the final stage RCA is implemented. The RCWTM is depicted in Fig. 6, which uses minimum number of HAs with a small increase in the number of FAs.

The RCWTM is implemented using PFAL is as shown in Fig. 7, where the FAs and HAs are labeled and the unlabeled circuits are the buffers. AND gates are used to multiply the input bits. The FA, HA, and AND gates are designed using PFAL. The logic circuit operates at 0 V and 1.8 V, which are the lowest and highest voltages, respectively. The inputs provided to the circuit are pulse varying between 0 and 1.8 V. The RCWTM is implemented using PFAL. Different power clocks are provided at each stage and the output of one stage is provided as input to the next stage.

The RCWTM is verified by providing various combinations of inputs. The RCWTM is implemented using PFAL. Different phases of power clocks are provided at each stage and the output of one stage is provided as input to the next stage. The RCWTM is verified by providing various combinations of inputs. Figure 8 illustrates the output waveforms of the RCWTM implemented in PFAL logic for the inputs A = "11001110" and B = "11010011", output of RCWTM is "1010100111001010". The PFAL based RCWTM output is displayed in the 5th phase due to the use of different phases of power clocks in each stage and in case of static CMOS circuit, the output is displayed in the 1st phase as a single power supply is used for the entire circuit.

Table 1 provides a comparison of power dissipation of adders designed using static CMOS and PFAL. The PFAL based HA and FA dissipates 99.99% less power

Fig. 6 Snapshot of CMOS based 8 × 8-bit RCWTM



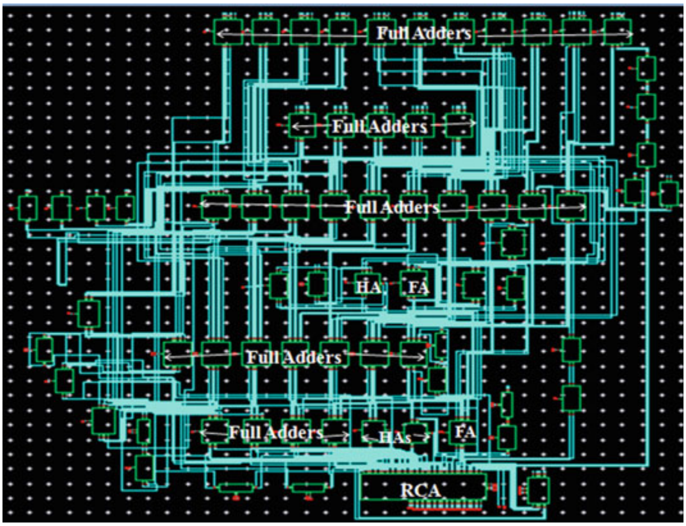


Fig. 7 Snapshot of PFAL based 8 × 8-bit RCWTM

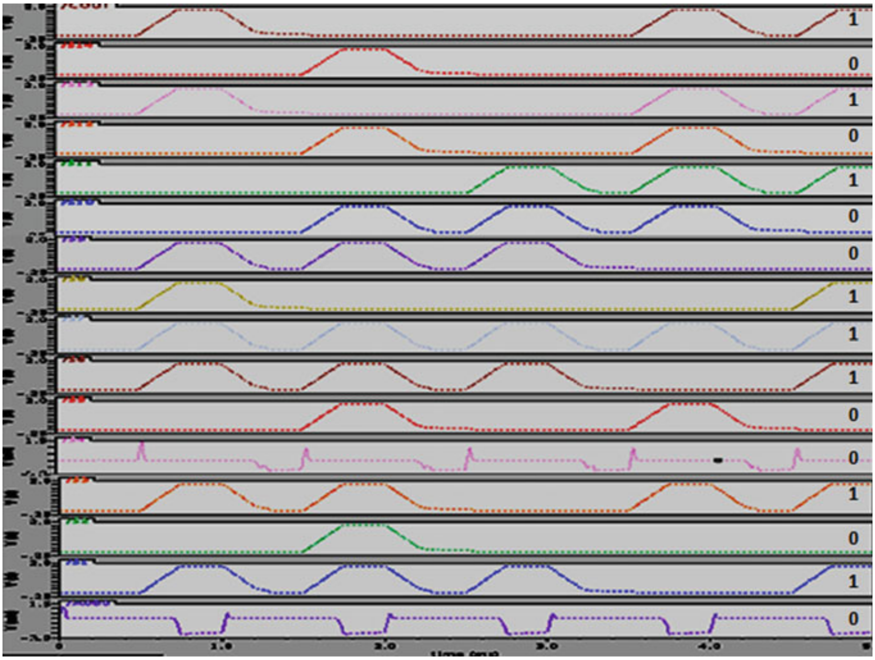


Fig. 8 Snapshot of output waveform of PFAL based RCWTM

Table 1 Power analysis of circuits at 1 kHz

Adders	Power dissipation in CMOS (W)	Power Dissipation in PFAL (W)
Half Adder	1.39 μ	0.091 n
Full Adder	2.2 μ	0.124 n

than the CMOS based design. Table 2 lists the comparison between CWTM and RCWTM w.r.t. power dissipation and area. The power dissipation in both CWTM and RCWTM is almost same, but RCWTM occupies less area by the use of less number of gates. The number of HAs in the RCWTM has reduced by 52.94% and the number of FAs has increased only by 2%. This proves that RCWTM is efficient than the CWTM in terms of area. The PFAL based circuits dissipate less power as

Table 2 Power and Area Analysis of CWTM and RCWTM at 1 kHz

Static CMOS based multipliers	Power dissipation (W)	Number of gates
8 × 8 CWTM	22.58 μ	64 (47 FAs and 17 HAs)
8 × 8 RCWTM	22.60 μ	56 (48 FAs and 8 HAs)

Table 3 Power Analysis of RCWTM at various frequencies

Frequency (Hz)	Power dissipation (W) of 8 × 8 RCWTM
1 K	22.62 μ
10 K	23.22 μ
100 K	29.60 μ
1 M	87.3 μ
50 M	667.6 μ
100 M	1.854 m
250 M	5.591 m

Table 4 Power Analysis of WTM of the Previous and Proposed Work at 250 MHz

Multiplier architectures	Power dissipation (mW)
8 × 8 WTM [13]	9.64
Proposed 8 × 8 RCWTM	5.591

Table 5 Power Analysis of CMOS and PFAL based RCWTM

Multiplier	Power dissipation in CMOS (W)	Power dissipation in PFAL (W)
8 × 8 RCWTM	22.62 μ	4.11 μ

compared to the static CMOS based circuits, which proves PFAL to be an efficient technique with respect to power dissipation.

Table 3 lists the power dissipation of CWTM and RCWTM designed using static CMOS logic at different frequencies. It is inferred from Table 3 that with the increase in frequency, power dissipation also increases. The power dissipation is almost the same for CWTM and RCWTM at all frequencies. Table 4 provides the comparison of power dissipation of the WTM in [13] and the proposed RCWTM. It is evident from Table 4 that the power dissipation is 42% less using the proposed RCWTM as compared to the WTM in [13]. The RCWTM implemented using PFAL is a low power approach which minimizes the power dissipation as compared to CMOS design. Table 5 lists power dissipation of CMOS and PFAL based RCWTM. The power dissipation of CMOS based RCWTM is 22.62 μW and PFAL is 4.11 μW. The use of PFAL in the RCWTM circuit minimizes the power dissipation by 81.8%.

5 Conclusions and Future Scope

Power is a crucial criterion in the design of circuits, though area and delay are equally important. Adiabatic circuits prove to be efficient in minimizing the power dissipation in the circuits. Multipliers are an integral part of the digital systems. WTM is efficient with respect to power, area, and delay. The RCWTM reduces the area of the circuit compared to CWTM by reducing the number of gates by 12.5% though the power dissipation in both the circuits is almost same. The CWTM and RCWTM are simulated at different frequencies, and it is evident that the power dissipation increases with the increase in frequency. The power dissipation of WTM in [13] has been minimized in the proposed work. The RCWTM realized in static CMOS logic dissipates 42% less power as compared to [13]. The RCWTM designed using PFAL is a new approach which dissipates 81.8% less power than the CMOS design.

In the future, the multiplier can be implemented using other adders instead of conventional FA and HA. The designed multiplier can be used in DSP applications. The designed adder and multiplier along with the shifter can be used in the implementation of FIR filters.

References

1. Biradar, V.B., Vishwas, P.G., Chetan, C.S., Premananda, B.S.: Design and performance analysis of modified unsigned braun and signed Baugh-Wooley multiplier. In: Proceedings of the International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques, pp. 1–6 (2017)
2. Bennet, B., Maflin, S.: Modified energy efficient carry save adder. In: Proceedings of the International Conference on Circuits, Power and Computing Technologies, pp. 1–4 (2015)
3. Kaur, J., Sood, L.: Comparison between various types of adder topologies. Int. J. Comput. Sci. Technol. **6**, 62–66 (2015)
4. Alam, S., Ghimiray, S.R., Kumar, M.: Performance analysis of a 4-bit comparator circuit using different adiabatic logics. In: Innovations in Power and Advanced Computing Technologies, pp. 1–5 (2017)
5. Premananda, B.S., Chandana, M.K., Shree Lakshmi, K.P., Keerthi, A.M.: Design of low power 8-bit carry select adder using adiabatic logic. In: Proceedings of the International Conference on Communication and Signal Processing, pp. 1764–1768 (2017)
6. Chinnapparaj, S., Somasundareswari, D.: Incorporation of reduced FA and HA into WTM and improved CSA for digital FIR filter. Int. J. Circuits Syst. **7**, 2467–2475 (2016)
7. Mandloi, A., Agrawal, S., Sharma, S., Shrivastava, S.: High-speed, area efficient VLSI architecture of Wallace-Tree Multiplier for DSP-applications. In: Proceedings of the International Conference on Information, Communication, Instrumentation and Control, pp. 1–5 (2017)
8. Jaiswal, K.B., Nithish Kumar, V., Seshadri, P., Laksminaryan, G.: Low power Wallace Tree Multiplier using modified full adder. In: Proceedings of the 3rd International Conference on Signal Processing, Communication and Networking, pp. 1–4 (2015)
9. Vijay, A.J., Krishnamurthy, M.: Design of area and power aware reduced Wallace Tree Multiplier. Int. J. Adv. Res. Electron. Commun. Eng. **6**, 1241–1244 (2017)
10. Waters, R.S., Swatzlander, E.E.: A reduced complexity Wallace multiplier reduction. IEEE Trans. Comput. **59**, 1134–1137 (2010)
11. Bhati, P., Rizvi, N.: Adiabatic logic: an alternative approach to low power application circuits. In: Proceedings of the IEEE International Conference on Electrical, Electronics, and Optimization Techniques, pp. 4255–4260 (2016)
12. Kumar, S.D., Thapliyal, H., Mohammad, A.: FinSAL: A Novel FinFET based secure adiabatic logic for energy-efficient and DPA resistant IoT devices. In: Proceedings of the IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 37, pp. 110–122 (2018)
13. Kumawat, P.K., Sujediya, G.: Design and comparison of 8 × 8 Wallace Tree Multiplier using CMOS and GDI technology. IOSR J. VLSI Sig. Process. **7**, 2319–4197 (2017)

Low-Power 8-Bit Adiabatic Barrel Shifter for DSP Applications

Nagesh Nazare, B. S. Premananda, Pradeep S. Bhat and R. J. Nayana

Abstract The CMOS technology is recognized for high levels of integration and low power dissipation. Portable devices usually operate on batteries with minimal charge storage capacity. Power consumption is the limiting factor for the functionality offered by portable devices that operate on batteries. This has made researchers explore for new techniques to recover the energy from the circuit. One of the efficient ways to reduce the power consumption of the circuits is to design the circuit using adiabatic logic. Shifters are one of the basic elements of digital signal processors and are used for manipulating the data. Shifters are used to efficiently perform division and multiplication operation on unsigned integers by powers of two. There exist two types of shifters, viz. combinational and sequential Shifters. The paper focuses on the design of power efficient 8-bit combinational shifter, namely barrel shifter using positive feedback adiabatic logic (PFAL). Also, a comparison is performed with the conventional static CMOS barrel shifter using multiplexer logic and pass transistor logic. Cadence Virtuoso is used for the implementation with CMOS 180 nm technology, and Cadence Spectre simulator is used for simulation and functional verification. The analysis of the logic circuits infers that PFAL-based barrel shifter consumes the least power. The power values infers that PFAL technique is advantageous when applied to low-power digital devices operated at relatively low frequencies.

Keywords Adiabatic logic · Barrel shifter · DSP · Spectre · PFAL

N. Nazare (✉) · B. S. Premananda · P. S. Bhat · R. J. Nayana
R.V. College of Engineering, Bengaluru, India
e-mail: nagesh.nazare0831@gmail.com

B. S. Premananda
e-mail: premanandabs@gmail.com

P. S. Bhat
e-mail: psbhat1996@gmail.com

R. J. Nayana
e-mail: nayanarjambhe@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_14

1 Introduction

From decades, concentration has been towards designing IC with more speed. But today, as more and more devices are becoming portable, low-power design techniques have become very vital. Some of the common goals of a low-power IC designer are to reduce the supply voltage, circuit complexity and frequency of operation. The above-mentioned parameters can reduce the power consumption to some extent but are usually limited by the threshold voltage and noise margin requirements. However, the new research is towards recycling of energy from the circuit, back to the source [1]. One of the promising techniques to minimize the power is adiabatic logic technique, which can be used in the design of digital, as well as analog circuits.

Adiabatic circuits have two types of losses, viz. adiabatic losses and non-adiabatic losses. Non-adiabatic losses are independent of frequency of operation, and hence cannot be reduced. Adiabatic losses depend on the operating frequency and by using efficient design techniques can be minimized to an extent. Adiabatic logic techniques are of two kinds, viz. partial/quasi-adiabatic circuits (QAC) and fully adiabatic circuits (FAC) [2]. Quasi-adiabatic logic circuits have non-adiabatic losses when there is a flow of current through the non-ideal switches. Non-adiabatic losses are significantly constrained in case of FACs. There are several adiabatic logic families which have their peculiar advantages and disadvantages [3].

The power analysis between conventional CMOS and adiabatic PFAL digital circuits infers that the power dissipation is greatly reduced due to recycling process in PFAL. Efficient charge recovery logic (ECRL) is also one of the efficient adiabatic design techniques [4]. From the plot shown in Fig. 1 [5], it can be noted that the adiabatic energy dissipation increases with increase in the frequency. Non-adiabatic energy dissipation is independent of frequency and losses due to leakage dominate at low frequencies. Hence, it is very critical to find the optimum operating frequency for adiabatic circuits.

The energy dissipation in case of adiabatic circuits is given by Eq. (1), where R is the resistance of the circuit consisting of the on resistance of transistors in the charging path, T represents the charging and discharging time, C_L represents the load capacitance and V represents the voltage of the supply. Theoretically, the energy

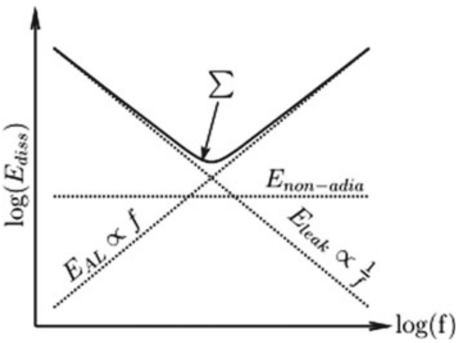


Fig. 1 Energy losses versus frequency plot [5]

dissipated can be limited by increasing the switching time T [2].

$$E_{\mathrm{diss}} = \left(\frac{RC_{\mathrm{L}}}{T}\right) C_{\mathrm{L}} V^2 \tag{1}$$

Shifting the binary data is one of the basic requirements in many applications that include arithmetic operations, bit-indexing and variable-length coding. Digital signal processors and embedded processors use barrel shifters to shift or rotate data in a single clock cycle [6]. The computer operation requires shifting of data, for address generation to arithmetic functions. There are two kinds of shifters, viz. sequential and combinational shifter. Sequential circuits have a problem of excess area consumption for registers [7]. An accurate clock is also required by the registers. Routing the clock path and meeting the timing requirements is another issue faced by the sequential shifters. Hence, combinational shifters are preferred where area and power are critical in circuits.

In this paper, a combinational (barrel) shifter using multiplexer, which has the characteristic of 'N' data inputs and 'N' data outputs with '$\log_2 N$' control inputs, is designed using both static CMOS and PFAL technique. PFAL is one of the quasi-adiabatic logic families which satisfy our requirement. It is robust and compatible to static CMOS design flow. PFAL has four-phase power clock which can be generated efficiently. The design is implemented in Cadence Virtuoso Design Environment with 180 nm CMOS technology with Spectre simulator. The design parameters are as listed in Table 1.

Width of the pMOS transistors is chosen to be 2.75 times that of nMOS transistor which was derived based on DC analysis of a CMOS inverter circuit, such that the inverter has equal rise and fall time. pMOS is slower than nMOS, and hence to meet the timing requirements, width of pMOS is increased, in turn matching with the device resistance of nMOS. Length of both the transistors is kept same.

The paper is organized as follows: Sect. 2 describes the design and implementation of adiabatic multiplexer, and barrel shifter using conventional and adiabatic logic. Further, left-shift barrel shifter using pass transistor logic is also described. In Sect. 3, results have been discussed along with the waveforms obtained during simulation. The conclusions derived and future scope for the work have been provided in Sect. 4.

Table 1 Design parameters

Specification	Value
Technology	180 nm
Power-supply	1.8 V
L_{PMOS}	180 nm
W_{PMOS}	2.75 µm
L_{NMOS}	180 nm
W_{NMOS}	1 µm

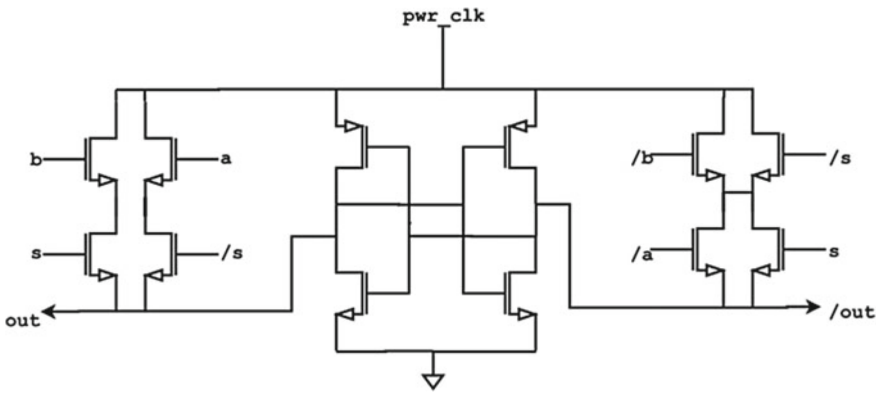


Fig. 2 Circuit of 2:1 PFAL multiplexer

2 Design and Implementation

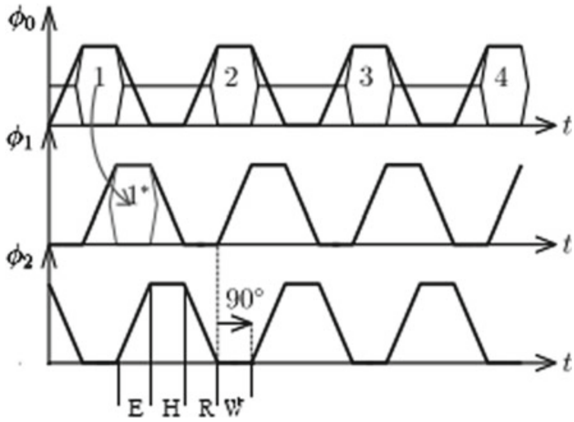
2.1 PFAL Multiplexer

PFAL is a dual-rail circuit classified under quasi-adiabatic circuits. It has a property of providing decent robustness against the technical variations. It consists of dual nMOS trees containing complementary logic functions. In PFAL, a latch is formed consisting of two inverters connected back-to-back which avoid the logic-level degradation at the two complementary output nodes [8]. A total of ten nMOS and two pMOS transistors are required to design a 2:1 PFAL multiplexer (MUX) as shown in Fig. 2. Basic components such as NOT, AND and OR gates required for designing larger adiabatic systems such as adders and multipliers are illustrated in [9]. Four-phase power-clock scheme is used to drive the circuit. There are four phases in every power-clock cycle, as shown in Fig. 3. The outputs are assessed from the steady input signals in the evaluate (E) phase. Outputs during the hold (H) phase are kept constant to provide the succeeding gate with a steady input signal. Power is recycled back in the recover (R) phase. As generation of symmetric signals is more efficient, a wait (W) phase is introduced [5].

2.2 Barrel Shifter

Barrel shifter is one of the functional parts of a microprocessor CPU. The control inputs are used to specify the number of shifts in either direction. Circular shift, arithmetic shift and logical shift are three possible operations. Barrel shifters are used in the digital signal processors to manipulate the data. An array of MUX is used to implement the barrel shifter [10]. In this implementation, each MUX output is

Fig. 3 Four-phase power clock scheme



connected to the input of the succeeding MUX. The control lines of left-shift barrel shifter run vertically, whereas processed input data runs diagonally upward till the last array, where the output is obtained [11]. The shifting of the data in the barrel shifter can be as per the user requirement, which can be controlled using the control lines. The wide usage of it can be attributed to its execution of multi-bit shifts in a single operation.

a. **MUX-based Static CMOS Barrel shifter**

The 'M' most significant bits are removed from the MSB region and the remaining (N-M) bits can be shifted to the left 'M' places. The void created in the LSB region is then filled with zeros. Left-shift barrel shifter is a command corresponding to a multiplication by 2^M. Left-shift barrel shifter using MUX is implemented in conventional CMOS technique, as shown in Fig. 4. A_0 A_1, …, A_6 A_7 are the input bits, where A_7 represents the MSB bit and A_0 represents the LSB bit. Sh_2 Sh_1 Sh_0 are the control inputs. The outputs are represented by OUT_0 OUT_1, …, OUT_6 OUT_7. The first stage of MUX can shift the data to maximum of 1-bit, whereas the second and third stages can shift data by 3-bit and 7-bit, respectively.

b. **MUX-based Adiabatic Barrel shifter**

Adiabatic left-shift barrel shifter implemented using PFAL is shown in Fig. 5. Cascaded adiabatic logic circuits have inherent micropipelining scheme. Four-phase power-clock ϕ_1, ϕ_2, ϕ_3 with 90° phase shifts are used to implement the barrel shifter circuit [12].

The binary data is transferred from one stage to other in a pipeline fashion. Table 2 shows the operation of the barrel shifter for the given control inputs. Adiabatic barrel shifter has high throughput, i.e., after initial three clock phases, output is obtained at every phase. Cadence Virtuoso implementation of adiabatic left-shift barrel shifter is shown in Fig. 6. Similarly, a right-shift barrel shifter is designed and functionally verified.

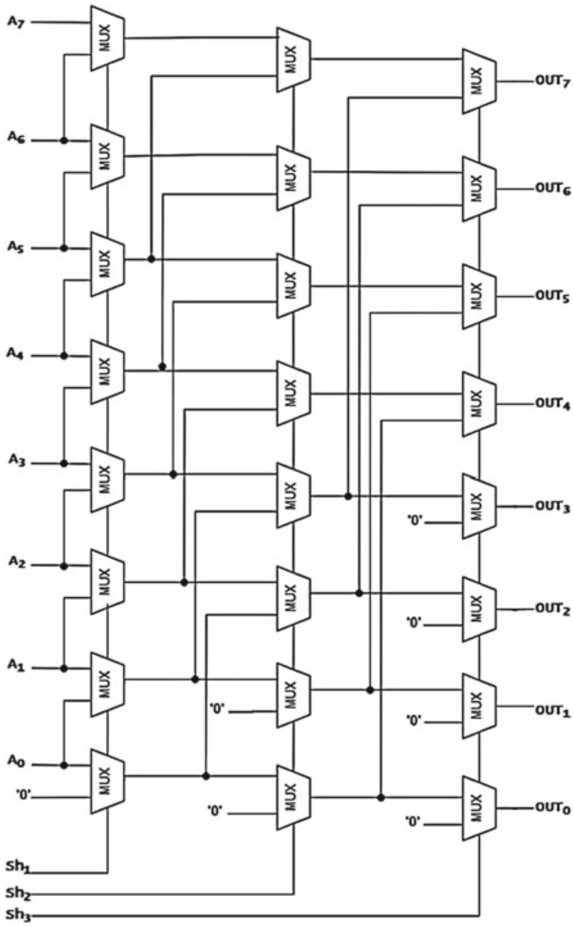


Fig. 4 CMOS MUX-based left-shift barrel shifter

c. **Pass Transistor Logic Barrel Shifter**

Pass transistor logic is another way of designing a barrel shifter. It consists of an array of 8 × 8 nMOS, in which the word length of the data is equal to the number of rows, and the maximum shift width is equal to the number of columns. The control lines run diagonally through the array as shown in Fig. 7. The advantage of this type of shifter is that the data passes through only one transistor gate array. The propagation delay is independent of the shifter size or shift value. But, the parasitic capacitance at the input of buffer increases with shifter size. Output of pass transistor circuits are degraded logic ‘1’ which requires insertion of buffers to overcome the degradation of logic values.

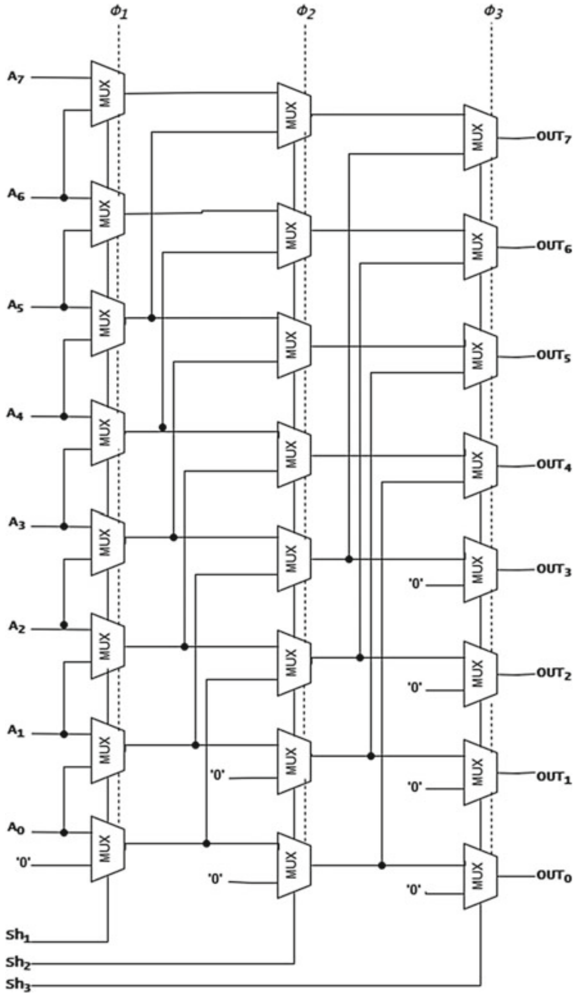


Fig. 5 PFAL MUX-based left-shift barrel shifter

Table 2 Operation of left-shift barrel shifter

Left shifts	0	1	2	3	4	5	6	7
Sh_3	0	0	0	0	1	1	1	1
Sh_2	0	0	1	1	0	0	1	1
Sh_1	0	1	0	1	0	1	0	1

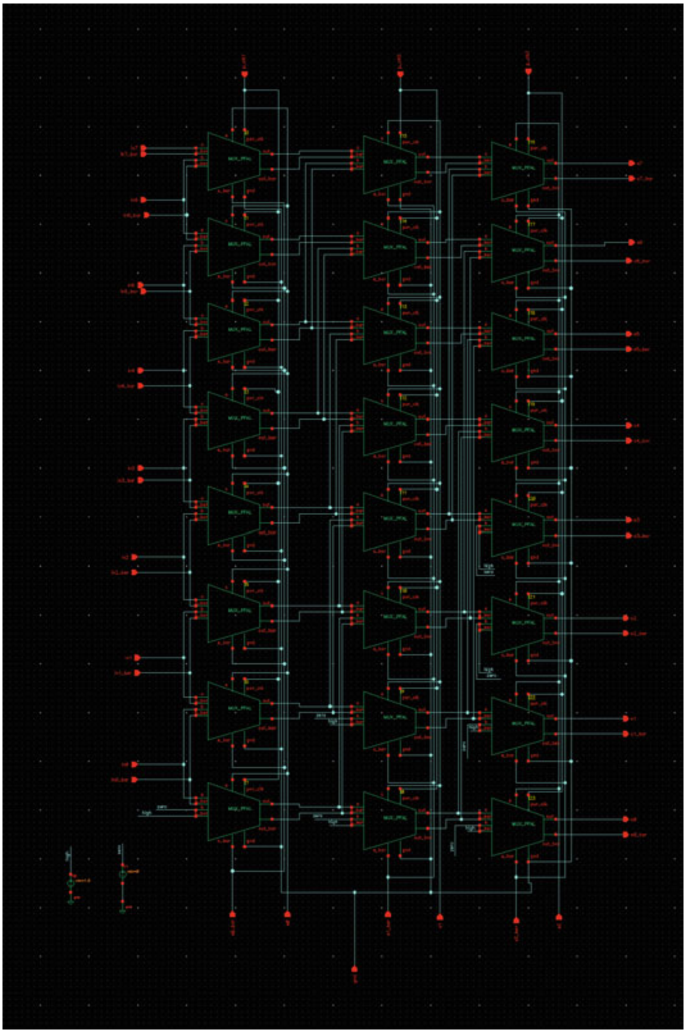


Fig. 6 Cadence implementation of PFAL barrel shifter

3 Results and Discussion

The schematic entry of the left-shift and right-shift barrel shifter is performed in Cadence Virtuoso®. CMOS 180 nm technology is used for the design with the logic level '1' as 1.8 V and logic level '0' as 0 V. Adiabatic 8-bit left-shift barrel shifter is obtained by cascading three stages of MUX, which are driven by power-clocks ϕ_1, ϕ_2, ϕ_3, respectively. Similarly, 8-bit right-shift barrel shifter is designed and implemented in Cadence Virtuoso and functionally verified. The simulation results are discussed only for left-shift case. The functional verification of the design is performed in

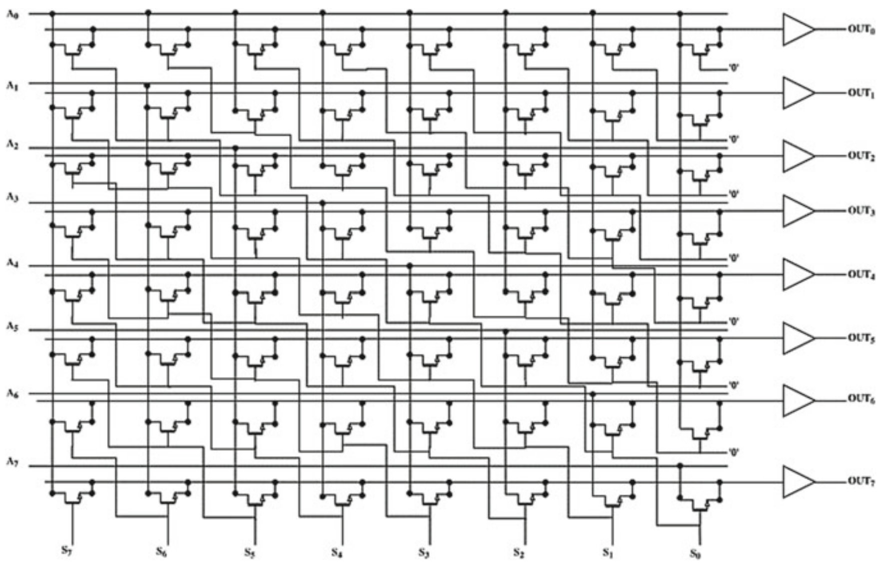


Fig. 7 Pass transistor left-shift barrel shifter

Spectre simulator. Both left-shift and right-shift barrel shifter are functionally verified for various test cases. To illustrate, the input to shifter is considered as '11100111', as shown in Fig. 8. Figure 9 shows the snapshot of the output for the above input with control inputs $Sh_2Sh_1Sh_0$ = '011', and output is shifted left by three bits and the output obtained is '00111000'. Figure 10 shows the simulated output for the same input that is '11100111' with the control input $Sh_2Sh_1Sh_0 = 100$, four bits are shifted towards left, hence the output obtained is '01110000'. The transistor count for each barrel shifter is tabulated in Table 3. Pass transistor barrel shifter uses the

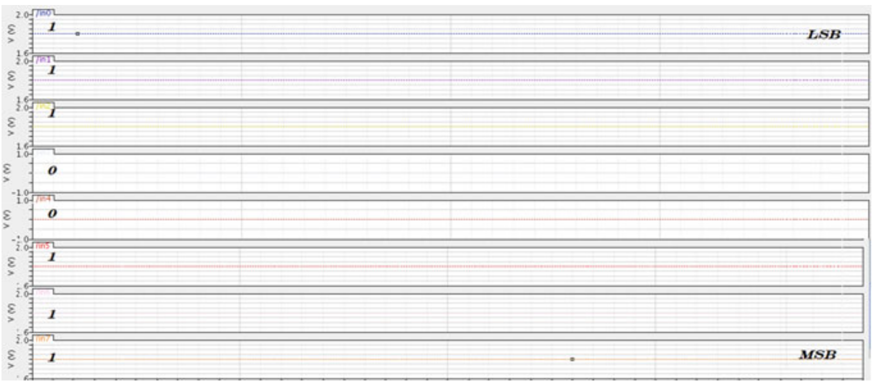


Fig. 8 Input data '11100111'

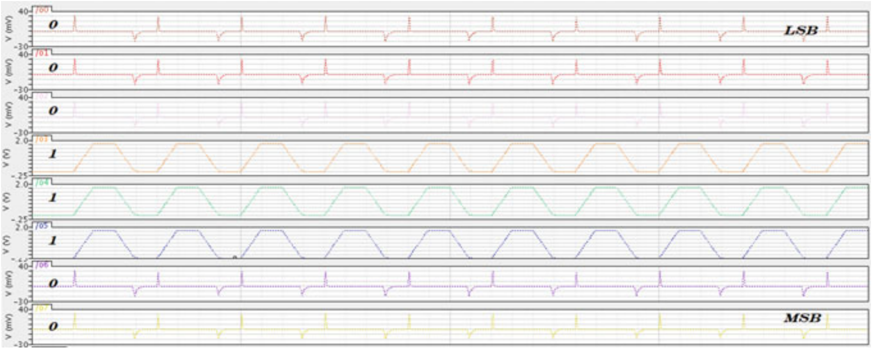


Fig. 9 Output waveform ‘00111000’ for $Sh_2Sh_1Sh_0$ = “011”

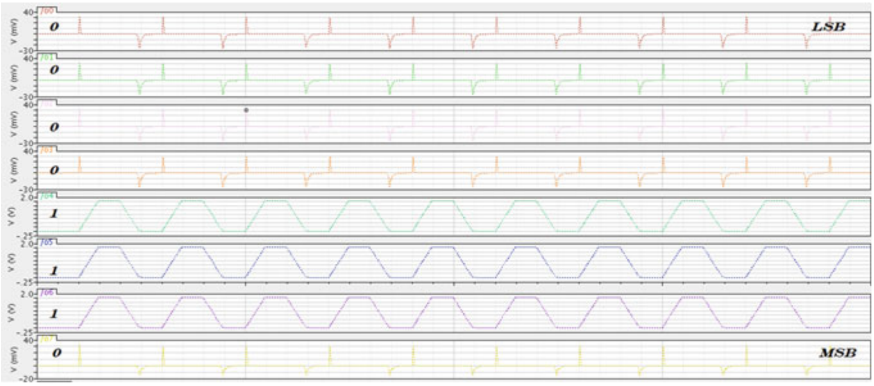


Fig. 10 Output waveform ‘01110000’ for $Sh_2Sh_1Sh_0$ = ‘100’

Table 3 Transistor count of 8-bit barrel shifter

8-bit barrel shifter topology	Number of transistors
Static CMOS MUX-based barrel shifter	144
Adiabatic MUX-based barrel shifter	288
Pass transistor barrel shifter	96

least number of switches but is dominated by wiring. Adiabatic barrel shifter contains more number of transistors, for the same shifter size.

The power dissipated by the left-shift barrel shifter at different frequencies is tabulated in Table 4. The frequency of adiabatic barrel shifter is decided by the four-phase power clock. As the power clock has large rise time and fall time, there is slow transition from one state to other. Hence, less power is dissipated by the adiabatic circuit. It is more practical to plot power dissipation in logarithmic scale than in linear scale, as discussed and shown earlier in Fig. 1 [5]. Figure 11 illustrates the power

Table 4 Power consumption of various left-shift barrel shifters

Frequency	Static CMOS MUX-based left-shift barrel shifter	Adiabatic MUX-based left-shift barrel shifter	Pass transistor left-shift barrel shifter (μW)
1 kHz	4.59 μW	16.5 nW	32.41
10 kHz	45.9 μW	49.2 nW	32.41
100 kHz	459.5 μW	321.1 nW	32.41
1 MHz	4.59 mW	3.315 μW	40.4
10 MHz	45.7 mW	34.52 μW	50.56
100 MHz	438.7 mW	472 μW	196.4

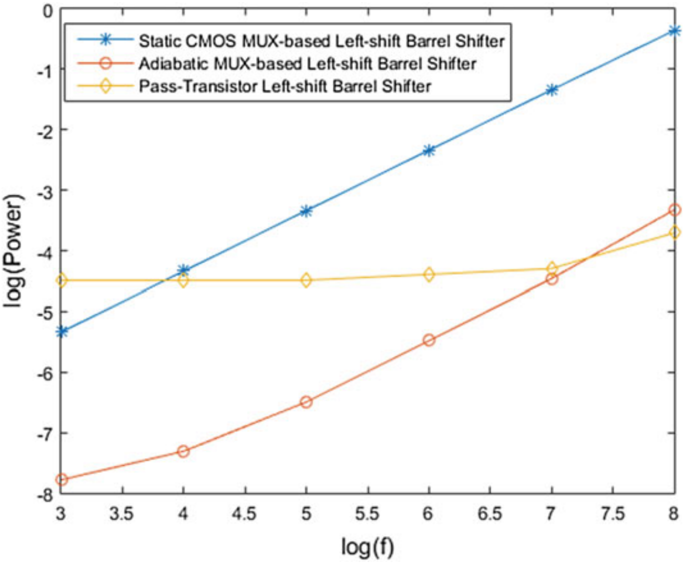


Fig. 11 Power dissipated by left-shift barrel shifter

dissipation plot for the data provided in Table 4. It can be noted that the power of adiabatic barrel shifter is the least and is influenced by the frequency of the power clock. Also, the rise time and fall time of the power clock need accurate design or else the power dissipation abruptly increases due to the increase in the number of transistor count. Power saving factor (PSF) gives a reasonable comparison of adiabatic circuits and static CMOS circuits with respect to power and is given by Eq. (2). It represents the amount of excess power dissipated by the static CMOS circuit with respect to adiabatic circuit [4]. Table 5 lists the PSF for MUX-based left-shift barrel shifter calculated using Eq. (2). It can be observed that at an operating frequency of 100 kHz, a PSF of 1431 is obtained.

Table 5 Power saving factor for MUX-based left-shift barrel shifter

Frequency	1 kHz	10 kHz	100 kHz	1 MHz	10 MHz	100 MHz
PSF	278.18	932.92	1431.01	1384.6	1323.8	929.44

$$\mathrm{PSF} = \frac{\text{Power dissipated by CMOS MUX based Barrel Shifter}}{\text{Power dissipated by Adiabatic MUX based Barrel Shifter}} \tag{2}$$

4 Conclusions

The key factor in recent circuit design is to improve the performance and portability of the device without increase in the power consumption. Adiabatic logic is one of the promising techniques to reduce power consumption. 2N-2P, PFAL, IPGL and ECRL are some of the prominent adiabatic techniques present. PFAL is recognized for efficiently preventing logic-level degradation on the output nodes, using a latch to recycle the power. Embedded and digital signal processors (DSP) use barrel shifters as one of the main component in the processing unit. In this paper, the 8-bit left-shift and right-shift barrel shifter is designed using three different logic techniques. Results from the designed circuits illustrate that adiabatic logic-based barrel shifter is more advantageous in terms of power, but a trade-off has to be made in terms of area. It is observed from Tables 4 and 5, that using adiabatic circuit's power dissipation can be limited to a great extent, under certain conditions and is always better than the static CMOS based circuits. Higher-order barrel shifters can be implemented with adiabatic logic in a similar way.

These adiabatic barrel shifters along with the adiabatic adders and multipliers can be further used in the implementation of larger sub-systems such as FIR filters, MAC unit in DSP and ALU. Optimization in terms of area and latency of PFAL-based circuits lies in the future scope of this work. Hence, it can be concluded that proposed adiabatic-based barrel shifter is very promising and power efficient in low-power design regime.

References

1. Bhati, P., Rizvi, N.Z.: Adiabatic logic: an alternative approach to low power application circuits. In: The Proceedings of IEEE International Conference on Electrical, Electronics, and Optimization Techniques, Chennai, vol. 1, pp. 4255–4260 (2016)
2. Teichmann, P.: Introduction to Adiabatic logic. Springer Series in Advanced Microelectronics, vol. 34. Springer, Dordrecht (2012)
3. Sivakumar, R., Jothi, D.: Recent trends in low power VLSI design. Int. J. Comput. Electr. Eng. **6**, 509–523 (2014)

4. Goyal, S., Singh, G., Sharma, P.: Variation of power dissipation for Adiabatic CMOS and conventional CMOS digital circuits. In: The Proceedings of 2nd IEEE International Conference on Electronics and Communication Systems, vol. 2, pp. 162–166 (2015)
5. Teichmann, P.: Fundamentals of Adiabatic logic. Springer Series in Advanced Microelectronics, vol. 34. Springer, Dordrecht (2012)
6. Verma, R., Mehra, R.: Area efficient layout design analysis of CMOS Barrel Shifter. Int. J. Sci. Res. Eng. Technol. 84–89 (2015)
7. Yedidya, H., Lee, R.B.: A new basis for shifters in general-purpose processors for existing and advanced bit manipulations. IEEE Trans. Comput. **58**, 1035–1104 (2009)
8. Premananda, B.S., Chandana, M.K., Shree Lakshmi, K.P., Keerthi, A.M.: Design of low power 8-bit carry-select adder using Adiabatic logic. In: The Proceedings of IEEE International Conference on Communication and Signal Processing, pp. 1764–1768 (2017)
9. Nazare, N., Bhat, P.S., Jambhe, N.R., Premananda, B.S.: Design and analysis of 16-bit parallel-prefix Adiabatic adders. In: The Proceeding of 3rd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, Bengaluru (2018)
10. Anantharaman, K., Manikandan, A.V.M.: Adiabatic techniques for energy-efficient Barrel Shifter design. In: VLSI Design: Circuits, Systems and Applications. Lecture Notes in Electrical Engineering, vol. 469. Springer (2018)
11. Tatajee, D.A., Rajesh, K.: Semi-custom implementation of area efficient Barrel shifter. Int. J. Adv. Comput. Eng. Appl. **1**, 112–114 (2012)
12. Nikhil, G.V., Vaibhav, B.P., Naik, V.G., Premananda, B.S.: Design of low power Barrel Shifter and Vedic-Multiplier with Kogge-Stone adder using reversible logic gates. In: The Proceedings of IEEE International Conference on Communication and Signal Processing, pp. 1690–1694 (2017)
13. Singh, R., Rajesh, M.: Power efficient design of Multiplexer using Adiabatic logic. Int. J. Adv. Eng. Technol. **6**, 246–254 (2013)

High-Speed Loop Unrolled Grain Architecture in Reconfigurable Hardware

Paresh Baidya, Rourab Paul and Suman Sau

Abstract Crypto core finds implementaions in software, Application Specific Integrated Circuit (ASIC), and Field Programmable Gate Array (FPGA). The crypto software programs achieve very less throughput, whereas ASIC implementations serve reasonably cost-effective design with efficient performance. In ASIC platform, once the design is implemented, it is not possible to alter those circuit connection, whereas FPGA is a flexible solution which can be reconfigured in the field. For huge number of gate applications, FPGA is cost-effective compared to ASIC and software implementation. This submission explores FPGA design spaces of Grain crypto hardware core. The article implements seven loop unrolled architectures which are implemented in Xilinx Zynq FPGA using VHDL language in Vivado Design Suit. The results show that three basic systems' parameters such as resource usage, power consumption, and throughput of proposed implementation are satisfactorily optimized compared to existing literature.

Keywords FPGA · Grain · Cryptography · Coprocessor · RISC processor · Loop unrolled

1 Introduction

In the new century, hardware implementation in cryptography is needed because of requirement of high speed, low cost, and security communication combined with physical security. Almost all the crypto algorithms involve considerable mathematical

P. Baidya (✉) · R. Paul · S. Sau
Department of Computer Science and Engineering, Siksha 'O' Anusandhan
Deemed to be University, Bhubaneswar, Odisha, India
e-mail: pareshbaidya@soa.ac.in

R. Paul
e-mail: rourabpaul@soa.ac.in

S. Sau
e-mail: sumansau@soa.ac.in

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_15

complexity which is handled well in software platform in respect of algorithmic flexibility in the event multi-ciphering applications are required. The hardware implementation of such algorithms provides relatively better throughput, but does not usually provide the algorithmic flexibility required for multi-ciphering applications in the sense that it is not possible to change algorithm during runtime at the same silicon spaces. The solution to achieve algorithmic flexibility in hardware is to store each crypto algorithm at different silicon spaces, and during runtime, the relevant algorithm is called for execution. The earlier FPGA was made of configurable logic blocks (CLB) that is programmable and is named as programmable logic (PL) environment. One can program applications in PL environment to run them concurrently. It may be noted that the hardware implementation for multi-ciphering algorithms was possible in earlier FPGAs, but it used to occupy considerable silicon spaces and consumed significant power during runtime. Subsequently, a feature of hardware reconfiguration is introduced in FPGA. With already present inherent spatial parallelism, the FPGA provides better throughput and the hardware reconfigurable feature has restored the algorithmic flexibility in hardware. In FPGA with reconfigurable features, the algorithmic flexibility is achieved, the power consumption is reduced, the prices are lowered, and these are coupled with high physical security and higher marketing speed. Very recently, the hardware reconfiguration in FPGA technology is replaced by dynamic partial hardware reconfiguration and added a hard core processing system (PS) where software programs are executed concurrently with hardware programs being executed in PL with an exclusive aim to increase throughput. In such combined PL and PS environment, the hardware designer can considerably reduce silicon resources and power consumption maintaining reasonably high throughput making the environment ideal for efficient crypto applications in resource constrained system [1]. In our research exploration, we tried to optimize performances of crypto algorithms in hardware by adopting suitable approaches in concurrence with the growth of FPGA-based embedded system. The growth of such systems can be categorized in three generations: The first one is composed of a complete programmable logic (PL) block in which the various types of softcore embedded processors are available in library and programming activities that are undertaken using high-level language. A user can undertake an application program along with the necessary I/O block in a suitable softcore processor or can write the I/O block using a processor and the application separately in the PL block using HDL or can write the complete softcore embedded processor incorporating the application in the PL block itself using HDL. In the second generation, a hard core embedded processor to be programmed with high-level language, named as processing system (PS), is made available along with the PL block to be used for hardware implementation of different applications using HDL. In the third generation, in addition to the PS block, the programmable reconfiguration facilities are made available in the PL block including the softcore embedded processor. While undertaking research explorations in concurrence with the growth of technology, the performances of crypto algorithms are being optimized in hardware by adopting various approaches. For example, throughput can be optimized by pipeline architecture, loop unrolling, and several parallelism techniques, resource may be optimized by decreasing parallelism, and power also has direct correlation with resource consumption. The solution of these bottlenecks is not singular; for different applications, different solutions are being explored. Even though FPGA with dynamic reconfigurable feature is an optimal platform for crypto applications, there is still a serious trade off between the hardware designer and the designer of the crypto algorithm. For a hardware designer, efficient design is necessary when a resource uses problem within a hardware environment of limitations and the goal is to find a favorable balance between silicon area, operational throughput, power consumption, and the state of design followed by implementation of a system. The crypto designer, on the other hand, is interested in the crypto qualities of the algorithm, such as randomness of cipher texts and resistance of the cipher texts to various attacks. The divergence of requirements of both the sides once reasonably converges, the crypto designer would look for a performance metric comparing the hardware and software issues of different crypto algorithms without being able to appreciate the constraints of the system required for the hardware designer, while the hardware designer would not be able to fully comprehend the crypto qualities. For a best design, it is necessary to find a correlation between the stipulated issues of hardware design and the performance metrics of a crypto algorithm. It seems that the job is complex. Another issue is relevant in the sense that the mathematically secured crypto algorithm may become vulnerable while implementing in hardware, since it might be possible to extract secret information by introducing faults in a controlled fashion due to which fault detection techniques turn out to be an important issue related to hardware implementation. The real challenge for solution of crypto hardware design is to maintain a stipulated balance among all issues. Our proposal in this paper implements the Grain in FPGA with limited hardware environments and consists of two shift registers with bit size 80, one is linear and other is nonlinear. In this paper, we described briefly our proposal in each section. In Sect. 2, we described related work. In Sect. 3, we design the grain hardware. In Sect. 4, we described the grain hardware core. In Sect. 5, we have written grain acceleration. In Sect. 6, we design the processor to grain core communication. In Sect. 7, we describe FSL interface design. In Sect. 8, we give the result and implement the grain in FPGA. The conclusion is finally given in Sect. 9.

2 Related Work

There are two types of symmetric key encryption algorithms, block ciphers (AES [2], DES, RC5, IDEA, etc.) and stream ciphers (RC4 [3], Grain [4], Salsa, etc.). Both types have many commercial uses and applications, such as electronic banking, electronic mail, Internet network service, and messaging network. The key exchanges ensure the authentications. Crypto algorithm should be optimized interms of resources, throughput, and power consumption. Many researchers implemented AES [1], ECC, RSA [5], and RC4 [3] in both of software and hardware. In this proposal, we described the Grain in FPGA, providing a higher security than other well-known ciphers.

3 Grain Hardware

The Grain stream cipher [4] is the lightest stream cipher among all others, and it is presented by Martin Hell, Thomas Johansson, and Willi Meier in eSTREAM in 2004. The Grain is such type of stream cipher which is based on bit oriented and specially for limited hardware environments. It consists of two shift registers, one linear feadback and other one is nonlinear feedback, and at time t, the states of the shift registers are defined with $\{l_i, \ldots, l_{i+79}\}$ and $\{n_i, \ldots, n_{i+79}\}$. From 12 states of update function, Grain that gives output a single bit z_i at each instant time is stated in the following equation.

$$\begin{aligned} z_i = {} & n_i + 1 + n_{i+2} + n_{i+4} + n_{i+10} + n_{i+31} + n_{i+43} + n_{i+56} \\ & + h(l_{i+3}; l_{i+25}; l_{i+46}; l_{i+64}; n_{i+63}) \end{aligned} \tag{1}$$

the following Fig. 1 shown, h(.) is a function with degree 3.A primitive feedback polynomial for new LFSR is defined by
f(y) = $1 + y^{18} + y^{29} + y^{42} + y^{57} + y^{67} + y^{80}$ the updated equation is.

$$l_{i+80} = l_{i+0} + l_{i+13} + l_{i+23} + l_{i+38} + l_{i+51} + l_{i+62} \tag{2}$$

The nonlinear feadback register updated with a function g(.) of degree six whose output is masked with the low-order LFSR bit:

$$\begin{aligned} n_{i+80} = l_i + g(n_i; n_{i+9;} n_{i+14}; n_{i+15}; n_{i+21}; n_{i+28}; n_{i+33}; n_{i+37}; \\ n_{i+45}; n_{i+52}; n_{i+60}; n_{i+62}; n_{i+63}) \end{aligned} \tag{3}$$

Upto 16 clocks, in Grain, the bit which enters in the registers remains unused. Since, in the state update functions, the highest register indices n_{i+63} and l_{i+64} have occurred, it permits the device to execute the cipher sixteen times in parallel for speeding up the cipher to increase the gate count. During the key generation and key initialization, the state of Grain update function maintains the inversibility. Interesting thing is that we can recover the key which is used in cipher if we find the state of the function at some time t. The functions which are used in Grain method is shown in Fig. 1 pictorially which will clear to us.

4 Grain Hardware Core

The Grain hardware core uses two storage blocks with 80 bit width for equations $f(x)$ and $g(x)$, respectively. The $nfsr$ and $lfsr$ take key and a 64 bits IV with 16 bit padded '1', respectively. As shown in Eqs. 3 and 2, the two functions $g(x)$ and $f(x)$

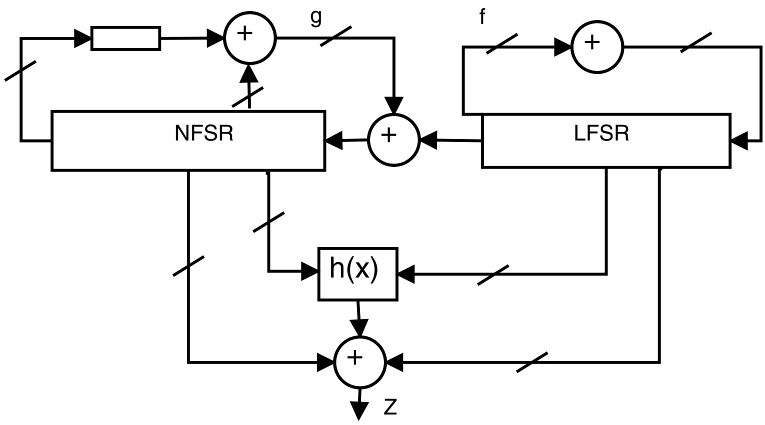


Fig. 1 Grain functionality

are executed which are taken to the most significant bits of these two shift registers with a $xored\ z$ value as stated in Eq. 1. After completing 160 clocks by 160 times the output $z\ xored$ with the plain text.

5 Grain Acceleration

The standard Grain [6] gives the output 1 bit at each clock after the 160 initial clock time stand and it consumes resource and power. If we go through more resource and power, grain can be accelerated by 16 times using loop unrolling method. It is observed that the register shifted from msb to lsb, and the used msb in $g(x)$, $f(x)$, and $h(x)$ as shown in Eqs. 3, 2, and 1, respectively, is in 63rd position. Hence, Grain can be accelerated 16 times by loop unrolling process which can generate upto 16 bit per clock.

6 Processor to Grain Core Communication

Bus is possible communication between processor and coprocessor. The processor data and coprocessor data can be passed through various available bus architectures. In Xilinx FPGA [7] platform, the communication between main processor and co-processor can be implemented with three available bus architectures.

1. Processor local bus (PLB)
2. Fast simplex bus (FSL)
3. Advanced extensible interface bus (AXI)

6.1 FSL Bus

For the high performance in FPGA, FSL bus starts to communicate between two devices in FPGA. The Xilinx MicroBlaze processor that uses FSL interface is an interface that transfers the data to the register from the hardware as well as from hardware to register in the running time on the FPGA. The features of FSL are stated below:

- 32-bit wide interface.
- Configurable FIFO depths of FSL is 1–8192 which use shift registers named as SRL16.
- The clock of FSL is synchronous or asynchronous which is based on MicroBlaze system clock.
- Simple software interface acting as application peripheral interface (API) written in C instructions are generated automatically by the xilinx tool.
- Blocking and non-blocking software instructions are used for data and control.
- Return from interrupt resume the FSL instruction.
- The FSL is compatible with SRL16 shift register and dual port LUT RAM or Block RAM based FIFO implementation.

7 FSL Interface Design

In our implementation, we use FSL bus architecture [8]. In our architecture, two dynamic inputs and one output are used. The Microblaze soft core can accommodate sixteen FSL interface buses. Out of the sixteen buses, eight can be used as inputs to the user core and the rest are for the outputs. Fig. 2 shows that FSL bus connects Microblaze processor core with a coprocessor. For a requirement, needed external

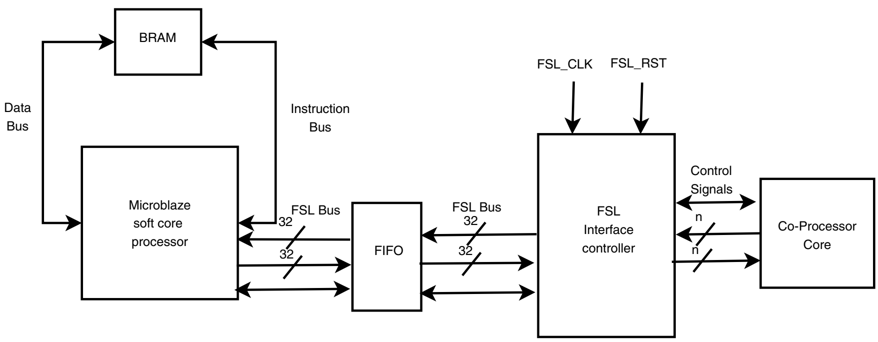


Fig. 2 A figure of a coprocessor attached with the Microblaze in various ways

signals can connect directly with the coprocessor. For sending and receiving the data to the hardware, FSL used C-macros in the FPGA.

The RC4 [3], AES [2], Grain [4] , and RSA [5, 9, 10] algorithms are written in VHDL and unlike [3, 5, 9] other articles [2, 10] did not implement crypto coprocessor with processor. In our proposal, Microblaze processor communicates with the coprocessor using FSL bus. The Grain hardware is placed inside the coprocessor. In Microblaze processor [11], the FSL interface communicates by FIFO based and it also supports control and data communication port. The data that pass through the FIFO depend on the clock time. In general, depth of FIFO is taken as data vector length. When FIFO depth size is less than the data input size, then the speed of data processing may vary. A few signals of the FSL bus protocol are shown in Fig. 2. Here, in Fig. 2, a main processor named Microblaze sends 32 bit data at a time to FSL bus. The FSL bus has two communication modes,

(1) Normal mode communication and (2) burst mode communication. A FIFO which supports asynchronous and synchronous read write operation is implemented inside FSL bus because there might be a requirement of clock domain crossing between main processor and coprocessor. The FSL interface controller handles the bus and coprocessor data management. FSL interface controller has six main ports toward coprocessor, namely

1. **Master Data Buffer**: which sends the n (n >32 and multiple of 32) bit width data to coprocessor.
2. **Data Load**: *data load* acknowledged to the coprocessor when the data are ready in *master data buffer*.
3. **Slave Data Buffer**: It sends n bit data from coprocessor to FSL interface controller through *slave data buffer*.
4. **Data Ready**: After completing the operation, coprocessor sends a ready acknowledgment signal to FSL interface controller.
5. **Coprocessor Clock**: It is the most important signal for coprocessor. The *clock* might be taken directly from FPGA clock resource or via this wrapper. As per requirement, you can send the clock signal through FSL interface manager. In few cases, like clock gated technology to reduce dynamic power consumption and PR clock control, you might have a necessity to control the coprocessor clock. It is to be noted that Xilinx suggests to stop the clock signal during partial bit configuration process.
6. **Coprocessor Reset**: Reset pin from wrapper side for coprocessor is also an optional like clock input.

8 Result and Implementation

7 Grain has been proposed as different approaches, namely Grain 1 LU, Grain 2 LU, Grain 4 LU, Grain 5 LU, Grain 8 LU, Grain 10 LU, and Grain 16 LU. Grain 1 LU has 1 loop unrolling factor, Grain 2 LU has 2 loop unrolling factor, and so on, and

Table 1 The comparison table of the proposed grain

Name of algorithm	Slice/gate in count #	Power (mw)	Throughput rate Gbps	Critical path (ns)	Board name
Proposed Grain 1 LU	361	90.99	0.223	1.689	Zynq
Grain 1 LU [4]	1435	–	0.047	-	MAX 3000A
		–	0.195	–	MAXII
		–	0.2755	–	Cyclone
Proposed Grain 2 LU	375	91.27	0.466	1.704	Zynq
Grain 2 LU [4]	1607	–	0.096	–	MAX 3000A
		–	0.412	–	MAXII
		–	0.5625	–	Cyclone
Proposed Grain 4 LU	399	88.12	0.931	1.749	Zynq
Grain 4 LU [4]	1950	–	0.191	–	MAX 3000A
		–	0.617	–	MAXII
		–	0.852	–	Cyclone
Proposed Grain 5 LU	426	88.78	1.164	1.754	Zynq
Proposed Grain 8 LU	463	87.7	1.86	1.734	Zynq
Grain 8 LU [4]	2636	–	0.234	–	MAX 3000A
		–	1.156	–	MAXII
		–	1.695	–	Cyclone
Proposed Grain 10 LU	496	91.41	2.328	1.772	Zynq
Proposed Grain 16 LU	591	98.44	3.72	1.743	Zynq
Grain 16 LU [4]	4008	–	2.078	–	MAXII Cyclone
Grain 16 LU(128) [1]	5840	–	6.25	–	Virtex-5

Grain 16 LU has 16 loop unrolling factor. We implemented the 7 Grain variants in zynq 7z020 Xilinx FPGA board with ISE 14.4 Design Suite. The resource, power, and throughput of seven design approaches and existing Grain implementation are shown in Table 1.

9 Conclusion

The conflicting relation between three basic parameters of embedded systems are power, throughput, and resource implying the need of optimization which is again a relative consideration with the changing requirement of the user. The goal of this paper is to implement Grain algorithm in coprocessor to achieve higher throughput and low power consumption. Here, 16 loop unrolling method is used to accelerate the Grain which is proposed in FPGA. In future, the proposed Grain core will be implemented as a coprocessor along with a main processor where main processor will handle all the interface processings and the high-speed coprocessor will execute crypto algorithms such as Grain. The parallel functionality of main processor and coprocessor will increase over all system throughput.

References

1. Aumasson, J.-P., Dinur, I., Henzen, L., Meier, W., Shamir, A.: Efficient FPGA implementations of high-dimensional cube testers on the stream cipher grain-128. IACR Cryptology ePrint Arch. **2009**, 218 (2009)
2. Liberatori, M., Otero, F., Bonadero, J.C., Castineira, J.: AES-128 cipher. high speed, low cost fpga implementation. In: 2007 3rd Southern Conference on Programmable Logic, 2007. SPL '07, pp. 195–198, Feb 2007
3. Paul, R., Dey, H., Chakrabarti, A., Ghosh, R.: Accelerating more secure RC4 : implementation of seven FPGA designs in stages upto 8 byte per clock (2016). CoRR, arXiv:1609.01389
4. Hell, M., Johansson, T., Meier, W.: Grain: a stream cipher for constrained environments. IJWMC **2**(1), 86–93 (2007)
5. Sahu, S.K.: FPGA implementation of RSA encryption system. J. Comput. Appl. **19**(9), 10–12 (2011)
6. Paul, R., Saha, S., Zaman, J.K.M.S.U., Das, S., Chakrabarti, A., Ghosh, R.: A simple 1-byte 1-clock RC4 design and its efficient implementation in FPGA coprocessor for secured ethernet communication (2012). CoRR, arXiv:1205.1737
7. In: NSA Approved Defense-Grade Spartan®-6Q FPGA in Production for Highest Level Cryptographic Capabilities Strengthens Xilinx Secure Leadership. Xilinx, Inc. website, pp. 138 –148, 31 Aug 2011
8. Sau, S., Paul, R., Biswas, T., Chakrabarti, A.: A novel AES-256 implementation on FPGA using co-processor based architecture. In: Proceedings of the International Conference on Advances in Computing, Communications and Informatics, ICACCI '12, New York, NY, USA, pp. 632–638. ACM (2012)
9. Paul, R., Chakrabarti, A., Ghosh, R.: Multi core SSL/TLS security processor architecture and its FPGA prototype design with automated preferential algorithm. Microprocess. Microsyst. **40**(C), 124–136 (2016)
10. Paul, R., Saha, S., Pal, C., Sau, S.: Novel architecture of modular exponent on reconfigurable system. In: 2012 Students Conference on Engineering and Systems (SCES), pp. 1–6, Mar 2012
11. Microblaze Xilinx: Microblaze processor reference guide. Xilinx, Microblaze. http://www.xilinx.com/tools/microblaze.htm

Cloud Computing, IoT, and Bigdata

Automated Red Lesion Detection: An Overview

Arati Manjaramkar and Manesh Kokare

Abstract Incidences of diabetes are increasing worldwide. An eye complication associated with uncontrolled diabetes is diabetic retinopathy. If not treated, diabetic retinopathy can be vision threatening. The microaneurysms and dot hemorrhages are the only prominent clinically observable symptoms of DR. Their timely detection can help ophthalmologists in treating abnormalities efficiently and limit the disease severity. So detecting red lesions in early stage has become an indispensable task today. This paper gives an overview of earlier proposed algorithms and methods. It also compares these algorithms based on their performance for supporting the researchers by providing the gist of these algorithms. The standard retinal image databases are also compared and discussed.

Keywords Automated detection · Diabetic retinopathy · Microaneurysm · Hemorrhage · Red lesion · Fundus image

1 Introduction

1.1 Motivation

Vision loss as a result of diabetic retinopathy (DR) has a negative impact on the patient's life. So, the diagnosis of DR in its primary stage is significant when cure is possible. But DR is a silent disease, and the patient realizes its presence when the retinal changes have progressed to the extent where treatment is nearly impossible. The only method for detecting DR in its early phase and keep track of its progress is by screening. Many countries do DR assessment manually, which is a time-consuming,

A. Manjaramkar (✉) · M. Kokare
SGGS Institute of Engineering and Technology, Nanded 431606, India
e-mail: akmanjaramkar@sggs.ac.in
URL: http://www.sggs.ac.in

M. Kokare
e-mail: mbkokare@sggs.ac.in

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_16

error-prone, tiring job and causes clinical overload. An ophthalmologist is now in need of automated red lesion detection system that can help in grading and treat conditions accordingly.

1.2 Background

If a person suffers from diabetes, DR is an ultimate fact over a period of 10 years, even if the diabetes is in control. It is infrequent that there is no effect of DR on the patient's retina. Diabetes over a period can injure retinal blood vessels. Alike injury caused to retinal vessels is referred as DR. DR can be arranged in categories like non-proliferative DR (NPDR) and proliferative DR (PDR). NPDR termed background or pre-proliferative retinopathy is an initial phase of DR. In NPDR, no indications of the disease are sensed and sufferer will have normal vision (20/20 vision). The fragile retinal vessels start leaking blood or fluid; this leakage causes retinal swelling or forms deposits termed exudates.

NPDR contains microaneurysms (MAs), hemorrhages (HMs), and hard exudates (EXs). The MA and HM being frequently arranged in one cluster are called red lesions. MA becomes visible as distinct, small, and circular red objects and is often 10–100 μm in diameter. MA appears red in fundus images and dark in red-free images. Individual MAs may leak resulting in dot HM, edema, and an exudate. Dot HMs are difficult to differentiate from MAs as they are similar in appearance. Dot HMs are rarely larger than 200 μm. Since we cannot differentiate them clinically, the term dot hemorrhage/ microaneurysms (H/Ma) is used often. The newly grown weak blood vessels bleed into retinal layers resulting in HMs. This bleeding is not a vision threatening until it occurs in or around the macula (Fig. 1).

This paper is an extensive overview and comparison of earlier proposed algorithms for red lesion detection. The remainder of the paper is structured as follows: Sect. 2 elucidates generalized procedures. Sections 3 and 4 give an extensive description of the methods used in MA and HM detection, respectively. Section 5 discusses the publicly available standard retinal image databases. Section 6 describes the

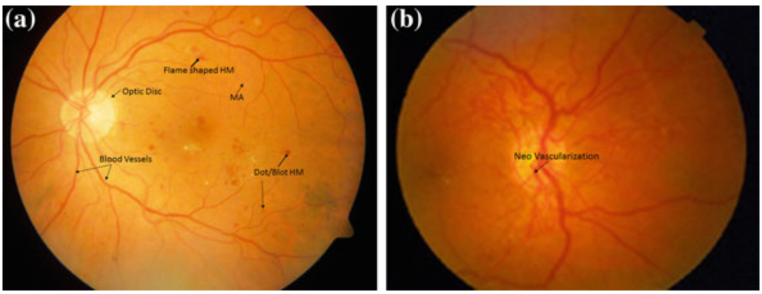


Fig. 1 Diabetic retinopathy types **a** NPDR and **b** PDR

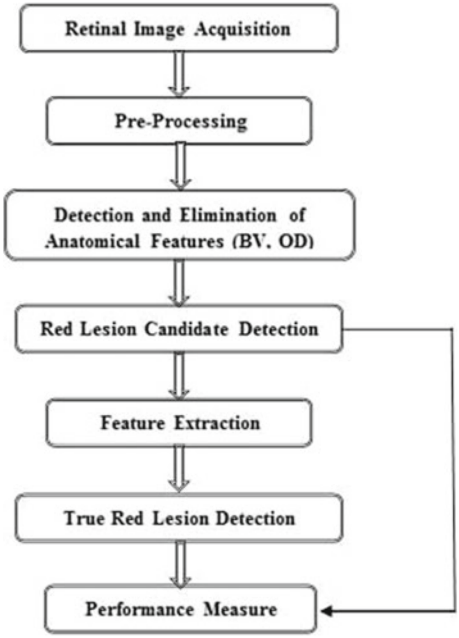


Fig. 2 Generalized procedure for red lesion detection

performance evaluation criteria for the red lesion detection algorithms. Section 7 compares the performances of red lesion detection algorithms. Conclusions and directions for future research are presented in Sect. 8.

2 General Red Lesion Detection Procedure

Several researchers have presented retinal image analysis algorithms. Figure 2 represents general system construction for detecting the red lesions in retinal images.

a. Retinal image acquisition
b. Pre-processing
c. Detection and elimination of anatomical features
d. Detection of red lesions.

3 Microaneurysm Detection

MAs are the earliest clinically visible symptoms of DR; they are saccular bulges in the capillary vessels. Although capillary vessels are not able to be seen in color FI (fundus images), MAs come into view as small, dark reddish, round spot in FI

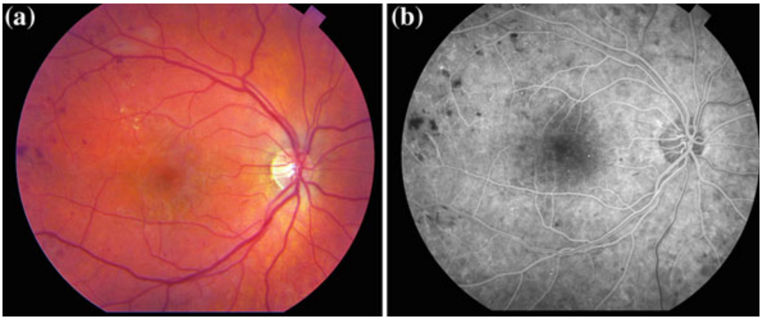


Fig. 3 Images of patient with NPDR **a** color fundus image and **b** fluorescein angiogram

(Fig. 3a), whereas MAs in fluorescein angiograms (FA) appear as small white dots as shown in Fig. 3b. It is difficult to find and separate out MAs from the background noise. In addition, few MAs take up the irregular shape due to cluster formation or getting attached to vessels. Thus, automated MA detection in the color FI is a challenging task.

3.1 Mathematical Morphology

a. MA detection in FA

Mathematical morphology is commonly used for MA detection. Automated MA detection by Baudoin et al. [2] is one of the earliest reviewed journal papers. They proposed a morphological approach for MA detection in FAs. A top-hat transformation was used for differentiating the blood vessel (BVs) and MAs.

Spencer et al. [20] presented an image-processing method for detecting MAs in digitized FA. In the pre-processing stage, image correction was done to correct the vignetting effect due to digitization. MAs were segmented from vasculature using bilinear top-hat linear structuring element. Region growing was applied to each candidate. Simple shape and color intensity features were given as input to a rule-based classifier. Thirteen images were studied for deciding rule filter parameters. Application of the filters on all thirteen images reported a highest sensitivity of 82.4% at a specificity of 85.6%.

b. MA detection in FI

Walter et al. [24] presented a four-step mathematical morphology algorithm—first, pre-processing step; second automatic threshold and grayscale diameter closing; third feature extraction step; and finally, kernel density estimation for classification to get true MAs automatically. For training the algorithm, 21 illustrated images were used. Compared with manual grading, 88.5% sensitivity was achieved. However, the second step of diameter closing and thresholding usually causes deprivation of significant feature points which were pivotal for later processing and classification.

3.2 Filtering

a. MA detection in FA

Hipwell et al. [6] presented a technique for detecting MAs in red-free images. As a pre-processing step, input images were shade corrected. A top-hat transformation was then used to remove vessels and other distractions. Further classification of candidate MAs was retained by applying a Gaussian matched filter and was based upon 13 different specifications such as shape, intensity, circularity, viz... which were obtained on 102 train-set images having a varied retinopathy degree. The system reports 81% sensitivity and 93% specificity, which was achieved only when doubtful HMs and MAs debarred.

b. MA detection in FI

Zhang et al. [25] developed a multiscale correlation filtering (MSCF) method for MA detection. For locating MAs in retinal images, a two-level hierarchy is used. First-level (coarse-level) candidate MAs is detected by MSCF. Five values of sigma in Gaussian kernels were used for calculating each pixel's correlation coefficient. In next level (fine level), 31 feature thresholds were used to classify them as true MAs. Prior expert knowledge was used for deciding each feature threshold. But considering all features can introduce redundancy or irrelevant features, thereby reducing classification accuracy and computational complexity. The algorithm was evaluated on two public datasets—ROC [15] and DIARETDB1 [9]. The proposed method reported an sensitivity of average value 0.713.

3.3 Pixel Classification

a. MA detection in FI

Niemeijer et al. [16] presented an hybrid algorithm for detecting red lesions. Previous work of Spencer and Frame was merged to have two novel contributions. Firstly, red lesions were detected using pixel classification technique. The pixel classification technique distinguished foreground and background objects. Then connected blood vessel structure is separated leaving aside candidate red lesions. Secondly, several new features are added to Spencer–Frame. All these features were used to classify true candidates by K-NN. The system reported a sensitivity and specificity of 100% and 87%, respectively.

Kande et al. [8] also used pixel classification in combination with mathematical morphology to detect MA and HM. Red lesions were detected by the matched filter.

Top-hat transformation is used for suppressing enhanced blood vessels. To further classify red lesions, SVM is used. 100% sensitivity with 91% specificity is attained by this approach.

3.4 Miscellaneous Approaches

a. MA detection in FA

Tavakoli et al. [23] proposed an MA detection algorithm using multi-overlapping windows and Radon transform. In a pre-processing step, the average filter and top-hat transformation were used for background removal. After pre-processing, images were split as subimages segmenting the vascular by applying the Radon transform on every subimage. The MAs were identified and quantified by using Radon transform and proper threshold.

b. MA detection in FI

Ram et al. [18] proposed an MA detection strategy based on clutter rejection. The target occurrence probability is much less than clutter. A stepwise strategy is designed with a set of features to reject certain classes of clutter and maintain true positives. True positives left after the final clutter rejection stage are assigned a score based upon its similarity with true MA. It performs well in image-level detection as compared to the lesion level.

Quellec et al. [17] proposed a template matching supervised MA detection method in wavelet subbands. Three image modalities, color photographs, green channel, and FA, were used. Probable location of MA is detected by matching templates in wavelet domain. Proposed methods execution was enhanced by the wavelet lifting strategy. According to the modality of image, MA detection sensitivity of 89.62%, 90.24%, and 93.74%, respectively, was achieved.

Seoud et al. [19] proposed a technique to detect red lesions using shape features set called Dynamic Shape Features (DSF). Firstly, to handle various image resolutions, spatial calibration is done. Then, CFI is pre-processed by color normalization and contrast equalization. The third step is automated optic disk detection and elimination. In the fourth step, candidates are identified from a pre-processed image. For each identified candidate, color feature and DSF features were extracted in the fifth step. Finally, lesions and non-lesion pixels were classified by random forest. Results show that the method outperforms several earlier proposed approaches based on pre-lesion and per-image levels. In the ROC, the method ranks fourth with an FROC score of 0.420.

Manjaramkar and Kokare [13] have proposed statistical geometrical features for MA detection. A feature set is proposed build on statistical geometrical properties of connected components, which can smoothly discriminate lesion and other pixel [10]. For analysis and validation, DIARETDB1, ROC, STARE, and MESSIDOR datasets were used. The used dataset has non-identical image attributes as well as camera settings. Best results were obtained for per-image evaluation on DIARETDB1 dataset reporting 88.09% sensitivity with 92.65% specificity.

4 Hemorrhage Detection

HM occurs due to abnormal bleeding inside retinal blood vessels. A retinal HM can be caused by a disease which affects the person's circulatory system; such diseases are hypertension and diabetes. HMs vary in shape and size. HM type and appearance are dependent on its location within the retina. The dot/blot HMs are located in the outer plexiform layer and are common in DR. Flame-shaped HMs are located in the posterior pole and usually resolve in a short period of around six weeks. HMs have a less distinct border and variable size: small to large. Automated HM detection is not clearly documented as automated MA detection, apart from dot HMs which are normally identified in the automated MA detection of FI due to its resemblance. Quantitative analysis binds them according to an approach used for HM detection.

4.1 Morphological Processing

In Fleming et al. [4], blot hemorrhage candidate lesions were detected by examining maximal of multiple linear top hats (MMLTH) applied on the transposed image. MMLTH detected larger blobs at several scales by replicating at several structuring elements. The vascular candidates were identified with threshold and skeletonizing approach. The SVM classifier reported a sensitivity and specificity of 98.60 and 95.50% for blot HM detection, respectively.

4.2 Filters

A three-phase red lesion detection method was proposed by Marino et al. [14]. In the first phase, correlation filters with several resolutions were applied on green channel to get the primary candidate lesions. These primary candidates are inputted to second phase; region growing segmentation process rejects the points which does not satisfy size criteria of red lesion. Lastly, the first-phase candidates are sieved by four filters: shape, intensity, correlation mean, and finally, the last filter which discards regions on the blood vasculature by ridge lines. A sensitivity of 78.5% was reported by this method. Limitation of this method was that it was not tested on a larger image set and more validation was required.

4.3 Miscellaneous Approaches

Tang et al. [22] proposed a splat-based feature classification algorithm for detecting HMs in CFI. The aggregated and non-aggregated splat features were extracted for detecting HMs. The proposed system was able to mark various lesion shapes efficiently regardless of their texture and size and was able to extract large and irregular HMs. Image-level HM detection achieved an area under the curve of 0.87 for the MESSIDOR dataset.

Manjaramkar et al. [12] proposed an connected component clustering technique using maximally stable extremal regions (MSER) to detect incidences of HMs of varied characteristics from fundus image. The proposed technique has three prominent phases: HM candidate generation, feature extraction, and HM detection. The large, blot, flame, vessel-connected HMs were accurately detected by proposed technique. The evaluation on DIARETDB1 and MESSIDOR dataset reported best sensitivity and specificity of 96.45 and 97.64 at image level, respectively.

5 Retinal Image Databases

Many researchers in red lesion detection have used images originating either from screening programs, local eye hospitals, or originating from their own sources. Few researchers have used publicly available retinal image databases. These publicly available databases have become a precious tool for researchers. They provide necessary data for testing new methods and allow quantitative performance comparisons between existing approaches. Some online available retinal image databases are DRIVE, STARE, DIARETDB0, DIARETDB1, MESSIDOR, ROC. Table 1 gives comparison of online obtainable retinal image standard datasets.

Table 1 Comparative analysis of online obtainable retinal image standard datasets

Dataset	Blood vasculature	MA	HE	Image format	No. of images in dataset
STARE [7]	√			PPM	81
DRIVE [21]	√			TIFF	40
DIARETDB1 [9]		√	√	PNG	89
MESSIDOR [3]				TIFF	1200
ROC [15]	√			JPEG	100

√ Ground truth available

6 Performance Measures

The evaluation of the algorithm is done by using several performance measures. Sensitivity (SN), specificity (SP), accuracy, and receiver operating characteristics are the most used ones in red lesion detection systems. Sensitivity is the probability that a test will indicate disease among those with the disease, and specificity is a measure of those without the disease will have a negative test result. High values of SN and SP are considered better for diagnosis. SN and SP can be given as: Sensitivity (SN) = TP/TP + FN, Specificity (SP) = TN/TN + FP, and Accuracy (ACC) = (TN+ TP)/(TN + TP + FN + FP), where true positive (TP) measures the proportion of abnormal FI correctly identified as abnormal, true negative (TN) measures the proportion of normal FI found as normal, false positive (FP) measures the proportion of normal FI found as abnormal, and false negative (FN) measures the proportion of abnormal FI found as normal.

7 Comparison of Red Lesion Detection Algorithms

This section summarizes and compares experimental results of red lesion detection algorithms. Their comparison in terms of the algorithm used, a number of images used for evaluation, dataset used, and sensitivity and specificity are summarized in Table 2 and are bound with approaches used in the detection. The graphical comparison of sensitivity and specificity is shown in Fig. 4.

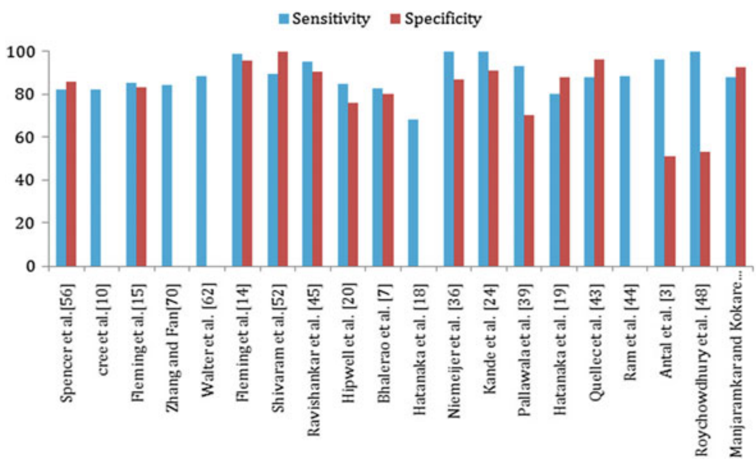


Fig. 4 Graphical comparison of sensitivity and specificity

Table 2 Comparative summary of algorithms for red lesion detection

Reference	Data source	Images used	Sensitivity	Specificity
Mathematical morphology				
Spencer et al. [20]	Digitized FA	–	82	86
Fleming et al. [5]	Grampian diabetes retinal screening program	1677	85.4	83.1
Walter et al. [24]	Color FI (pupil dilated)	115	88.5	–
Filtering				
Zhang et al. [25]	ROC and DIARETDB1	111	–	–
Pixel classification				
Niemeijer et al. [16]	ROC	–	100	87
Kande et al. [8]	Regional hospital, STARE, DIARETDB0, DIARETDB1 databases	89	100	91
Miscellaneous				
Quellec et al. [17]	Green channel fundus	995	87.90	96.20
Ram et al. [18]	DIARETDB1, ROC and custom-dataset CRIAS	427	88.46	–
Antal et al. [1]	ROC, DIARETDB1 2.1 and private eye hospital database	258	96	51
Manjaramkar and Kokare [13]	DIARETDB1, STARE, Messidor, ROC[a]	1726	88.09	92.65
Manjaramkar and Kokare [11]	DIARETDB1	89	92.9	99.99

[a] train set only

8 Conclusions and Future Research Directions

The presence of red lesions in fundus images is early sign of DR. Several automated DR detection algorithms have been developed and proposed by researchers in this field. But none of them has been found to be accurate and effective for clinical use, indicating room for improvement in the performance of automated red lesion detection system. This paper can be a resource for new researchers who are interested in the automated detection of a red lesion in retinal images. This paper will help the researcher in developing more accurate and efficient algorithms. It can be concluded from this overview that automated red lesion detection system has a huge potential for wider acceptance in real life.

References

1. Antal, B., Hajdu, A.: An ensemble-based system for microaneurysm detection and diabetic retinopathy grading. IEEE Trans. Biomed. Eng. **59**(6), 1720–1726 (2012)
2. Baudoin, C., Lay, B., Klein, J.: Automatic detection of microaneurysms in diabetic fluorescein angiography. Revue d'épidémiologie et de santé publique **32**(3–4), 254–261 (1983)
3. Doe, R.: Messidor: methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology techno-vision project @ONLINE (Aug 2015). http://messidor.crihan.fr/
4. Fleming, A., Goatman, K., Williams, G., Philip, S., Sharp, P., Olson, J.: Automated detection of blot haemorrhages as a sign of referable diabetic retinopathy. In: Proc. Medical Image Understanding and Analysis (2008)
5. Fleming, A., Philip, S., Goatman, K., Olson, J., Sharp, P.: Automated microaneurysm detection using local contrast normalization and local vessel detection. IEEE Trans. Med. Imaging **25**(9), 1223–1232 (2006)
6. Hipwell, J., Strachan, F., Olson, J., McHardy, K., Sharp, P., Forrester, J.: Automated detection of microaneurysms in digital red-free photographs: a diabetic retinopathy screening tool. Diabet. Med. **17**(8), 588–594 (2000)
7. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Trans. Med. Imaging **19**(3), 203–210 (2000)
8. Kande, G.B., Savithri, T.S., Subbaiah, P.V.: Automatic detection of microaneurysms and hemorrhages in digital fundus images. J. Digit. Imaging **23**(4), 430–437 (2010)
9. Kauppi, T., Kalesnykiene, V., Kamarainen, J.K., Lensu, L., Sorri, I., Raninen, A., Voutilainen, R., Uusitalo, H., Kälviäinen, H., Pietilä, J.: The diaretdb1 diabetic retinopathy database and evaluation protocol. In: BMVC, pp. 1–10 (2007)
10. Manjaramkar, A., Kokare, M.: A rule based expert system for microaneurysm detection in digital fundus images. In: 2016 International conference on computational techniques in information and communication technologies (ICCTICT), pp. 137–140. IEEE (2016)
11. Manjaramkar, A., Kokare, M.: Decision trees for microaneurysms detection in color fundus images. In: 2017 International conference on innovations in green energy and healthcare technologies (IGEHT), pp. 1–4. IEEE (2017)
12. Manjaramkar, A., Kokare, M.: Connected component clustering based hemorrhage detection in color fundus images. Int. J. Intell. Eng. Syst. **11**(2), 143–151 (2018)
13. Manjaramkar, A., Kokare, M.: Statistical geometrical features for microaneurysm detection. J. Digit. Imaging **31**(2), 224–234 (2018)
14. Marino, C., Ares, E., Penedo, M., Ortega, M., Barreira, N., Gomez-Ulla, F.: Automated three stage red lesions detection in digital color fundus images. WSEAS Trans. Comput. **7**(4), 207–215 (2008)
15. Niemeijer, M., Van Ginneken, B., Cree, M.J., Mizutani, A., Quellec, G., Sánchez, C., Zhang, B., Hornero, R., Lamard, M., Muramatsu, C., et al.: Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. IEEE Trans. Med. Imaging **29**(1), 185–195 (2010)
16. Niemeijer, M., Van Ginneken, B., Staal, J., Suttorp-Schulten, M.S., Abràmoff, M.D.: Automatic detection of red lesions in digital color fundus photographs. IEEE Trans. Med. Imaging **24**(5), 584–592 (2005)
17. Quellec, G., Lamard, M., Josselin, P.M., Cazuguel, G., Cochener, B., Roux, C.: Optimal wavelet transform for the detection of microaneurysms in retina photographs. IEEE Trans. Med. Imaging **27**(9), 1230–1241 (2008)
18. Ram, K., Joshi, G.D., Sivaswamy, J.: A successive clutter-rejection-based approach for early detection of diabetic retinopathy. IEEE Trans. Biomed. Eng. **58**(3), 664–673 (2011)
19. Seoud, L., Hurtut, T., Chelbi, J., Cheriet, F., Langlois, J.: Red lesion detection using dynamic shape features for diabetic retinopathy screening (2015)

20. Spencer, T., Olson, J.A., McHardy, K.C., Sharp, P.F., Forrester, J.V.: An image-processing strategy for the segmentation and quantification of microaneurysms in fluorescein angiograms of the ocular fundus. Comput. Biomed. Res. **29**(4), 284–302 (1996)
21. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M., Van Ginneken, B., et al.: Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging **23**(4), 501–509 (2004)
22. Tang, L., Niemeijer, M., Reinhardt, J.M., Garvin, M.K., Abràmoff, M.D.: Splat feature classification with application to retinal hemorrhage detection in fundus images. IEEE Trans. Med. Imaging **32**(2), 364–375 (2013)
23. Tavakoli, M., Shahri, R.P., Pourreza, H., Mehdizadeh, A., Banaee, T., Toosi, M.H.B.: A complementary method for automated detection of microaneurysms in fluorescein angiography fundus images to assess diabetic retinopathy. Pattern Recognit. **46**(10), 2740–2753 (2013)
24. Walter, T., Massin, P., Erginay, A., Ordonez, R., Jeulin, C., Klein, J.C.: Automatic detection of microaneurysms in color fundus images. Med. Image Anal. **11**(6), 555–566 (2007)
25. Zhang, B., Wu, X., You, J., Li, Q., Karray, F.: Detection of microaneurysms using multi-scale correlation coefficients. Pattern Recognit. **43**(6), 2237–2248 (2010)

Enhanced Performance of Android Application Using RecyclerView

N. Sabiyath Fatima, D. Steffy, D. Stella and S. Nandhini Devi

Abstract In android application, list view is most commonly used. List view is used to arrange the list of items in a sequence manner as scrollable items. List view consumes more memory and CPU usage. Because of this, performance of an android application is decreased. In order to increase the performance of the application, RecyclerView is used in the proposed android system. The main objective of this paper is to improve the performance of android application which handles large number of datasets using RecyclerView. The paper discusses the working functionality and performance analysis of RecyclerView technique using quiz application. RecyclerView used in this application is to list all quizzes to the user. The performance of RecyclerView is compared with ListView widgets based on various metrics like CPU usage and memory allocation for each item. The RecyclerView implementation results in the performance enhancement of an android application.

Keywords Android · ListView · Memory · RecyclerView

1 Introduction

Mobile apps are becoming more powerful and popular. There is no one in the world without mobile. Because of this reason, this application is developed as an android application. To conduct test manually, a lot of preparations are needed. This application reduces those preparations. The user can able to test his/her knowledge in the

N. Sabiyath Fatima (✉) · D. Steffy · D. Stella · S. Nandhini Devi
B.S.A. Crescent Institute of Science and Technology, Chennai, India
e-mail: sabiyathfathima@crescent.education

D. Steffy
e-mail: steffydavies29@gmail.com

D. Stella
e-mail: stelladavies29@gmail.com

S. Nandhini Devi
e-mail: Nandhini@crescent.education

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_17

specific domain by taking up the quizzes available in the application. They can select their own domain from the list available. They can update their profile details. All these domains and quizzes are loaded from the server. Since quizzes are loaded from the server, the application requires the Internet connection. Internet becomes a major part of this application. Nothing can be done without the use of the Internet. Here comes the network usage and performance of the application. So, this application should provide 100% performance with less usage of Internet. RecyclerView is used in this application to improve the performance of this application. This widget is used to list all quizzes that are loaded from the server. Because RecyclerView provides more performance than ListView.

2 Related Works

Mobile application development is one of the emerging trends in IT. Researchers have done a lot of work related to development of the mobile applications.

Wong et al. [1] proposed the interactive user-friendly voting and quiz system along with companion app which can be used in classrooms. This proposed system improves teaching and learning experience. This companion android application requires scanning the QR code which is displayed in Web system in order to access the questions which have been created by the teacher with the help of the system.

Meenakshi et al. [2] proposed an application that can be used for testing the knowledge level of the user, especially employee of the software company who needs to be specialized in single domain. This proposed application reduces paper work done by the teachers to prepare question paper or conducting examinations for the students.

Lakshmi et al. [3] discussed a cloud database which can be used for developing Web and mobile applications. Parse database is the cloud database which is most popular among other cloud databases. This service provides three products. Parse core handles saving of data and the integration of data. Push handles sending push notifications for the application. Parse analytics takes care of app data.

Kathuria et al. [4] discussed the three types of mobile applications: native, Web and hybrid. These apps can be downloaded from application store. These apps are developed particularly for specific platform. They will utilize all the features of that particular device. Native apps can be used in offline mode. Web apps are nothing but a Web site which looks like native application.

Gupta et al. [5] discussed the different components present in the android application framework. Activities, content providers and service are some o the components that receive and act on messages sent to all applications and messages.

Holla et al. [6] described an approach which can be followed while developing an android application. This approach is called layered approach. This approach can be used for Web development also. This work discussed the security measures related to android platform. Some of them are explicitly defined application permissions and application signing or trusting relationships.

3 Proposed Methodology

The proposed design improves the application performance. Because this application developed using RecyclerView reduces CPU usage. Because of this, performance of the application is increased. In this application, user can able to login using his/her credentials into application. User can able to update his/her profile details. User can able to take the quiz by choosing one of the domains from the list. All quizzes and domains are listed in the applications which are loaded from the server. User can able to save the quiz. He/she can able to resume the quiz whenever they want. There is a timer for each quizzes, and the user has to answer all questions within time. If the user failed to submit the quiz before time ends, then user will be prevented from taking up the quiz. The score will be displayed when he/she submits the quiz. The architecture of the proposed android application is shown in Fig. 1.

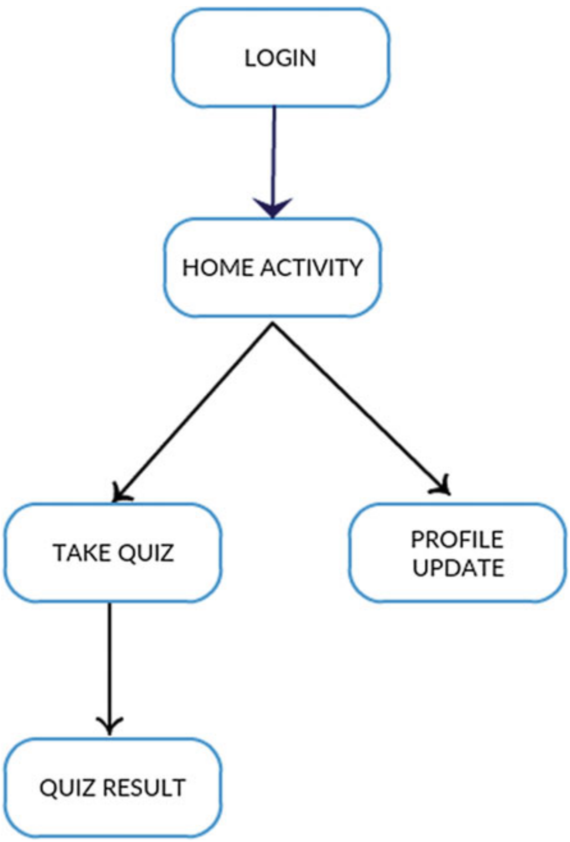


Fig. 1 Architecture of proposed android application

3.1 Application Model

See Fig. 1.

3.2 Application Overview

To ensure the maximum performance of the application, RecyclerView reusing algorithm is implemented in this proposed system. Initially, the user will first login into the system. His/her login credentials will be sent to the server after encoding the username and password. After successful login, the user can able to view the quizzes and domains that are loaded from the server in the home activity. The user has to select any one of the quiz from the list. Now, the user can able to answer the questions available in the quiz. User can able to save the quiz by clicking save button. He/she can able to resume the quiz y clicking the resume button. Timer gets started when user clicks start button. User has to submit the quiz before time ends. Otherwise, the quiz will get automatically submitted which will prevent the user from continuing the quiz. The score of the user gets displayed in the score page. Since all quizzes are loaded from the server, Internet connection is important. User can also able to update his profile details such as his name, email ID and mobile number. RecyclerView reusing algorithm is implemented in the proposed design. Because of this algorithm, application performance is increased.

4 RecyclerView Reusing Algorithm

The main idea of RecyclerView is to provide listing functionality. As name indicates, RecyclerView recycles its items or each view. This recycling process is done by the ViewHolder. Because of this ViewHolder pattern, there is no need to call findViewById method frequently. All references will be stored in memory. This significantly improves the performance. The two main entities of the RecyclerView are LayoutManager and ItemAnimator. LinearLayoutManager helps us to create vertical as well as horizontal list in android application. ItemAnimator helps us to add animations while adding or removing items from the view. Architecture of RecyclerView is shown in Fig. 2.

Advantages

- Supports animations
- Supports both horizontal and vertical views
- ViewHolder pattern improved performance of the application
- Supports both grid and list.

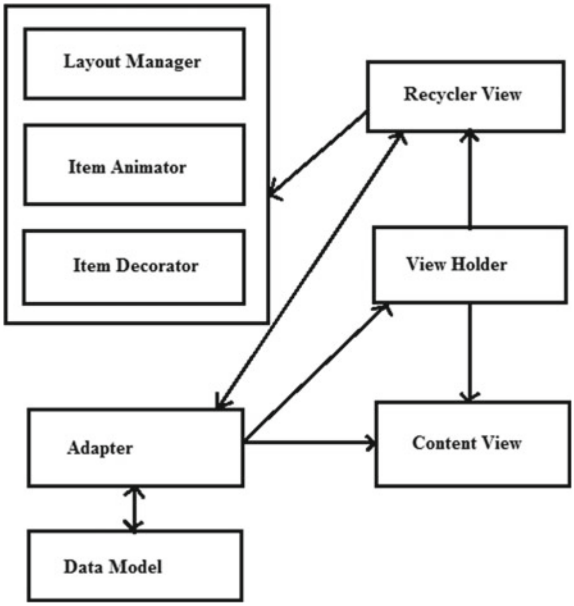


Fig. 2 Architecture of RecyclerView

Disadvantages

- There is no OnItemClickListener.

RecyclerView Architecture

See Fig. 2.

Architecture Components

LayoutManager
This component allows the developer to decide how items will be scrolled or arranged on the screen. There are three types of layout manager: (i) LinearLayoutManager, (ii) GridLayoutManager and (iii) StaggeredGridLayoutManager.

ItemAnimator

This component allows the developer to do animations in RecyclerView.

ItemDecorator

This component allows the developer to do sectioning in RecyclerView.

Adapter

This component presents items on the application screen which are pulled from data model. It is responsible for creating view for each item and replaces new content whenever the old content is unavailable.

ViewHolder

This component is a class which consists of references of the UI components whenever the system shows on the screen.

Algorithm

Step 1: Create custom Adapter class to display custom items in a view
Step 2: Inflate the item layout using getView method
Step 3: Create ViewHolder to store the reference to view in a memory
Step 4: Override the onCreateViewHolder method to inflate the view
Step 5: Override the onBindView method to bind the data to the view.
Step 6: Finally, ItemView will be generated.

The performance of this android application is computed based on memory allocation. The allocation of memory is calculated based on the number of items (N) and memory allocated (M).

Using RecyclerView, memory allocated is calculated as

N Items = M memory allocated

Consider a scenario in which the number of items is 20 and the memory allocated is 10. If the screen loads first ten items, allocated memory will be used for these ten items. Then, for the next ten items, same memory will be reused.

Using ListView, memory allocated is calculated as

N Items = $N * M$ memories allocated

Consider a scenario in which if allocated memory is 10 mb, then 10 * 100 mb will be allocated for 100 items.

4.1 System Implementation

This quiz application is designed and developed using XML and Java. The platform used to develop this application is Android Studio 3.0.1. In order to develop application, Java Development Kit and Android SDK are needed. Backend server is needed where the quizzes are stored and retrieved.

Various modules of the application are as follows:

Login Module

This module will discuss the login process. This application will allow the user to login via social media and their credentials. The login success message will be displayed to the user. Here, the user can able to reset their password via forgot password link. When user clicks the link, it will be redirected to forgot password screen. When the user types registered email ID, then reset password link will be sent to that email ID to change password.

Profile Module

Profile module allows the students to update their personal information and credentials. Basic details include first name, last name, profile name, primary email ID, secondary email ID, time zone, country, state and mobile number. Credential includes their account password.

User Module

This module will discuss the user activities. The user activities include the new quizzes, completed quizzes, incomplete quizzes and in-progress quizzes. In progress is nothing but the saved quizzes. Saved quizzes are displayed on their dashboard (home activity). Other quizzes including saved quizzes will be displayed in my quizzes section.

Assessment Module

This module will discuss the assessments. The assessment result will be displayed once assessment is completed. All quizzes will be loaded from the server. RecyclerView is implemented in this module. This RecyclerView reduces the CPU usage consumed by the android application. When user opens the quiz, the details of the quiz will be displayed to the user. When user clicks start button, the quiz will get started. When user swipes left, the next question will be displayed and when user swipes right, the previous question will be displayed. Timer will be enabled in all quizzes, and it will get started once quiz started. The user has to save or submit the quiz before the time ends. When user saves the quiz, it will be displayed in their dashboard. When user submits the quiz, the score will be displayed to the user immediately. The positive and negative messages will be displayed to the user based on user score.

5 Comparison Between RecyclerView and ListView

5.1 RecyclerView

RecyclerView is an advanced version of ListView. The main aim of RecyclerView is to provide listing functionality. This can be used when there is a large number of datasets.

5.2 ListView

ListView is used to display a list of items. Items will be inserted from Adapter which pulls content from array. This is subclass of AdapterView class. Architecture of ListView is represented in Fig. 3.

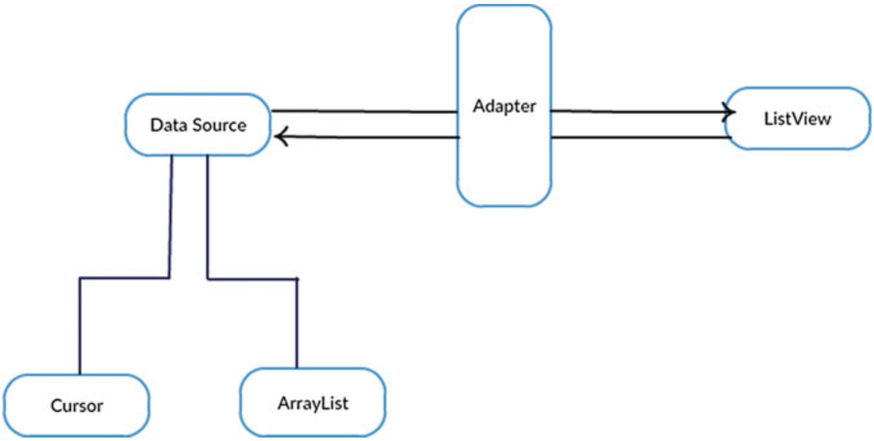


Fig. 3 Architecture of ListView

5.3 ListView Architecture

Adapter

Adapter in ListView is acting as a bridge between UI and the data source from which data is pulled to fill UI.

Data Source

This is a source where data will be retrieved. The source can be either array or ArrayList.

RecyclerView and ListView are compared using various characteristics such as ViewHolder, memory, view types, animations, ItemDecorator and performance. Comparison made between RecyclerView and ListView is shown in Table 1.

Table 1 RecyclerView versus ListView

Characteristics	RecyclerView	ListView
ViewHolder	This view has inbuilt ViewHolder	ViewHolder has to be implemented explicitly
Memory	This view consumes less memory	This view consumes more memory
View types	This view supports both vertical and horizontal list views	This view supports vertical list view
Animations.	Implementing animations is very easy	Implementing animations is very hard
ItemDecorator	This view has helper class like ItemDecorator	This view provides divider height parameter to add divider
Performance	Performance of the app is increased	Performance of the app is less

6 Results and Performance Analysis

6.1 Memory Usage

The performance of the application will be reduced if the application uses memory efficiently. If the application uses more memory, the mobile displays out of memory error. It is observed from the table that memory consumed by the ListView is 1 Mb more than the RecyclerView. RecyclerView consumes 3% less than the ListView. This is because memory allocated for each item will be reused by RecyclerView for all the items displayed in the view, whereas the memory will be allocated for all items separately. RecyclerView does not store invisible items in memory. Memory consumed by ListView and RecyclerView is are computed and shown in Table 2.

The graphical representation of Table 2 is shown in Fig. 4.

Table 2 Effect of varying memory usage for RecyclerView and ListView

Time (min)	RecyclerView (Mb)	ListView (Mb)
1	16	17
3	19	21
6	22	23
9	22	25
12	24	26

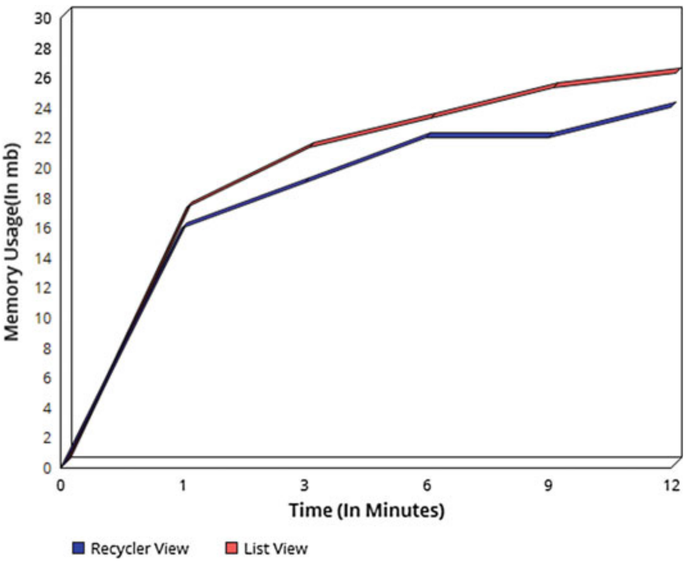


Fig. 4 Performance of android application in terms of time versus memory usage

6.2 CPU Usage

This metric is very important in case of mobile. Because more CPU consumption leads to more battery consumption which will reduce the battery power of the user mobile. It is observed from the table that CPU consumed by the ListView is 8% more than the RecyclerView. This is because RecyclerView will save CPU resources. View is stored in memory, so there is no need to inflate new views all the time. CPU consumed by ListView and RecyclerView values is computed and shown in Table 3.

The graphical representation of Table 3 is shown in Fig. 5.

Table 3 Effect of varying CPU usage for RecyclerView and ListView

Time (min)	RecyclerView (%)	ListView (%)
1	15	23
3	19	25
6	21	28
9	25	33
12	26	37

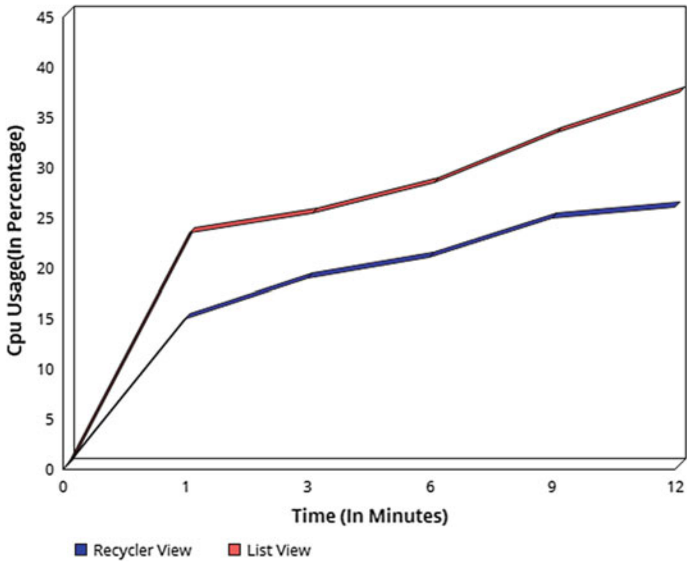


Fig. 5 Performance of android application in terms of time versus CPU usage

7 Conclusion and Future Work

In this paper, the proposed RecyclerView reusing algorithm is implemented using quiz application. The main objective of this algorithm is to improve the performance of the application. In this algorithm, instead of allocating different memory for each item, it reuses the allocated memory for all items that are displayed in the view. The computed results show that the RecyclerView produces more performance of the application than the ListView.

RecyclerView can be extended further by adding the OnItemClickListener method which will help the developers to easily write their own code for user click action on an item. This work can be further extended by considering various parameters such as network activity and efficiency.

References

1. Wong, A.K.Y., Cheok, S.-M., Tang, S.-M.: A user-friendly voting and quiz system for classroom use with connected devices. In: IEEE International conference on consumer electronics-China (ICCE-China) (2016)
2. Meenakshi, T., Salma Sulthana, S., Hema Sairam, G., Ayyappa Swami Y.: Quiz application. Int. J. Emerg. Trends Eng. Res. **4**, 259–262 (2016)
3. Lakshmi, B., Sai Durga, V., Anji Reddy, K.: Mobile quiz through wi-fi on android platform. Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) **4**(4) (2015)
4. Kathuria, A., Gupta, A.: Challenges in android application development: a case study. Int. J. Comput. Sci. Mob. Comput. IJCSMC **4**(5), 294–299 (2015)
5. Gupta, P., Kumar, M., Sharma, M.: Architecture for mobile quiz application using android application framework. IPASJ Int. J. Comput. Sci. (IIJCS) **2**(11) (2014)
6. Holla, S., Katti, M.M.: Android based mobile application development and its security. Int. J. Comput. Trends Technol. **3**(3), 486–490 (2012)

Personalized Web Search

Hema Sai Desu, Phalani Paladugu, Santoshi Sameera Adibhatla, Sushma Swaroopa Sorda and K. S. Sudeep

Abstract World Wide Web (WWW) serves as a platform for retrieving a variety of information. This high volume of information leads to the information overload on the Internet. When a query is made, most search engines return same results to all users without considering the context of the user's requests. For example, 'virus' in the context of one user can be malicious software, but for a biologist, it is a microorganism. This gives rise to the need for personalized web search. It is one of the growing concepts in web technologies. It is all about giving the right content to the right person at the right time. With the use of the user's browsing history, we can personalize the results and improve the user's visit experience. Through this paper, we emphasize on techniques to rank the pages in the order of relevance to the context of the user to achieve better results.

Keywords Web search · Personalization · Browsing history · User context · User profile · Browsing history · Rerank · Stop words · Url

H. S. Desu · P. Paladugu (✉) · S. S. Adibhatla · S. S. Sorda · K. S. Sudeep
National Institute of Technology Calicut, Kozhikode, India
e-mail: mailofphalani@gmail.com

H. S. Desu
e-mail: desuhema1996@gmail.com
URL: http://www.nitc.ac.in

S. S. Adibhatla
e-mail: adibhatlasameera@gmail.com

S. S. Sorda
e-mail: sushmaswaroopa@gmail.com

K. S. Sudeep
e-mail: sudeep@nitc.ac.in

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_18

1 Introduction

Personalization is a method to retrieve the most relevant documents from the large set of results based on the user's context. When different users make the same query, a typical search engine displays the same results irrelevant to the user's context. This may not be useful for users who require different information. Personalized web search is a favourable solution to solve this problem. With this technique, the search engine can intelligently treat specific queries differently based on the context and prioritize search results accordingly. In order to predict such information needs, many novel adaptive systems have been proposed which give more relevant information to user [1]. In these systems, user has to register personal information such as age, gender, location, and interests. Few systems also recommend based on user ratings and reviews. These types of registration, feedback, or ratings can become time consuming and users prefer easier methods. Therefore, through this project, we build a model, which creates user profile based on browsing history of user, and use this profile to give relevant results based on the user's context.

1.1 Literature Survey

Search engines return the same results to all users without considering information needs of the user. Liang [2] identifies that users may need information specific to him and the use of personalized web search could help them. Different approaches like Rocchio method, K-nearest method, and support vector machines have been used to build user profile among which K-nearest method was found to be better than others in terms of its efficiency. There are several search systems which provide users with relevant information. Sugiyama et al. [1] reviews some of them like hyperlink-based personalized web search, personalized websites, and recommender systems. In hyperlink-based personalized web search, hyperlink structure of web is used, and personalized page ranks are given which are better compared to global page rank algorithm. Link topology, the structure, and contents of web pages are often used in the construction of a personalized website. Recommender systems alleviate 'information overload' by providing recommendations with help of user ratings, feedback previously recorded, etc. These search systems have some shortcomings. Hyperlink-based personalized search systems do not clarify whether their search results actually satisfy each user's information need [1]. In personalized websites, the load on users becomes high. They have to manually register their preferences, and if their interests change, preferences have to be modified again. In recommender systems, users' ratings for items are key factors in achieving better recommendations. Since many users are unwilling to rate items, the accuracy of recommender systems might be poor [1]. Therefore, the search system should exactly capture each user's preference without any user effort, in order to provide more relevant information for each user effectively. Another approach to personalized web search was suggested by the students of Stanford University [3] using a point-wise approach. Point-wise algorithms typically rely on either regression or classification methods. This approach was based on the classification of the test URLs into one of the three classes—0 (missed /irrelevant/skipped), 1 (relevant), 2 (highly relevant). High Relevant: URL clicked and time spent on that is more than 300 units. Relevant: URL clicked and time spent is between 50 and 300 units. Irrelevant: URL clicked and time spent is less than 50 units or URLs missed/skipped [3]. The following algorithms [3] have been used to train the model to classify the URLs into the above classes:

(1) Random Forest.
(2) Gradient boosted trees.

1.2 Motivation

The information requirements are different for different user [1]. When the user performs a web search, traditional search engines give the results, in which both relevant as well as irrelevant information will be displayed together [1]. So the user takes time to find the result which is relevant to his/her requirements which reduce the user experience and thus arises the need for personalization. With the help of personalized web search, the user can make his search efficient and faster.

2 Problem Definition

When the user performs a web search, he/she may get some irrelevant results. Therefore, in order to reduce the wastage of time of the user, we need to personalize the results before displaying them. This can help the user to quickly find the information relevant to his/her query. In our proposed system, separate user profiles are used for each user, to retrieve the relevant information to address the query in his context. We take the user's browsing history as input, to make the personalized user profile. When the user enters a query in the user interface, his profile will be used to rerank the search results and display the personalized results as the output.

3 Design

In our implementation of personalized web search, we will use the concept of user profile construction and reranking of web pages [4]. The three main components in the design include user interface, servlets, and database.

3.1 User Interface

The main functionalities of the user interface include login and logout of the user, entry of search queries, and displaying the personalized search results to the user.

3.2 Servlets

In case of a request for dynamic web pages by a client, the server redirects the request to the web container present in it, which hosts the servlets. A servlet is a Java file which can take the request from the client, process the request, and provide the response in the form of an HTML page. The web container will have a deployment descriptor named web.xml which maps each request with its corresponding servlet. Some of the web containers are Apache Tomcat, Glassfish, JBoss, etc. (Fig. 1).

In our project, we used servlets to handle background processes like validation of user, redirecting search query given in the user interface to the search engine, reranking search results obtained from the search engine, storing session details of the user, extracting keywords from web pages and assigning weights to them, and storing keywords to the database for constructing the user profile.

3.3 Database

We primarily used two tables, one is USER_DETAILS table and the other is USER_KEYWORDS table. The logical schema of these tables is shown in Fig. 2. The attributes of USER_DETAILS

1.USER_ID(Integer)
2.USER_NAME(String)

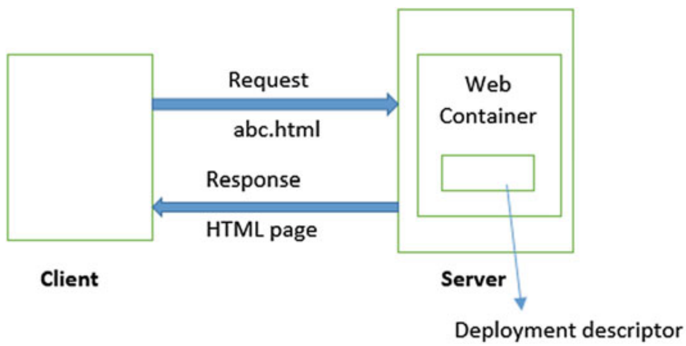


Fig. 1 Servlet

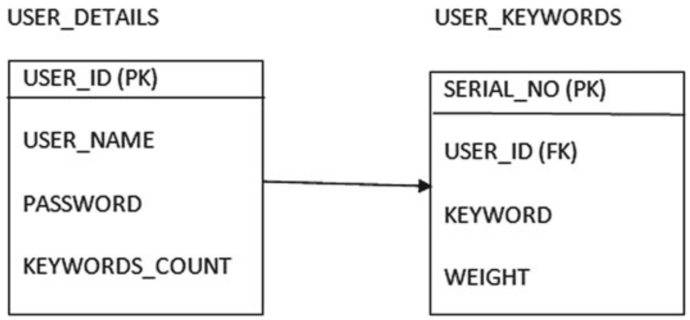


Fig. 2 Logical schema

3.PASSWORD(String)
4.KEYWORDS_COUNT(Integer)
During the creation of user account, each user will give a username and a password. This username will be stored in USER_NAME column, and password will be stored in PASSWORD column. USER_ID is a system-generated unique value given to each user. This acts as a primary key. KEYWORDS_COUNT contain the number of keywords present for each user in USER_KEYWORDS table.

The following are the attributes of USER_KEYWORDS table:
1.SERIAL_NO(Integer)
2.USER_ID(Integer)
3.KEYWORD(String)
4.WEIGHT(Integer)

SERIAL_NO is a system-generated unique value for each tuple. This acts as a primary key. USER_ID acts as a foreign key to the table. For each user, keywords are extracted from the web pages visited by him. Each keyword of a user is stored in a separate tuple along with its weight. Here, WEIGHT signifies the importance of a keyword (Fig. 2).

3.4 Steps in the Design Process

Login

The login page contains Login/SignUp options. If the user does not have an account, he has to register using the SignUp option. He has to provide a new user name and password. The uniqueness of username is tested, and these details are entered into the USER_DETAILS table. If the user already has an account, he can directly login using the username and password which were registered already. Once the user clicks on the login button, a code written in the corresponding servlet is executed to check the validity of these details by comparing these values against the tuples present in

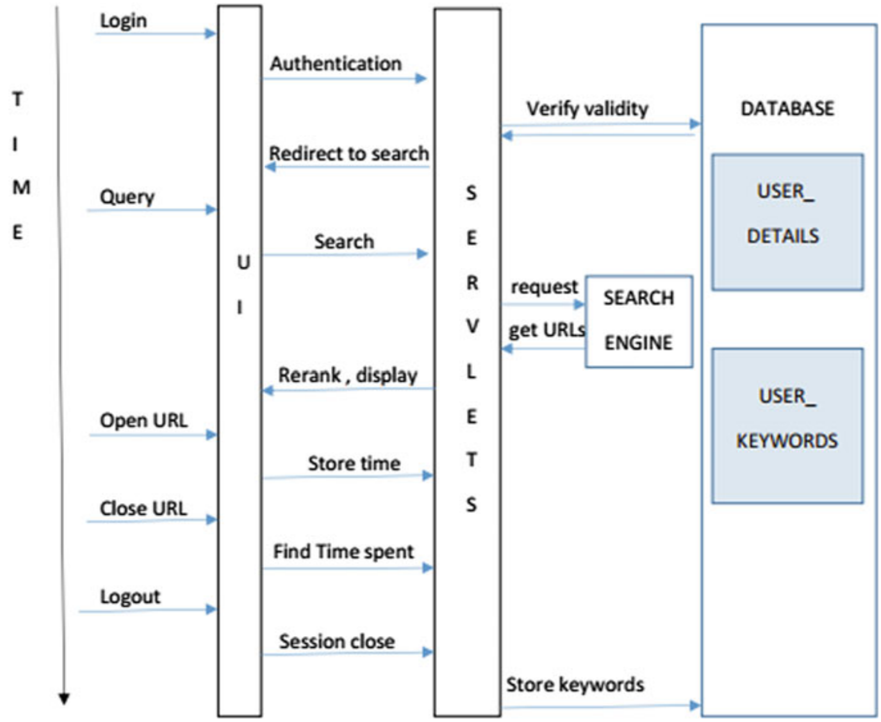


Fig. 3 Design diagram

the USER_DETAILS table. If the details provided by the user are valid, the login page will be redirected to the search page. At this point, a HttpSession is initiated. A session usually corresponds to one user, which persists for a specified time period. In this case, the session is closed once the user logs out. If the details are not valid, the login page is displayed again (Fig. 3).

Search

The user will be able to enter his search query in the user interface. This query will be redirected to a search engine with the help of a servlet. The results given by the search engine are reranked by using the user profile. These personalized results will be displayed back on the user interface.

Reranking

Reranking is done by using affinity function. Affinity function is calculated for each web page returned by the search engine, and these web pages will be ranked in the descending order of their affinity values. The affinity function [5] is a mathematical formulation for measurement of similarity between user interesting keywords in the user profile and words in search result pages. Let the web page for which we calculate the affinity function be W_j. From [5], it is evident that relevance and interest are the two important factors for calculating affinity function of each URL. The following equation calculates the affinity function for finding the similarity between search results and user profile. The simplest version of affinity function using these two factors as mentioned

$$\text{Affinity Function}(W_j) = \text{relevance} + \text{interest}$$

where relevance is the similarity between query word vector that is submitted to the search engine and words of the page W_j.

$$\text{Relevance} = \sum_{i=1}^{N} x_i$$

where N is the number of words in the query vector,

$x_i = 1$ if QWV_i is present in W_j

else $x_i = 0$.

QWV_i is the ith word in the query word vector (QWV).

Interest parameter is used to find the similarity between keywords in user profile and web page words.

$$\text{Interest}(W_j) = \sum_{i=1}^{N} x_i$$

where N is the number of keywords in the user profile

x_i = 1 if the word i in the user profile is present in the web page W_j

else $x_i = 0$.

We identified the algorithm for reranking as follows:

```
n -> no of web pages
    for < i= 1 to n> do
      affinity=Relevance+Interest;
    end for
Sort web pages{descending order of affinity}.
```

Browsing

The personalized results are displayed to the user, with explicit open and close buttons for each URL. The user has to click on the open button before visiting the actual URL. This is used to keep track of the opening time of a URL. To close the URL, the user has to click on an explicit close button. This is used to keep track of the closing time of a URL. At this point, the difference between the previously recorded time for the same URL, i.e. the opening time of the URL, and the current time, i.e. the closing time, is found out to calculate the time spent on that particular URL(i.e. web page) and it is stored. Also, the user will be prompted to rate the relevance of the URL's content to him. This procedure is repeated for all the URLs browsed by the user in that particular session.

Updating the User Profile on Logout

Once the user clicks on the Logout button, session has to be ended. But before ending the session, the user profile has to be updated using this session history. The updation procedure is as follows. Each URL browsed, along with the time spent, is considered. The content of each URL should be extracted, and it should be preprocessed to remove unimportant words which include articles, prepositions, conjunctions, etc [6]. For all the words obtained after preprocessing, frequency and term density [5] will be calculated. Let the web page which is currently being processed be denoted by W_i. Frequency represents the number of occurrences of the word j in W_i. Term density of the word j is obtained by dividing its frequency with the total word count of W_i. Total Word Count of W_i is the sum of number of occurrences of all terms in W_i after preprocessing.

$$\text{TermDensity} = \frac{\text{Frequency}}{\text{TotalWordCount}}$$

As the time spent on the web page also plays an important role in identifying the relevance of the web page to the user, this factor should also be considered. But the time spent which is stored cannot be used as such, since the length of the web pages varies. If the length of the web page is more, automatically the time spent on it will be more. This does not mean that the web page is more relevant to the user. To overcome this problem, time spent is normalized [5], i.e.

$$\text{NormalizedTime}(W_i) = \frac{\text{ActualTimeSpent}}{\text{Size of } W_i}$$

where Size of W_i is the total count of all words in W_i before preprocessing.

Now, the weight of the keywords obtained after preprocessing has to be calculated. Weight is calculated with the help of weight function. We defined weight function as the product of term density and normalized time.

$$\text{WeightFunction}(j) = \text{TermDensity}(j) * \text{NormalizedTime}(W_i).$$

There will be a limit in the number of keywords that can be present in the USER_KEYWORDS table for each user. We set the limit as 40. Now, extract all the keywords belonging to the particular user along with their weights from the USER_KEYWORDS table. Combine these keywords with the keywords obtained after preprocessing W_i and sort them in the descending order of their weights. Consider only top 40 keywords. First, all the keywords corresponding to the user must be deleted from the USER_KEYWORDS table. Then, the top 40 keywords from the sorted keyword list should be inserted back into the table. The updated user profile can be used for reranking in future.

We identified the algorithm for updating the user profile in the following way.

```
n->no of web pages
list1->null
list2-> null
for <i= 1 to n> do
     Calculate Normalized time
     k->no of words in web page i
     for <j= 1 to k> do
           Calculate WeightFunction
           list1.Add(keyword,weight)
     end for
end for
list2.AddFromDatabase(keyword,weight)
Append list2 to list1
Sort keywords{descending order of weight}
Insert the top 40 keywords into the database.
```

The overall algorithm for personalization is summed up as follows:

```
Login
Search
  Reranking of search results
  Browse
Logout
  Updation of user profile
  Session invalidation.
```

3.5 Result Analysis

In the screenshots above, we displayed the search results of two different users. One user is a biology student, and another user is a computer science student. When both of them searched for the word ‘worm’, from Fig. 4, we can infer that the biology student got all the results related to biological worms, and from Fig. 5, we can see that the computer science student gets the results related to computer worm (a type of malware) at the top.

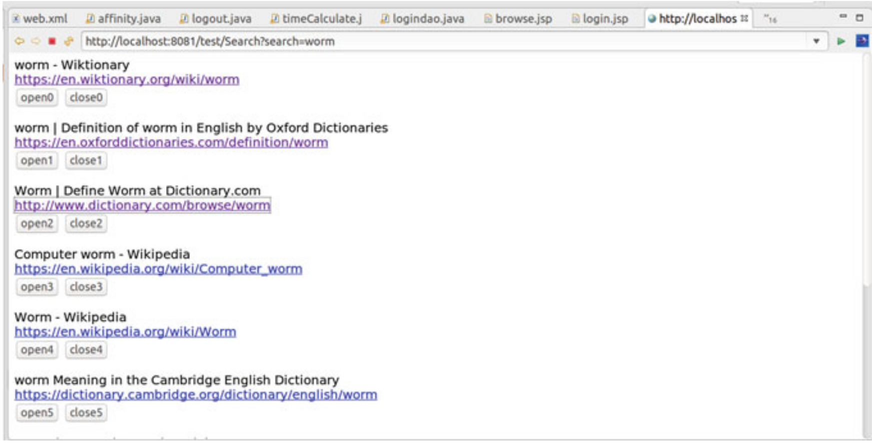


Fig. 4 Our search results for biology student

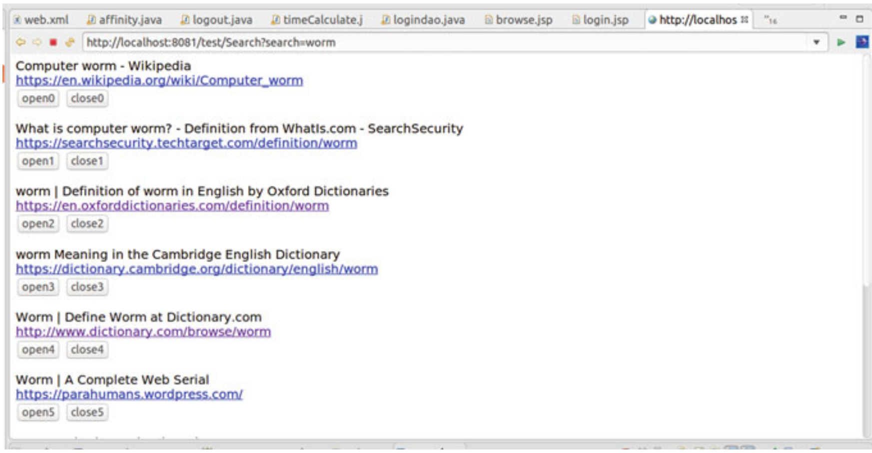


Fig. 5 Our search results for computer science student

4 Conclusion

From the results obtained, it can be concluded that the user's browsing history plays a significant role in personalization. The factors considered for calculating interest of the user like time spent on each URL and frequency of each word present in the web page have a major impact on the construction of user profile.

5 Future Work

User profile construction can be made more efficient by using natural language processing techniques for keyword extraction. Interface can be made more user-friendly. Currently, our system reranks only the results of the first page returned by Google search engine. This can be further extended to reranking the results of all the other pages returned. More measures can be taken to improve the security of the system.

References

1. Sugiyama, K., Hatano, K., Yoshikawa, M.: Adaptive web search based on user profile constructed without any effort from users. In: Proceedings of the 13th International Conference on World Wide Web, pp. 675–684 (2004)
2. Liang, C.: User profile for personalized web search. In: International Conference on Fuzzy Systems and Knowledge Discovery, pp. 1847–1850 (2011)
3. Mavilodan, D., Jaisinghani, K., Bansal, R.:Personalized Web Search
4. Joshi, C., Jaiswal, T., Gaur, H.: An overview study of personalized web search. Int. J. Sci. Res. Publ. 3 (2013)
5. Bhosale, M., Jagdale, N., Gawali, S., Dhotre, V.: Personalised web search using browsing history and domain knowledge
6. ranks.nl. Stopwords. https://www.ranks.nl/stopwords. Accessed 19 Feb 2018

Cooperative Game Theory-Based Request Distribution Model in Federated Cloud Environment

Sahil Kansal, Harish Kumar and Sakshi Kaushal

Abstract Cloud computing has emerged as the most important paradigm of distributed computing in the last decade. Increasing demand of the cloud users and inadequacy of resources along with the cloud providers has encouraged the federation of multiple cloud providers. In federation, multiple rational providers coordinate for the execution of the user's tasks. In federated cloud, the distribution of the user's request for resources and revenue among the providers has evolved as the most challenging task. Existing works aim at the profit maximization of the providers while discounting the importance of fair distribution. This paper presents cooperative game theoretic model for the fair distribution of revenue and resources among the providers, and analyzes the variation of power index with the change in the user's request. For analysis of the proposed model, comparative evaluation is performed with the existing distribution model. An experimental result demonstrates that the proposed model distributes the resource request and revenue more fairly that the existing state of art.

Keywords Federated cloud · Cooperative game theory · Shapley Shubik Power Index · Request distribution

1 Introduction

Cloud computing is distributed computing service model where the resources like memory, storage, processing power and bandwidth are provided as services to the end users [1]. Owing to its significant resource, provisioning features like on-demand

S. Kansal (✉) · H. Kumar · S. Kaushal
University Institute of Engineering and Technology, Panjab University, Chandigarh 160014, India
e-mail: sahil.kansal12@gmail.com

H. Kumar
e-mail: harishk@pu.ac.in

S. Kaushal
e-mail: sakshi@pu.ac.in

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_19

self-service, rapid elasticity, pay-as-use made it widely accepted by the organizations globally. Increasing demand of the consumers makes it difficult for the single providers to meet all the resource requirements. To satisfy the user's request, providers are coordinated with other providers to form the pool of shared unlimited resources [2]. These different interconnected clouds are federated clouds that cooperate with each other to fulfill the extensive requirements of the consumers. It gives an opportunity to the cloud providers to utilize their unused resources by forming the union of cloud resources and generating the addition revenue to the cloud providers.

In federated cloud, each provider possesses its own pricing and provisioning policies. These policies vary from one cloud provider to another provider. The providers in federated cloud are ready to share their resources at lower prices than their on-demand and reserved service instances. For every user's request that is arriving at the federated cloud, the major issue lies in the distribution of request for resources and the revenue among the participating providers [3–5]. The distribution is required to be fair and stable such that the providers are incentivized to contribute in the federation and do not like to break the coalition. This paper presents the game theoretic framework that aims for the fair distribution of the user's resource request and revenue among the providers in federation.

Different pricing and distribution models have been proposed by different researchers. Marian and Yong present resource pricing algorithm for multiple resources types. The author proposed mechanism design that follow different properties like individual rationality, incentive compatibility. These design incentives the providers to participate in the federated cloud [6, 7]. Different price sharing methods, i.e., (i) α-based sharing (ii) proportional revenue sharing and (iii) Shapley value, are presented by Zant, Amigo and Gagnaire in [5]. Authors had accessed the stability of these models by comparing it with standalone single cloud provider, but in the cooperative game itself, the providers tend to form the small coalitions, i.e., coalition structure, and the situation arises that instead of forming grand collation providers inclined to form smaller stable coalitions, thus forming stable CS-core. Mashayekhy, Nejad and Grosu proposed profit maximization cooperative game theoretic model where the cloud providers tend to dynamically form the coalitions to maximize their profits [8]. Hassan, Al-Wadud and Fortino present a socially optimal resource and revenue sharing mechanism that intends to form the federation of the providers so as to maximize the total profit of the providers and to promote the individual's profit. The authors have used the shadow pricing to govern the prices of the virtual machines (VMs) and forming the coalition of only those providers that will generate more profit [4]. But in the long term, the user's requirements sometimes exceed the available resources with the small coalition and required to form the bigger coalitions. In the existing works, the authors have considered the profit as an incentive for the providers to form federation. Our work aims at the fair distribution of the user's request such that all the providers in the federation get the chance to serve the user's part of request instead of only some providers forming the coalition and serve the user's request. As in the cloud federation, the providers get the part of marginal contribution from the overall profit. So, it is very essential that the distribution should be fair. The major contribution in this paper can be summarized as follows:

- Cooperative game theoretic model is proposed for the distribution of the user's request and revenue in the federated cloud computing environment such that the distribution is fair and cloud providers remain incentivized to be part of federation.
- It is showed that the federated cloud under the given assumption satisfies certain conditions to provide fair game theoretic solution.
- The variation of the behavior of the proposed solution is analyzed under variable user's request, and comparative analysis is done with the proportional request distribution.

The rest of the paper is organized as the following. Section 2 presents the problem statement and the proposed model, followed by experimental results and discussion in Sect. 3. Section 4 concludes the work with future directions.

2 Request Distribution Model

2.1 Federated Cloud

Federated cloud is the pool of interconnected cloud where the single provider can extend the accessibility of resources of another cloud. As the number of organizations and users are moving on cloud due to the agility and pay-per-use characteristics of cloud computing, single provider lacks the availability of the resources to meet the increasing requirement. So, multiple cloud providers are sharing their resources to satisfy the user's resource requests. Providers in federation share controlled agreement among each other. The cloud request broker governs the distribution of the user's request and the revenue among the provider in the federation [3, 4]. The major challenge faced by the third-party broker in the federated cloud is the distribution of the user's request and corresponding revenue among the cloud providers such that the allocation is fair that will incentive the providers to work in federation. Different symbols used in this paper are presented in Table 1.

2.2 Problem Statement

Problem definition: Given the user's request $\mathrm{Ur}(q)$ and N cloud providers forming the federation, how to attain the fair allocation of the user's request among the providers such that the allocation achieves the equilibrium point. Given the request can be executed, if all the providers take the part in federation.

The use of grid computing and cloud computing is encouraged by the scientific and experimental applications that require high computation power. But presently,

Table 1 Symbols used

Notations	Description
Ur(R)	User's request
TR	Total resources
$v()$	Valuation function
w_i	Resources with provider i
q	Quantity of resources requested by the user
$R_{\max}$	Maximum resources that can be provided by the federation
x_i	Revenue portion earned by ith cloud provider

increasing demand for the cloud services makes it difficult for the single providers to meet the stochastic demand of the users and organizations. It is required that they combine and coordinate with multiple providers [6]. The federation of the cloud providers depends on the size of the various cloud providers and type of the application load, and providers tend to form the federation with the providers with the comparable sizes [7]. On-demand service instances are highly dynamic where the user's requirements fluctuate on hourly basis. For federation, multiple providers offer their resources, and user' demand fluctuates from R to Rmax. So, the challenge for the providers is how to allocate the user's request, i.e., Ur(q), among the providers such that every provider is getting some part of the resources, and the distribution is fair; i.e., providers are getting the profit according to their marginal contribution. The following presents the desired definition for the game theoretic modeling of the user's request distribution.

Definition 1 (*Monotone Game*) The cooperative game said to be monotone if $v(S) \leq v(T)$ for every pair of coalition $S, T \subseteq N$ where $S \subseteq T$; i.e., valuation of the providers increases by forming the coalition.

By ensuring definition 1, provider's value of the utility function increases by forming the federated cloud.

Definition 2 (*Unanimity Game*) The game is called unanimity game if $\sum_{i \in N} w_i \geq q$; i.e., the sum of resources provided by all the players in federated cloud must be greater than the required user's request.

The unanimity game is achieved when sum of all the resources, contained with all the multiple providers that are participating in the federated cloud, is greater than the user's request.

Definition 3 (*Individual Rationality*) The players in the game are rational, i.e., self-interested. In individual rationality mechanism, the utility function of the players increases by forming federation instead of avoiding it.

By ensuring definition 3, valuation of each provider's utility will increased and thus incentives the providers to form federation.

Definition 4 (*Efficient*) The game is efficient if the payment function involves all the players in the federation.

This states that whatever the payoff is obtained that is distributed among all the cloud providers that are participating in the cloud federation. All the providers will be involved in the revenue distribution.

$$\text{i.e.,} \quad \sum_{i \in N} x_i = v(N) \tag{1}$$

Definition 5 (*Simple Game*) The game is simple game if the game is unanimity game, monotone game and its utility function, i.e., valuation function, follows:

- 1 for winning coalition
- 0 for loosing coalition.

In the weighted voting game cooperative models, the players form small coalition with the players such that the user's requests Ur(q) are satisfied. The coalitions that can fulfill Ur for v(C) is 1, and the collation which cannot fulfill for v(C) is 0.

2.3 Proposed Model

This work is the extension of the provider selection model for federated cloud, presented in [9, 10], where the providers for forming the federation select the providers using NSGA-II and other multiple objective decision-making (MODM) techniques, i.e., SPEA2. The providers that can form the federation are selected using MODM methods and further the allocation of the user's request is done in federated cloud. But as the user's demand fluctuates, especially in case of on-demand service instances, the fair distribution of resource requested becomes difficult. This section presents modeling of the federated cloud distribution problem and corresponding cooperative game theoretical model.

Assumption 1 The user's request, i.e., Ur $> \frac{1}{2}$TR where TR is the total resources obtained by forming the federation.

Theorem 1 *The cooperative game in the federated cloud is simple game.*

Lemma 1 *The valuation of the utility function of the coalition in federated cloud is either 1 or 0.*

$$\begin{pmatrix} 1 \ if \ \sum \Pr s_i \geq \mathrm{Ur}_i \\ 0 \ if \ \sum \Pr s_i < \mathrm{Ur}_i \end{pmatrix}$$

Proof In federation, multiple providers coordinate to form federated cloud. Each provider is sharing certain amount of computation resources like $P1, P2, P3, P4, P5$

and $P6$ are the different providers forming the coalition. When the user's request Ur arrives, it can be fulfilled by the different combination of providers $\{P1, P3, P4\}$ or $\{P2, P5, P6\}$ and so on more combination of the providers. The valuation of the utility function is 1, when the request is fulfilled by the coalition. The value is 0 for the coalition that is not satisfying the user's request.

Lemma 2 *The cooperative game among the providers in the federated cloud is unanimity game.*

Proof Multiple providers form the federated cloud to fulfill the excessive demand of the resources [11, 12], and only those requests are accepted which can be executed by the federated cloud. Let $R1$, $R2$, $R3$, $R4$, $R5$ and $R6$ are the amount of resources provisioned by the providers $P1$, $P2$, $P3$, $P4$, $P5$ and $P6$, respectively. Ur (q) is the user's request where q is the resources required by the user. The user request is accepted by the federated cloud if following Eq. (2).

$$\sum_{i \in N} R_i \geq Q \tag{2}$$

Equation (2) follows that sum of all the resources contained by the federated cloud is more than the requested, thus holds the unanimity game.

Lemma 3 *The game in the federated cloud is monotone game.*

Proof Let C is the coalition of P*1* and $P2$, $v(P1)$ is valuation function of $P1$, $v(C)$ is the valuation function of C, and providers form the federated cloud to increase the utility of unused resources [13]; this implies $v(P1) < v(C)$ and $P1 \subseteq C$, thus holds monotone game.

The results of Lemmas 1, 2 and 3 prove the Theorem 1 that cooperative game in the federated cloud is a simple game.

Theorem 2 *The game in federated cloud has non-empty core.*

Proof From the Assumption 1, it is clear that the game does not contain any veto power player. From [13], this holds that the game consists of non-empty *Core* and does not have any stable *CS-core*.

From the Theorem 2, it is clear that the game is super-additive game, and the game forms the grand coalition [13]. Theorem 1 states that the cooperative game in the federated cloud is a simple game. From Theorem 1 and Theorem 2, the fair and stable request and revenue distribution solution of the problem is the power indexes, *i.e.,* Shapley–Shubik power index (SSPI).

Definition 6 (*Imputation*) The Core of the game follows imputation if the game is (i) efficient (Definition 1) and (ii) individual rationality (Definition 2).

As user's request in the federated cloud varies from 1 to the maximum amount of resources that can be provided by the federated cloud collectively. So, all the providers are required to participate in federation to forms grand federation. This implies that every user's request is being distributed among all the providers in federation; thus, the game is efficient. Every provider in the federation is intended toward personal benefit, and by joining federation, the providers intend to increase its own utility. Thus, the federated cloud follows imputation. So, fair and stable distribution solution is attained using SSPI.

SSPI is the estimation of the power of each player in the game. Here, in the federated cloud, SSPI measures the distribution of the user's request and the revenue generated among the participating cloud providers in the federation. It is the fair solution to the simple games in the cooperative game theoretical model [13]. In this model, the sets of all winning coalitions are considered along with the sequence of their arrival in the collation and look for the pivotal player. Pivotal player is a player whose departure from the winning coalition will leave the coalition a losing coalition. The count of times the player is pivotal is evaluated. The SSPI of each player is calculated as, SS_k, as mentioned in Eq. (3).

$$\mathrm{SS}_k(\mathrm{Ga}) = \frac{(C-1)!(N-C)![V(C) - V(C)\backslash\{k\}]}{N!} \tag{3}$$

On the basis of each provider's SS_k, each provider's contribution in the user's request and price is calculated as shown in Eq. (4) and (5).

$$\mathrm{Pr}_i = \mathrm{SS}_i \times \mathrm{Ur}(q) \tag{4}$$

$$P_i = \mathrm{SS}_i \times C_s \tag{5}$$

C_s is the total cost charged by the federated cloud from the consumer. P_i is each provider's part in the Cs. Pr_i is each provider's part in the user's request. The proposed solution can be understood effectively by the following example.

2.4 Evaluation Parameter

Fairness: The fairness in this work is defined as:

$$\forall x_i \in X; \sum_{i=1}^{N} x_i^r \geq q, \sum_{i=1}^{N} x_i^c = C_s \tag{6}$$

X is the set of all the providers in the federation, N is the total number of providers, x_i^r is resources that can be offered by the provider x_i and x_i^c is the revenue generated by the provider x_i. All cloud providers participating in the federation should get part

Table 2 Market share of different cloud providers

S. no.	Cloud provider	Initial market share	Percentage of market share	Provider's (VMs)
1	Rackspace	5.00	38.16	45
2	SoftLayer cloud servers	3.20	24.42	30
3	Savvis VPDC	2.00	15.26	18
4	Linode cloud hosting	1.60	12.21	15
5	Microsoft azure virtual	1.30	9.92	12

Table 3 Providers' share using SSPI

S. no.	Providers	SSPI
1	Rackspace	0.4500
2	SoftLayer cloud servers	0.2000
3	Savvis VPDC	0.1167
4	Linode cloud hosting	0.1167
5	Microsoft azure virtual	0.1167

of user's resource request to get executed and each provider should get the portion in the revenue according to their marginal contribution in the user's request.

Example Five multiple providers are considered in Table 2. The initial weights of these providers are taken as percentage of the amount of resources contributed by the providers in the cloud federation. These weights are taken from the benchmark reports cloud harmony reports [14], and providers' VM is the amount of VMS by each provider. Different types of VM instances are provided by the providers for the sake of simplicity; only large instances of the VMs [10] are considered.

Let the total amount of VMs with federated cloud are 120; then, the each provider's number of VMs is given in Table 2. Ur(q) be the user's request for the 65 VMs. Then, each provider's power index using SSPI is given in Table 3. Using these indexes, each provider's share in the user's request and in the price is calculated using Eqs. (4) and (5).

3 Analysis and Discussion

For analyzing the effectiveness of the proposed price distribution model using SSPI, unidimensional resource requirements of the user are considered where the resource provisioning is done in the form of VMs. Different configurations of VMs are offered by the cloud providers as given in [10]. In this scenario, large VMs are considered.

Federation of five cloud providers is taken as shown in Table 2 with each provider's share in the market. For the analysis, it is assumed that the total number of VMs is 120, and each provider's contribution of VMs in the federation is 45, 30, 18, 15 and 12 as shown in Table 2. Experiments are performed on system with configuration as: 4 GB RAM, 500 GB hard disk and i5 processor with 3.30 GHz processing speed. Computation is done using optimization computation toolbox in MATLAB 2012b.

For demonstrating the user's request and price distribution using SSPI, different user's requests for VMs are considered as given in Table 3, and each provider's share is calculated using SSPI and using proportion proportional distribution (PD). Each provider's share in the federation cloud for the different values of user's requirement of VMs is given in Table 3. Each provider's power index (PI) shows variable behavior. For the values of q, i.e., 65 and 70, the PI values remain same. But with more increase in the value of q, $P1$'s and $P2$'s SSPI increases but $P4$'s and $P5$'s SSPI value decreases as this is the cooperative game with the transferable of utility (TU). From Table 3, it can be seen that the variation of each provider's SSPI depends more on provider's contribution in the winning game than each provider's contribution in the federation.

In the proportional distribution (PD), the user's request and the price is distributed according to the marginal contribution of the providers in the federation [7]. There is no coordination among providers. The combination of providers that can fulfill the requirements of the users is selected, and request is distributed according to the provider's proportion in the federation. Table 4 shows the comparative analysis of the SSPI and PD and fluctuating demands of the users. The on-demand user's request is varying from 95 to 65 VMs. As a result, each provider's marginal contribution is also varying. Results show fairness analysis of SSPI and PD. Table 4 calculates the power or the share of each provider using the SSPI and PD. It is shown that the for the request of 65 VMs, providers $P1$ and $P1$ together can provide 70 VMs and corresponding share of providers $P1$ and $P2$ in the request, and price will be 57% (approx.) and 42%, respectively, using PD leaving no share for provider $P3$, $P4$ and $P5$. On the other hand for the same user's request of 65 VMs, using SSPI, the request is distributed is among all the five providers. The user's request and corresponding revenue is distributed among all the providers which will be calculated using Eqs. (4) and (5), thus providing more fairness and leaving no reason for the provider's to leave the federation. Similarly for the other requests as given in Table 4, the request distribution is fairer using SSPI than the distribution using PD.

4 Conclusion and Future Directions

The paper presents cooperative game theoretic model for the fair distribution of user's request and revenue among the cloud providers in the federated cloud. The providers coordinate to form the federation to meet the extensive requirements of the users. Providers have different contributions in federation, and user's request can be satisfied by the combination of the different providers in federation. In such scenario, distribution of user's request and corresponding revenue among the federated

Table 4 Provider's share using SSPI and using PD with user's requirement of VMs

User's request (VMs)	Providers									
	$P1$		$P2$		$P3$		$P4$		$P5$	
	SSPI	PD	SSPI	PD	SSPI	PD	SSPI	PD	SSPI	PD
95	0.3000	0.3478	0.3000	0.2608	0.3000	0.1565	0.0500	0.1304	0.0500	0.0000
90	0.4667	0.3478	0.2167	0.2608	0.1333	0.1565	0.1333	0.1304	0.0500	0.0000
85	0.5000	0.4545	0.2500	0.3636	0.0833	0.1584	0.0833	0.0000	0.0833	0.0000
80	0.5000	0.4545	0.2500	0.3636	0.0833	0.1584	0.0833	0.0000	0.0833	0.0000
75	0.4167	0.4545	0.2500	0.3636	0.1667	0.1584	0.0833	0.0000	0.0833	0.0000
70	0.4500	0.5714	0.2000	0.4215	0.1167	0.1167	0.1167	0.0000	0.1167	0.0000
65	0.4500	0.5714	0.2000	0.4215	0.1167	0.1167	0.1167	0.0000	0.1167	0.0000

providers is a challenging task. For the accomplishment of said constraint, this paper presents fair cooperative distribution game theoretic solution. Experimental results indicate the effectiveness of the proposed model in the fair distribution of the user's resource request and revenue among providers. As the future work, the efficiency of the proposed model can be compared with other fair distribution models under multiple scenarios and can be further analyzed by considering the different VMs configurations.

References

1. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Zaharia, M.: A view of cloud computing. Commun. ACM **53**(4), 50–58 (2010)
2. Liaqat, M., Chang, V., Gani, A., Ab Hamid, S.H., Toseef, M., Shoaib, U., Ali, R.L.: Federated cloud resource management: review and discussion. J. Netw. Comput. Appl. **77**, 87–105 (2017)
3. Niyato, D., Vasilakos, A.V., Kun, Z.: Resource and revenue sharing with coalition formation of cloud providers: game theoretic approach. In: 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), pp. 215–224. IEEE (2011)
4. Hassan, M.M., Al-Wadud, M.A., Fortino, G.: A socially optimal resource and revenue sharing mechanism in cloud federations. In: IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 620–625. IEEE (2015)
5. El Zant, B., Amigo, I., Gagnaire, M.: Game theoretic approaches for revenue sharing in federated cloud. In: IEEE 3rd International Conference Cloud Networking (CloudNet), pp. 300–306. IEEE (2014)
6. Mihailescu, M., Teo, Y.M.: Strategy-proof dynamic resource pricing of multiple resource types on federated clouds. In: International Conference on Algorithms and Architectures for Parallel Processing, pp. 337–350. Springer, Berlin, Heidelberg (2010)
7. Mihailescu, M., Teo, Y.M.: Dynamic resource pricing on federated clouds. In: Proceedings of the 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, pp. 513–517. IEEE (2010)
8. Mashayekhy, L., Nejad, M.M., Grosu, D.: Cloud federations in the sky: formation game and mechanism. IEEE Trans. Cloud Comput. **3**(1), 14–27 (2015)
9. Hassan, M.M., Song, B., Huh, E.N.: A market-oriented dynamic collaborative cloud services platform. Ann. Telecommun. Annales des Telecommun. **65**(11), 669–688 (2010)
10. Amazon Web Services, [Online]. Available: https://aws.amazon.com/pricing/. Accessed 8 Dec 2017
11. Truong-Huu, T., Tham, C.K.: A game-theoretic model for dynamic pricing and competition among cloud providers. In: IEEE/ACM 6th International Conference Utility and Cloud Computing (UCC), pp. 235–238. IEEE (2013)
12. Guazzone, M., Anglano, C. and Sereno, M.: A game-theoretic approach to coalition formation in green cloud federations. In: 2014 14th IEEE/ACM International Symposium Cluster, Cloud and Grid Computing (CCGrid), pp. 618–625. IEEE (2014)
13. Chalkiadakis, G., Elkind, E., Wooldridge, M.: Computational Aspects of Cooperative Game Theory. Morgan and Claypool Publisher, CA (2011)
14. Cloud Harmony: Gartner Company, [Online]. Available: https://cloudharmony.com/. Accessed 2 Oct 2017

A Monarch Butterfly Optimization Approach to Dynamic Task Scheduling

Chouhan Kumar Rath, Prasanti Biswal and Shashank Sekhar Suar

Abstract Parallel processing has been employed many years in high-performance computing. Parallel processing is mostly used because it provides concurrency, maximizes load balancing, and minimizes system idle time and execution time which depend on various computer architecture. Task scheduling is a parallel processing technique which allocates the tasks to various processors. Mapping of heterogeneous tasks to a heterogeneous processor or core dynamically in a distributed environment is one of the active area of research in the field of parallel computing system. Here in this paper, monarch butterfly optimization (MBO) [1], a nature-inspired metaheuristic algorithm is implemented for a task scheduling problem. A monarch, a North American butterfly is best known for its migratory behavior in the summer season. There are mainly two processes to get the best solution. First, using a migration operator, a new generation is created. Second, they update their position by using a butterfly adjusting operator. The fitness value is evaluated and updated the population with a higher fitness value to satisfy the objective of the problem. Minimizing the cost and time of the scheduling strategy is the main objective of the proposed work.

Keywords Task scheduling · Parallel processing · Monarch Butterfly Optimization (MBO) · Genetic Algorithm (GA) · Evolutionary computation

C. K. Rath (✉) · P. Biswal · S. S. Suar
Sambalpur University Institute of Information and Technology, Sambalpur, Odisha, India
e-mail: chouhanrath@gmail.com

P. Biswal
e-mail: prasantibiswal66@gmail.com

S. S. Suar
e-mail: shashank.suar99@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_20

1 Introduction

Heterogeneous processors compute heterogeneous tasks, but some issues arise in the sequential allocation process which is not rectified properly but handled in the parallel distributed system. Most of the time, the task scheduling problem is considered as an NP-hard problem [2]. Scheduling of dependent or independent tasks is a propitious approach to get the effectiveness of computational requirements. Dynamic task scheduling has several advantages, such as efficient computing, lower computational cost, lower communication cost, and so on. Hence, current research is going on the dynamic task scheduling problem.

To make an efficient schedule is a challenge because most of the time parallelism or the distributed system fails due to excess communications and improper allocation of resources. So, it needs to make a schedule that generates a scheduling strategy which utilizes the resources properly in minimum execution time.

Genetic algorithm is a most popular optimization technique among all heuristic search algorithm. In scheduling problem, it applies evolutionary strategies which allow the user to find a good solution in a large search space [3–5]. Modified GAs like critical path genetic algorithm (CPGA), task duplication genetic algorithm (TDGA) [6], and other hybridized GAs are also developed to get the most efficient result. For the applications of task scheduling, different kinds of GAs have been proposed with the idealization of the genetic evolution process [7]. Some other evolutionary and nature-inspired optimization techniques such as particle swarm optimization (PSO), ant colony optimization (ACO), and artificial neural network (ANN) are also developed for the task allocation problem. Some modification in these optimization techniques like parallel orthogonal particle swarm optimization (POPSO), lazy ant colony optimization (LACO), and radial basis functional neural network (RBFNN) is to provide the efficient scheduler [8–12].

Here, tasks are allocated dynamically which allows the new tasks to arrive dynamic time interval. Monarch butterfly optimization (MBO) is used to make efficient scheduling strategy. MBO is considered because it is based on the intelligence of swarm behavior. It creates the search space according to the migration behavior. To measure the effectiveness of the optimization technique, it is compared with the most popular genetic algorithm. This paper presents a comparative study of both optimizations with similarities and dissimilarities.

Organization of this paper: Sect. 2 describes the problem definition. Details of monarch butterfly optimization and genetic algorithm are discussed in Sect. 3. Section 4 shows the experimental analysis and simulation results, and conclusion is explained in Sect. 5.

2 Problem Definition

It is difficult to make a schedule in a distributed computing system to handle dependent or independent tasks. Various scheduling methods are existing, but here dynamic allocation method is applied, which allows deterministic execution. In

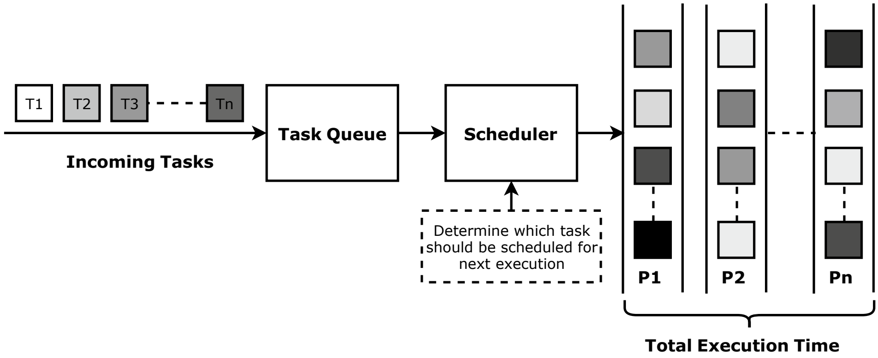


Fig. 1 Working model of dynamic task scheduling

dynamic scheduling, the processor does not have any prior knowledge about tasks to be executed. Figure 1 shows how the scheduling process works dynamically. The independent tasks ($T_1, T_2, \ldots, T_n$) are arriving and stored in a task queue which ensures that tasks are available in the pipeline and ready to execute. In the next stage, the scheduler determines which task should be scheduled for which processor. Processors are encountered with different computation cost and communication cost while tasks are executing. Computation cost is decided randomly based on the efficiency of the individual processor; whereas, the communication in between the processors is assumed as an identity. The scheduler provides the best scheduling strategy based on speedup, efficiency, complexity, and quality.

The objective function for the formulation of fitness value is

$$\sum_{i=1}^{n} U_i = \sum_{i=1}^{n} C_i + \sum_{i=1}^{n} M_i \tag{1}$$

where,
C_i = Communication cost of ith job
M_i = Make-span time of ith job
U_i = Utility function of ith job.

Here, the objective function refers to the utility function. The aim is to minimize the utility function subject to optimize the cost and time by scheduling the tasks to the appropriate processors.

$$\begin{aligned} \sum_{i=1}^{n} C_i &\le C_{\max} \\ \sum_{i=1}^{n} M_i &\le T_{\max} \end{aligned} \tag{2}$$

C_{max} = Maximum permissible cost.
T_{max} = Maximum permissible time.

3 Proposed Methodology

The task schedule process needs less execution time, and it should be adaptable in various environments. In this section, a swarm intelligence method monarch butterfly optimization (MBO) is applied. This section describes step by step procedure of MBO and how it is applied in the task scheduling problem.

3.1 *Monarch Butterfly Optimization (MBO)*

A monarch butterfly (a swarm from North America) is well known for its migratory behavior. It flies thousands of miles for reproduction. The female monarch butterfly lays eggs for the new generation during these movements.

MBO is an optimization technique based on the social behavior of the swarm (monarch butterfly) called swarm intelligence (SI) method [1]. The MBO is able to enhance the search space efficiently and able to identify the best optimal solution quickly because of the movements of the individual which are large step size. In MBO, each swarm is a solution in the search space called a population. According to Algorithm 1, the total population is located in two Land1 and Land2. Here, the migration behavior of the monarch is addressing the optimization technique. A new population is generated by using the migration operator in both lands. They update their positions using an adjusting operator. To keep the population size unchanged, the original population is replaced with a newly generated population having better fitness. In this way, the fittest monarch butterfly moves automatically to the next generation. Figure 2 shows the schematic presentation of MBO.

Protocols of MBO (Mapping with GA)

- Monarch butterflies are considered as tasks.
- Initial population can be generated by assigning the tasks to different processors randomly. The population is divided equally for both Land1 and Land2.
- The selection of monarch butterflies is based on the fitness value available in the current state.
- Migration operation is considered a crossover.
- The updating operator refers to the mutation operator.
- Discarding the population is based on the fitness criteria.

3.2 *Population Initialization*

In the proposed work, the scheduling strategy is based on make-span minimization. The make-span is the summation of completion time of each task and communication cost. So, the tasks are assigned to a processor that results in minimum execution time.

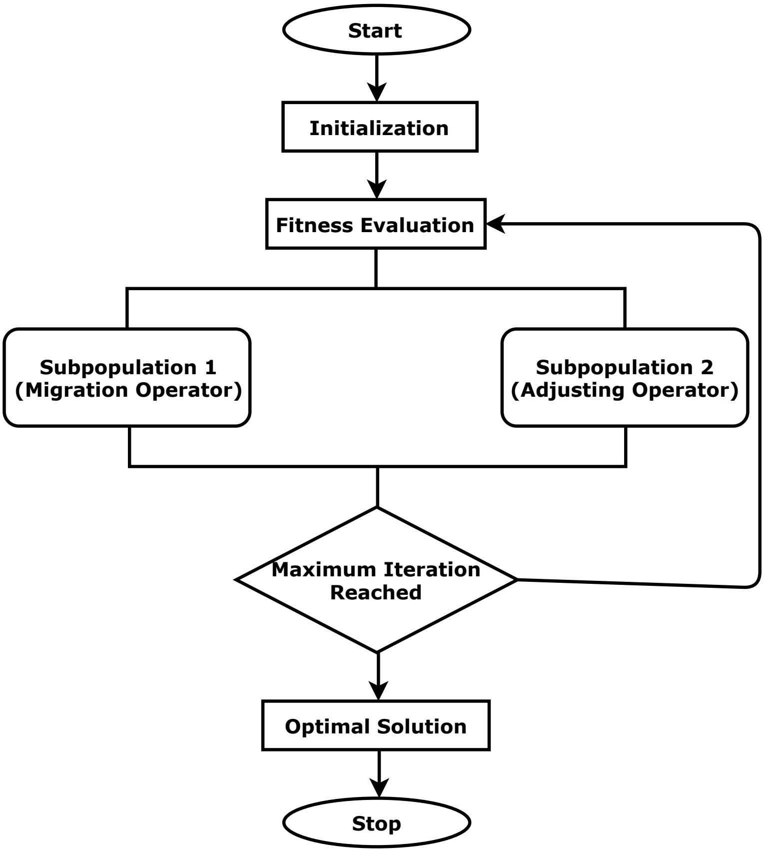


Fig. 2 Flowchart of monarch butterfly optimization

Algorithm 1: Monarch Butterfly Optimization

Step 1:Initialization;
Step 2:Fitness evaluation;
Step 3:**while** $i < MaxIteration$ **do**
    Sort all the butterflies by their fitness values.;
    Distribute over two networks (Land1 & Land2) according to Algorithm 3 & Algorithm 4;
Step 4: Combine both new generated populations with initial population.;
Step 5: Get the Optimal solution.;

Table 1 Example of initial population with ten tasks and four processors

Tasks	A	B	C	D	E	F	G	H	I	J
Butterfly 1	1	2	2	3	4	4	3	1	1	3
Butterfly 2	3	4	1	1	2	3	3	3	4	4
Butterfly 3	2	1	3	4	4	2	1	2	3	1
Butterfly 4	2	2	4	3	3	1	4	4	1	1

Table 1 is the example of the initial population considered as the positions of monarch butterflies. Randomly, the tasks are assigned on the basis of availability of processor because tasks are independent. For example, in butterfly 1 (Table 1), A is assigned to P1, B is assigned to P2, and so on (P1 and P2 are the processors). Computation cost is required if the new task is assigned to different processors. The fitness value is the sum of the total execution time and communication cost termed as make-span.

Algorithm 2 shows the distribution of random solutions over two different networks. According to their position, fitness value is evaluated using fitness function as in Eq. 1. Positions are updated step by step using Algorithm 4 until the objective function is satisfied. To keep the population unchanged, a new population generates by the migration operator and evaluates in both networks NW_1 and NW_2. They are migrated using shifting operator as defined in Algorithm 3.

Algorithm 2: Population Initialization

Data: Number of tasks, Number of processors, Communication cost, Computation cost
Assign all tasks with N number of processors;
Distribute tasks $J_i \in NW_1$ and $J_j \in NW_2$, Where $J_i + J_j = J$ Total number of tasks;
if $S_{i+1} < S_i$ **then**
  S_i is discarded from the population;
else
  S_i will move to the next generation and becomes static object with best fitness value;
end
NW_1 & NW_2 are Land1 and Land2 respectively.

3.3 Shifting Operator

As mentioned before, monarch butterflies choose two different lands to create their new generation. They generally migrate from Land1 to Land2 during the month of April to find a place to lay their eggs and return during the month of September. With the inspiration of this migration process, it observed that the monarch butterflies stay at Land1 for 5 months and Land 2 for 7 months. According to Algorithm 3, the distribution of the total population in Land1 and Land2 depends on 'r' value. Where

$r = m\times$ peri, 'peri' is the migration period, and 'm' is a random value drawn from a normal distribution. Using the migration ratio 'p,' the direction of the butterfly is decided and balanced their distribution. If 'p' is more, then most monarch butterflies are selected from Land1. If 'p' is less, then most monarch butterflies are selected from Land2. This indicates which subpopulation plays an important role to generate the new population.

Algorithm 3: Shifting Operator

```
for i=1 to J (for all monarch butterflies) do
    m=rand (randomly generate a number);
    r = m × peri;
    peri is the migration period.;
    if r ≤ p then
        Maximum elements from Land 1 will be selected (sub-population1);
    else
        Maximum elements from Land 2 will be selected (sub-population2);
```

3.4 Adjusting Operator

Monarch butterflies update their place by using the adjusting operator. It needs updating while rand $> BAR$, here BAR is the butterfly adjusting rate, and walk step is calculated by a levy flight function (a standard benchmark function). It also updates the walk step size by calculating the weight value α.

Algorithm 4: Adjusting Operator

```
for i=1 to J(for all monarch butterflies) do
    Calculate the walk step using Leavy Flight.;
    m=rand (randomly generate a number);
    if m ≤ p then
        Generate elements with best fitness value.;
    else
        Select a monarch butterfly from sub-population2 randomly;
        if rand > BAR then
            X_j^{t+1} = X_j^{t+1} + α + (Leavy(X_j^t) − 0.5);
            α =Step size calculation.;
            Bigger α =Long step and Smaller α =Short step;
```

3.5 Scheduling Using Genetic Algorithm

For optimization problems, the most popular genetic algorithm is used. It is famous for its natural selection and the process that drives biological evolution [4, 5]. GA uses possible solutions to create a population. Each solution is considered as a chromosome. This technique generates a new generation by evaluating their fitness using a fitness function. To create a new ideal solution in a search space, it uses the generic operator. Here, the procedure for dynamic task scheduling using GA is as follows:

1. Population initialization with randomly generated chromosomes.
2. Fitness value evaluation using a fitness function (objective function).
3. **Selection**: Select random population having the highest fitness value for the crossover and mutation process.
4. **Crossover**: Random point crossover is done to form a new offspring.
5. **Mutation**: Mutation process is applied for random changes to individual parents to form children.
6. Update the existing population with the new population based on the fittest chromosomes.
7. Step 2, 3, and 4 will continue until the most iteration value is reached.
8. Finally obtains the optimal solution.

4 Result and Analysis

In this section, the simulation result presents the performance evaluation of both algorithms. Here, make-span is the criteria to measure the efficiency. To measure the effectiveness of MBO, it compared with GA because both the algorithms nature-inspired and evolutionary optimization technique. For experimental analysis, the parameters are as follows:

- Size of tasks (J) = 20.
- Number of processors (N) = 5.
- Iteration = 100, 200, 300, 400, and 500.
- Computation cost is a random number in between 0 and 10.
- Communication cost for each processor is set at 2.
- Population size = 10, 20, 50, 100, and 200.

Parameters taken for both the optimization are same. Figure 3 shows a comparative analysis of the execution time in both GA and MBO with the observation that MBO gives better performance than GA. The fitness value is the sum of communication cost and computation cost. It is observed that the fitness value of MBO is much less than GA which satisfies the goal of the experiment. In GA, the fitness value varies from 110 to 180 where it varies from 70 to 120 in case of MBO.

Table 2 illustrates the effect of population size on average fitness value. For the experiment, the value of population size is 10, 20, 50, 100, and 200 and observed

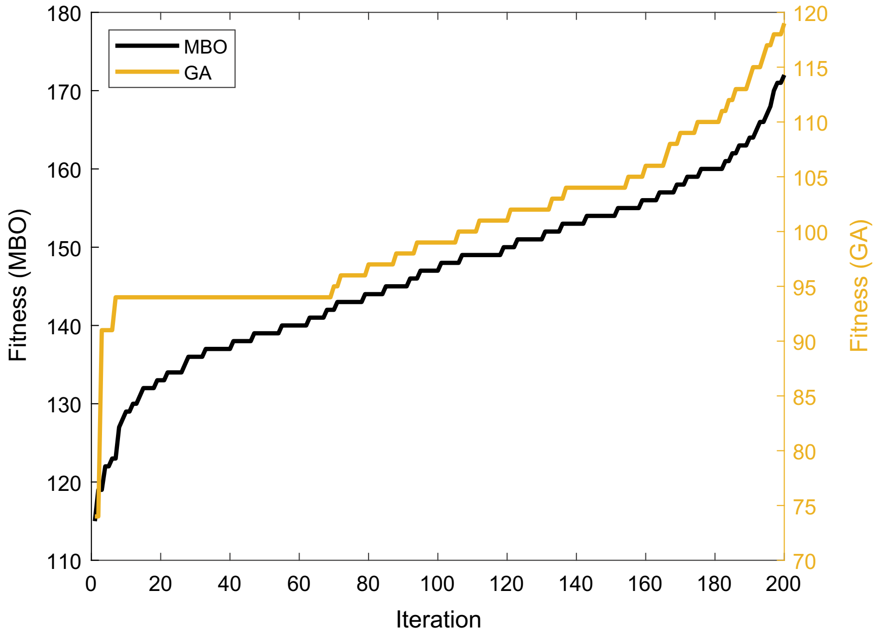


Fig. 3 Comparison of MBO and GA for the fitness value

Table 2 Effect of population size over make-span

Population size	Initial	MBO	GA
10	135.2000	116.9000	132.4000
20	134.4000	106.8500	130.9500
50	134.1000	98.6600	132.2400
100	124.6200	85.0700	122.0100
200	132.2050	76.9150	129.9500

that the results of MBO are continuously decreasing with the increase in population size for a fixed iteration time. Table 3 also indicates the average fitness value with different iteration time for a fixed population size 200. Hence, both the comparison tables (Tables 2 and 3) indicate that make-span is much more reduced by MBO.

5 Conclusion

This paper introduces the use of monarch butterfly optimization in task scheduling problem. Experimental results, for the scheduling problem, show that MBO gives better performance compared to GA. The shifting operator of MBO is acting like the

Table 3 Effect of maximum iteration over make-span

Iteration	Initial	MBO	GA
100	149.6650	121.9950	148.0450
200	149.6650	120.1500	146.3150
300	149.6650	105.9350	142.6400
400	149.6650	103.9250	141.7250
500	149.6650	102.8000	140.9200

crossover operator which enhances the search space. Similarly, updating operator is also performing like the mutation operator which leads to search a new scheduling strategy. Both algorithms are comparable because both are population-based, nature-inspired, and evolutionary metaheuristic optimization technique. For the future work, it may allow using a dependent task graph to test the optimization technique. More task scheduling problems such as load balancing, processor use, and resource allocation may consider for an effective scheduler. To do more efficiently, some hybrid techniques may embed with MBO.

References

1. Wang, G.G., Deb, S., Cui, Z.: Monarch butterfly optimization. In: Neural Computing and Applications (2015)
2. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman & Co., New York, NY (1979)
3. Page, A.J., Naughton, T.J.: Dynamic task scheduling using genetic algorithms for heterogeneous distributed computing. In: 19th IEEE International Parallel and Distributed Processing Symposium, pp. 189a–189a
4. Page, A.J., Keane, T.M., Naughton, T.J.: Multi-heuristic dynamic task allocation using genetic algorithms in a heterogeneous distributed system. J. Parallel Distrib. Comput. **70**(7), 758–766 (2010)
5. Holland, J.H.: Adaptation in Natural and Artificial Systems. MIT Press, Cambridge, MA, USA (1992)
6. Omara, F.A., Arafa, M.M.: Genetic algorithms for task scheduling problem. J. Parallel Distrib. Comput. **70**(1), 13–22 (2010)
7. Beyer, H., Schwefel, H.: Nat. Comput. Kluwer Academic Publishers, Dordrecht (2002)
8. Sivanandam, S.N., Visalakshi, P.: Dynamic task scheduling with load balancing using parallel orthogonal particle swarm optimisation. Int. J. Bio-Inspired Comput. **1**(4), 276–286 (2009)
9. Tripathy, B., Dash, S., Padhy, S.K.: Dynamic task scheduling using directed neural network. J. Parallel Distrib. Comput. **75**, 101–106 (2014)
10. Bakshi, T., Sarkar, B., Sanyal, S.K.: An evolutionary algorithm for multi-criteria resource constrained project scheduling problem based On PSO. Procedia Technol. **6**, 231–238 (2012)
11. Li, X., Zhang, J., Yin, M.: Animal migration optimization: an optimization algorithm inspired by animal migration behavior. Neural Comput. Appl. **24**(7–8), 1867–1877 (2014)
12. Tiwari, P.K., Vidyarthi, D.P.: Improved auto control ant colony optimization using lazy ant approach for grid scheduling problem. Future Gener. Comput. Syst. **60**, 78–89 (2016)

Automation of Traffic Violation Detection and Penalization

C. Valliyammai, J. Sridharan and A. Ramanathan

Abstract Violation of traffic rules in two-wheelers might risk the life of persons traveling. Finding the violation of traffic rules requires a large number of police personals at several locations. In this paper, we have proposed a system that directly finds traffic violations done by two-wheelers using trained convolutional neural networks (CNN) without any human assistance. CNN uses hidden layers to perform operations, which allows applying a wide variety of filters in each layer. CNN produces better accuracy when compared with traditional image processing techniques. Violated vehicle number plates are found using optical character recognition (OCR) techniques and penalized according to the violation.

Keywords Deep learning · Traffic violation · Video processing · CNN · OCR

1 Introduction

In today's world, road and transport have become an indispensable part for any person. The road crashes result in the loss of lakhs of lives and also major injuries to crores of people every year. According to a report published by the World Health Organization (WHO) in 2004, more than million people were killed and 50 million injured in traffic collisions on the roads around the world each year. In India, about eighty thousand people are killed in road crashes every year. In a survey conducted by Times of India, reports reveal that 98.6% of motorcyclists who died in accidents did not wear a helmet. These lives could have been saved if people have worn a helmet while they are traveling. Motorcycles are meant for only two to travel at a time, but

C. Valliyammai (✉) · J. Sridharan · A. Ramanathan
Department of Computer Technology, Madras Institute of Technology Campus, Anna University, Chennai, Tamil Nadu, India
e-mail: cva@annauniv.edu

J. Sridharan
e-mail: jjsridharan@gmail.com

A. Ramanathan
e-mail: ramanathanarun14@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_21

most of the times more than two are traveling in motorcycles which provides an imbalance to the rider of the motorcycle. Finding the mistakes done by motorcyclists and warning about the mistakes will help to reduce accidents to a great extent. Three major mistakes are not wearing a helmet, traveling with more than two persons and over speeding in motorcycles.

The proposed system makes use of CNN to find motorcycles, persons, helmets and license plate in a frame of the video. Once a violation of rules by the motorcyclist is found, then the corresponding motorcycle rider is identified using the rider's license plate number by OCR and penalizes the rider accordingly. This paper is organized as follows: Sect. 2 explores the literature survey which briefs the existing work in violation detection using traditional algorithms. Section 3 describes the proposed method with CNN algorithms. The implementation details and results are explored in Sect. 4. Section 5 provides the conclusion of the work carried out and future enhancements.

2 Literature Survey

2.1 Automatic Traffic Violation

A framework was proposed using light of radio-frequency identification (RFID) technology to manage the traffic problems [1]. The significant segments are a camera that is utilized to track and take photos of the vehicle, RFID tag, RFID reader and a database server which contains the points of interest of the red signal violators and utilizes GSM mechanism to send a notification. One of the drawbacks here is the expense of usage, each user needs to have an RFID tag, and RFID readers ought to be available in huge quantity; whereas in the proposed method, everything is done through processing of frames from live video. Another disadvantage of this technique is that it checks only on speed violation of vehicles and no other violation because it relies on RFID. This weakness was resolved by distinguishing infringement through video processing by utilizing two cameras, one of which is of high resolution for license plate imaging, and another, a low-resolution camera for processing of movement [2]. As it is self-evident, it almost doubles the cost at every place where the camera is to be installed. Another drawback here is that the background subtraction techniques were used for vehicle detection, which gives mediocre accuracy when compared with state-of-the-art techniques like CNN.

2.2 Object Detection

Region proposal technique provides a quality solution to solve the problems of redundant windows [3]. It makes good utilization of the object information present in an image, such as edge, color and texture, which ensures the detection algorithm has a relatively high recall even with a fewer number of windows. The quality of region proposals is higher when compared with sliding windows, and it decreases the time complexity of the detection process substantially.

Region convolutional neural network (R-CNN) generates 2000 candidate windows per image during inference. Each region will extract the features through the CNN, making the feature computation process time-consuming, approximately 50s per image. Spatial pyramid pooling network (SPP-Net) was proposed to overcome the existing deficiencies in R-CNN [4]. SPP-Net runs the convolutional layer for single time to get the feature map on the entire image instead of running it over each region as in R-CNN which reduces the inference time by a large margin, approximately 2400 times.

The real-time object detection region proposal computation is still a bottleneck in the fast R-CNN; the frames per second of the detection and the accuracy of the objects are directly impacted by high-quality region proposals. The region proposal network (RPN) makes the object detection close to be real time [5]. RPN is a fully convolutional network (FCN) that can predict the location of the objects and probability of a particular object within that location simultaneously. The main idea of RPN is using CNN to generate region proposal with a high recall directly.

2.3 Vehicle Speed Calculation

A speed prediction algorithm is proposed that has microscopic and macroscopic model properties [6]. These properties are based on the gas-kinetic traffic modeling approach which predicts the speed based on actual traffic data. For different applications, adjustment and alignment are essential which do not make the system stable [7]. The speed prediction was based on artificial neural network (ANN) for urban traffic network with four signs such as stop sign, traffic light area, free flow and turning; then for each signal, ANN model was trained [8]. The additional factors like climate and driver conduct impacts are not considered in this algorithm while anticipating vehicle speed and the definite segmentation.

2.4 Text Extraction and Recognition

Many techniques have been proposed over the years for text extraction and recognition. Neumann and Matas proposed a method using the localization of character and character segmentation. Characters are detected as extremal regions that are followed by the formation of text line, character recognition and finally selection of sequence [9]. The efficient method in character segmentation convolves Laplacian in the frequency domain with wavelet sub-bands at different levels for improving low-resolution character detection, and stroke width transform (SWT) returns the pixels where it thinks there might be a letter [10]. The text position is based on the interval between the nearest neighbor clusters of text regions which helps for the segmentation of characters.

3 Proposed Work

Given a live stream of the road scene, the motto is to accomplish the task of identifying and penalizing the violations committed by two-wheelers. The violations are found by extracting the following features:

- Identify the presence of motorcycle.
- Recognize the presence of helmet within the region of interest.
- Identifying the existence of people within the region of interest.
- Discover the license plate.
- Localize the characters within the license plate.
- Recognize the characters within the license plate.

3.1 System Architecture

Object Detection. The main objects to be detected in this problem statement are motorcycle, helmet, persons and license plate. The dataset for all these objects is collected manually from Google Images. CNN is used to train and classify these four classes. The classification provides the probability of finding a particular class and the location of these objects in the frame. Helmet, person and license plate detection are not applied to the entire frame, and these three object classifications are done to the region where the motorcycle is identified, which helps to identify the violation of a particular vehicle.

License Plate Text Extraction and Analysis. As shown in Fig. 1, from the image, the license plate is localized using object detection and text extraction to recognize the license plate number which is necessary for speed calculation and penalization process. The text within the license plate is localized using the binarization technique, and the localized text is given to the OCR classifier which is developed using CNN.

Speed Calculation. Instead of using speed sensors to calculate the speed, the position of the license plate between successive frames is processed in order to identify the speed of the vehicle. The difference in pixel positions of the boundary of the license plate with respect to the frame of the live stream is calculated for each vehicle. Depending on the camera angle, a hard-coded value is assigned which maps the pixel difference to actual road distance difference. This difference divided by the frames per second (FPS) parameter of the camera gives the vehicle speed.

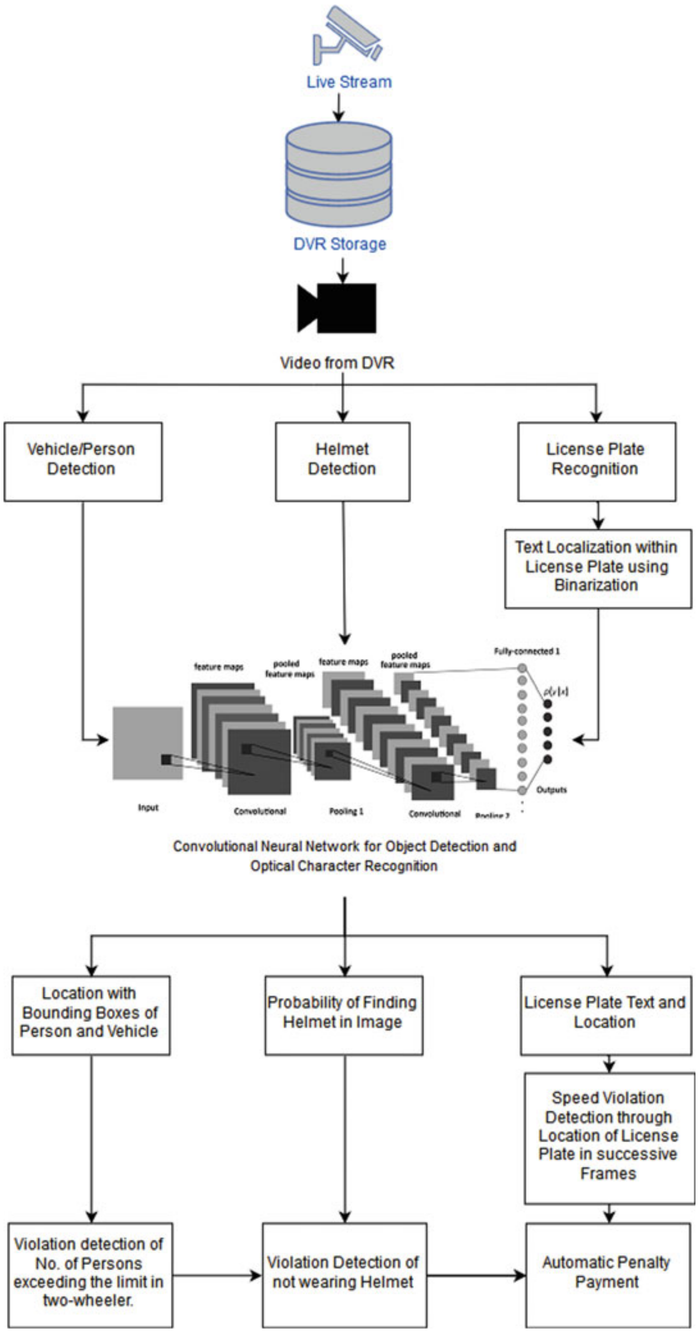


Fig. 1 Proposed system architecture

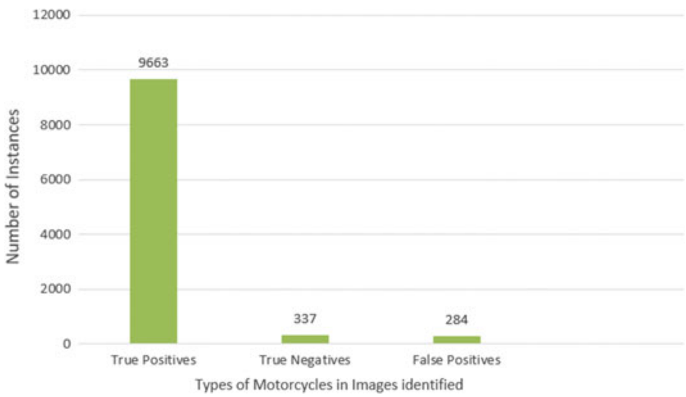


Fig. 2 Motorcyclists detection

3.2 Traffic Violation Detection Algorithm

Each frame of the live stream is subject to processing. Algorithm 1 shows that the first stage of processing is to check whether the frame contains a motorcycle or not; if the frame does not contain motorcycle, then that frame is skipped for further processing else that frame is subject to license plate localization, text localization and text extraction and speed calculations. For each identified motorcycle, if the helmet is not detected for the rider else the number of persons in that motorcycle is greater than two or the speed of the motorcycle exceeds 30, then an amount is deducted from the motorcycle owner by identifying the license plate number.

Algorithm 1: Traffic Violation Detection

```
Input   : Livestream
Output  : Deduction of an amount from violated persons

Procedure TrafficViolationDetection(Live_Stream)
begin
   while(!EOF):
      curframe ← readframe(Live_Stream)
      if IsVehicleNotPresent(currFrame)
         continue
      else if IsMotorCyclePresent(currFrame)
           totalmotorcycles←Location of each motorcycle
      end if
           for each motorcycle in totalmotorcycles
              Plate ← Plate_Localize(motorcycle)
              Curr_Co-ord ← Text_Localize(Plate)
              License_Number ← Text_Extract(Curr_Co_ord)
              Count ← People_Count(motorcycle)
              Speed ← Curr_Co_ord - Prev_Pos
              if ( IsHelmetNotPresent(motorcycle) or
                   Count > 2 or Speed > 70 )
                 MoneyDeduction(License_Number)
              end if
           end for
           Prev_Pos ←Curr_Co-ord
   end while
end
```

3.3 Object Detection Algorithm

A CNN follows the architecture of You Only Look Once (YOLO v2) which is utilized for identifying motorcycles, persons, helmets and license plates [11]. YOLO v2 is chosen because it has a higher FPS processing capability than the other state-of-the-art object detection models like YOLO [12], faster R-CNN [13] and single shot multibox detector (SSD) [14]. The dataset used for motorcycle class consists of 100,000 images of motorcycles and helmets, and persons in these 100,000 images are used for the helmet and person classes. License plate dataset is constructed by taking 50,000 images of Indian license plate with different fonts. CNN gives the probability of these four classes in a frame and its associated boundary coordinates. If the probability of finding a motorcycle in a particular frame is greater than 82%, then that frame is passed on for further processing and analyzing, else the frame is skipped.

3.4 Text Localization and Analysis Algorithm

After recognizing the license plate, it is binarized and then a sliding window of with minimum size slightly less than that of a character in the license plate and maximum slightly greater than the size of a character in the license plate. This is shown in Algorithm 2. This window slides across the plate which crops license plate into small images and then checks for the white pixel count in the current window. If the count is greater than 20, the top-left and bottom-right of the current window are stored. This finds all the characters in the license plate. These character images are resized to 32×32 and then given to OCR for further processing. A two-layered CNN is used for the purpose of recognizing the text from the license plate. The dataset used for training this model is the characters manually cropped from the Indian license plate images which are used for license plate detection. The input image to this CNN model after training is the 32×32 character images obtained from the text localization process.

Algorithm 2: Text Localization

```
Input   : License Plate image
Output  : Co-ordinates of individual characters
Procedure TextLocalization(License_Plate_Img)
  begin
     color_img ← open(License_Plate_Img)
     gray_img ← RGB2GRAY(color_img)
     bin_img ← Binarize(gray_img)
     for cropped width in range(0, width/2):
        for cropped height in range(0, height/2):
           for x in range(width/2, width - cropped_width):
              for y in range(0, height - cropped_height):
                 cropped_img ← crop(y:y+cropped_height,
                                    x:x+cropped_width)
                 top_left ← cropped_img(x,y)
                 bottom_right ← cropped_img(y+cropped_height,
                                    x + cropped_width)
              end for
           end for
        end for
     end for
     overlap←is_overlap(top_lt,btm_rt,rect_top_lt,
             rect_btm_rt)
     if overlap:
        top_lt, btm_rt ← max_area(top_lt, btm_rt,
                          rect_top_lt, rect_btm_rt)
        if top_left == top_lt & bottom_right == btm_rt:
           skip
        else:
           rectangle.append(top_left, bottom_right)
        end if
     end if
     return rectangle
  end
```

4 Implementation Details and Result

4.1 Helmet Detection

The accuracy of helmet detection is tested in 10,000 images, where the number of motorcycles is greater than at least 1 in all images. There were a total of 28,171 helmet instances. Out of these 28,171 instances, 27,073 instances are rightly identified, and 695 instances were wrongly determined as helmets; an example of one such case is where the rider is hairless which is similar to open helmets with color that of our skin tone. Helmet detection gives a precision of 97% which is shown in Fig. 3. A

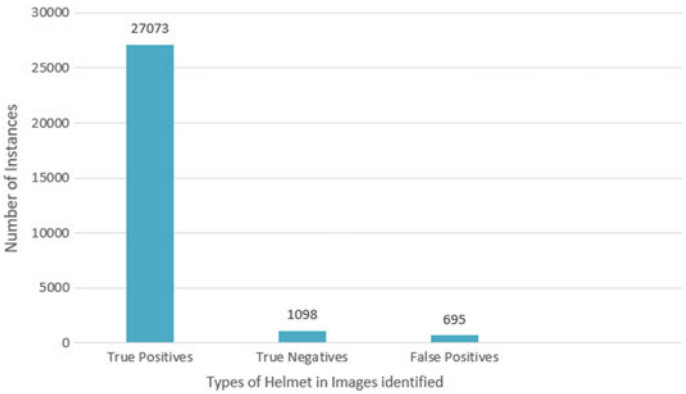


Fig. 3 Helmet detection

similar analysis is carried out for the other three classes where motorcycle detection achieves a precision of 97.1% which is shown in Fig. 2, person detection achieves a precision of 95%, and license plate detection achieves a precision of 94.3%. To obtain convergence quickly, momentum optimizer and definite weight initialization instead of random are applied for object detection model. The weight initialization of the network is done in a way that the neuron activation functions are not starting out in saturated or dead regions. Initialization of the weights is made with random values that are not 'too small' and not 'too large.' When the local minimum is reached, the momentum is pretty high, and it does not know to slow down at that point due to the high momentum which could cause it to miss the minimum entirely and continue moving up. Figure 4 shows the training loss over 100 K iterations of the multi-class CNN object detection.

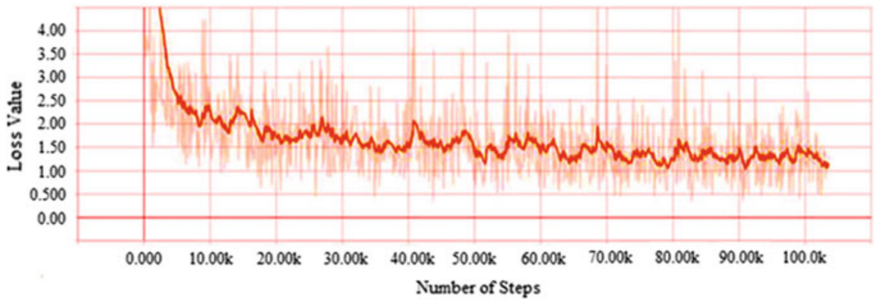


Fig. 4 Loss in the multi-class CNN object detection training

Fig. 5 Overlapping rectangular boxes

Fig. 6 Final localized characters in license plate

4.2 *Text Localization*

Changing sliding window size up to half of width and height recognizes the same region in the license plate. This has to be removed as it represents the same region of space where it already found some digits. Current rectangular box overlaps with previously found boxes. This overlapping problem can be eliminated by comparing the boundary values of the current box with already found rectangular boxes whether it overlaps or not. If yes, the outer rectangular box has to be removed for example in Fig. 5, it has already found 'D' and 'L' in separate, but it again finds the rectangular box with a width containing both 'D' and 'L.' So, to remove outer rectangle, compare the area of the box containing both 'D' and 'L' with the box containing 'D' and another box containing 'L.' Then skip the maximum area rectangular box which brings the localized array containing individual boxes with 'D' and 'L.'

As shown in Fig. 5, the sliding window identifies the space inside A as a character. In this situation, this small area will be sent to the OCR as a region if this problem is not addressed. This is avoided by excluding the localized boxes if their area is not greater than 20. After solving all the problems, the coordinates of the bounding boxes as shown in Fig. 6 are given to OCR to find the license plate number.

4.3 *Text Recognition*

The OCR of text in license plate gives the accuracy of 99.1% when the CNN model with a filter size of 5 $\times$ 5 and valid padding is trained for ten epochs with a batch size of 128 and learning rate of 0.001, where 50,000 images with different characters and different fonts from the license plate are used for training and 5000 images are used for testing purposes. For introducing nonlinearity, ReLU activation function is used in the intermediate layers, and for our dataset, it is seen that having a sigmoid layer in the output layer has a better accuracy than ReLU.

5 Conclusion and Future Work

The automatic detection of traffic violation is proposed through tuning various computer vision techniques to our specific needs like Tesseract optical character recognition for license plate recognition and CNN for identifying on-road objects like vehicles, persons and for helmet detection. By using these techniques, the violations like over speeding, riders of a two-wheeler not wearing a helmet and number of persons exceeding the permitted count of a two-wheeler are identified, and the violators are penalized automatically by reducing money from their account where the users are identified through their vehicles license plate number. The proposed method uses only one high-resolution camera to find all the violations lively, and it automatically deducts money. Through this work, the users will be cautious about their driving, and henceforth, less probability of accidents taking place. The system can be further extended by alerting the nearest police station in case any of the violation is found to cease the vehicle. This can be done by finding the nearest police station geo coordinates of the camera with respect to the location of the camera.

Acknowledgements The authors gratefully acknowledge CTDT, Anna University for providing financial support to carry out this research work under the student innovative projects scheme. The authors also gratefully acknowledge the Web site: https://www.trickspagal.com/2017/08/trace-vehicle-number-with-owner-name-and-address.html for providing the license plate image to test the proposed system. The authors also thank Big Data Analytics Laboratory of MIT campus, Anna University for infrastructure and support.

References

1. Aarthy, D.K., Vandanaa, S., Varshini, M., Tijitha, K.: Automatic identification of traffic violations and theft avoidance. In: Second International Conference on Science Technology Engineering and Management, pp. 72–76 (2016)
2. Marikhu, R., Moonrinta, J., Ekpanyapong, M., Dailey, M., Siddhichai, S.: Police eyes: real world automated detection of traffic violations. In: 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, pp. 1–6 (2013)
3. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. Int. J. Comput. Vis. **104**(2), 154–171 (2013)
4. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
5. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
6. Park, J., Murphey, Y.L., McGee, R., Kristinsson, J.G., Kuang, M.L., Phillips, A.M.: Intelligent trip modeling for the prediction of an origin–destination traveling speed profile. IEEE Trans. Intell. Transp. Syst. **15**(3), 1039–1053 (2014)
7. Liu, R., Xu, S., Park, J., Murphey, Y.L., Kristinsson, J., McGee, R., Kuang, M., Phillips, T.: Real time vehicle speed prediction using gas-kinetic traffic modeling. In: IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems, pp. 80–86 (2011)

8. Neumann, L., Matas, J.: Real-time lexicon-free scene text localization and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **38**(9), 1872–1885 (2016)
9. Park, J., Murphey, Y.L., Kristinsson, J., McGee, R., Kuang, M., Phillips, T.: Intelligent speed profile prediction on urban traffic networks with machine learning. In: 2013 International Joint Conference on Neural Networks, pp. 1–7 (2013)
10. Liang, G., Shivakumara, P., Lu, T., Tan, C.L.: Multi-spectral fusion based approach for arbitrarily oriented scene text detection in video images. IEEE Trans. Image Process. **24**(11), 4488–4501 (2015)
11. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp. 6517–6525 (2017)
12. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
13. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 1137–1149 (2017)
14. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer, Cham (2016)

Secure Big Data Deduplication with Dynamic Ownership Management in Cloud Computing

Srinivas Gonthireddy and Syam Kumar Pasupuleti

Abstract Due to the huge size of data, traditional database systems cannot handle the big data. Cloud is the suitable platform for storing and processing the big data, but moving data into cloud brings privacy problem. Privacy of the data can be protected by storing encrypted data in the cloud. However, encrypted data introduces a new challenge for cloud that is data duplication, which is most important for big data storage, because storing redundant copies waste the lot of storage space and complicates the data management. Recently, several schemes were proposed to address the secure data deduplication. However, existing schemes have security weakness and do not support data access controls. In this paper, we propose a secure big data deduplication with dynamic ownership management in cloud based on identity-based encryption and proxy re-encryption. Through the security and performance analysis, we prove that our scheme is secure and efficient.

Keywords Big data · Cloud computing · Privacy · Deduplication · Identity-based encryption · Proxy re-encryption

1 Introduction

Digital world grows in exponential speed, which produces large volumes of data called big data, which describes in 4Vs [1] as volume, velocity, variety, and veracity. Volume describes the amount (size) of data coming from various resources like social networking, business transactions, etc. Velocity refers to at what speed the data is generating from the digital world. Variety refers to various formats of data, like text files, photographs, audios, and videos. Veracity refers to the trait of deserving trust and confidence of the data. Due to the huge size of data, traditional database systems

S. Gonthireddy
School of Computer and Information Science (SCIS), University of Hyderabad, Hyderabad, India
e-mail: srinivascse02@gmail.com

S. K. Pasupuleti (✉)
Institute for Development and Research in Banking Technology (IDRBT), Hyderabad, India
e-mail: psyamkumar@idrbt.ac.in

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_22

cannot handle the big data, so cloud computing is the best option for storing and processing the big data because cloud computing provides various resources (e.g., computing, storage) through the Internet to users based on their need.

However, the cloud is not trusted by users because storing big data in the cloud has one significant challenge, such as privacy [2]. Although the attacks on data at CSP cannot be avoidable, it leads to data privacy leakage [3]. Privacy of the data can be achieved by encrypting data before uploading to cloud [4], but achieving deduplication on encrypted data is a challenging task. Then, same data gives different ciphertexts, because we use different encryption keys, which create trouble for the cloud server to verify whether data is same or different. For example, user A encrypts a file M using his public key d_A and stores it in cloud as ciphertext CT_A. Similarly, user B encrypts same file M using his public key d_B and it is also stored in cloud as CT_B. If both ciphertexts CT_A and CT_B are different, how it is possible to detect data duplication over ciphertexts in the cloud. Even if duplicates are detected, then how to give the access rights to both users can get the data based on their private keys is a challenging task.

To address these problems, several schemes [5–17] were proposed to ensure the deduplication over encrypted data based on different encryption techniques. The schemes [5–9] guarantee the data privacy with deduplication but do not support ownership verification. The schemes [10–17] were proposed to achieve deduplication over encrypted data with support of ownership verification. However, these schemes have security loopholes and do not support fine-grained access controls.

In this paper, we propose a secure big data deduplication with dynamic ownership management in cloud computing scheme to save cloud storage space and ensure the privacy of data. The main contributions our scheme as follows:

- In our scheme, we use identity-based encryption to ensure the data privacy and access control.
- Our scheme also uses hashing technique to check whether the duplicates are present or not in the cloud storage.
- Our scheme uses dynamic data ownership protocol to verify the ownership of the user.
- We also proposed the proxy re-encryption scheme to give the data access among multiple users.
- Through the security and performance analysis, we prove our scheme is secure and efficient, respectively.

The remaining paper is organized as: "Sect. 2 gives description about related work. In Sect. 3 defines system model and then preliminaries and its security model. In Sect. 4, we give the description about our proposed scheme. In Sects. 5 and 6, we describe security analysis and performance measures. We describe conclusion in Sect. 7."

2 Related Work

In past decade, several deduplication schemes over encrypted data were proposed to address the secure deduplication in the cloud. We have categorized these schemes into two types.

2.1 Deduplication Over Encrypted Big Data

To ensure the deduplication of encrypted data, Douceur et al. [5] proposed a message-locked encryption (MLE) for data duplication storage, in which the data is encrypted on the client side. In convergent encryption (CE) schemes, the user computes a key K ← H(M) from the data file (M). By using K, the user computes ciphertext CT ← E(K, M) using deterministic symmetric encryption algorithm. After that, the user uploads ciphertext CT to the cloud. If another user wants to encrypt the same file using same encryption algorithm, then same ciphertext has been generated because both encryption techniques are same. Thus, the second user also uploads ciphertext to the cloud, which causes data duplication. Meye et al. [6] proposed two servers inter-deduplication and intra-deduplication, which are used to resist the attacks on data identifier manipulation. This scheme also does not support data sharing among multiple users like previous scheme [5]. Puzio et al. [7] introduced a ClouDedup scheme to ensure the security of CE. But it does not deal with the data deletions; i.e., the after deletion of data from the cloud, still it can be accessible by other data holders because they know their private key. However, all these deduplication schemes may leak the data because of CE. The attacker can recover the plain text M derived from $S = \{M_1, M_2, \ldots, M_n\}$ off-line encryptions because the CE is deterministic and keyless.

To overcome these problems, Bellare et al. [8] first proposed a DupLess scheme to avoid the brute-force attacks. In this scheme, "the clients encrypt their data by using key server (KS), which is different from storage server (SS) to get the privacy and also clients authenticate themselves to the KS without disclosing any information." Whenever the attacker accesses the KS, it remains inaccessible from the user's view, so this scheme is highly secure. Liu et al. [9] constructed a proxy re-encryption scheme, but it cannot handle data management after deduplication (e.g., ownership verification and deletion) and it cannot assess the scheme performance. However, all these schemes achieve data privacy but do not support ownership verification.

2.2 Data Ownership Verification

Data ownership verification plays a key role to identify the correct user while designing deduplication system, in which Halevi et al. [10] introduced "proof of ownership

(PoW)" scheme for data deduplication at client side. "They applied a hash function on the original file and applied Merkle tree on pre-processed data for getting information about verification." This scheme identifies deduplication at client side and reduces network bandwidth. Di Pietro et al. [11] verified PoW scheme by selecting random bit positions of a file, but these schemes cannot give the data privacy. Ng et al. [12] used PoW scheme to manage ciphertext duplication, in which Merkle trees are used to check the deduplication. Data blocks are used to find the leaf value in Merkle tree, in which interactive proof protocols challenge one leaf at a time. For getting higher security, the protocol is executed several times, which causes more overhead computed.

Yang et al. [13] used a cryptographically efficient and secure scheme for checking data ownership, in which the client owns the entire file and proves to the server without sending the file. By checking dynamic ownership verification, the data holder has an access to "dynamic portions of the data file for getting proof of possession about the file." And it reduces the computational cost and also minimizes the communication cost of both CSP and data holders. This scheme mixes dynamic coefficients with randomly selected indices to get unique proof in every challenge. Yuan and Yu [14] proposed a "secure and efficient data integrity auditing for data deduplication in cloud storage based on authentication tags and homomorphic linear authenticators." However, in this scheme, deduplication applies over two files and authentication tags of those two files, so it cannot support big data duplication.

Wu et al. [15] constructed a scheme Index Name Server (INS) to reduce several workloads not only data storage, server load balancing, and optimized node selection but also minimizes "file compression, chunk matching, real-time feedback control, and busy level index monitoring." However, this scheme does not support deduplication over encrypted data. Yan et al. [16] proposed a scheme, to challenge the data ownership verification by using key pairs of elliptical curve cryptography, in which the scheme requires certificates to maintain expiration time of the public keys.

3 Problem Statement

In this section, we define the system model, preliminaries, definitions, and goals.

3.1 System Model

The proposed system model contains four types of entities as shown in Fig. 1, such as Cloud Service Provider (CSP), Users (U), Key Generation Center (KGC), and Authorized Party (AP).

1. *CSP*: who provides storage for large volumes of data to users based on demand but it cannot be fully trusted.

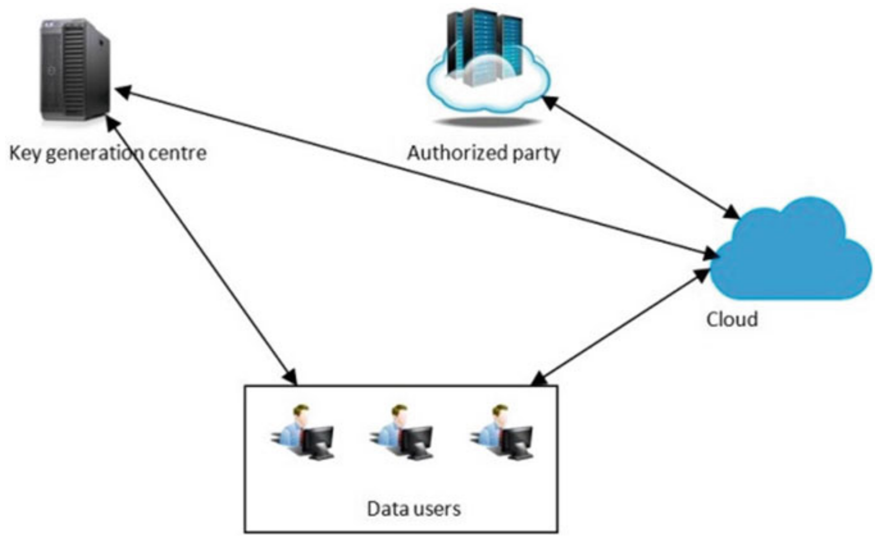


Fig. 1 System model

2. *Data users* who upload a data file into the cloud and access the data from cloud.
3. *KGC* is fully trusted by data users, and it generates public keys and private keys for encryption.
4. *The authorized party (AP)* is fully trusted by data users (u_i), which verifies whether the data user is authorized or not by using ownership verification protocol.

3.2 Preliminaries

In this section, we give basic description of technical backgrounds: bilinear pairing and identity-based encryption.

3.2.1 Bilinear Pairing

The bilinear pairing denoted by $PG = (G_1, G_2, G_T, e, g_1, p)$. "Let G_1, G_2 be elliptic groups and G_T be multiplicative group, where all groups are of same prime order p, g_1 is a generator of G_1 and g_2 is a generator of G_2. Let $e: G_1 \times G_2 \rightarrow G_T$ be a map with the following properties."

Bilinearity: $e(g_1^a, g_1^b) = e(g_1, g_2)^{ab}$ for all $g_1, g_2 \in G_1$ and $a, b \in Z_p$

Non-degeneracy: $e(g_1, g_2) \neq 1$

Computability: "There exist an efficient algorithm to compute" $e(g_1, g_1)$ for all $g_1, g_2 \in G_1$.

3.2.2 Identity-Based Encryption

The IBE consists of four algorithms:

Setup: This algorithm generates master public key and master secret key
Key Gen: It generates private key based on identity
Encrypt: It generates ciphertext for plaintext
Decrypt: This algorithm takes ciphertext and it outputs a message.

3.3 Definitions

In this section, we introduce the following algorithms to construct our scheme.

Setup (λ) → (*mpk, msk*): This algorithm is run by the user. It takes security parameter λ as input and generates master key pair (mpk, msk) as output.
Signature generation (*H, M*) → H_1: This algorithm is run by the user. It takes message M as input and produces hash value as output.
Deduplication check (S_i) → (*Positive/Negative*): The deduplication check algorithm is run by cloud. It takes signature as input and produces positive/negative as output.
Key generation (ID_1, *mpk, msk*) → d_{ID}: It is run by the key generation center (KGC), which is a trusted authority. It takes user identity (ID) and master key pair as input and produces a private key d_{ID} as output.
Encryption (ID_1, *M, mpk*) → CT_1: It is run by data user, and it takes master public key mpk, message M, identity ID as input and produces ciphertext CT_1.
Data Ownership Verification (ID_i, *R*, S_i,) → {0,1}: This algorithm is run by authorized party (AP), it takes the identity, signature, and random number R as input, which returns 1, and the owner is authorized otherwise unauthorized.
Re-key generation (*mpk, sk*, ID_2) → $rk_{ID1 \rightarrow ID2}$: It is run by KGC, and it takes mpk, identity of user 2, and secret key of user 1 as input and generates re-encryption key $rk_{ID1 \rightarrow ID2}$ with respect to user 2.
Re-encryption ($rk_{ID1 \rightarrow ID2}$, CT_1) → CT_2: It is run by proxy, which takes re-encryption key $rk_{ID1 \rightarrow ID2}$ and ciphertext CT_A and produces the ciphertext CT_2 as output.
Decryption (CT_2, d_{ID}) → *M*: It is run by user, and it takes ciphertext CT_2 and private key d_{ID} as input and generates message M as output.

3.4 Design Goals

We proposed a secure deduplication scheme to achieve the following goals:

Data Deduplication: Data deduplication eliminates the redundant data and reduces storage overhead.
Privacy: The data stored in the cloud should not disclose any information to unauthorized users.
Ownership Verification: Ownership verification indicates that whether the user is authorized or not based on identity of the user.

4 Construction of Proposed Scheme

In this section, we construct the proposed secure big data deduplication scheme, which consists of system setup, data upload, data deduplication, dynamic ownership verification, and data retrieval.

4.1 System Setup

In this algorithm, the KGC generates public parameters and master secret key as follows:

Pairing group PG = (G_1, G_T, g, e, p, H) and select random $r, s, \alpha \in Zp$
Compute $u = g^r, g \in G_1$
The parameters of mpk = (g, u) and msk = (r, s).

4.2 Data Upload

Before uploading the file into cloud, the user needs to perform the following algorithms.

- **Signature generation**: The user u_1 generates a signature of a file using signature generation algorithm:

$$\text{Compute } S_1 = H(M)^r$$

- **Data deduplication check**: After generation of signature S_1, the user u_1 sends it along with the identity (ID_1) to the CSP. The CSP checks u_1's signature with existing signatures in cloud. If the result is negative, the CSP asks the user u_1 to upload encrypted data. Then, the user u_1 encrypts data file M using encryption algorithm.

- **Encryption**: In encryption algorithm, the user first selects a random number r and computes $u = g^r$ and $y = H(ID_1)$. Then

$$\text{Compute } C_1 = M * e(u, y^{\alpha})$$
$$C_2 = g^{\alpha}$$
$$C_3 = y^{\alpha}$$

After encrypting the data, user uploads the ciphertext $CT = (C_1, C_2, C_3)$ to the cloud.

4.3 Data Deduplication

After user u_1 uploads data M to the CSP, another user u_2 is trying to store the same data M in the cloud by sending a package (S_2, ID_2). Then, CSP checks the deduplication exists or not by using deduplication check algorithm. So the result is positive, which causes data deduplication. Then, CSP forwards the same package to the authorized party (AP) for ownership verification.

4.4 Dynamic Ownership Verification

After receiving the package from CSP, the AP verifies whether the owner is authorized or not by sending a random number R to the user $(u2)_2$. Then, u_2 calculates a new signature $Y = H(H(M, R))$ and sends back to the AP. Then, AP calculates $H(S_2, R)$ and gets a new signature $Z = H(S_2, R)$. If the both signatures are same $(Y = Z)$, it indicates that user u_2 is authorized, otherwise unauthorized. If the user is authorized, then AP forwards the result (S_2, ID_2) to the CSP.

- **Proxy Re-encryption**: The CSP forwards the result to the proxy, then proxy request KGC for re-encryption key. Then, KGC computes $k = H(ID_2)$, $u = g^r$ and generates re-encryption key $rk = (rk_1, rk_2) = (k, e(u, k))$.
After generation of re- key, KGC forwards it to the proxy. Then, proxy gets the ciphertext CT from the cloud and re-encrypts using re-encryption key as follows. After re-encryption proxy forwards new ciphertext $CT' = (C_{11}, C_{12}, C_{13})$ to the cloud.

$$\begin{aligned} C_{11} &= M * e(u, y^{\alpha}) * rk_2 * e(u, 1/rk_1) \\ &= M * e(g^r, H(ID_1)^{\alpha}) * e(g^r, H(ID_2)^{\alpha}) * e\,(g^r, 1/H(ID_2)^{\alpha}) \\ &= M * e(g^r, H(ID_1) \end{aligned}$$

$$C_{12} = g^{\alpha}$$
$$C_{13} = y^{\alpha}$$

4.5 Data Retrieval

In the data retrieval, user retrieves the data from the cloud using two algorithms: key generation and decryption.

- **Key generation**
  We initially compute $y = H\,(ID_1)$
  The private key $d_{ID} = (d_{ID1}, d_{ID2}, d_{ID3}) = (y^r, y^s, g^s)$
- **Decryption**: The decryption algorithm decrypts the data using own private key d_{ID}. The whole deduplication process is given in Fig. 2.

$$\begin{aligned} M &= C_{11} * e(d_{ID3}, C_{13})/e(C_{12}, d_{ID1}) * e(C_{12}, d_{ID2}) \\ &= M \end{aligned}$$

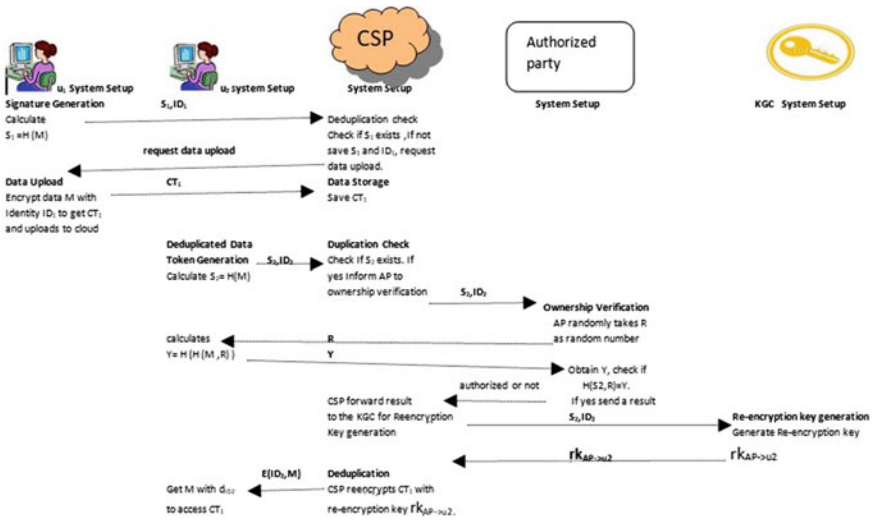


Fig. 2 Process for secure big data deduplication

5 Security Analysis

In this section, we prove the correctness, data privacy, data deduplication, and user authentication.

5.1 Correctness

The correctness of our scheme can be proven as follows.

$$\begin{aligned}
M &= C_{11} * e(d_{ID3}, C_{13})/e(C_{12}, d_{ID1}) * e(C_{12}, d_{ID2}) \\
C_{11} &= C_1 * rk_2 * e(u, 1/rk_1) \\
&= M * e(g^r, H(ID_1)^\alpha) * e(g^r, H(ID_2) * e(\, g^r, 1/H(ID_2)) \\
&= M * e(g^r, H(ID_1)^\alpha) \\
C_{12} &= C_2, C_{13} = H(ID_1)^\alpha \\
&= M * e(g^r, H(ID_1)^\alpha)\ * e(g^s, H(ID_1)^\alpha)/e(g^\alpha, H(ID_1)^r) * e(g^\alpha, H(ID_1)^s) = M
\end{aligned}$$

5.2 Data Privacy

Suppose the user u_i, file M is encrypted by using identity-based encryption. Whenever, the adversary requests for data file M, first sends identity (ID) to the KGC, which provides private key d_{ID}. After getting the private key, adversary will send his identity (ID) to the cloud. Then, the cloud checks whether the ID is matched with any existing ID or not; if the check is positive, then the cloud sends the ciphertext associated with that identity. Otherwise, the cloud sends a message as unauthorized response. Whenever multiple users are trying to store same data into the cloud, the cloud checks the signatures of the each user and sends it for user authorization. If the authorization is success, then the proxy re-encrypts the ciphertext with re-encryption key. Another way of protecting M is whenever an authorized user wants to get the file, send his identity to the cloud. The cloud checks present user identity (ID) with existing IDs. If the check is positive, then the cloud forwards ID and signature to the AP. The AP challenges the user based on dynamic ownership verification protocol; after getting successful verification, the cloud sends ciphertext to the user. Then, user applies decryption algorithm, which takes CT and d_{ID} as input and produces original data file M as output. Since data is encrypted by using identity-based encryption and decrypted by using private key d_{ID}, there is no way for adversary to access the data from the cloud.

5.3 Data Deduplication

Whenever two users u_1 and u_2 are trying to store same data into the cloud, the deduplication of data can be checked by signatures of the files ($S_1 = S_2$). By verifying signature of file, we can detect and eliminate the data deduplication. If the malicious user gets the signature from the CSP, then the data file M is not disclosed because the malicious user does not know about any hashing algorithm, which the authorized user is used. Hence, we ensure the deduplication of the data in the cloud.

5.4 User Authorization

Authorization means checking whether the user is authorized or not. Suppose, whenever the user u_i requests the data to the cloud corresponding to his identity (ID_i), then the cloud checks whether the identity (ID_i) is matched with any existing IDs of other users or not. If the match is positive, then the cloud sends a pair containing signature and identity (S_i, ID_i) to the AP. After receiving the pair, the AP sends a random number R to the user, and the user calculates $Y = H(H(M, R))$ and sends it back to AP. Now, AP calculates a new signature $Z = H(S_2, R)$. If $Y = Z$, then the user is authorized; otherwise, the user is unauthorized. To get the plain text from ciphertext, users must have private key d_{ID} and users pass the ownership verification. Even though some malicious users try to get the private key, but it is impossible to cross the ownership challenge because they cannot get the signature from the CSP. Hence, without ownership verification success no one can get the ciphertext.

6 Performance Analyses

In this section, we analyze the performance of the proposed scheme based experiment result. The experiment is carried out on the machines with the configuration of intel(R) Core(TM) i5-6500U CPU @2.50 GHz and 8 GB RAM with windows 10 64-bit operating system, and we also used PyCharm IDE [18] and charm crypto [19] library. In our scheme, we used signature generation and verification algorithms for big data deduplication checking. We also analyze the experimental results of encryption and decryption algorithms using different input file sizes.

6.1 Signature Generation

Here, we evaluated the time for signature generation on different file sizes as shown in Fig. 3. From Fig. 3, we can observe that even if we give the 100 MB file, then

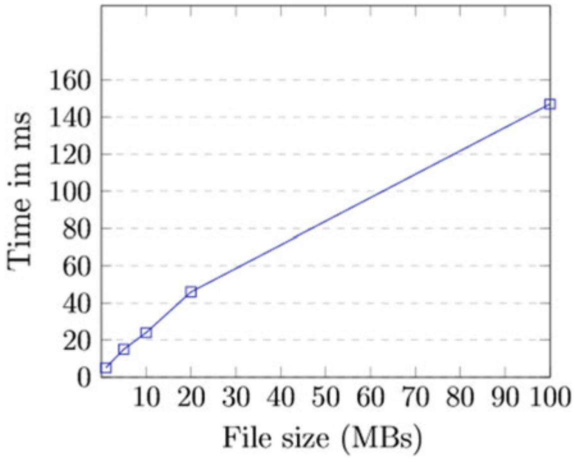


Fig. 3 Signature generation

the signing and verifying algorithm takes less than 200 ms. The file size is directly proportional to time; if the file size increases, then time also linearly increases.

6.2 Signature Verification

Here, we evaluated the time for signature verification algorithm based on different file sizes as shown in Fig. 4. From Fig. 4, we can observe that even if we give the 100 MB file then the signing and verifying algorithm takes less than 200 ms. The file size is directly proportional to time; if the file size increases, then time also linearly increases. So, the hashing algorithm is better for deduplication check in the cloud.

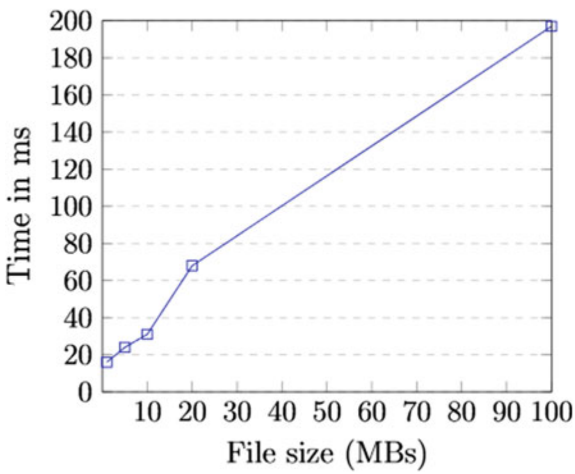


Fig. 4 Signature verification

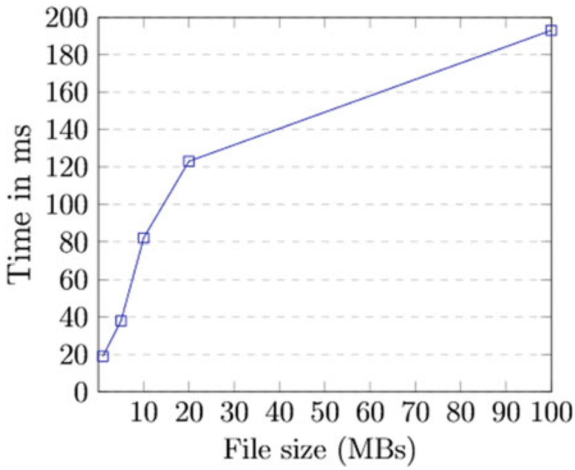


Fig. 5 Encryption

6.3 Encryption

Here, we measured the times for encryption operations as shown in Fig. 5. From Fig. 5, the size of the public key parameters does not influence any algorithm that is present in our scheme. If the file size is 100 MB, then the encryption algorithms take less than 200 ms, which indicate that our scheme has less time for encryption of big data.

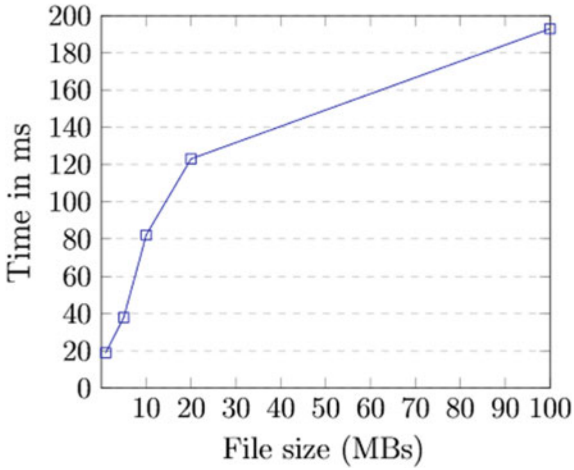
6.4 Decryption

Here, we measured the times for decryption operations as shown in Fig. 6. From Fig. 6, the size of the private key does not influence any algorithm that is present in our scheme. If the file size is 100 MB, then the decryption algorithms take less than 200 ms, which indicate that our scheme has less processing load for decryption of big data. So, our scheme has less accessing time among multiple users.

7 Conclusion

In this paper, we proposed a secure big data deduplication with dynamic ownership management scheme in cloud computing. In this scheme, we used signatures for deduplication check, which identifies the redundant data or duplicate data. We also proposed identity-based encryption, which ensures the data privacy. So, no adversary can be decrypted the data without users private key. In addition, we also proposed dynamic ownership scheme to verify whether the user is authorized or not and the

Fig. 6 Decryption



PRE scheme provides data access controls among the multiple users. From the security analysis, we proved that our scheme is provably secure against adaptive chosen ciphertext attack. Through the performance analysis, we showed that our scheme is efficient and suitable for big data deduplication.

In future work, we implement our scheme based on attribute-based encryption fine-grained access controls. We also extend this scheme to support the integrity of data along with privacy and deduplication.

References

1. http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters
2. Wei, L., Zhu, H., Cao, Z., Dong, X., Jia, W., Chen, Y., Vasilakos, A.V.: Security and privacy for storage and computation in cloud computing. Inf. Sci. **258**, 371–386 (2014). https://doi.org/10.1016/j.ins.2013.04.028
3. Wu, T., Dou, W., Hu, C., Chen, J.: Service mining for trusted service composition in cross-cloud environment. IEEE Syst. J. **PP**(99), 1–12 (2014). https://doi.org/10.1109/jsyst.2014.2361841
4. Liu, C., Yang, C., Zhang, X., Chen, J.: External integrity verification for outsourced big data in cloud and iot: a big picture. Future Gener. Comput. Syst. **49**, 58–67 (2015)
5. Douceur, J.R., Adya, A., Bolosky, W.J., Simon, P., Theimer, M.: Reclaiming space from duplicate files in a serverless distributed file system. In: Proceedings 22nd International Conference on Distributed Computing Systems. IEEE, pp. 617–624 (2002). https://doi.org/10.1109/icdcs.2002.1022312
6. Meye, P., Raipin, P., Tronel, F., Anceaume, E.: A secure two- phase data deduplication scheme. In: Proc. HPCC/CSS/ICESS. IEEE, pp. 802–809 (2014). https://doi.org/10.1109/hpcc.2014.134
7. Puzio, P., Molva, R., Onen, M., Loureiro, S.: ClouDedup: secure deduplication with encrypted data for cloud storage. In: 2013 IEEE 5th International Conference on Cloud Computing Technology and Science. IEEE, pp. 363–370 (2013). https://doi.org/10.1109/cloudcom.2013.54

8. Bellare, M., Keelveedhi, S., Ristenpart, T.: DupLESS: server aided encryption for deduplicated storage. In: Proceedings of 22nd USENIX Conference Security, pp. 179–194 (2013)
9. Liu, C., Liu, X., Wan, L.: Policy-based deduplication in secure cloud storage. In: Proceedings of Trustworthy Computer Services, pp. 250–262 (2013). https://doi.org/10.1007/978-3-642-35795-4_32
10. Halevi, S., Harnik, D., Pinkas, B., Shulman-Peleg, A.: Proofs of ownership in remote storage systems. In: Proceedings of 18th ACM Conference on Computer and communications security. ACM, pp. 491–500 (2011). https://doi.org/10.1145/2046707.2046765
11. Di Pietro, R., Sorniotti, A.: Boosting efficiency and security in proof of ownership for deduplication. In: Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security. ACM, pp. 81–82 (2012). https://doi.org/10.1145/2414456.2414504
12. Ng, W.K., Wen, Y., Zhu, H.: Private data deduplication protocols in cloud storage. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing. ACM, pp. 441–446 (2012)
13. Yang, C., Ren, J., Ma, J.: Provable ownership of file in de-duplication cloud storage. In: Proceedings of IEEE Global Communication Conference. IEEE, pp. 695–700 (2013). https://doi.org/10.1109/glocom.2013.6831153
14. Yuan, J., Yu, S.: Secure and constant cost public cloud storage auditing with deduplication. In: Proceedings of IEEE International Conference on Communications and Network Security. IEEE, pp. 145–153 (2013). https://doi.org/10.1109/cns.2013.6682702
15. Wu, T.Y., Pan, J.S., Lin, C.F.: Improving accessing efficiency of cloud storage using deduplication and feedback schemes. IEEE Syst. J. **8**(1), 208–218 (2014). https://doi.org/10.1109/JSYST.2013.2256715
16. Yan, Z., Ding, W., Yu, X., Zhu, H., Deng, R.H.: Deduplication on encrypted big data in cloud. IEEE Trans. Big Data **2**(2), 138–150 (2016)
17. Yan, Z., Zhang, L., Ding, W., Zheng, Q.: Heterogeneous Data storage management with deduplication in cloud computing. IEEE Trans. Big Data (2017)
18. https://www.jetbrains.com/pycharm/download/
19. https://jhuisi.github.io/charm/genindex.html

Characterization of Range for Smart Home Sensors Using Tsallis' Entropy Framework

Sujit Bebortta, Amit Kumar Singh, Surajit Mohanty and Dilip Senapati

Abstract The deployment of sensor nodes (SNs) in smart homes induces multipath transmission of signals in indoor environments (IEs). These paths occur due to the presence of household utility objects, which produce several reflected communication paths between the sender and the receiver. In order to determine the effective position of a target SN in a wireless sensor network (WSN), certain localization schemes are required in conjunction with trilateration methods. Therefore, it is worthwhile to estimate the uncertainties in the range for time of arrival (TOA) localization of SNs subjected to non-line of sight (NLOS) conditions. In this article, we provide a technique to characterize the variations in the range corresponding to the TOA based on the well-known Tsallis' entropy framework. In this model, the non-extensive parameter q characterizes the variations in the localization range caused due to multipath components. In this context, we optimize the Tsallis entropy subject to the two moment constraints (i.e., mean and variance) along with the normalization constraint. Our proposed model is in excellent agreement with the synthetic data in contrast to the mixture model. This paper also provides a new approach for estimating the parameters corresponding to the mixture model and the proposed $q-$Gaussian model by minimizing the Jensen–Shannon (JS) symmetric measure between the two models and the synthetic data.

Keywords $q-$Gaussian distribution · Tsallis entropy · Mixture model · Wireless sensor networks · Localization · JS measure

S. Bebortta · D. Senapati (✉)
Department of Computer Science Ravenshaw University, Cuttack 753003, India
e-mail: senapatidillip@gmail.com

A. K. Singh
Department of Computer Science Ramanujan College, University of Delhi,
New Delhi 110019, India

S. Mohanty
Department of Computer Science and Engineering DRIEMS Tangi,
Cuttack, Orissa 754022, India

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_23

1 Introduction

In the recent times, the development of the smart homes has drastically increased with the availability of various smart sensors and energy-efficient resources. Several smart devices like the smartphones, televisions, HVAC (heating, ventilation, and air conditioning) systems, etc., have contributed to the growth of these smart homes. These technologies employ several sensors for acquiring data from varying locations. So, it is quite difficult to infer the actual location of a target sensor node (SN) due to the ubiquity of scattering objects and multiple transmission paths. Thus, for transmission in indoor environments (IEs), the ultra-wideband (UWB) technology is considered to be the most reliable approach for facilitating communication between the smart devices [1–4, 26]. The UWB technology has marked a prime development in wireless communication industries, with the ability to support high data rate transmission over a short range. It also consumes less power which makes it an energy-efficient choice for personal area networks (PANs). The time of arrival (TOA)-based ranging is used distinctively with the UWB technology for localizing the sensors in a wireless sensor network (WSN). The TOA ranging can make complete utilization of the high bandwidth and appropriately manages the time delay resolution of the UWB technology [14, 15].

In IEs, the data transmission path between the source and the sink can be located either by using a line of sight (LOS) or non-line of sight (NLOS) propagation. In smart homes, the NLOS propagation indicates a lack of direct communication path between the source and the sink, due to the presence of obstacles, $viz.$, furniture and walls. A major concern with indoor complex networks is the transmission of data over the UWB channel in NLOS environments. The execution of a TOA-based UWB communication relies on accessibility of a direct signaling path between the source and the sink [10]. Thus due to the ambiguity in the existence of an explicit data path between the source and the sink, the TOA localization can exhibit significant variations in its range.

In this paper, we employ a straightforward approach to characterize the variations in range by using the well-known Tsallis' entropy framework. This framework was proposed to depict the consolidation of the power-law behavior and tail fluctuations of dynamical systems in statistical mechanics [5–9]. We first consider the localization of SNs by using the UWB-based TOA approach to infer the location of an SN. In IEs, the uncertainties in range caused due to multiple paths can be characterized by using the proposed model, i.e., the $q-$Gaussian distribution which is determined by maximizing the Tsallis entropy. Here, we consider a network of sensor nodes supporting multipath transmission; thus, a prior information about the location of a set of SNs is essential in order to infer the location of other SNs in the network. Therefore, by using an efficient sensor localization techniques, the initial cost associated with the deployment of SNs, as well as the cost for powering each node in a WSN, can be reduced greatly. As we are considering a multipath UWB communication between the SNs, the prime source of error is due to the presence of multiple paths between the transmitter and the receiver. Therefore, due to the presence of these errors originating from various points in a WSN, the estimates obtained using the TOA localization get degraded [10, 16]. So, it becomes quite challenging and inevitable to characterize the variations in range for multipath communication channels. Here, we employ maximum Tsallis' entropy framework subjected to the normalization constraint, and the first two moment constraints, which yields the $q-$Gaussian distribution. The different values of the non-extensive parameter q characterize the variations in range. Beck and Cohen [18] proposed a superstatistics framework for modeling complex systems, by averaging over the environmental fluctuations. Thus by following [18], we design a mixture model for the variations in range by incorporating the fluctuations caused due to multipath components.

We also present a robust technique for estimation of the parameters corresponding to the PDFs of the mixture model and the proposed model, in order to characterize the synthetic data.

The rest of this paper is as follows: Sect. 2 illustrates the problem statement. In Sect. 3, we derive the $q-$Gaussian distribution using the maximum Tsallis' entropy framework. The parameter estimation techniques corresponding to the mixture model and our proposed model are discussed in Sect. 4. Section 5 deals with the characterization of range for different values of q. Finally, in Sect. 6 we provide the conclusion and future works.

2 Problem Statement

When a signal travels from the sender (or transmitter) to a receiver, then its propagation time can be estimated by using the TOA technique. Generally, TOA estimation is used along with UWB technology for an energy-efficient ranging in IEs with multiple propagation paths. We use localization of SNs to quantify the signal obtained from a heterogeneous set of sensors. Figure 1 illustrates the deployment of several SNs in a smart home. These SNs are positioned abundantly throughout the smart home and can be used to sense diverse indoor environments.

Figure 2 shows the existence of multiple transmission paths in ubiquity of several interfering objects. These signals undergo scattering and reflection due to interference of several objects present in an indoor environment. The interfering objects scatter the waves while being transmitted from the transmitter (Tx) to the receiver (Rx) which leads to difficulty in identifying the actual location of a target node.

We can model the UWB channels having multipath components by utilizing the impulse response [11], i.e.,

$$h\left(t\right) = \sum_{l=0}^{m} a_l \delta\left(t - \tau_l\right), \tag{1}$$

where a_l is the amplification factor for the lth multipaths and τ_l represents the time delay of the channel.

Fig. 1 Deployment of multiple sensors in a smart home environment

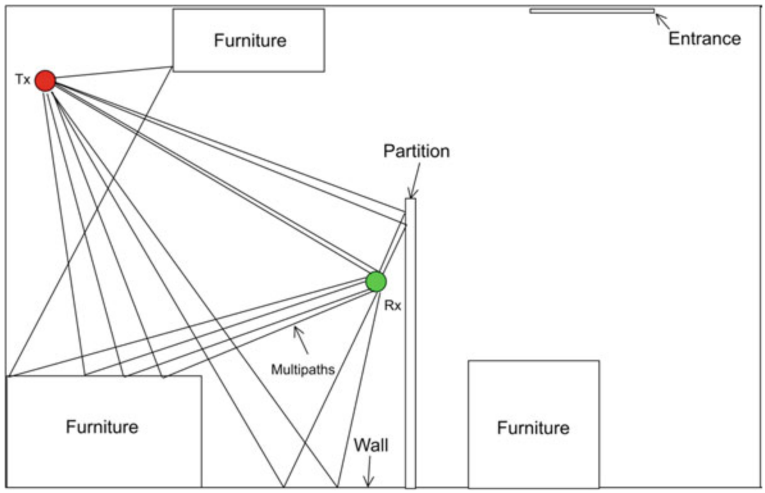


Fig. 2 Variations in range under the presence of multiple paths in IEs

Therefore, using Eq. (1), the received signal corresponding to the TOA estimate can be characterized using [13] as

$$r_{i,l}\left(t\right) = \sum_{i=0}^{n} \sum_{l=0}^{m} a_{i,l} s\left(t - \tau_{i,l}\right) + n_{i,l}\left(t\right), \tag{2}$$

where $s\left(t\right)$ represents the transmitted signal and $n_{i,l}\left(t\right)$ denotes the additive white Gaussian noise. Thus, it can be observed that under NLOS conditions these channels reflect dense multipath effects and time delays which may affect the overall performance of the system.

Let us consider the sensor nodes to be deployed in an indoor environment. If we consider the TOA ranging for determining the location of the sensor nodes, then their range can be given as

$$r_i = d_i + n_i\,, \quad i = 0, 1, 2, \ldots, n, \tag{3}$$

where r_i is the measure of the range between a source and receiver for ith sensor nodes and d_i is the distance between the source and the ith sensor and can be stated as

$$d_i = ||x - x_1||_2 = \sqrt{(x - x_i)^2 + (y - y_i)^2}, \tag{4}$$

here, $(x, y) \in X$ is a set of remote source locations and $(x_i, y_i) \in X_i$ is the set of recognized coordinates of the ith sensor, and n_i is the additive white Gaussian noise, such that $n_i \sim N\left(0, \sigma_i{}^2\right)$.

Thus if we characterize the range in terms of a Gaussian distribution [19], then its probability density function (PDF) can be expressed as

$$f(r_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \mathrm{e}^{-\frac{(r_i - d_i)^2}{2\sigma_i{}^2}}, \tag{5}$$

where $r_i \sim N\left(d_i, \sigma_i{}^2\right)$, with $\sigma_i{}^2$ representing the variance corresponding to r_i.

However, the variations $\left(\sigma_i{}^2\right)$ in the range r_i can be characterized by inverse Gamma distribution [17, 18, 20]. So, the parameter σ_i can be considered to be a random variable such that

$$\beta_i = \frac{1}{\sigma_i{}^2} \sim Gamma\,(a, b)\,, \tag{6}$$

where a and b are the shape and scale parameters, and β_i is the inverse of the variations caused in the range for ith SNs.

Let $g\,(\sigma_i)$ be the variational range distribution and be defined as

$$g\,(\sigma_i) = \frac{1}{b^a \Gamma\,(a)} \beta_i{}^{a-1} \mathrm{e}^{-\frac{\beta_i}{b}}, \quad \beta_i \geq 0. \tag{7}$$

The PDF for the gamma distribution in compliance with the synthetic data is illustrated in Fig. 5.

The unconditional distribution of range is obtained by averaging over the gamma distribution [17, 25],

$$p\,(r_i) = \int_0^\infty f\,(r_i|\sigma_i)\, g\,(\sigma_i)\, \mathrm{d}\sigma_i, \tag{8}$$

where $f(r_i|\sigma_i)$ is the conditional probability distribution of r_i for a given σ_i.
Now, we have

$$p(r_i) = \int_0^\infty \sqrt{\frac{\beta_i}{2\pi}} e^{-\frac{\beta_i (r_i - d_i)^2}{2}} \frac{1}{b^a \Gamma(a)} \beta_i{}^{a-1} e^{-\frac{\beta_i}{b}} \frac{1}{2} (\beta_i)^{-\frac{3}{2}} d\beta_i. \tag{9}$$

So, the probability distribution corresponding to the variations in range observed in an indoor multipath environment is given as

$$p(r_i) = \frac{\Gamma(a-1)}{2\sqrt{2\pi} b^a \Gamma(a)} \left[\frac{1}{b} + \frac{(r_i - d_i)^2}{2}\right]^{-(a-1)}. \tag{10}$$

The above mixture model along with the synthetic data is shown in Fig. 6.

3 Proposed Model

In order to obtain accuracy in inferring the location of SNs in an IE, we need to characterize the variations in the measurement of the range observed due to multipath components. If we consider r_i to be a random variable, then by applying Tsallis entropy [6, 7, 21–23], the uncertainties in the range obtained in Eq. (3) can be characterized through the non-extensive parameter q and is expressed as

$$S_q(r_i) = \frac{1 - \int_{-\infty}^{\infty} \left[f_q(r_i)\right]^q dr_i}{q-1}. \tag{11}$$

Here, we investigate the probability distribution corresponding to the variations in range for smart home sensors by maximizing the Tsallis entropy framework subjected to the normalization constraint and the moment constraints.
Thus, the constraint for normalization can be given as

$$\int_{-\infty}^{\infty} f_q(r_i) dr_i = 1. \tag{12}$$

The first two constraints for the moments of r_i can be specified as

$$\int_{-\infty}^{\infty} f_{es}(r_i) r_i dr_i = \mu, \tag{13}$$

and

$$\int_{-\infty}^{\infty} f_{\mathrm{es}}(r_i-\mu)^2 \mathrm{d}r_i = \sigma^2, \tag{14}$$

where μ and σ^2 indicate the mean and variance.

Therefore, the probability distribution in Eqs. (13) and (14) corresponding to any probability f_q can be expressed as

$$f_{\mathrm{es}}(r_i) = \left[f_q(r_i)\right]^q \Big/ \int \left[f_q(r_i)\right]^q \mathrm{d}r_i, \tag{15}$$

where $f_{\mathrm{es}}(r_i)$ denotes the escort probability and is used to represent the asymptotic decay of q moments. Since we are using a parameter q, so the type of expectation used for defining the variations in the range is called the $q-$expectation. Thus, by considering the Lagrangian, for any stationary SN, we can optimize the range of the received signal as

$$\begin{aligned} L\left(f_q, r_i\right) = {} & \frac{1-\int_{-\infty}^{\infty} f_q(r_i)^q \mathrm{d}r_i}{(q-1)} + \omega_1\left(1-\int_{-\infty}^{\infty} f_q(r_i)\,\mathrm{d}r_i\right) \\ & + \omega_2\left(\mu-\int_{-\infty}^{\infty} r_i f_{\mathrm{es}}(r_i)\, d_i\right) + \omega_3\left(\sigma^2-\int_{-\infty}^{\infty} f_{\mathrm{es}}(r_i-\mu)^2 \mathrm{d}r_i\right), \end{aligned} \tag{16}$$

where ω_1, ω_2 and ω_3 are the Lagrangian parameters. Now by employing the Euler–Lagrange equation, we have

$$\frac{\partial L}{\partial f_q(r_i)} - \frac{\mathrm{d}}{\mathrm{d}r_i}\left(\frac{\partial L}{\partial f_q{}'(r_i)}\right) = 0 \tag{17}$$

Thus, by using Eq. (17) in association with Eqs. (11)–(16), we have

$$f_q(r_i) = \frac{1}{z}\left(1-\frac{1-q}{(3-q)\,\sigma^2}(r_i-\mu)^2\right)^{1/(1-q)}, \quad 1<q<3, \tag{18}$$

where $z = \sigma\sqrt{\frac{(3-q)\pi}{q-1}}\frac{\Gamma\left(\frac{3-q}{2q-2}\right)}{\Gamma\left(\frac{1}{q-1}\right)}$ and is called the normalization constant. From Eq. (18), we obtain the maximum value for the entropy probability distribution corresponding to the variations in range of the TOA estimate.

4 Parameter Estimation

In this section, we discuss the parameter estimation for the mixture model obtained in Eq. (10) and the proposed $q-$Gaussian distribution obtained in Eq. (18). The shape parameter 'a' and scale parameter 'b' of the mixture distribution can be estimated through the probability distribution of 'σ_i'. The random variable σ_i is defined as

$$\sigma_i\left(r_i\right) \equiv \left[\frac{1}{N}\sum_{\tau=1}^{N}\left|v_\tau\left[t:t+\tau\right]\right|\right]^2, \quad i=1,2,\ldots,n, \tag{19}$$

where N is the number of samples per time stamp (τ) and v_τ represents the variations in the range within the interval $[t, t+\tau]$.

The estimate of non-extensive parameter 'q' can be expressed through the generalized Jensen–Shannon (JS) symmetric measure. Thus, the symmetric JS divergence measure can be defined as

$$\begin{aligned}\mathrm{JS}\left(f_q\left(r_i\right)||g\left(r_i\right)\right) = &\frac{1}{2}\mathrm{KL}\left(f_q\left(r_i\right)||\frac{f_q\left(r_i\right)+g\left(r_i\right)}{2}\right)\\ &+\frac{1}{2}\mathrm{KL}\left(g\left(r_i\right)||\frac{f_q\left(r_i\right)+g\left(r_i\right)}{2}\right)\end{aligned} \tag{20}$$

where $KL(.)$ is the generalized Kullback–Leibler measure. This measure provides the distance between the proposed distribution $f_q(r_i)$ and the empirical distribution $g(r_i)$ [7, 24]. Thus, the KL measure is given as

$$\mathrm{KL}\left(f_q\left(r_i\right)||g\left(r_i\right)\right) = \sum_{k=1}^{n} f_{q_k}\left(r_i\right)\log_2\left|\frac{f_q\left(r_i\right)}{g_k\left(r_i\right)}\right| \tag{21}$$

The optimal value corresponding to the non-extensive parameter q is attained at which the symmetric JS divergence is minimum.
Thus, the optimal value of the $\hat{q}$ estimate provides a closed agreement between the synthetic data and the proposed theoretical distribution. Therefore, the optimal value of q is estimated as

$$\hat{q} = \underset{q\in(1,3)}{\arg\min}\left\{\mathrm{JS}\left(f_q\left(r_i\right)||g\left(r_i\right)\right)\right\} \tag{22}$$

5 Results and Discussion

In this section, we plot the PDF for the maximum entropy obtained in Sect. 3 for the different values of q, viz., $q=1.5$ (proposed model), $q=1.37$ and $q=1.45$. It can be observed that at $q=1.5$, our model successfully captures the variations in range

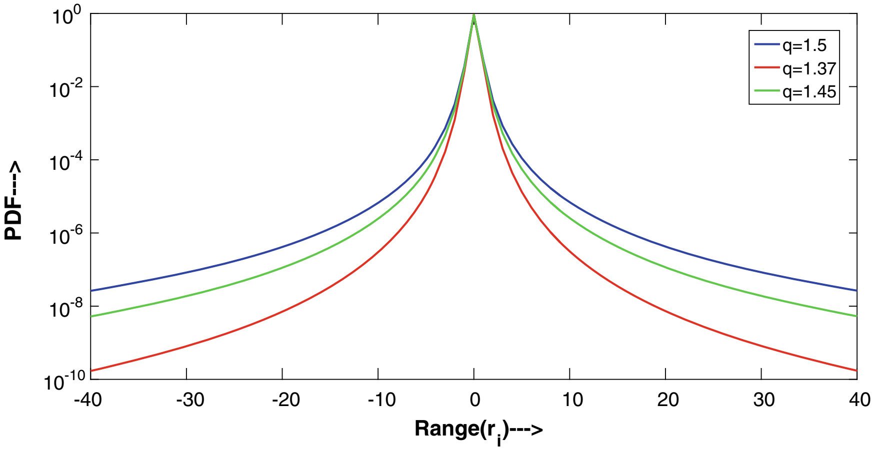


Fig. 3 $q-$Gaussian PDFs for different values of q, viz., $q = 1.5$ (proposed model), $q = 1.37$ and $q = 1.45$

caused due to indoor multipath propagation. It is worth noting that for $q = 1.5$ (as shown in Fig. 3), the model exhibits the long memory behavior and extensively entails the extreme fluctuations in the range. In order to validate our model, we generate synthetic data using MATLAB incorporating multipath components. Figure 4a shows the trend plot for the synthetic data with a sample size of 1000, and Fig. 4b shows its respective histogram. We transformed the variations in range for indoor environments to the gamma distribution, obtained in Eq. (7), and its compliance with the synthetic data is shown in Fig. 5. In IEs, the presence of dense multipath and the ubiquity of NLOS conditions causes degradation in the estimation of TOA. From Fig. 6, it can be observed that our proposed framework captures the dense multipath components of the generated signal even at the tail end. From the above discussions, we can draw an inference that when $q = 1.5$, the model provides a better characterization for the uncertainties in the TOA estimation. Further, we can obtain greater precision for estimating the range of SNs, using TOA with resolutions for providing more accuracy in temporal and multipath components.

6 Conclusion and Future Work

In this paper, we discussed the UWB-based TOA localization and variations in its range for indoor environments. We characterized the variations in range for TOA localization by using Tsallis' entropy framework. The maximized Tsallis entropy subjected to the first two moment constraints (i.e., mean and variance) along with the normalization constraint is derived. It is observed that our model captures the variations in the localization range in the existence of dense multipaths in indoor

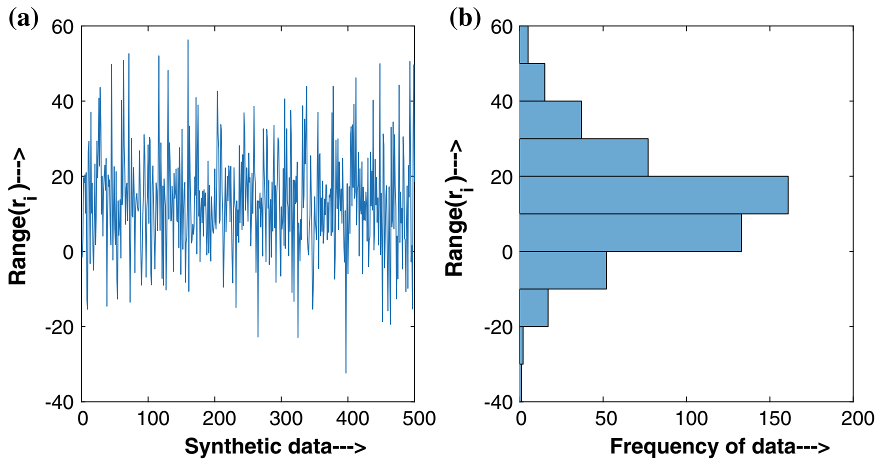


Fig. 4 **a** Representation of the trend plot with sample size 1000. **b** Representation of histogram corresponding to the trend plot

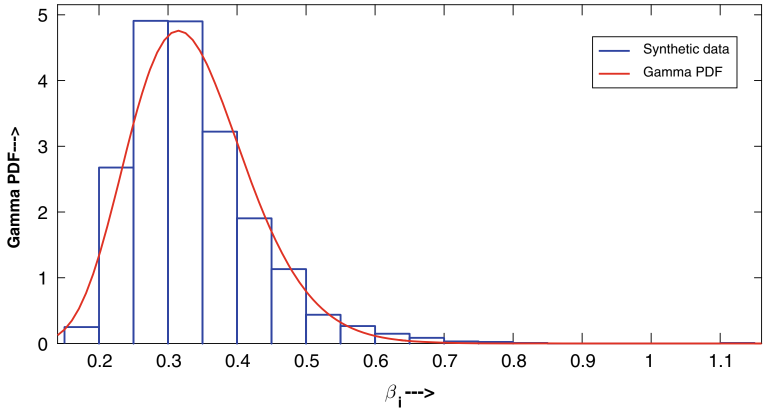


Fig. 5 Representation of the gamma PDF in convergence with the synthetic data for $a = 9.04432$ and $b = 0.0285463$

environments. Our proposed $q-$Gaussian model operates well over the synthetic data in contrast to the mixture model. It is observed that the $q-$Gaussian distribution provided a better characterization of the variations in the localization range for $q = 1.5$. This paper also presented a new approach for estimation of parameters corresponding to the mixture model and our proposed model, by minimizing the JS symmetric measure between the two models and the synthetic data.

In the future, we will provide an intrinsic stochastic differential equation (SDE) model and its corresponding Fokker–Planck equation, to entail the variations in the localization range for densely positioned indoor environments.

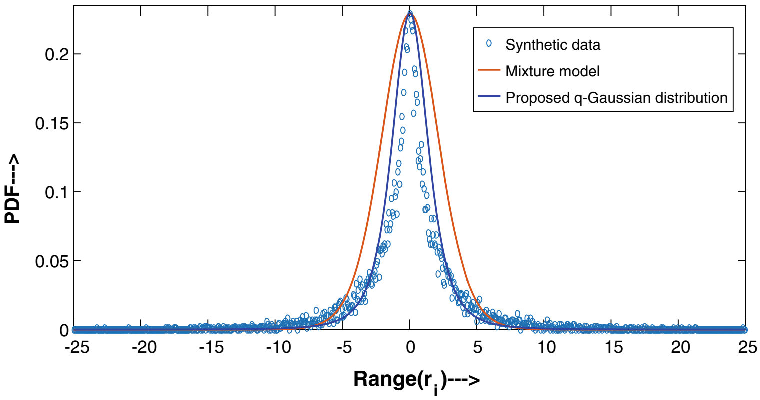


Fig. 6 Fitting of PDFs for the mixture model for $a = 9.04432$ and $b = 0.0285463$ along with the proposed $q-$Gaussian model over the synthetic data for uncertainties in range, obtained for $q = 1.5$ with $\mu = 0.0428$ and $\sigma = 1.4138$.

References

1. Porcino, D., Hirt, W.: Ultra-wideband radio technology: potential and challenges ahead. IEEE Commun. Mag. **41**(7), 66–74 (2003)
2. Oppermann, I., Hämäläinen, M., Iinatti, J. (eds.): UWB: Theory and Applications. Wiley & Sons (2005)
3. Rajput, N., Gandhi, N., Saxena, L.: Wireless sensor networks: apple farming in northern India. In: 2012 Fourth International Conference on Computational Intelligence and Communication Networks (CICN). IEEE (2012)
4. Gezici, S., et al.: Localization via ultra-wideband radios: a look at positioning aspects for future sensor networks. IEEE Signal Process. Mag. **22**(4), 70–84 (2005)
5. Tsallis, C., et al.: Nonextensive statistical mechanics and economics. Physica A Stat. Mech. Appl. **324**(1–2), 89–100 (2003)
6. Tsallis, C., Non-extensive Statistical Mechanics: Construction and physical interpretation. In: Nonextensive Entropy Interdisciplinary Applications, pp. 1–52 (2004)
7. Senapati, D., Karmeshu: Generation of cubic power-law for high frequency intra-day returns: maximum Tsallis entropy framework. Digit. Signal Process. **48**, 276–284 (2016)
8. Singh, A.K.: Karmeshu: Power law behavior of queue size: maximum entropy principle with shifted geometric mean constraint. IEEE Commun. Lett. **18**(8), 1335–1338 (2014)
9. Singh, A.K., Singh, H.P., Karmeshu: Analysis of finite buffer queue: maximum entropy probability distribution with shifted fractional geometric and arithmetic means. IEEE Commun. Lett. **19**(2), 163–166 (2015)
10. Lee, J.-Y., Scholtz, R.A.: Ranging in a dense multipath environment using an UWB radio link. IEEE J. Sel. Areas Commun. **20**(9), 1677–1683 (2002)
11. Ghavami, M., Michael, L., Kohno, R.: Ultra wideband signals and systems in communication engineering. Wiley & Sons (2007)
12. Abe, S., Bagci, G.B.: Necessity of q-expectation value in nonextensive statistical mechanics. Phys. Rev. E **71**(1), 016139 (2005)
13. Saleh, A.A.M., Valenzuela, R.: A statistical model for indoor multipath propagation. IEEE J. Sel. Areas Commun. **5**(2), 128–137 (1987)
14. Pietrzyk, M.M., von der Grün, T.: Experimental validation of a TOA UWB ranging platform with the energy detection receiver. 2010 International Conference on Indoor Positioning and Indoor Navigation (IPIN). IEEE (2010)

15. Alavi, B., Pahlavan, K.: Modeling of the TOA-based distance measurement error using UWB indoor radio measurements. IEEE Commun. Lett. **10**(4), 275–277 (2006)
16. Zhou, Y., et al.: Indoor elliptical localization based on asynchronous UWB range measurement. IEEE Trans. Instrum. Meas. **60**(1), 248–257 (2011)
17. Gerig, A., Vicente, J., Fuentes, M.A.: Model for non-Gaussian intraday stock returns. Phys. Rev. E **80**(6), 065102 (2009)
18. Beck, C., Cohen, E.G.D.: Superstatistics. Physica A Stat. Mech. Appl. **322**, 267–275 (2003)
19. Zhao, W., et al.: Outage analysis of ambient backscatter communication systems. IEEE Commun. Lett. **22**(8), 1736–1739 (2018)
20. Beck, C., Cohen, E.G.D., Swinney, H.L.: From time series to superstatistics. Phys. Rev. E **72**(5), 056133 (2005)
21. Rajput, N.K., Ahuja, B., Riyal, M.K.: A novel approach towards deriving vocabulary quotient. Digit. Sch. Humanit. **33**, 894–901 (2018)
22. Riyal, M.K., et al.: Rank-frequency analysis of characters in Garhwali text: emergence of Zipf's law. Curr. Sci. **110**(3), 429 (2016)
23. Schäfer, B., et al.: Non-Gaussian power grid frequency fluctuations characterized by Lévy-stable laws and superstatistics. Nat Energy **3**(2), 119 (2018)
24. Osán, T.M., Bussandri, D.G., Lamberti, P.W.: Monoparametric family of metrics derived from classical JensenShannon divergence. Phys. A Stat. Mech. Appl. **495**, 336–344 (2018)
25. Papadimitriou, K., Sfikas, G., Nikou, C.: Tomographic image reconstruction with a spatially varying gamma mixture prior. J. Math. Imaging Vis. 1–11 (2018)
26. Kaiwartya, O., Abdullah, A.H., Cao, Y., Rao, R.S., Kumar, S., Lobiyal, D.K., Isnin, I.F., Liu, X., Shah, R.R.: T-MQM: testbed-based multi-metric quality measurement of sensor deployment for precision agriculture—a case study. IEEE Sensors J. **16**(23), 8649–8664 (2016)

Call Data Analytics Using Big Data

V. N. S. Manaswini and K. V. Krishnam Raju

Abstract One of the important strategies that can increase the business success rate is "customer monitoring" in which customer service representatives (CSRs) measure the satisfaction of customers. Here, the only way to know about customer's experience (CX) is conducting the surveys which are either through phone calls or sending emails. The main challenge of customer monitoring is to know about customer's experience, and the customer service representatives record their phone calls and analyze those recorded calls by converting them into text files. They maintain a large amount of memory to store a huge number of audio files and text files. This results in data tempering, corruption of data, unauthorized access to tables, columns and rows and burden of managing the data. In this paper, we replace the recorded files with direct phone calls. Now, we can convert the phone calls to text files with the help of speech-to-text (STT) algorithm, then analyze a huge amount of text files using Hadoop MapReduce Framework and apply the text similarity algorithms for getting better results to improve the business.

Keywords Customer monitoring · Speech-to-text algorithm · Hadoop MapReduce · Text similarity algorithms

1 Introduction

Big data [1] is used to describe a large amount of data (>1 TB) that cannot be stored and processed using database management software tools and some other traditional methods. Let us take some examples of big data which are maintaining a large amount of data such as Google Search index (GSI), the database of Twitter and Facebook user's profiles and the data of e-commerce Web site's product list. Due to the large amount of data, it is very difficult to store in database management system and also

V. N. S. Manaswini (✉) · K. V. Krishnam Raju
Computer Science Engineering, S.R.K.R.Engineering College, Bhimavaram 534204, India
e-mail: manaswini.vemuri@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_24

big data is distributed where it takes data from multiple sources and processes them for better solution. Big data is widely classified [2] into three types of data those are as follows.

Structured data: The data which is in the form of ordinary format is known as structured data which is stored and processed in the database table with rows and columns. But only 5–10% of information is used by structured data.

For example: Employee details in database table.

Semi-structured data: The data is in the form of semi-structured that means data does not store directly in relational database but with some process we can store data in database even though some data does not store in database.

For example: XML and JSON documents are semi-structured documents, and NoSQL databases are considered as semi-structured.

Unstructured data: The data is in the form of unstructured that means data consists in different formats. Nowadays, unstructured data is everywhere. Everyone is sharing information in the form of unstructured data like social media, blogs, through email, chatting, and some additional ways are YouTube, other video and audio sources and through phone calls.

Let us consider one live example social media in which users share information in any ways like sending messages, uploading images or videos, giving feedbacks in Web sites, etc.

Basically, big data is mainly useful for storing and analyzing the unstructured data. Big data analytics is nothing but analytical techniques are applied on big data sets for that we will use some technologies like Hadoop MapReduce and streaming. Big data is in many forms of unstructured data like social media data, transaction data, system-to-system data, biometric data and human-generated data [3]. In this paper, we mainly work on human-generated data that consists of emails, documents, voice recordings and data through phone calls, surveys and so on that create and share data with others. Here, customer service representatives analyze the data directly from customer calls.

For converting live calls to text files, we are using one speech recognition library tool known as Sphinx4 tool where it is an open source [4] and it works on Java platform. It is based on hidden Markov model (HMM). The main advantage of this tool is identifying the sounds accurately, and under time stamping it aligns speech to text.

Text analytics [5] is also known as text mining after converting the live calls into text files. With the help of text analytics, we convert the unstructured data into some meaningful data; then, it seems to be the good result. With the usage of some techniques, we will get accurate data such as traditional machine learning techniques, natural language processing, information retrieval and knowledge management. It is having various applications such as security applications, sentiment analytics, software applications, social media applications, business and marketing analytics (BMA) and predictive analytics.

2 Related Work

Shim et al. [3] proposed the system in which they introduced phonetic search technology on speech for better performance. The main drawback of this paper is they maintained a large number of recorded files.

Elgendy and Elragal [6] aimed to analyze some various analytical techniques and tools which are used in big data for analyzing the text accurately, and with the help of big data storage and management they store the large amount of unstructured data by using ETL and ELT tools and manage the data using data mining techniques and online analytical tools.

Karakus and Aydin [7] proposed a distributed system for call monitoring system, and they tried to know their agent's performance by using some text similarity techniques. The main issue of this system is they considered recorded files to analyze the call data by converting them into text file but with that recorded files some storage issues were raised. So, in our paper we replaced recorded calls with live calls, so by this we reduce the storage issues.

Chowdhury et al. [8] acquired the call data records for their project which are used from telco operator. They analyze the call data. With basis of call data sets, they find which subscriber is related to that data with the help of MapReduce function. With the help of this paper, we analyze the customer's data and get the report under the basis of data.

Pal and Saha [9] proposed that how to detect slang words using Synset and concept models from supervised learning and how to remove the slang words from text; by this, they used some modules. This paper is useful to our paper on detecting the slang words.

3 Big Data Analytics (BDA)

For improving the business growth, the customer service representatives (CSRs) try to satisfy the customers. So, they analyze a huge amount of customer's call data using big data analytics [6]. It includes text analytics, natural language processing, machine learning, data mining techniques for getting better result which is helpful to satisfy the customers. To analyze the large amount of data, we are using Apache Hadoop technology. It is most powerful and popular big data tool and an open-source platform that provides data management progress, and it is distributed from one server to many client systems.

The Hadoop ecosystem [10] consists of various components, tools and accessories which are provided by the Apache Software Foundation (ASF); these components increase the function ability to Hadoop.

According to Fig. 1, we will discuss each and every component of Apache Hadoop software in detail.

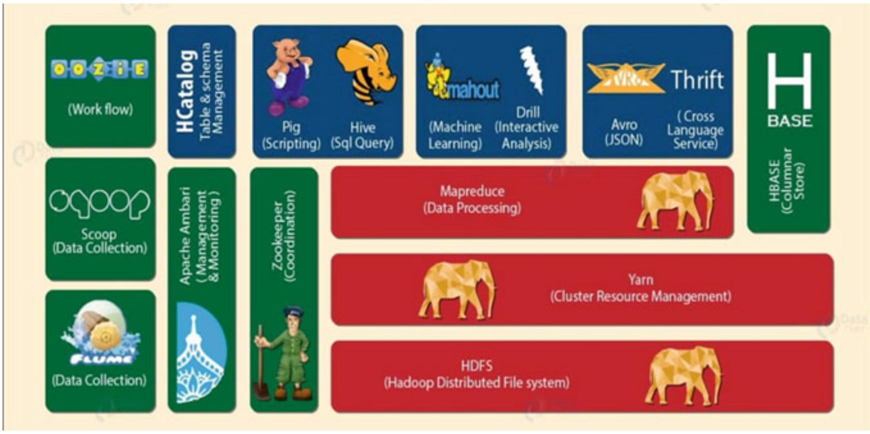


Fig. 1 Hadoop ecosystem components

Hadoop distributed file system (HDFS)

HDFS is one of the main components of Hadoop ecosystem. It is the main storage system of Hadoop which stores input and output files. HDFS is having two main components those are as follows.
Name node: It is also known as "master node" where it stores metadata and it contains number of blocks, location of block, rack of block, type of the data node and so on.
Data node: It is also known as "slave node" where it stores the actual data in HDFS and it performs read and write operations which depend on client's request. Data node is having replica block; it consists of two files in which data is stored in first file and second file is used for records of block's metadata; it includes checksums, start-up, handshaking process.

MapReduce

MapReduce tasks are controlled by master node which are split into two phases those are Map and Reduce. In Map phase, the input data is divided in the form of key–value pairs and the output is sorted with the help of key. In Reduce phase, it combines the values of input data to get final result.

MapReduce is implemented in different programming languages. This framework is having features like fault tolerance, scalability, data-based optimization and automatic parallelization.

Yet Another Resource Negotiator (YARN)

Hadoop YARN is related to resource management in Hadoop. It is also called as operating system of Hadoop which is used to manage and monitoring the workloads. It is also involved in multiple data processing engines such as batch processing and real-time streaming.

Hive

Apache Hive is an open-source data warehouse (DWH) system which is used for generating query and analyzes a large number of data sets that are stored in Hadoop files.

Pig

Apache Pig is one of components in Hadoop ecosystem. It is a high-level language and is used for querying and analyzing a large number of data sets that are stored in Hadoop distributed file system (HDFS). It is also similar to SQL which retrieves the data, applies filters to data and stores that data in required format.

HBase

Apache HBase is a scalable distributed and NoSQL database which is used to store structural data in database tables where it contains large number of rows and columns which is built on top of Hadoop distributed file system (HDFS). There are two components: HBase Master and region server.
HBase Master: It is not related to the actual data storage, but it balances the load across all region server. It is administrator and can handle DDL operations.
Region server: It is a worker node which can handle some requests like write, read, update and delete from clients. It runs on every node in Hadoop distributed file system (HDFS) data node.

HCatalog

HCatalog is a storage management layer for Hadoop system. It supports some components like Hive, MapReduce and Pig. It is used to store the data in any format by the user.

Avro

It is an open-source system that provides data exchange and data serialization features. It can support either through independent or together in nature.

Thrift

Thrift is a software framework which is scalable and used for cross-language service development. It is an interface for remote procedure call (RPC).

Apache Drill

It is a large-scale data processing which includes both structured and semi-structured data. It is having low-latency distributed query engine that is designed on large amount of data.

Apache Mahout

It is also an open source which is creating machine learning algorithm and data mining library. After storing the data in HDFS, Mahout provides some tools to find meaningful patterns in large data sets. Some of the algorithms are clustering, collaborative filtering, classifications and frequent pattern mining.

Apache Sqoop

It imports the data from different sources into related components like HBase, HDFS or Hive and also exports the data from Hadoop to other sources. It works with relational database management system (RDBMS) such as Oracle, MySQL, Microsoft SQL Server, PostgreSQL, Teradata, IBM Db2 and Netezza.

Flume

Flume effectively gathers aggregate and moves a huge amount of data from its starting point and sending it back to HDFS. It follows fault-tolerant and reliable mechanisms. With the help of flume, we can get data from various servers into Hadoop.

Ambari

It is a management platform for managing, provisioning, securing and monitoring Hadoop cluster.

ZooKeeper

It is a centralized service for maintaining, naming, configuration info, distributed synchronization and providing group services. It manages and coordinates a big cluster of machines.

Oozie

Oozie is a workflow scheduling system for controlling Hadoop jobs. It combines more number of jobs sequentially into one unit of work.

4 Call Data Analytics

In this paper, we propose a new customer monitoring system to analyze the customer's call data which was conducted between customer service representatives (CSRs) and customers in the customer care center. The main view of proposed system is observed in detail by Fig. 2.

The architecture of this system (Fig. 2) is mainly partitioned into three blocks, namely conversion of call data and data storage, analysis of call data and reports on customer's experience.

4.1 Conversion of Call Data and Data Storage

For analyzing the call data first, we convert the direct phone calls to text files with the help of Sphinx4 tool which is worked on Java platform.

After conversion of data, we store the call data by using Apache HBase component. It is a NoSQL database that is built on the top of Hadoop. Then, that result is stored in HBase tables in XML format.

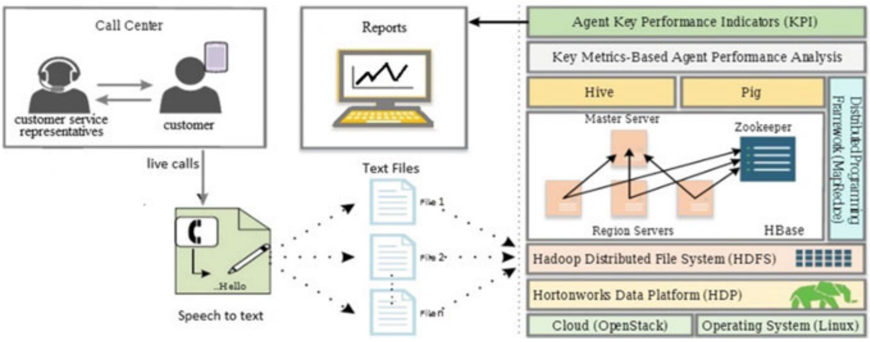


Fig. 2 Architecture of call data analytics

4.2 Analysis of Call Data

For analyzing the call data, customer care center generally uses key metric-based process. For this process, they analyze the call data under some metrics those are:

Slang word detection

The customer may use slang words [9] during the conversation between representatives and customer. In such case, representatives should warn the customers. Sometimes, representatives should end the conversation with customers too. For detecting the slang words first, we have to create a list of English slang words which are stored in HBase tables. Then, compare these conversation words with slang word list and check whether it is slang word or not. In this paper, Synset and concept methods are used to analyze the slang words.

Detecting text similarity

The main goal of text similarity [11] is to find the similarity between two or more words/phrases/documents. Similarity of words is finding either through meaning of word or surface closeness. The text similarities are mainly in two ways those are lexical similarity and semantic similarity.

Lexical similarity

Lexical similarity is based on surface closeness where large amount of the people prefers the lexical similarity due to the surface, but it is not so accurate than semantic similarity. Let us consider one example that is finding the similarity between two sentences.

For example: “the dog ate the cat” with “the cat ate the dog food.”

In this example, so many words are similar but an actual meaning is different from another sentence.

Generally, lexical similarity is used in information retrieval, redundancy removal and clustering areas. According to Fig. 3, lexical similarity is divided into two ways, namely character-based and statement-based.

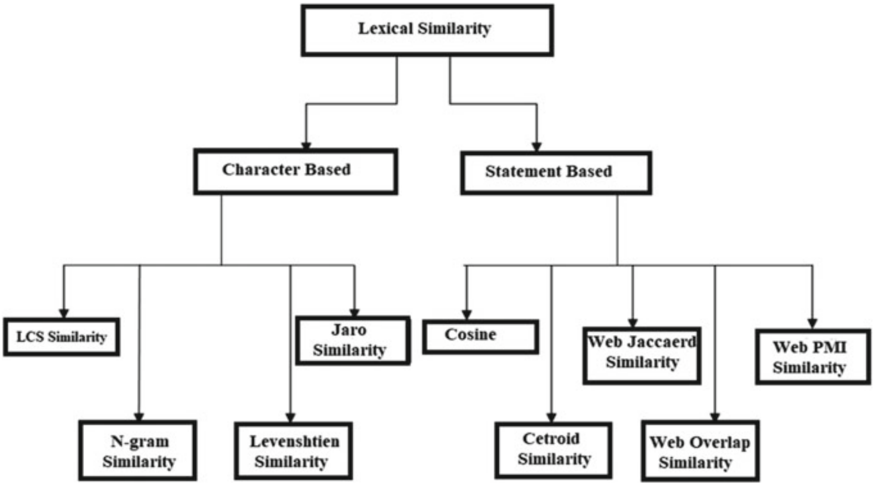


Fig. 3 Lexical similarity

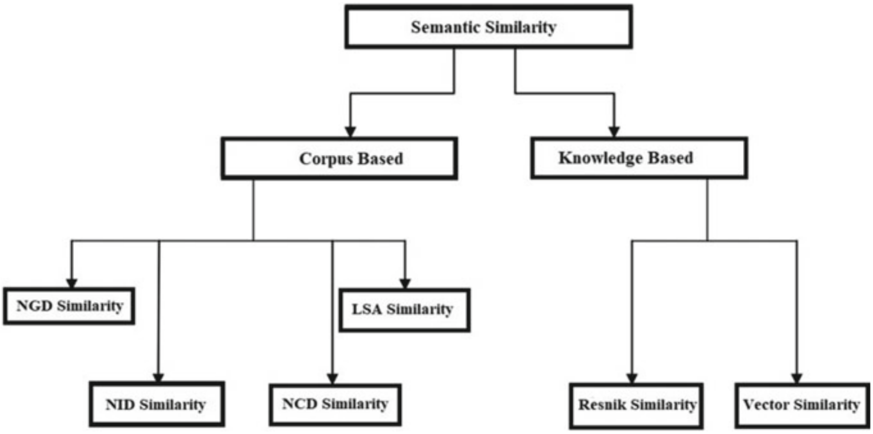


Fig. 4 Semantic similarity

Semantic similarity: It is based on meaning of words rather than word-by-word matching. According to Fig. 4, semantic similarity is divided into two ways, namely corpus-based and knowledge-based.

In this paper, we are using N-gram and Jaccard coefficient for getting accurate result.

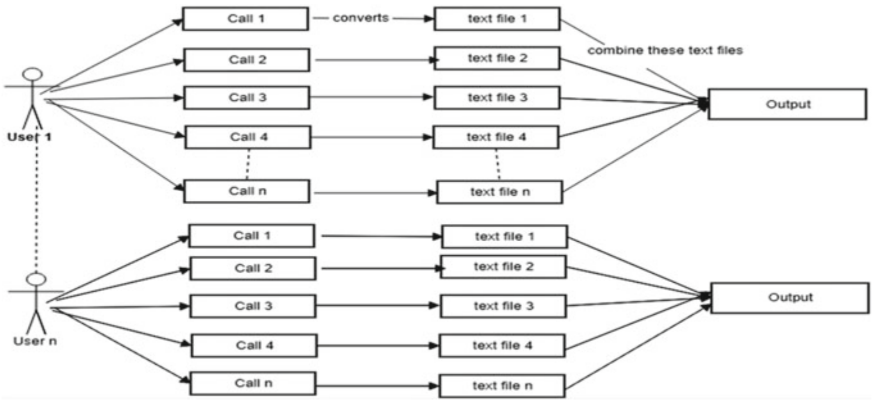


Fig. 5 MapReduce function for call data analytics

4.3 Reports on Customer's Experience

After finding the similarity of text, we are implementing Hadoop MapReduce function for getting final report. First, Map phase divides the input data into key–value pairs and then the output of map tasks is sorted with related key. The Reduce phase merges the value into final report.

In Fig. 5, we store all text files and then produce key–value pairs; then according to key-based, we combine all the text files, analyze those text files and then reduce the file by merging the values into final report.

5 Implementation

Algorithm 1: This algorithm converts the speech to text using Sphinx tool on Java platform. The time complexity of an algorithm is O(n).

Input: Speech using microphones

Output: Text document

1. Speaking words/phrases with microphones.
2. Sphinx4 library tool detects the speech from live and recognizes that sound.
3. Then save that output in the form of text.
4. Now compare the output file whether it is expected or not.
5. If it is expected, one then saves that file.
6. Else go to step 2.

Algorithm 2: This algorithm is describing how to remove the unwanted words from text, then find similarity of text and make a meaningful text.

Input: Input text

Output: Meaningful text

1. Eliminate the stop words from text file.
2. Now text file consists of meaningful text.
3. Check whether file is having suspicious words or not.
4. If text contains suspicious words, then apply related text similarity algorithm.
5. Else, no need to apply any algorithm.
6. If the text is free from unwanted words.
7. Stop
8. Else go to 4.

Algorithm 3: This algorithm is used to count each word in the file where file is distributed and chunked in Hadoop cluster. After completing the map function, it enters into reducer phase; else it gets an error and automatically execution will have stopped.

1. First, we store the call data set to HDFS.
2. Then, we will create one map function to analyze the call data. In this algorithm, map function will perform reading each line and then find customer's intention.
3. Combiner combines all the call data from mapper's output. This will result in key–value pairs as subscriber number and word frequency.
4. Later with help reducer, it sums all the values with basis of key. This result in subscriber number and total frequency of words in text.

In Fig. 6, we observe the input text which is chunked and distributed. Later, the data is to be shuffled and sorted; finally, it is reduced by reducer and gets an aggregated result.

Result Analysis

Here in Fig. 7, we are comparing the accuracy of phone calls and audio files with the help of Sphinx tool and we may found better accuracy result with direct calls.

In Table 1, we are analyzing the data using N-gram analysis for getting a better result (Fig. 8).

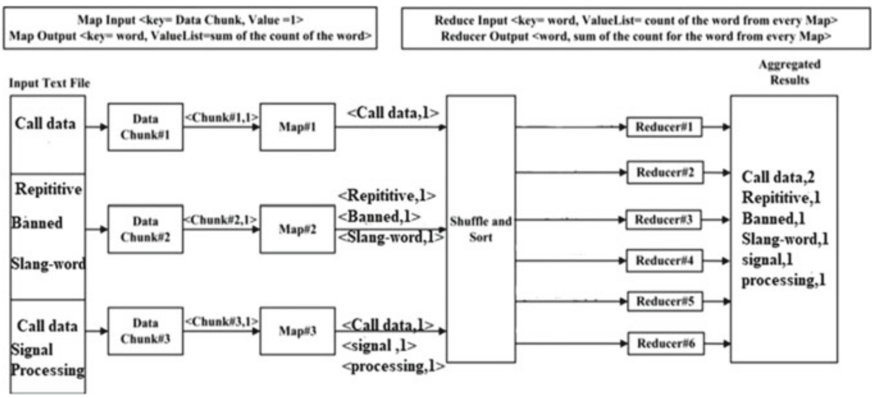


Fig. 6 MapReduce functionality

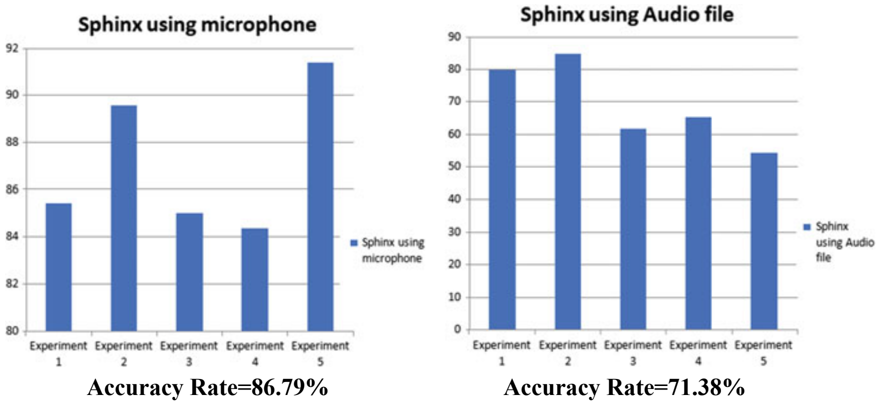


Fig. 7 Comparison between sphinx live calls and audio files

Table 1 Number of words generated for N-gram analysis

N-gram analysis in words	10.1 MB	21.2 MB	31.3 MB	40 MB	50 MB
N = 1	69,347	106,133	145,246	174,658	210,128
N = 2	800,921	1,334,591	2,024,822	2,591,408	3,220,503
N = 3	1,155,566	2,247,198	3,377,150	4,335,267	5,412,955

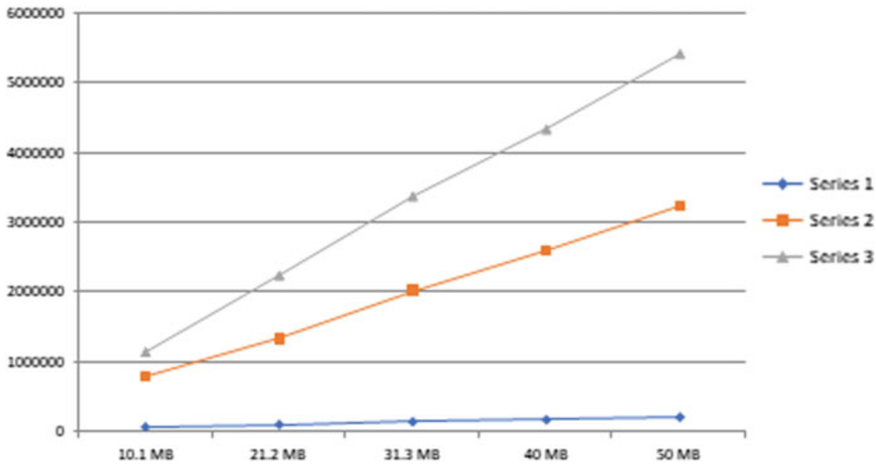


Fig. 8 Experimental details using Table 1 with results for words using N-gram analysis

6 Conclusion

In this paper, we proposed a new system for customer's satisfaction where customer service representatives analyze their customer's call data with the help of Apache Hadoop platform and we choose Sphinx4 tool for converting live speech to text file for better performance. With the help of text similarities, we analyze the data from text file and get aggregate result. The main issues of existing system are converted speech is stored in the form of recorded files by this process, and we may get some storage issues. Another one is lack of call record data. So, we dynamically implement the OpenStack on Hadoop which is related to cloud infrastructure for processing large amount of data. In the future, we are trying to improve the process for more accurate data.

Acknowledgements The call data used in the study is taken from Airtel Customer Care Center in Hyderabad.

References

1. https://techterms.com/definition/big_data
2. https://jeremyronk.wordpress.com/2014/09/01/structured-semi-structured-and-unstructured-data/
3. Shim, J.P., Koh, J., Fister, S., Seo, H.Y.: Phonetic analytics technology and big data: real-world cases. Commun ACM (2016)
4. Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., Woelfel, J.: Sphinx-4: a flexible open source framework for speech recognition (2004)
5. https://www.predictiveanalyticstoday.com/text-analytics/
6. Elgendy, N., Elragal, A.: Big data analytics: a literature review paper. In: ICDM 2014, LNAI 8557, pp. 214–227 (2014)
7. Karakus, B., Aydin, G.: Call center performance evaluation using big data analytics. In: ISNCC, IEEEXplore (2016)
8. Chowdhury, T.R., Arumugam, A.S.P., Lee, J.: Analysis of call data record (CDR) using Hadoop cluster. In: ASEE-NE'15 (2015)
9. Pal, A.R., Saha, D.: Detection of slang words in e-Data using semi-supervised learning. Int. J. Artif. Intell. & Appl. **4**(5) (2013)
10. https://data-flair.training/blogs/hadoop-ecosystem-components/
11. Pradhan, N., Gyanchandani, M., Wadhvani, R.: A review on text similarity technique used in ir and its application. Int. J. Comput. Appl. **120**(9) (2015)

Lightweight, Scalable and Secure Middleware for Service-Centric Context-Aware Applications

Veeramuthu Venkatesh, Pethuru Raj and P. Balakrishnan

Abstract Wireless sensor networks (WSNs) play a vital role in the Internet of things (IoT)-based application. IoT environment contains monitoring and collaboration from a wide number of remote or local sensor actuators, a device such as smartphones, robots and computers. In this paper, an efficient REST-based lightweight scalable and security approach proposed for providing secure data exchange without compromising the public quality issues. The communication scheme purely centred on the Extensible Messaging and Presence Protocol (XMPP) protocol. XMPP supports push-based notification that will be more proper functionality for any event-driven model for any sensor-based application. Based on these features, XMPP protocol, which integrated with the proposed approach, assures a scalable real-time event notification scheme. The framework is also adopting the light and portable Representational State Transfer (REST) web support features and consequently integrates sensor and intelligent things with the web. Here, the use of convention JavaScript Object Notation (JSON) format which will be the alternative for the verbose XML for data exchange. It also provides an encryption and data authentication required during transmission. The JSON is used along with the XMPP to support remote web services over wireless sensor networks. Further, the proposed approach is used to realize a secured event-driven smarter environment. The proposed approach validated by simulating the number of concurrent requests, and this measures their corresponding CPU and memory utilization along with the additional overhead because of encryption. The proposed approach consumes 54% less CPU and five times more memory (i.e. because of encryption) than the traditional HTTP requests. Besides, the proposed encryption mechanism REST-JSON with Bcrypt consumes 7% more time

V. Venkatesh
Embedded Systems Design Laboratory (ESDL), SASTRA Deemed University, Thanjavur 613401, India

P. Raj
Reliance Jio Cloud Services (JCS), Bangalore 560025, India

P. Balakrishnan (✉)
SCOPE, Department of Analytics, VIT University Vellore Campus, Vellore, India
e-mail: baskrish1977@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_25

than REST-JSON MD5/SHA1 to serve the requests and provide improvised security against dictionary attacks which is common in MD5/SHA1.

Keywords RESTful web service · Notification service · XMPP · Scalability · Security

1 Introduction

There is a wide variety of application in the area of wireless sensor networks [1] especially for governing and monitoring the application environment through physical sensor like temperature, humidity, moisture, humidity, etc. However, the wireless sensor networks are rigorously restricted on the communication, processing and power supply capabilities. This places serious limitations on the design and function of sensor networks. As a result, techniques seemed to be extremely improved and centred on proprietary communication technologies. On the other side, it is evident that the digital world shifted towards the Internet protocol (IP) e-mails and services. Hence, there is a need to solve the emerging problem like how to assimilate sensor networks within the IP-related framework. An immediate solution is to connect via any gateways from registered sensor networks to IP infrastructure. Until currently, this was regarded a paradise due to the IP computation and interaction overhead, but with the help of recent technologies, integrated circuits there is a path between IP and sensor networks in the form of connection establishment with fewer overheads. The trustworthiness [2] of the consequences needs the communication model and simulation tool to be very secure, by offering strong self-reliance in information transfer using a suitable protocol. In this work, the lightweight protocol is used especially for the private confidentiality purpose. Information integrity and availability are the salient features that applied to web services to provide secure information exchange between the server and client. To do that, REST is highly suitable and also a lightweight framework for any secure application.

There are several classification methods are available for polling or to push data to the users and many servers the renowned comet web system style makes use of polling and long-polling technology to permit the server to push data to the consumers. Imputable recurring polling needs and long-held contacts it will deplete lots of system resources, comprising connection establishment and also computing power for the resources. So, the scalability, reliability and compatibility functions will be naturally decreased. The next will be the Pub/Sub method that needs extra hardware to push the data to the clients; it indirectly increases the cost of the overall method. To solve the above-said problem, the alternate method will be Web Socket method to reduce the cost value. Web Socket's disadvantage is that it needs an extra Web Socket server to deal with all demands which may become a bottleneck in the application environment. The proposed framework will support the XMPP protocol [3, 4] and also Server-Sent Event (SSE) side for push and web services using RESTful web services. For any smart application, XMPP and REST middleware enable the bridge between any sensors and IP-enabled devices to create a sophisticated application. Apart from this, the proposed security mechanism, REST-JSON with Bcrypt is sturdy against dictionary attacks than the REST-JSON MD5/SHA1 mechanism at the expense of additional overhead (7%) in request processing. The rest of the paper is orchestrated as follows. The important research works which are closely associated with the proposed work are given in Sect. 2. Then, the proposed architecture and their components are explained in Sect. 3. The simulation environment, experiments and their results are discussed in Sect. 4. Finally, the important insights are concluded in Sect. 5.

2 Related Work

In [5], proposed RESTful interfaces are improving the interoperability and enhancing the availability of data offered by sensor networks. Furthermore, this system contributes value to unique sensor data to support in the decision-making procedure. Pautasso and Wilde [6] proposed how to integrate BPM and push to exhibit business processes and pragmatic in a RESTful way. Based on this method, loose coupling and excellent scalability can be achieved. Mesbah and van Deursen [7] has come out with a component—a push-based framework that can synchronize the actions both on the server and the client expeditiously. By using SPIAR architectural design client connections, user-perceived latency, information coherence can be taken out. Pohja [8] proposed XMPP function to reduce network load and improves execution effort and communication overhead. Eugster et al. [9] proposed methods to publish/subscribe. Moreover, it is possible to carry out a complete decoupling of the interacting organizations, area and synchronization. Both are with regards to connections and implementations. Serme et al. [10] proposed RESTful services equivalent to WS-Security and also offered to centralize protection metadata in headers to prevent support interruption and to incorporate other field protection: headers, factors, etc. Choudhary et al. [11], Lin et al. [12] proposed a web support protection design centred on HTTPI method over SOAP, with the protection goal: client/server verification and reliability on the concept, without privacy.

3 The Proposed Data Transfer Model

The polling, long-polling and loading systems all use HTTP reactions for transporting data from the server to the consumer. Thus, they depend on getting HTTP requests from the consumer, to become able to deliver responses. Such systems normally work fine, if the information transportation period regularity is known. Also, the systems think that the customer does not need to deliver information to the web server very often, and the systems do not provide proper means for bi-directional transportation of information. Web Socket method provides a solid bi-directional information transportation connection among a web browser and a web server. HTTP only will be used for discussing and allowing the Web Socket relationship but not for transporting the actual data. To overcome these issues, the best approach for data transfer is Server-Sent Events (SSE). SSE is just like the long-polling procedure, except it does not deliver only one perception per connection. The client drives a request server keeps an association until new information is ready. Then, it gives the message back to the client while still maintaining the connection open so that it can be used for another perception once it becomes available. SSE also offers standard JavaScript client API used the most modern web browser. These benefits conclude that the SSE as a potential push-based transport mechanism to realize near real-time notification services. Similarly, XMPP is a push-based application protocol that securely delivers notification among the communicating entities. Hence, the proposed approach uses SSE as a transport protocol and XMPP as an application protocol to realize REST-based lightweight secure, scalable notification middleware architecture.

3.1 System Architecture

The system architecture discusses the different technologies targeting the integration of sensor networks with application environment shown in Fig. 1.

3.2 Raspberry Pi

The Raspberry Pi is a transportable, powerful and handheld personal computer. The on-board panel dimensions are only 85 mm, and girth is only 56 mm. Pi can be used anywhere for any embedded application of a PC. They are high-resolution video, excel spreadsheets, word-processing, activities and more. Raspberry Pi has more extensive application range, such as music equipment, parent pointers to weather stations, peeping birdhouses with infra-red-based cameras, lightweight web server, smart home automation server and furthermore. It allows people of all age groups to discover processing, learns to a system and understands how computer systems work. Raspberry Pi on-board design B+ affords more GPIO, extra USB than design B. It also enhances less energy consumption, audio circuitry and micro SD card. The design B+ is the higher-spec variant of Pi model compared with any one of the B like the model number of GPIO. The header related with GPIO has gone up to 26 pins. It has four high-speed USB 2 ports, comparatively improved hot plug and ended with current behaviour.

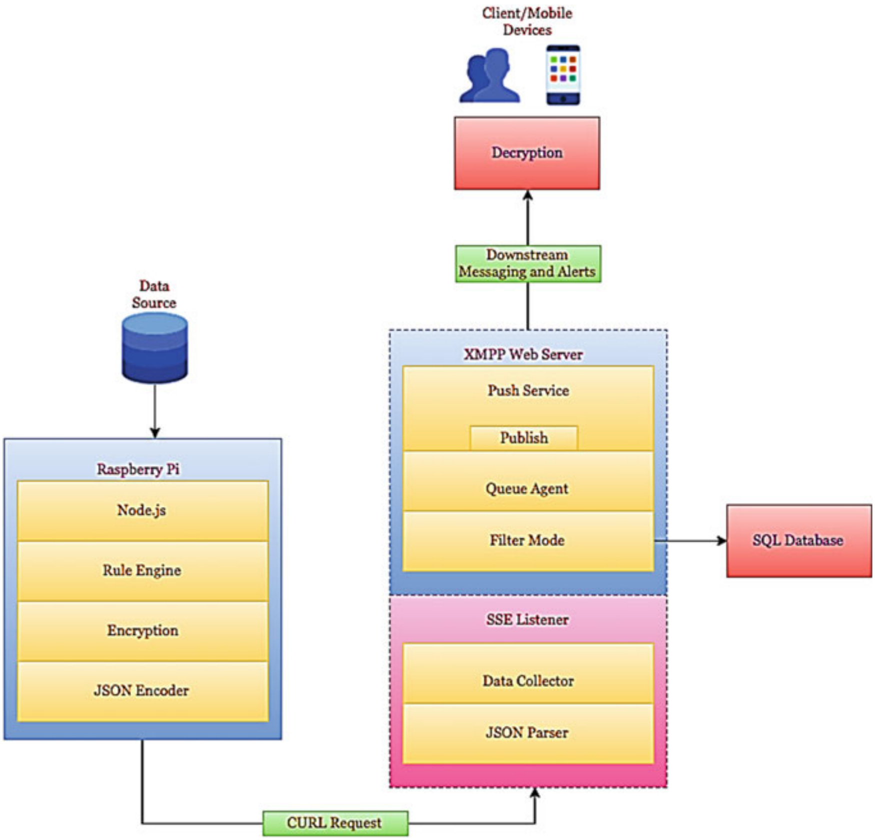


Fig. 1 Lightweight, scalable and secure middleware architecture

3.3 Node.js

Node.js developed because concurrency is difficult in many server-side development 'languages' and often brings to low efficiency. It is particularly more fast and efficient with basic online principles like HTTP, DNS and TCP. Also, JavaScript was a well-known terminology, making Node.js instantly available to the whole web design group. Node.js runs using a single thread, using non-blocking I/O calls, enabling it to support thousands of concurrent requests without running into the cost of thread context switching. Node.js has an event-driven architecture, more suitable for asynchronous I/O. This will improve the throughput and scalability in Web programs with many input/output functions, as well as for real-time Web programs. Node.js is primarily used to develop network-related applications programs such as web servers, creating it just like PHP. The significant distinction between Node.js and PHP is that many features in PHP prevent until completion, while features in Node.js are developed to be non-blocking (commands perform in similar and use callbacks to indication finalization or failure). Designers can make extremely scalable web servers without using threads, by using a simple kind of event-driven development that uses callbacks to indicate the finalization of a task.

3.4 Rule Engine Module

The main purpose of the rule engine module is dealing with the rules sent by any users with the help of any one of the browsers available in the node engine. This can be implemented by the Node.js method.

3.5 Security Strategy for Data Transfer Using Bcrypt

A lightweight crypt algorithm required for securing the information which is transferred through the resource-constrained environment. Hash function-based crypt algorithm will satisfy the above-said problem. There are several algorithms like MD5 and SHA1 are some lightweight used for secure data transfer. The problem with this algorithm is more vulnerable to dictionary/Rainbow Attack, Brute Force Attack. The main advantage of choosing a hash method is to make the variable length input to fixed output and also lightweight. The main purpose of choosing Bcrypt method is used to produce tunable cost parameter. This method uses the parameters like salt, key and cost as input. In this approach, the number of iteration is based on a cost function. However, it can be of the hardware and its specification.

3.6 JavaScript Object Notation (JSON)

JSON has obtained extensive support across services and customer platform. It is an alternative to XML because of purely lightweight. So, it can easily adopt for any web application, and it is developed to be a data exchange terminology which is individual understandable and simple for computer systems to parse and use JavaScript application. JSON is straight reinforced inside JavaScript and is the most effective application to JavaScript applications, thus offering important performance over XML, which needs extra libraries to retrieve information Document Object Model (DOM). JSON is approximated to parse up to one hundred periods quicker than XML.

3.7 Listener Module

The role of the listener is to collect messages from the clients. Whenever there is any specific message or any explicit header format/event from the browser, the listener has to establish a connection based on XMPP between the server and client appropriately. Once the SSE connection set up is established, the notification service will forward the message to the exact client, and if there is any data matching occurs, the client-specific conditions also will be forwarded.

3.8 Filter Module

A filter is an element which will figure out the same rules for the collected information from the user. The filter will decide whether the information has to be a push to the user or it has to be discarded. The filter makes use of the XMPP to push the information to the client whenever it is needed based on matching rules.

3.9 Query Agent

In any application paradigm, the user instantly makes a query to the network devices for any particular requirements. Queries are usually used in aggregation with the data-driven standards in order to express the precise data abstraction demands. Here, the data agent is used to pull the message from the message queues, litigate them and then push to the storage service. The role of this agent is to handle the function of the HTTP protocol. This can be implemented based on Node.js JavaScript Platform. This is considered as an important module in the notification service between server and client. The main job of this module is to listen to any one of the client requests periodically. This is considered as a common mechanism to deal with HTTP protocol request if any.

3.10 Push Function

Every time, push function is called by the filter module whenever it gets the new data is similar to the client request. Then, it responses back with the capacities of HTTP header containing, content-type, cache-control and connections. The connections remain well precise in formats. The headers stay distinct in SSE event source. When the client obtains the reply constituents, it will know about how to litigate the content constructed on the substance of headers.

3.11 Server-Sent Events (SSE)

The polling, long-polling and loading systems all use HTTP reactions for transporting data from the server to the consumer. Thus, they depend on getting HTTP requests from the customer, to become able to deliver responses. Such systems typically work fine, if the information transportation period regularity is known. Also, the systems think that the customer does not need to deliver information to the web server very often, meaning the systems do not provide good means for bi-directional transportation of information. Web Socket method provides a fixed bi-directional information traffic connection established between a web browser and a web server. HTTP will only be used for discussing and allowing the Web Socket relationship but not for transporting the actual data. To overcome these issues, the best approach for data transfer is Server-Sent Events (SSE). SSE is just like the long-polling procedure, except it does not deliver only one perception per connection. Web application designers can avail the Web Socket process using the HTML5-based Web Socket API integration.

4 Results and Discussion

4.1 Simulation Environment

The primary objective of the proposed approach is to realize a scalable and secured notification service. Our approach experiments the following information (XMPP and SSE) with security (MD5/SHA1 and Bcrypt) mechanisms: REST-JSON-MD5/SHA1 and REST-JSON-Bcrypt. Finally, the proposed notification service uses XMPP for scalability along with Bcrypt security, thereby ensuring secured delivery of notifications (REST-JSON-Bcrypt).

4.2 CPU Utilization Graph

In Fig. 2, it is evident that the proposed XMPP-SSE consumes 54% less CPU to encrypt a standard text using MD5/SHA1 (green line) than the traditional HTTP (dark blue line) requests. However, the MD5/SHA1 is prone to dictionary attack and can be easily compromised. To alleviate this, Bcrypt algorithm is integrated into XMPP (XMPP-Bcrypt-SSE) to encrypt a standard text. But, the Bcrypt algorithm has an exponential response time. Hence, an optimal cost function can be chosen to save time—indicating a resource consumption rate approximately equal to the consumption (blue line) of XMPP-SSE with improved security.

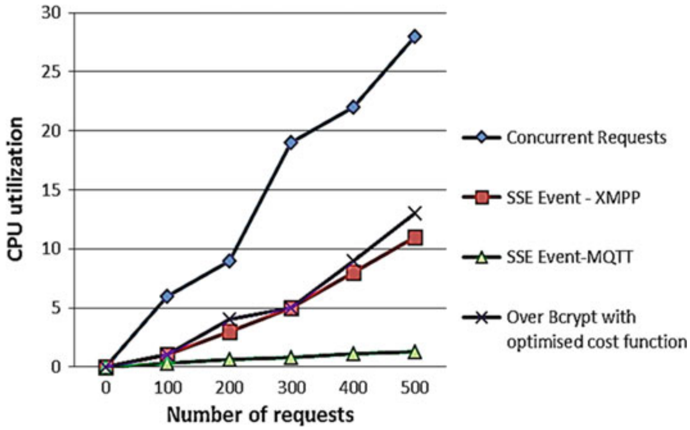


Fig. 2 CPU utilization graph

4.3 Memory Utilization Graph

The length of the encrypted Bcrypt result will be of fixed length, depending on the cost function. This results in higher memory consumption for XMPP-Bcrypt-SSE (Fig. 3). Than that of normal usage, because despite the length of the unencrypted text, the output hash will be of a long base-64-encoded value of length 53 characters, for a two-digit cost function. The memory utilization of the standard SSE channel would be relatively lesser than that of the encrypted channel (XMPP-MD5/SHA1-SSE) because of the raw data of length 3–8 characters. The memory utilization for concurrent requests (HTTP) will remain low, irrespective of the number of claims

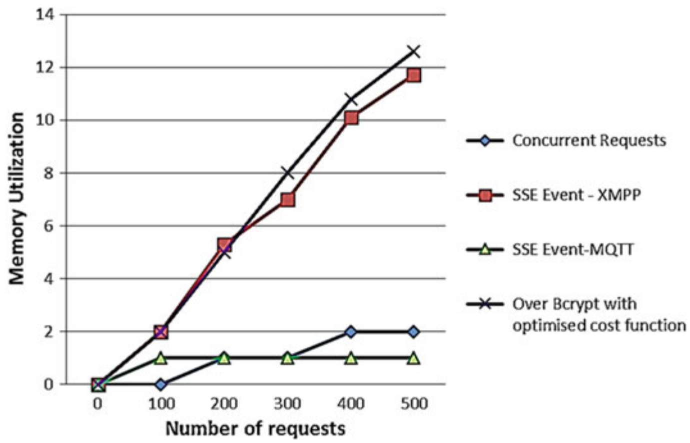


Fig. 3 Memory utilization graph

because of the high memory capacities of modern computers. It is the CPU utilization that shoots up because of the amount of computation that is required for the encryption to take place.

4.4 Encryption + SSE

Since nine (out of 100 requests) users are seeking encrypting service, Fig. 4 contains only nine users in the *X*-axis. On a comparison between MD5/SHA1 (REST-JSON-MD5/SHA1) and Bcrypt (REST-JSON -Bcrypt), heavily outweighs MD5/SHA1 in both performance and security. The cost function passed as a parameter to the Bcrypt function increases the computation time for the encryption/decryption. Moreover, the result of a Bcrypt function is not always the same. This makes it impossible for a dictionary attack, unlike MD5/SHA1 encryption algorithms. An optimal cost parameter can be chosen to use the advantages of the Bcrypt algorithm, and at the same time, it outweighs the functionalities of MD5/SHA1, using the same or even efficient consumption of computing resources.

Finally, it is compared with MQTT a lightweight protocol shown in Fig. 5. The experiment is carried with the mosquito server is hosted in AWS. The aws-iot client is created from Raspberry Pi board and is connected to mosquito server in AWS.

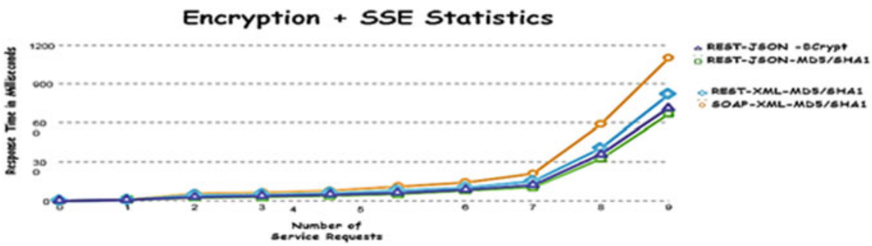


Fig. 4 Encryption + SSE statistics

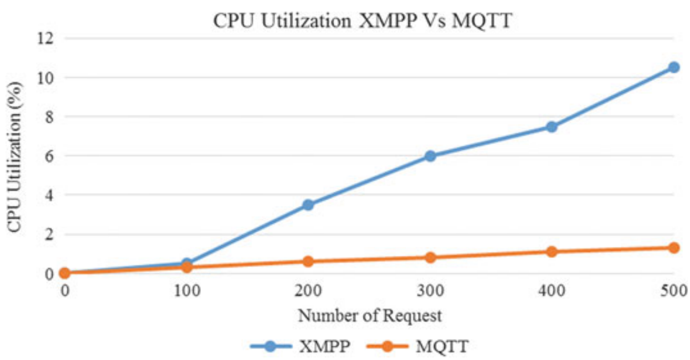


Fig. 5 CPU utilization XMPP versus MQTT

5 Conclusion

This work proposes a secure server notification on RESTful architecture. The proposed service eradicates the necessity of any middleware for existent server-side push frameworks and decreases the growth of complication and provides a secure platform for web applications. From the results, it concludes that the proposed notification service is well suited for a web application. Also, the security method using XMPP and SSE shows the efficiency of secured JSON message in relation with communication, size and time. In future, we have planned to propose an attribute-based encryption to reduce the time, memory and CPU utilization, which only encrypts the selected attributes based on the user preferences.

References

1. Alemdar, H., Ersoy, C.: Wireless sensor networks for healthcare: a survey. Comput. Netw. (2010)
2. Chen, H., Perrig, A.: Security and privacy in sensor networks. IEEE Comput. **36**(10), 103–105 (2003)
3. Hornsby, A., Belimpasakis, P., Defee, I.: XMPP-based wireless sensor network and its integration into the extended home environment. In: 13th IEEE International Symposium on Consumer Electronics, pp. 794–797 (2009)
4. Bendel, S., Springer, T., Schuster, D., Schill, A., Ackermann, R., Ameling, M.: A service infrastructure for the Internet of things based on XMPP. In: 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 385–388. San Diego, CA (2013)
5. Trilles, S., Belmonte, O., Diaz, L., Huerta, J.: Mobile access to sensor networks by using GIS standards and restful services. Sens. J. IEEE. **14**(12), 4143–4153 (2014)
6. Pautasso, C., Wilde, E.: Push-enabling RESTful business processes. In: Proceedings of the International Conference on Service-Oriented Computing. ICSOC'11, pp. 32–46 (2011)
7. Mesbah, A., van Deursen, A.: A component and push-based architectural style for Ajax applications. J. Syst. Softw. **81**(12), 2194–2209 (2008)
8. Pohja, M.: Server push with instant messaging. In: Proceedings of the 2009 ACM symposium on Applied Computing, pp. 653–658. ACM (2009)
9. Eugster, P.T., Felber, P.A., Guerraoui, R., Kermarrec, A.: The many faces of publish/subscribe. ACM Comput. Surv. **35**(2):114–131 (2003)
10. Serme, G., Oliveira, A.S., Massiera, J., Roudier, Y.: Enabling message security for RESTful services. In: Proc. IEEE ICWS, pp. 114–121 (2012)
11. Choudhary, P., Aaseri, R., Roberts, N.: HTTPI based web service security over SOAP. Int. J. Netw. Secur. Appl. 55–65
12. Lin, J., Yu, W., Zhang, N., Yang, X., Zhang, H., Zhao, W.: A survey on internet of things: architecture, enabling technologies, security and privacy, and applications. IEEE Internet Things J. **4**(5), 1125–1142 (2017)

Security Issues in IoT and their Countermeasures in Smart City Applications

Debabrata Singh, Bibudhendu Pati, Chhabi Rani Panigrahi and Shrabanee Swagatika

Abstract Smart city can be taken as the biggest implementation of Internet of things (IoT) applications. Smart city technologies promote cloud-based and IoT-based services in which real-world user interfaces use smart phones, sensors, RFIDs, etc. The IoT ecosystems cannot properly communicate between them. So it needs a bridge between them to fill the global common approach. Implementation of this can be done with the help of information and communication technologies (ICT) which may result in enhancing quality, interconnection, and the performance of various urban services. It highlights the need for cloud computing and IoT technologies and both have a major impact on how we build and deploy smart applications for it. There are still numerous difficulties that remain with regard to its deployment due to its varying requirements. In this paper, authors tried to converge the two domains that are cloud computing and IoT for smart city application deployment and also present the identified security threat types and their countermeasures in the context of smart city.

Keywords IoT · Cloud computing · ICT · Smart city · Cyberattacks · Security

D. Singh (✉) · S. Swagatika
Department of CSE, ITER, S'O'A, Bhubaneswar, India
e-mail: debabratasingh@soa.ac.in

S. Swagatika
e-mail: shrabaneeswagatika@soa.ac.in

B. Pati · C. R. Panigrahi
Department of CS, Rama Devi Women's University, Bhubaneswar, India
e-mail: patibibhudendu@gmail.com

C. R. Panigrahi
e-mail: panigrahichhabi@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_26

1 Introduction

IoT is a system of interrelated processing gadgets that are furnished with special identifiers and have the capacity to exchange information over a system without expecting human-to-human or human-to-PC collaboration. According to the IEEE report [1], IoT contains three main important components such as sensors, networks, and Internet connections. The most potential technology which integrates sensors, human interfaces, networks, and other valuable information is used in IoT to develop various smart city applications [2]. With the help of IoT technology, a smart city application can transmit all the captured data from all sensors in real time [3]. In a smart city, there may be a chance of occurrence of various types of problems such as crime, traffic congestion, robbery, waste management. Generally, the local government now tries to solve those identified problems by utilizing IoT technology in smart cities [4]. The smart city concept is generally based on smart mobility, smart living, smart economy, smart people, smart governance, and smart economy. In a smart city, various IoT applications are used such as smart lighting, GPS tracker, CCTV on a public area, traffic control management system, GPS tracking for ambulance, city touch technology, garbage truck, and automatic water sensors [5]. Over the Internet, the software and hardware resources can be delivered through cloud computing. The different innovative services can act as a bridge between the cloud and IoT [6]. On the other hand, smart city IoT concept envisions a new generation of devices (i.e., sensors both virtual and physical) that are connected to the Internet and provided different services for value-added applications in smart city [7].

Smart IoT incorporates heterogeneous systems and allows open access for the development and creation of advanced digital services [8]. Aggregation of heterogenous systems inside the cloud is defined as the Cloud of Things (CoT). It happens due to an extremely large variety of devices and link layer technology [9, 10]. In smart IoT, though machine-to-machine technology is the first phase, it enables new applications and bridges diverse technology by connecting physical objects together in support of making intelligent decisions. In smart IoT applications, we must think about business models and value chain for different categories of IoT [11]. It can also be divided into three categories based on usage and client base: *industrial IoT*, *commercial IoT*, and *consumer IoT*. Nowadays, it is a big challenge for any smart city to assure security through IoT. A smart city considers a number of factors while introducing new technologies still it has a series of concerns and challenges in terms of security and privacy. It is found that out of different types of attacks in IoT, threats percentage from external hackers are 37%, malware is 23%, internal hackers are 28%, and DoS attacks are 20% [12]. In IoT security services, maximum of 10% of IoT products provide security, so it is highly required to increase the security issues as much as possible for any smart city application.

Remaining part of the manuscript is structured as follows: In Sect. 2, various IoT applications in smart city are presented. In Sect. 3, the security issues with IoT applications in smart city are highlighted. Section 4 describes the security requirements for cloud platforms with respect to implementation of IoT applications. Section 5 deals with threat types and their countermeasure in context to smart city applications. Section 6 concludes the paper.

2 IoT Applications in Smart City

IoT innovation is making smart city simpler structures with an inheritance framework to spare vitality and enhance their suitability. The objective of this smart city model is to represent how the components of smart city are communicating to each other through IoT by building vitality administration frameworks.

A smart city is comprised of several heterogeneous and homogeneous components that use various available technologies such as Wi-Fi, ZigBee, 4G/5G mobile communication. These devices are connected with each other through Internet. In a smart city environment, the various components of the smart city use different services (on demand requirements) as depicted in Fig. 1 and are related to various smart city applications such as

Smart roads: Ultra-modern highways with intelligent warning messages and diversions to avoid traffic jams and accidents.
Smart parking: To avoid traffic jams and monitoring, the parking space depending upon the space availability.

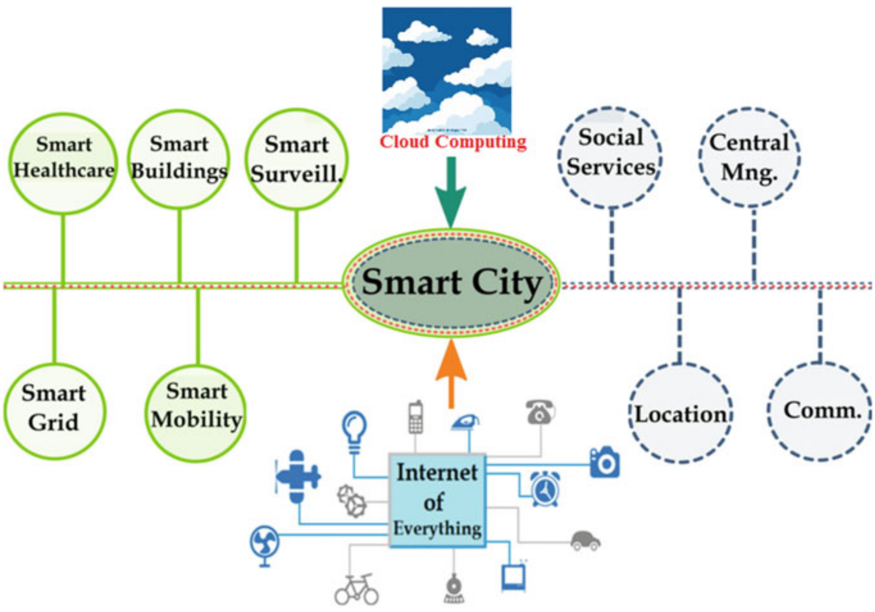


Fig. 1 Vaious IoT applications in smart city

Structural health: Impact on health through noise and sound monitoring, material conditions of building and malls, vibration level, bridges, and historical monuments.
Traffic congestion: Monitoring the pedestrian levels, vehicles to optimize walking route as well as driving routes.
Smart lighting: Lighting the city by weather adaptive, automatic off/on, intelligent lighting through advanced lights.
Waste management: Awareness of cleanliness in the city, optimize the trash collection, disposal methods, and detection of rubbish levels in containers.
Water management: It uses ultra-morden technologies, i.e., NFC and SCADA [13], etc., in water treatment plants.
Smart transport: It uses multi-way communication through modern transport system.
Smart phone detection: To detect and alert, the problems associated with smart city through any smart phones connected via Internet or Wi-Fi.
Noise urban maps: Sound monitoring in bar areas controls over the noise and the pollution centers.
Electromagnetic field areas: Used for measurement of the energy radiate cell station, industrial areas, and Wi-Fi routes.
Data management, visualization, and sharing: All data must be transparent to all the viewers, and it should be sharable and auto-control.

In the present scenario of urbanization, it is seen that it has come up with an exceptional efficient and social development which is unique in relation to different eras. The urban areas currently have the most pieces of the total population and an expanding offering of the world's most gifted, instructed, inventive, and entrepreneurial personnel [14]. More than 50% of the people on the planet now live in cities. So when we think about smart city, we must think of the city where we get all fully connected homes, smart communication and all other smart facilities. The city must have lighting, bus, train, and metro schedules, traffic lights, and all things must be suitably integrated to provide benefits like safety, maximizing comfort, and other essential real-time requirements. As indicated by the United Countries, this number will increase to 70% under 50 years [1]. This purported city development or rising of urban life is driving the city framework into a feeling of anxiety at no other time seen as the interest for essential administrations increments and is exponentially overburdened [5]. Urban areas are winding up progressively, engaged innovatively as their center frameworks (i.e., instruction, open security, transportation, vitality and water, medicinal services, and administrations) are instrumented and interconnected, empowering better approaches to manage gigantic, parallel, and simultaneously used. In the urban arranging field, this characterizes insightfulness. Another perspective of smart city application as presented in Fig. 1 points out three main characteristics of a smart city, i.e., (i) instrumented, (ii) interconnected, and (iii) intelligent.

Instrumented: Collection of set of sensor nodes by physical or social in a city can be instrumented. The core system of the cities can be controlled by these sensors and can have access to reliable information as well as the real-time system, which directly incorporates with IoT system.
Interconnected: The arrangements of different frameworks and their interconnection between data and sources with the system lead to the interconnectedness property. The relation between interconnected and instrumented frameworks makes an association with the physical world with the present reality.
Intelligent: The best utilization of data from various frameworks and sensors is instrumented and interconnected that offer a superior life to the subject.

Offering only one or any blend from those three features of IoT makes a situation where an imperative part will miss. The smart cities require a cloud environment for business intelligence and time management of government services [10]. Basically, these strategies are based on three principles, i.e., communication, cooperation, and combined (C3) [15].

3 Security Issues with IoT Applications in Smart City

A smart city application considers a number of factors while introducing new technologies and raises various challenges in terms of security and privacy. In this section, the various types of possible attackers in smart city applications, security issues with IoT devices, IoT infrastructure, and IoT architectural layers are presented.

3.1 IoT Attackers in Smart City Applications

The five major types of IoT attackers in smart city applications include:

1. *Amateur hackers*: These are the people who utilize hacking techniques on Internet of Everything (IoE) networks for limited sessions. They usually do these activities for fun, and they do not pose a major threat, e.g., hobbyists, enthusiasts, script kiddies, etc.
2. *Petty criminals*: They use hacking techniques for gaining personal benefits. A scenario can be, where a burglar disables security systems by remote logging into it via the less secure router, e.g., low-level cyber criminals.
3. *Cyber espionage groups*: Organized crime groups or syndicate groups, i.e., Black Vine, greenBug, etc., pose a great threat. They find and use the less secure network access points to exploit and perform cybercrimes at a very large scale.
4. *Terrorists/hacktivists*: Professional, non-state actors such as Oxblood Ruffin or political hacktivists sometimes use these vulnerabilities and pose threat in national and international levels.

5. *State-sponsored attackers*: Sometimes traditional adversarial states, e.g., Russia, China, allow foreign espionage via state-sponsored sabotage and create catastrophic situations using vulnerabilities in IOE networks.

3.2 Security Issues with IoT Devices

The biggest challenging factor in IoT is the security of the various IoT devices [16]. In IoT applications, most of the application data are either person concern, industry, and enterprise. For any kind of vulgarity, theft, and tampering of these applications, data must be secured and confidential. Especially, the security over the communication channel is most important. Although the IoT focuses on these devices, several issues exist with respect to scalability, response time, and availability of applications. Some of these security challenges with IoT devices include:

1. *Data privacy and security*: While transmitting data seamlessly, data must be secured from theft and hidden from the hackers.
2. *Insurance concerns*: Installation of IoT devices for any application to collect the data to take a decision by all the insurance companies.
3. *Lack of general standards*: Though many IoT companies and standards for IoT devices exist, the most challenging factor for them is the support for unauthorized devices in the IoT system.
4. *Technical concern*: Due to excess use of IoT devices, the traffic generated by these devices is also increasing. So, it must need a larger network capacity, which requires a large amount of data to harvest and store.
5. *Security attacks and system vulnerability*: IoT system mainly focuses on different security challenges to design proper guideline for security of a network and different security frameworks, i.e., system security. To handle IoT applications, both application and network security are required which help in securing communication among different IoT devices [12, 16].

3.3 Security Issues in IoT Infrastructure

The key security aspects of IoT Infrastructure include the security of information technology components and are discussed in this section. These include 1. *confidentiality*, 2. *integrity,* and 3. *availability*. These components are commonly referred as confidentiality, integrity, availability (CIA) triad. In order to achieve CIA triad, the following security requirements can be applied to IoT infrastructure.

(a) *Verification*: This procedure checks to guarantee that a client certification is substantial, so clients with invalid accreditations won't be permitted to get the hidden data. The easiest method to utilize confirmation is with the assistance of usernames and passwords. However, as hacking systems are advancing step by step, it is critical to guarantee that refined verification strategies are set up. One such confirmation instrument that is utilized is called multifaceted validation. After that, just two elements specified above might be utilized for verification, and all things were considered, and it is called two-factor confirmation.
(b) *Authorization*: Approval is a procedure which guarantees that a particular client has rights to perform particular tasks on a particular question. This is for the most part by conceding distinctive sorts of consents to various kinds of clients in view of their part in a city government. For instance, a fire station official will simply have the capacity to use the information related to other city divisions like water.
(c) *Encryption*: It is the way toward changing over information into a configuration that cannot be deciphered effortlessly and specifically by unapproved clients. It is vital to guarantee that information put away in the IoT framework, and the information transmitted by means of the systems is in encoded shape. This is extremely useful to anticipate unapproved gathering of information by outsider operators. The way toward the changing over the information back to its unique frame is called decoding. A few encryption programming is accessible in the market. The larger part of the IoT applications will be assembled and sent on cloud stages [8]. Subsequently, all security worries of cloud stages will posture security dangers for IoT parts too.

3.4 Security Issues in IoT Architectural Layer

To achieve ubiquitous smart city manipulates, the data detected from the physical world, transmitted in the corresponding world, and involved the information from unauthorized access, disclosure information, and networking paradigm. Datta et al. [17] studied the various architectures, protocols, and security aspects of IoT that can be implemented in smart city applications. They focussed on the importance of IoT in today's world, and how this methodology can be used in building up a perfectly smart city. They also focussed on how with the use of IoT, and real devices can be connected to virtual ones and can be used for automatically operations over the Internet-as-a-service platform.

3.4.1 Architectural Layers of IoT System with Their Attacks

IoT architecture is mainly divided into four layers:

A. *Physical layer* manages the identity and gathering of items—particularly data by the sensor devices. The different attacks in this layer include (i) tampering, (ii) DoS, (iii) sensors as security threats
B. *Network layer* is used to send the data collected by the devices to the information processing system; the different attacks in this layer include:

   Security countermeasure: This includes active firewalls for traffic, passive monitoring to raise alarms, traffic admission control, bidirectional link identification.
   Man in the middle: This includes eavesdropping, routing attacks, and replay attack.
   DoS attacks: This includes exhaustion, collision, unfairness, spoofed routing information, selective forwarding, sinkhole attack, wormhole attack, Sybil attack, flooding, and node replacement.

C. *Data processing layer* is used to process, compute ubiquitous information, and take automatic decision accordingly. The types of attacks in this layer include login abuse, eavesdropping, inappropriate system use, DoS attacks, network intrusion, session hijacking attack, fragmentation attack, cloud access control, and database integrity issues.
D. *Application layer* is used to globally manage the application based on the data processed through the middleware. This layer includes threats and issues with the client application, issues with the communication channels and their client applications, issues with system integrality of the client application, minor modification leading to complex issues, multiuser access and concurrent editing of configuration, data access, and traceability.

The few security disputes of the application layer are:

Malicious code infusion: In this procedure, the exploitative programmer infuses malevolent code from an obscure area into the framework and attempts to take or control the information of the approved client.
Denial-of-benefit assault: In this assault, the programmer puts on a show a verified client, signs into the framework, and interferes with the typical working of the system. This is the greatest helplessness to the framework.
Phishing assault: This is an assault in which the e-mail of the high-positioning expert in the system is utilized for assault. The assailant picks up accreditations access of that casualty and harms information.
Sniffing assault: In this composition, the assailant brings a sniffer application into framework to constrain an assault on it, which could pick up arranged data, prompting debase framework and unions application and middleware layer to shape an incorporated security system.

4 Security Requirements for Cloud Platforms

Cloud computing permits sharing of the assets of ground-breaking servers in server firms with another gadget like cell phones and PCs. IoT permits obliged sensor gadgets to send and store their estimations in an inside portion open by various different gadgets. Distributed computing is performed and then the progressed investigation, and modern administrations utilize the sensor information from the IoT gadgets. Different cloud storage, systems, plan rule, and execution did not address the security. Kaur et al. [15] discussed various applications and the usage of IoT and cloud computing in building smart cities and highlighted the importance of enhancing the quality of urban life in the present scenario.

Cloud security engineering has three distinct layers: programming applications layer, stage layer, and foundation layer [17, 18]. Each layer has its own arrangement of security concerns. One of the principal problems of cloud is multitenancy. Multitenancy alludes to the way that cloud foundation, in light of the fundamental virtualization stage, and gives the highlight to benefit numerous autonomous customers (in habitants) utilizing same arrangement of assets. This subsequently builds the dangers for information privacy and trustworthiness. These dangers are particularly more extreme if there should be an occurrence of open cloud condition. This is on account of open cloud, and administrations can utilize by contending customers when contrasted with Private Mists, and furthermore, the number of cloud clients is substantially higher [19]. The approaches to overcome these problems, which emerge because of multitenancy are 1. virtual machine division, 2. database division, and 3. virtual machine reflection. With a specific end goal to guarantee security of different gadgets and stages, which are a piece of the IoT organization, it is basic to guarantee security to all these gadgets. Some such gadgets include:

Secure booting: When a gadget control is on, there ought to be a confirmation component to check that the product which keeps running on the gadget is a real one. This is finished with the assistance of cryptographically created advanced marks [12, 16]. This procedure guarantees that exclusive true programming which has been intended to keep running on the gadgets by the concerned gatherings will keep running on the gadgets.
Required access control instruments: Compulsory access control instruments ought to be incorporated with the working arrangement of the gadgets keeping in mind the end goal to guarantee that the different applications and parts will approach just to those assets, which they require for their working. This will guarantee that if an assailant can access any of those segments/applications, they will have the capacity to access just extremely restricted asset. This altogether diminishes the assault surface.
Gadget confirmation for systems: A gadget ought to get associated with some sort of a wired or remote system keeping in mind the end goal to start transmission of information. At the point when a gadget gets associated with a system, it ought to validate itself before it begins information transmission. For a few kinds of implanted gadgets (which work without manual intercession), the confirmation should be possible with

the assistance of certifications, which are kept in an anchored stock piling region of the gadget.
Device-particular firewall: For every gadget, there ought to be some sort of firewall, which will channel and analyze the information that is particularly sent to that gadget. It is not required for the gadgets to look at a wide range of movement, which crosses through the system as, that will be dealt with the system security apparatus [11].

5 Threat Types and Their Countermeasures in Smart City

When cyber security is neglected, smart cities potentially contain serious risks and authorities for residents. Some of the threat types and their countermeasures in smart city are presented in Table 1.

5.1 Services and Secure Solution in Smart Cities

In this section, some of the services and secure solution in the smart cities are identified.

1. *Secure boot and firmware integrity*: Secure boot prevents malicious nodes, hackers from replacing firmware, and different cyber attacks. Secure boot uses cryptographic signing technique codes to operate the authenticated device trusted party.

Table 1 Threat types and their countermeasures

Threat types	IoT devices	Countermeasures
Privacy, data, and identity theft	Parking garage, EV charging station, surveillances feeds	Authentication, encryption, and access control
Man in the middle	Wastewater overflow	Authentication, encryption, security lifecycle managements
Application level distributed denial of service (DOS)	Parking meters, plethora devices, etc.	Device identification, access control, security monitoring, and analysis [8]
Device hijacking	Smart meter, stealthily siphon energy, etc.	Device identification, access control, security lifecycle managements
Permanent denial of services (PDoS)	Parking meter replacement or reinstallation of updated software, etc.	Authentication, encryption, access control and application level DDOS protection, security monitoring, and analysis

2. *Security monitoring and analysis*: Collecting, capturing, and monitoring various data at different endpoints/locations on the system and analyzing to detect the possible potential threats as well as the security violations.
3. *Security life cycle management*: Life cycle management allows the service providers to control the security of IoT devices, i.e., during cyber disaster, new service without authorization.
4. *Mutual authentication*: The connected network with the smarty cities must be authenticated. The data source must not be a fraudulent source as well as a legitimate device, secure, and mutual authentication protect against various malicious attacks.

5.2 *Attacks in Different Layers of IoT with Their Countermeasures*

Different types of defense mechanism are adopted to prevent the attacks in different layers of IoT and are given in Table 2.

Table 2 Attacks in different layers of IoT with their countermeasures

Layer	Attack	Countermeasures
Physical	Jamming (networks)	Channel surfing, priority messages, spatial retreat
	Tampering	Tamper proofing, hiding
	Radio interference	Delayed disclosure of keys
	Unfairness	Small frames
	Exhaustion	Rate limitation
	Collisions	Error-correcting code
Network	Sinkhole	Geo-routing protocol
	blackhole, wormhole	Authorizations, monitoring, redundancy
	Misdirection	Egress filtering, authorization, monitoring
	Homing	Encryption
Transport	De-synchronization	Authentication
	Flooding	Client puzzles
Application	Reprogram	Authentication
	Overwhelm	Rate-limiting

6 Conclusion

In this work, authors have identified the IoT stages as a feasible solution to make cities smarter. IoT architecture, security issues, their attacks, and application layer disputes that can be proficiently considered with regard to smart city is presented. Along with the identification of IoT applications with different security challenges, the probable solutions for the development of novel services, contributing to make the cities more sustainable, are also identified. In this work, authors have also identified how cloud computing act as a bridge between the IoT and Internet of people through the Internet of services and the security issues associated with cloud computing are highlighted. Lastly, some potential threat types and their countermeasure in the context of smart city are identified.

References

1. Minerva, R., Biru, A., Rotondi, D.: Towards a definition of the Internet of Things (IoT). IEEE Internet Initiative **1**, 1–86 (2015)
2. Harmon, R.: Smart cities and the Internet of Things. pp. 485–494 (2014)
3. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M.: Internet of things: a survey on enabling technologies, protocols, and applications. IEEE Commun. Surv. Tutorials **17**(4), 2347–2376 (2015)
4. Latre, S., Leroux, P., Coenen, T., Braem, B., Ballon, P., Demeester, P.: City of things: an integrated and multi-technology testbed for IoT smart city experiments. In: IEEE 2nd International Smart Cities Conference Improv. Citizens Qual. Life, ISC2 2016—Proceedings (2016)
5. Arasteh, H., Hosseinnezhad, V., Loia, V., Tommasetti, A., Troisi, O., Shafie-khah, M., Siano, P.: IoT-based smart cities: a survey. In: 2016 IEEE 16th International Conference Environment Electrical Engineering (EEEIC), pp. 2–7 (2016)
6. Patti, E., Acquaviva, A.: IoT platform for smart cities: requirements and implementation case studies. In: 2016 IEEE 2nd International Forum on Research Technologies for Society and Industry Leveraging a better tomorrow, pp. 1–6 (2016)
7. Wang, E., Chow, R.: What can i do here? IoT service discovery in smart cities. In: IEEE International Conference Pervasive Computing Communication Workshops (PerCom Workshops) (2016)
8. Rathore, M.M., Paul, A., Hong, W.H., Seo, H.C., Awan, I., Saeed, S.: Exploiting IoT and big data analytics: defining smart digital city using real-time urban data. Sustain. Cities Soc. **40**, 600–610 (2018)
9. Shah, S.H., Yaqoob, I.: A survey: Internet of Things (IOT) technologies, applications and challenges. In: 2016 IEEE Smart Energy Grid Engineering, vol. i, pp. 381–385 (2016)
10. Banerjee, A., Pati, B., Panigrahi, C.R.: SC2: a selection-based consensus clustering approach. In: Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2016, p. 177 (2017)
11. Bibri, S.E.: The IoT for smart sustainable cities of the future: an analytical framework for sensor-based big data applications for environmental sustainability. Sustain. Cities Soc. **38**, 230–253 (2018)
12. Benzarti, S., Triki, B., Korbaa, O.: A survey on attacks in Internet of Things based networks. In 2017 International Conference on Engineering & MIS (ICEMIS), pp. 1–7. IEEE (2017)
13. Sarangi, M Singh, D.: Supply and monitoring the municipality drinking water through NFC and SCADA system. In: Applied Electromagnetics, Signal Processing & Communication (AESPC-2018), IEEE (2018)

14. Kim, T., Ramos, C., Mohammed, S.: Smart city and IoT. pp. 159–162 (2017)
15. Kaur, M.J., Maheshwari, P.: Building smart cities applications using IoT and cloud-based architectures. In 2016 International Conference on Industrial Informatics and Computer Systems (CIICS), pp. 1–5. IEEE (2016)
16. Ferraz, F.S., Sampaio, C., Ferra, C.: Towards a smart-city security architecture proposal and analysis of impact of major smart-city security issues. In: Informatics Center Federal University of Pernambuco Recife, Brazil (2015)
17. Datta, P., Sharma, B.: A survey on IoT architectures, protocols, security and smart city based applications. In: 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–5. IEEE (2017)
18. Khan, Z.A.: Using energy-efficient trust management to protect IoT networks for smart cities. Sustain. Cities Soc. **40**, 1–15 (2018)
19. Petrolo, R., Loscri, V., Mitton, N.: Towards a smart city based on cloud of things. In: Proceedings of the 2014 ACM International Workshop on Wireless and Mobile Technologies for Smart Cities, pp. 61–66. ACM (2014)

An Efficient Approach for Running Multimedia Applications Using Mobile Cloud Computing

Rajesh Kumar Verma, Chhabi Rani Panigrahi, Bibudhendu Pati and Joy Lal Sarkar

Abstract Multimedia applications are resource-intensive applications and processing them using the mobile devices is a challenging task because of the limited resources of the mobile devices. To solve this issue, in this work, a robust architecture has been proposed where different multimedia applications are processed using mobile cloud computing (MCC) which is also based on multi-cloud environment, in order to run the applications in an efficient manner. Also, we have stated the future roadmap which aims at coming up with a scalable architecture which will provide near real-time experience of multimedia applications to the user community.

Keywords Mobile cloud computing · Offloading · Multimedia

1 Introduction

MCC provides an efficient mechanism to utilize the resources of the mobile devices and also the cloud in case there is a need for extra computational power or storage which may be required to run the particular application(s) (like gaming, simulation, movies, etc.) [1–17]. By the end of the year 2018, nearly one-third of the world's population is predicted to own a smartphone, which will result in around 2.53 billion

R. K. Verma (✉)
Department of Computer Science and Engineering, Biju Patnaik University of Technology, Rourkela, India
e-mail: rajeshverma_chicago2004@yahoo.com

C. R. Panigrahi · B. Pati
Department of Computer Science, Rama Devi Women's University, Bhubaneswar, India
e-mail: panigrahichhabi@gmail.com

B. Pati
e-mail: patibibudhendu@gmail.com

J. L. Sarkar
Department of Computer Science and Engineering, Tripura University, Agartala, Tripura, India
e-mail: joylalsarkar@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_27

smart mobile devices (SMD) users across the globe [18]. This number is predicted to touch around 6.1 billion smartphone users which are approximately 70% of the world's population using smartphones [18]. The offloading process to the cloud from the MCC happens when there is a requirement of huge amount of computation and storage for a particular application [3, 19–22]. Multimedia, as the name suggests incorporates various different types of content of material (such as videos and speech) and it is present in varied format [23–25]. This variety of content types is complex in nature, and performing manipulations on these data types are not a straight-forward technique. The initial use of multimedia techniques was in movies and subsequently spread across to different fields like gaming, medical, education, computer simulations, home shopping, training, etc [25]. The advantage of multimedia is that it is far more superior and also enhances the impact, user experience, and knowledge through the information which is received by the user [21]. Multimedia is a complex field which includes new algorithms, graphics, cameras, SMDs, etc. Researchers have been focusing on the several problems encountered in the space of integration of MCC and multimedia applications [25]. In this paper, we have discussed efficient techniques for integration of MCC and multimedia applications. In order to resolve the multimedia problems, we can plan to use MCC where-in highly computational tasks can be offloaded to the cloud in order to save energy in the SMD.

The paper is organized as follows. Section 2 details the motivational aspect which talks about the importance of multimedia applications in the current world and how the proposed efficient architecture helps to tackle several limitations of mobile devices. In Sect. 3, the related work of various researchers has been discussed in this field. The system architecture is described in detail in Sect. 4, which contains several layers that help in running the multimedia applications smoothly and efficiently. The advantages have been also discussed in this section. Section 5 concludes the paper.

1.1 Motivation

As multimedia applications are resource intensive, it is very essential that we have a robust architecture to run them using MCC. The popularity of multimedia applications is rising rapidly, especially among the users of smart mobile devices (SMDs). It is expected that fast response be provided by the architecture for multimedia applications that the users run on their SMD. However, due to the limitations of computational power, storage, and battery power, these mobile devices may not be able to meet the requirements of the users. MCC technique helps to improve computational power as well as storage of the mobile devices by integrating it with the cloud [26–29]. The transmission between the mobile devices and the cloud is a very challenging task as a huge amount of mobile device energy is required for the offloading process. To ensure that applications are processed at a faster rate, the mobile devices are integrated with the cloud. The processing then takes place in the cloud which has a vast amount of resources used on a pay-per-use basis. The output is then returned from the cloud via MCC to the user.

1.2 Contribution

In this section, the authors have highlighted the contributions of their work.

- Firstly, an efficient architecture has been proposed to run the multimedia applications using MCC and cloud, as these types of applications require much more computational power and storage.
- Different layers in the architecture have been detailed. The processing of the multimedia applications will have layers such as—Multi-cloud, MCC, Gateway, RAN, and the devices and application layer.
- Advantages have been detailed for the proposed architecture as the mobile devices suffer from constraints such as limited battery power, computational capability, and storage. Also, in case of hostile environment when network is not present, this architecture ensures that communication among the mobile devices is present such that any emergencies can be tackled.

2 Related Work

MCC has been growing popular at a very rapid pace, as the number of mobile devices (like smart mobile devices (SMD), tablets, PCs, etc.) in this world are increasing at a very fast pace [30, 31]. The usage of mobile devices is growing at a rate greater than 50% year-on-year (Y-o-Y) [18]. MCC is an amalgamation of SMD and the cloud which are inter-connected with the help of the network components [3, 19–22]. Among the mobile devices, the most popular one is the SMD followed by tablets and laptops and gaming applications. Mitra et al. proposed a technique which mainly focuses on emergency situation, where authors introduced cloud probing service (CPS) and cloud ranking service (CRS). The CPS was used for the selection of available cloud and from that the best public cloud is selected based on the CRS service. An anchor point (AP) has been used for running CPS and CRS. Mitra et al. considered the network probing service (NPS) to select the best available network and Amazon Web services were used as a public cloud. An emergency vehicle is used and is served as a local cloud. The proposed work is suitable when the networks (3G or Wi-Fi) are available. People around the world spend 90% of their mobile usage time on the different mobile apps. The differences between a cloud and cloudlet are that the cloud refers to a huge computing setup with standard features like elasticity, pay-per-use, on-demand, etc., and cloudlet refers to a smaller computing environment achieved by leveraging the power of mobile devices and hence possess limited power but is elastic at the same time [22]. Multimedia is content that uses a combination of different content forms such as text, audio, images, animations, video, and interactive content [23, 24]. Multimedia can be recorded and played, displayed, interacted with or accessed by information content processing devices, such as computerized and electronic devices, but can also be part of a live performance [22].

Multimedia devices are electronic media devices used to store and experience multimedia content. In order for multimedia applications to work efficiently and providing a rich user experience, it is very necessary to integrate with MCC as the majority of the multimedia processing for these compute-intensive applications will happen in the MCC [22]. MCC has a large amount of computing power resources and they are made use during the processing. The process of offloading from the SMD to the MCC happens when the mobile device of the user does not have enough resources to process the multimedia application. Upon subsequent processing in the MCC, the results are returned back to the SMD. This may result in delay of sometime due to the communication time required; however, the speed and quality of processing are far more superior in the MCC [23]. Also, this demands for a good integration between the MCC and the multimedia application. The various techniques for integration include REST API-based integration and message queueing-based integration, etc [23].

3 System Architecture

In this section, the authors have proposed a robust architecture which helps to process the multimedia applications using MCC. In our previous work, we have proposed an efficient architecture named as MBA which facilitated the efficient processing of the big data applications using MCC [31]. Multimedia application requires huge volumes of data which are stored in the big data ecosystem (like hadoop). We have used the component as illustrated in Fig. 1 and also the flow chart of MBA approach is depicted in Fig. 2.

During the processing of these applications, different layers are used, which help to run the applications as shown in Fig. 3. The different components used help to better serve the multimedia applications being run by the mobile user. The cloud layer consists of the different types of cloud topologies listed below:

- The public cloud is open to all the users on a pay-per-use basis and the popular ones are Amazon, Microsoft Azure, Google, Oracle, etc.
- The private cloud is used only within a particular enterprise and cannot be accessed by everyone.
- The hybrid cloud is a combination of both the public and the private cloud and used to host the general applications on the public cloud, whereas the more important applications are hosted on the private cloud.

The MCC layer is where the initial computation takes place once the user has initiated the particular application (multimedia applications like gaming, virtual reality, etc.). Offloading happens from the MCC which is the local cloud to the

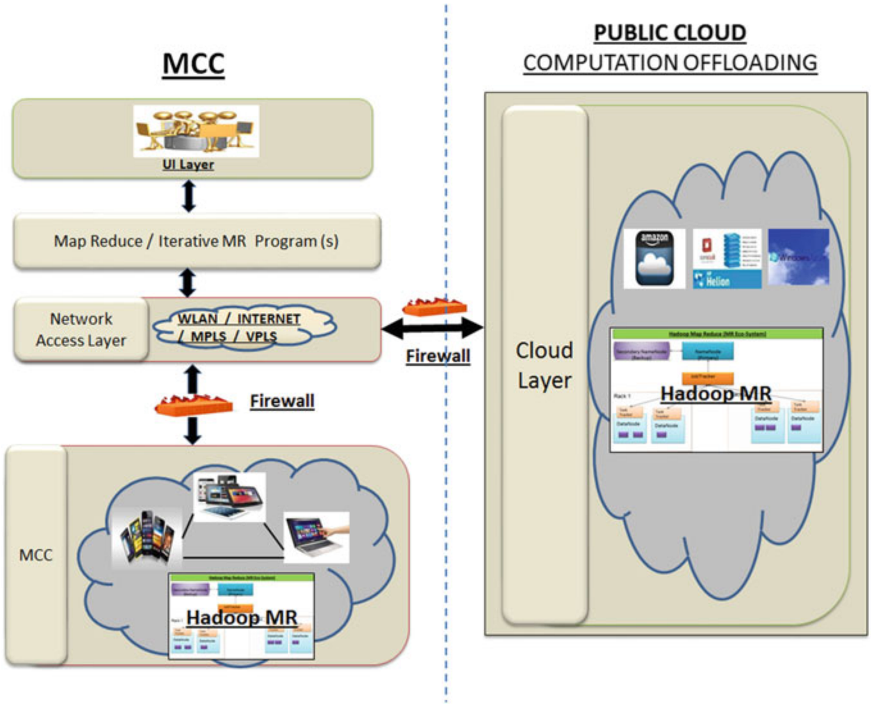


Fig. 1 Architecture of MBA computing

public/private/hybrid cloud when there is a requirement for more computational resources for various multimedia applications.

The gateway acts as the connection between the SMD's and the MCC. Packet data gateway (PDG) is connected to the local cloud (MCC) which in turn gets the data from the service gateway (SG). The SG collects the data from the various networks and acts as the bridge between the user and the MCC. The radio access network (RAN) provides input data to the SG by collecting it from the different SMD users. The input data comes from the various multimedia applications which are present in the SMD of the user.

The physical architecture is depicted in Fig. 4.

The initiation of the request starts with the mobile user. The user can start up any type of multimedia application like interactive TV, computer games or virtual reality. Upon initiation of the request, the RAN carries forward the request to the SG for further processing. The PDG aggregates the request from the various SG's and then sends it across to the MCC for computing. MCC tries to ensure that the computation is fully completed. However, due to the constraint of limited resources, MCC may not be able to fully process the request as the application may be demanding more amount of resources in terms of computational power, storage requirements, and network. In such a scenario, the resource-hungry applications are offloaded to the

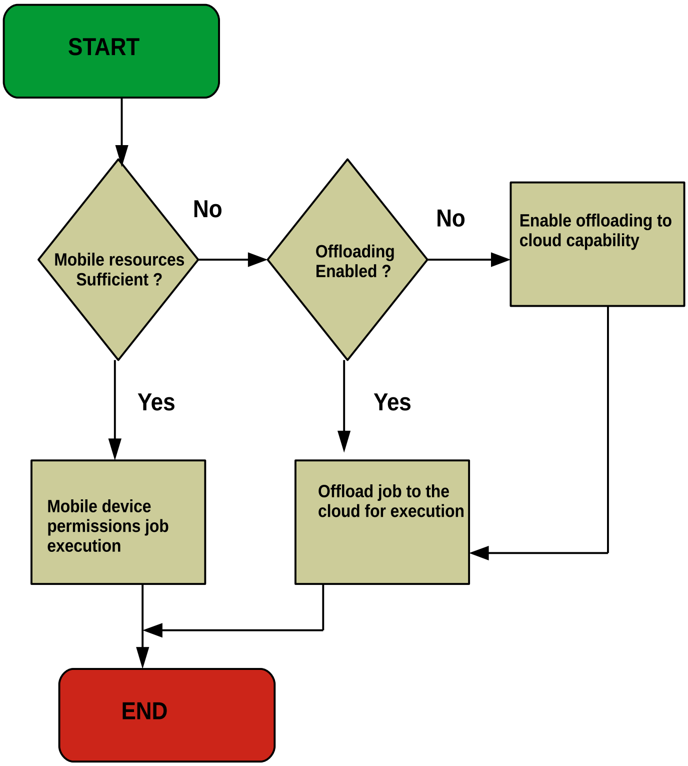


Fig. 2 Flow chart of MBA computing

cloud which has immense resources. The applications are then executed in the cloud and the results are returned back to the SMD user.

Definition 1 The offloading approach uses the following conditions:

$$\mathrm{Exec(me, ce, pc)} = \begin{cases} \mathrm{me}, & \text{if } e_c(m) \leq e_c(c) \\ \mathrm{ce}, & \text{if } e_c(m), e_c(\mathrm{pc}) > e_c(c) \\ \mathrm{pc}, & \text{if } e_c(c) > e_c(\mathrm{pc}) \end{cases}$$

where me, ce, and pc refer to the mobile, local cloud, and public cloud execution, respectively, and $e_c(m)$, $e_c(c)$, and $e_c(pc)$ are the consumption of energy required to process over the mobile device, local cloud, and public cloud execution, respectively.

The proposed architecture has several advantages as listed below.

– The use of offloading technique in this architecture helps to combat the limitations of the mobile devices such as limited storage capacity and also limited battery

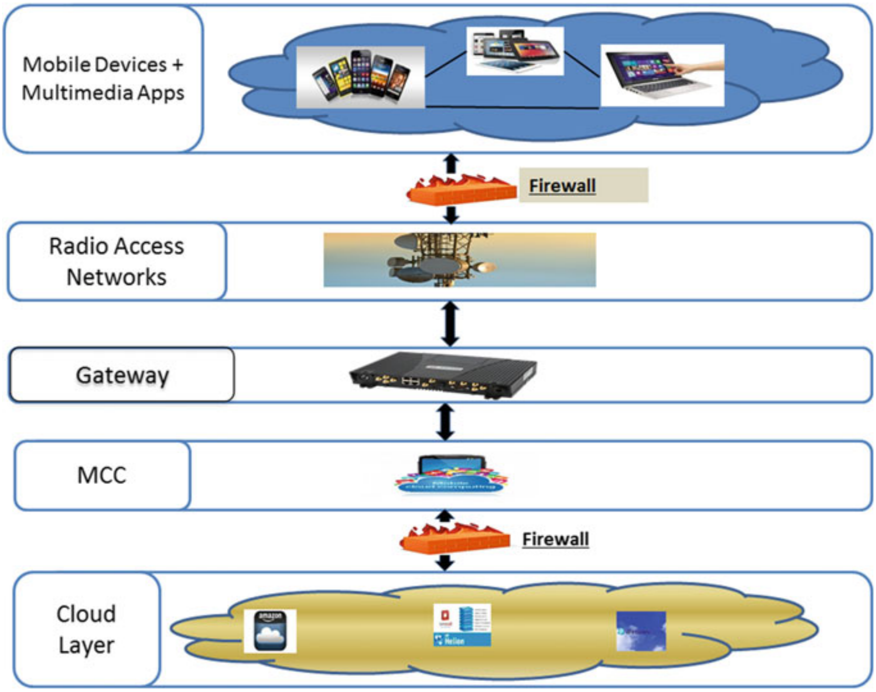


Fig. 3 Architecture layers of MCC and multimedia

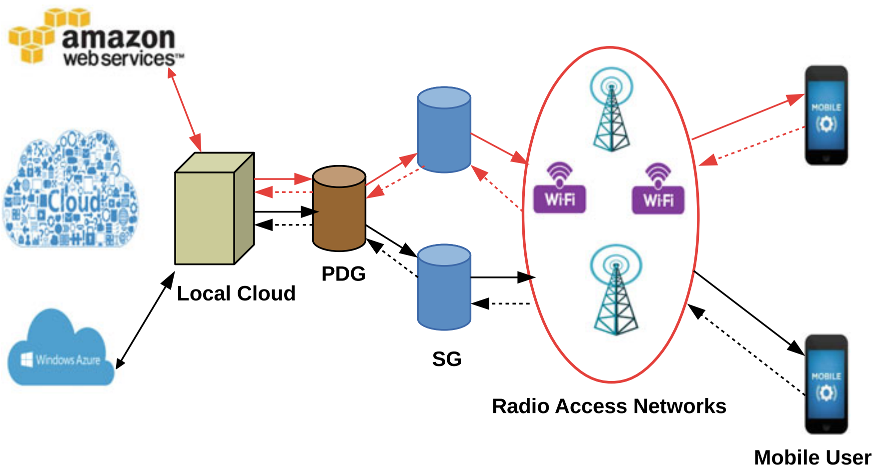


Fig. 4 Physical architecture of MCC and multimedia [32]

power. The data sensed via the embedded sensors of the smartphone is offloaded to the cloud and subsequently analyzed using various mining techniques (e.g., Bayesian algorithms, random forest, support vector machines, etc.), and can produce as output a classification results which is sent back to the mobile devices.
– It is capable to handle various types of multimedia applications in an efficient manner.
– In order to process the multimedia application faster, we plan to use the multi-cloud environment where based on the CPS service a best cloud will be selected using CRS service. We believe that, the network accessibility such as the problems related to low bit rates, high latency, packet losses, temporal disconnections, etc., and mobility issues such as roaming between different radio access technologies are also taken into considerations.

4 Conclusion

Multimedia applications use resources heavily for delivering rich user experience and hence there is a need to have a robust architecture for running such applications. In this paper, we have presented an architecture which will serve the purpose of running multimedia applications on the SMD of the user in a fairly decent manner. However, there is still lots of work to be done in the future as well which has been outlined in this paper. Multimedia applications are the most resource-intensive applications in terms of computational, communication, and storage power. Because of the limitations caused by limited resources, it is not always possible to deliver a rich user experience for different multimedia applications. Often high cost is required for such high quality of performance which incurs high cost of buying GPU-based machines and high-end storage or leveraging the cloud in order to deliver a near real-time experience to the user.

In the future, we need to come up with an architecture which can scale up to several thousands of users who can concurrently run their multimedia applications and get a very good near real-time user experience. Also, we need to ensure that the communication happens at a very fast rate so that near real-time output is visible to the user. This will be achieved by using faster algorithms and also high-end network equipment.

References

1. Abolfazli, S., Sanaei, Z., Gani, A., Shiraz, M.: MOMCC: market-oriented architecture for mobile cloud computing based on service oriented architecture. In: Proceedings of IEEE MobiCC'12, pp. 8–13 (2012)

2. Kumar, K., Liu, J., Lu, Y.-H., Bhargava, B.: A survey of computation offloading for mobile systems. Mob. Netw. Appl. **18**(1), 129–140 (2012)
3. Chun, B., Ihm, S., Maniatis, P., Naik, M., Patt, A.: Clonecloud: elastic execution between mobile device and cloud. In: Proceedings of the Sixth Conference on Computer Systems, pp. 365–372 (1999)
4. Schooley, B., Walczak, S., Hikmet, N., Patel, N.: Impacts of mobile tablet computing on provider productivity, communications, and the process of care. Int. J. Med. Inform. **88**, 62–70 (2016)
5. Cheong, S.H., Lee, K.-I., Si, Y.-W., U, L.H.: Lifeline: emergency ad hoc network. In: Proceedings of Seventh International Conference on Computational Intelligence and Security (CIS), pp. 283–289 (2011)
6. Garg, S.K. , Lyles, C.R. , Ackerman, S., Handley, M.A., Schillinger, D., Gourley, G., Aulakh, V., Sarkar, U.: Qualitative analysis of programmatic initiatives to text patients with mobile devices in resource-limited health systems. BMC Med. Inform. **16**(16), 1–112 (2015)
7. Martínez-Pérez, B., de la Torre-Díez, I., López-Coronado, M.: Privacy and security in mobile health apps: a review and recommendations. J. Med. Syst. **39**(1), 181:1–8 (2015)
8. Mitra, K., Saguna, S., Åhlund, C., Granlund, D.: M^2C^2: a mobility management system for mobile cloud computing. In: Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), pp. 1608–1613 (2015)
9. Misra, S., Chatterjee, S.: Social choice considerations in cloud-assisted WBAN architecture for post-disaster healthcare: data aggregation and channelization. Inf. Sci. **284**, 95–117 (2014)
10. Baccelli, F., Blaszczyszyn, B., Muhlethaler, P.: An Aloha protocol for multihop mobile wireless networks. IEEE Trans. Inf. Theory **52**(2), 421–436 (2006)
11. Fang, W., Yin, X., An, Y., Xiong, N., Guo, Q., Li, J.: Optimal scheduling for data transmission between mobile devices and cloud. Inf. Sci. **301**, 169–180 (2015)
12. Shu, P., Liu, F., Jin, H., Chen, M., Wen, F., Qu, Y., Li, B.: eTime: energy-efficient transmission between cloud and mobile devices. In: IEEE Infocom, pp. 195–199 (2013)
13. Panigrahi, C.R., Sarkar, J.L., Pati, B.: Transmission in mobile cloudlet systems with intermittent connectivity in emergency areas. Digit. Commun. Netw. (2017). https://doi.org/10.1016/j.dcan.2017.09.006
14. Kwak, J., Kim, Ye., Lee, J., Chong, S.: DREAM: dynamic resource and task allocation for energy minimization in mobile cloud systems. IEEE J. Sel. Areas Commun. **33**(12), 2510–2523 (2015)
15. Moltchanov, D.: Distance distributions in random networks. Ad Hoc Netw. **10**, 1146–1166 (2012)
16. http://www.missionmode.com/15-disaster-and-crisis-apps-for-android/. Accessed 27 Sept 2016
17. Fajardo, J.T., Oppus, C.M.: Implementation of an android-based disaster management system. In: Proceedings of the 9th WSEAS International Conference on Electronics, Hardware, Wireless and Optical Communications, pp. 126–130 (2010)
18. http://www.marketsandmarkets.com/PressReleases/mobile-cloud.asp
19. Ou, S., Yang, K., Liotta, A.: An adaptive multi-constraint partitioning algorithm for offloading in pervasive systems. In: Proceedings of IEEE International Conference on Pervasive Computing and Communications, pp. 116–125 (2006)
20. Huerta-Canepa, G., Lee, D.: An adaptable application offloading scheme based on application behavior. In: Proceedings of International Conference on Advanced Information Networking and Applications Workshops, pp. 387–392 (2008)
21. Gu, X., Nahrstedt, K., Messer, A., Greenberg, I., Milojicic, D.: Adaptive offloading inference for delivering applications in pervasive computing environments. In: Proceedings of IEEE International Conference on Pervasive Computing and Communications, pp. 107–114 (2003)
22. Mitra, K., Saguna, Ahlund, C.: A mobile cloud computing system for emergency management. IEEE Cloud Comput. **43**(4), 30–38 (2014)
23. Luo, H., Shyu, M.-L.: Quality of service provision in mobile multimedia—a survey. Hum.-Centric Comput. Inf. Sci. **1**(1), 1–15 (2011)

24. Yang, J., He, S., Lin, Y., Lv, Z.: Multimedia cloud transmission and storage system based on internet of things. Multimed. Tools Appl. 1–16 (2015)
25. Andersson, K., Granlund, D., Åhlund, C.: M4: multimedia mobility manager: a seamless mobility management architecture supporting multimedia applications. In: Proceedings of the 6th International Conference on Mobile and Ubiquitous Multimedia, pp. 6–13 (2007)
26. Panigrahi, C.R., Pati, B., Tiwary, M., Sarkar, J.L.: EEOA: improving energy efficiency of mobile cloudlets using efficient offloading approach. In: IEEE International Conference on Advanced Networks and Telecommunications Systems, pp. 1–6 (2015)
27. Pati, B., Panigrahi, C.R., Sarkar, J.L.: CETM: a conflict-free energy efficient transmission policy in mobile cloud computing. Int. J. Commun. Netw. Distrib. Syst. **20**(2), 129–142 (2018)
28. Pati, B., Sarkar, J.L., Panigrahi, C.R., Debbarma, S.: eCloud: an efficient transmission policy for mobile cloud computing in emergency areas. Progress in Intelligent Computing Techniques: Theory, Practice, and Applications, pp. 43–49 (2018)
29. Sarkar, J.L., Panigrahi, C.R., Pati, B., Trivedi, R., Debbarma, S.: E2G: a game theory-based energy efficient transmission policy for mobile cloud computing. Progress in Advanced Computing and Intelligent Engineering, pp. 1–8 (2017)
30. Panigrahi, C.R., Sarkar, J.L., Pati, B., Bakshi, S.: E^3M: an energy efficient emergency management system using mobile cloud computing. In: IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Bangalore, India, pp. 1–6 (2016)
31. Verma, R.K. , Pati B., Sarkar, J.L., Panigrahi, C.R., Mohapatra, S.D., M2C: an energy efficient mechanism for computation in mobile cloud computing. Progress in Advanced Computing and Intelligent Engineering, pp. 1–6 (2018)
32. Verma, R.K., Pati, B., Sarkar, J.L., Panigrahi, C.R.: mMedia: an efficient transmission policy for multimedia applications using mobile cloud computing. Int. J. Sens. Wireless Commun. Control (2018, in press). https://doi.org/10.2174/2210327908666180727130121

A Comparative Study of Task Scheduling Algorithm in Cloud Computing

Subhashree Mohapatra, Chhabi Rani Panigrahi, Bibudhendu Pati and Manohar Mishra

Abstract Nowadays, cloud computing plays an important role in both academia and industry. However, due to high demands from the clients and limited resources, it is required to transfer a few workloads to other data centers. This process helps to complete the applications submitted to the data centers even through more flexible and cheaper resources. As there are many tasks lined up at the server side and the resources are heterogeneous in nature, scheduling the tasks in multi-cloud environment is always a challenge. The loads should be processed in such a way that all the resources are used very efficiently. In this work, we have performed a comparative study on different task scheduling algorithms proposed in the literature in cloud computing.

Keywords Scheduling · Cloud computing · Makespan

1 Introduction

Cloud computing is a general term that refers to the utilization of different resources across the Internet as and when required. Cloud computing technology is being used by big firms to consume computer resources as utilities instead of having in-built infrastructure. The main aim of embedding this technology is to increase efficient usage of resources and decrease cost [1]. Self-service on-demand, elasticity and pay per use are few important benefits of cloud computing. It is designed in such a way

S. Mohapatra (✉)
S'O'A Deemed to be University, Bhubaneswar, India
e-mail: subhashreemohapatra68@gmail.com

C. R. Panigrahi · B. Pati
Rama Devi Women's University, Bhubaneswar, India
e-mail: panigrahichhabi@gmail.com

B. Pati
e-mail: patibibudhendu@gmail.com

M. Mishra
ITER, Siksha 'O' Anusandhan (SOA) Deemed to be University, Bhubaneswar, India
e-mail: manohar2006mishra@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_28

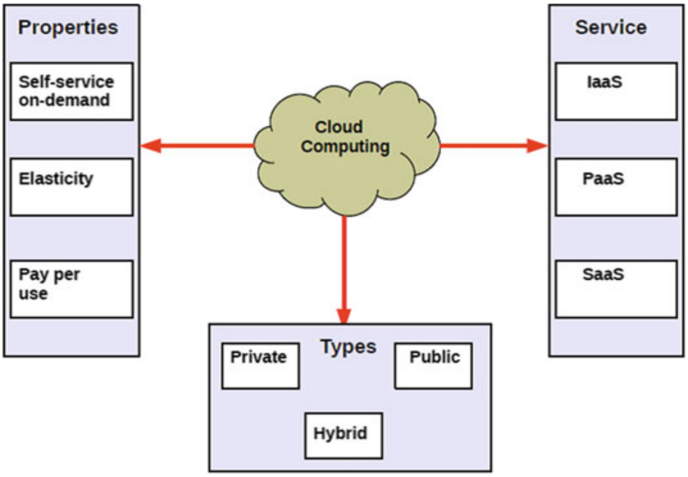


Fig. 1 Cloud computing infrastructure

so that it can provide services to the clients in three categories, namely software as a service (SaaS), infrastructure as a service (IaaS) and platform as a service (PaaS) [2–4].

Figure 1 shows the cloud computing infrastructure. Based on deployment of the cloud, it can be of three types which are private cloud, public cloud and hybrid cloud [5, 6]. As clients can make a demand to access a particular resource, there may be many tasks which have to be handled by the same resource. So, there is a need for proper management of the task execution. Scheduling is a procedure to find the sequence in which different tasks are executed for each machine [7]. Task scheduling in cloud computing is introduced as the procedure of assigning different tasks to resources according to predefined rules of resource use for a specific cloud environment [8]. The leading purpose of task scheduling is to minimize the makespan so that all resources are utilized in most efficient way [9]. There are two types of heuristic according to which the task can be mapped to different resources which are immediate mode and batch mode. In immediate mode heuristic, a task is allocated to a resource as soon as it is received by the mapper. In batch mode, tasks are put into a group that is examined for mapping at predefined intervals called mapping events [10].

1.1 Motivation

As the demand for cloud computing goes on increasing, it has become important to schedule the tasks properly. In cloud, many users demand for resources at the same time. Due to this, a scheduling decision has to be made in least time available.

Task scheduling procedure which can be accepted should have a good impact on the performance of the system [11]. For this, new and efficient strategies have to be developed and implemented. There are various heuristics already proposed regarding the issue of task scheduling. Before proposing any new methodology, it is important to know the benefit and drawback of the existing work.

1.2 Contribution

In this work, authors have studied various immediate as well as batch mode task scheduling strategies. Among the immediate mode heuristics, k-percent best, minimum completion time, switching algorithm, minimum execution time and opportunistic load balancing are studied. The scheduling algorithms studied under batch mode include Max-min algorithm, Min-min algorithm, Sufferage algorithm, RASA and TASA. Some of the variations of the existing batch mode methods have also been studied in this paper. All batch mode heuristics are reviewed in detail so as to have an analysis of the makespan achieved by the individual algorithms. Comparison of algorithms is also done to analyze which method is having the best performance.

The remaining part of the paper is categorized as mentioned. In the following sections, we present the different task scheduling algorithms already proposed and their comparison. Section 4 provides experimental setup. Section 5 gives the overall analysis of all the task scheduling algorithms. Finally, we conclude the paper in Sect. 6.

2 Immediate Mode Heuristics

Different immediate mode algorithms are described in this section. Described heuristics are (i) *minimum completion time*, (ii) *minimum execution time*, (iii) *switching algorithm*, (iv) *k-percent best* and (v) *opportunistic load balancing*. Table 1 presents a general overview of the above-mentioned techniques.

The MCT heuristic [10] takes into consideration the completion time of the tasks. It submits a task to a resource which has the earliest completion time. Sometimes so happens that some tasks are given to the resources which do not give the least execution time for them [12].

The MET heuristic [10] considers the execution time of the tasks. This heuristic allocates the tasks to the resource those have least execution time. The benefits of MET are that it allocates a task to a resource which executes the task in minimum amount of execution time, and also it is a simple heuristic.

The SA heuristic [10] is inspired from the stated reviews. MET algorithm can sometimes lead to imbalance of load throughout resources by allocating more number of tasks to few resources than to others. MCT algorithm goes for a load balancing by submitting tasks having earliest completion time. But when tasks enter in random

Table 1 General overview of immediate mode heuristics

Author	Algorithm	Time complexity	Assumption
Maheswaran et al. [10]	Minimum completion time	O (n)	Task is allocated to resource which completes the task within least completion time
Maheswaran et al. [10]	Minimum execution time	O (n)	Task is allocated to resource which completes the task within least execution time
Maheswaran et al. [10]	Switching algorithm	O (n)	Executes MET and MCT in cyclic alternately
Maheswaran et al. [10]	K-percent best	O (n log n)	Fulfills the objective of MCT and MET simultaneously
Maheswaran et al. [10]	Opportunistic load balancing	O (n)	Do not consider execution time of the task

collection, it is more feasible to apply MET technique for a better load balancing till it reaches a given threshold and then considering MCT to have a smooth balance of load across the resources.

A subset of resources is used in this method while mapping a task in KPB [10]. If T is the count of resources and N the set of tasks, then the subset is clustered by choosing the T * (N/100) best resources depending on execution times for the task, where 100/T <= N <= 100. The task having the earliest completion time is assigned to a machine.

The OLB heuristic [10] takes into consideration the machine which is next in the ready state. Without checking the execution time of a task, it is allotted to a resource that is ready next. If more than one machine is ready at a particular time, then randomly any of the resources can be selected.

3 Batch Mode Heuristics

In this section, the existing batch mode algorithms are outlined. These are: (i) *Min-min algorithm*, (ii) *Max-min algorithm*, (iii) *Sufferage heuristic*, (iv) *RASA and* (v) *TASA heuristic*. Some of the advanced versions of the mentioned algorithms are also described. In batch mode heuristics, set of tasks are mapped at previously defined time gaps. Table 2 gives a brief idea about these heuristics.

Table 2 Brief idea about batch mode heuristics

Author	Algorithm	Simulator used	Advantage
Freund et al. [13]	Min-min	Discrete event simulator	Gives a better makespan for longer task
XiaoShan et al. [14]	QoS-guided Min-min	Simulated grid environment	Represents QoS of a network by its bandwidth
Freund et al. [13]	Max-min	Discrete event simulator	Gives a better makespan for shorter task
Etminani et al. [23]	Min-min Max-min selective	GridSim	Uses resource utilization rate as a parameter
Elzeki et al. [15]	Improved Max-min	WorkflowSim	Selection basis of a task is maximum execution time rather than completion time
Bhoi et al. [16]	Enhanced Max-min	Java 6	Task having average or near to average execution time is selected for execution
Maheswaran et al. [10]	Sufferage	Discrete event simulator	The task which suffers most is executed first
Chiang et al. [18]	Advanced MaxSufferage	–	The task having the average expected completion time is executed first
Chaudhary et al. [19]	Deadline and Sufferage aware	CloudSim	Takes into account deadline of task, meta-task size, system volume for calculating execution time
Munnir et al. [17]	QoS Sufferage	Simulator with ETC matrix	Takes into account the network bandwidth as the QoS parameter
Parsa et al. [20]	RASA	GridSim	Reduces makespan by combining the advantage of Min-min and Max-min

(continued)

Table 2 (continued)

Author	Algorithm	Simulator used	Advantage
Dehkordi et al. [21]	TASA	CloudSim	Reduces makespan by combining the advantage of Min-min and Sufferage
Mittal et al. [22]	Optimized TSA	Java 7	Reduces makespan by combining the advantage of Min-min, Max-min and RASA

The Min-min heuristic [6, 13] selects a task for execution in reference to the completion time. It starts the process by finding the summation of execution time (ET_i) and the ready time (RT_i) for all tasks (T_i) in list. Then for every task, the machine (R_j) which has the minimum of earliest completion time (CT_i) is found out. Among these time values, the least one is selected. The machine which yields the minimum completion time is assigned with the chosen task [10]. Once a task is handed over to a machine, it is deleted from the task set. Completion time of each remaining task along with the ready time of selected machine is upgraded for the next round. This step is reiterated till every task has completed execution.

The QoS-guided heuristics [14] is same as Min-min technique, but it also considers the bandwidth as QoS of the network for selecting a task. In this method, the tasks having dissimilar level of QoS seek for resources. Some resources have high QoS, whereas other can have low QoS. So in this method, a task requesting for high QoS can be implemented in machine having high QoS and a task requesting for low QoS can only be executed in a resource having low QoS. If in any case the task has no QoS, request can be executed in either a resource having low or high QoS.

The Max-min heuristic [10, 13] is quite close to Min-min heuristic. It varies from Min-min heuristic by choosing the task with maximum earliest completion time for execution. It calculates expected completion time from the expected execution time and the ready time. Then from the list, the resources which have the minimum expected completion time are determined. Finally, the task which shows the maximum of the calculated value is assigned with the particular task. The process is iterated with the change of completion time and ready time of the machine, till all the tasks have completed execution.

The Min-min Max-min selective heuristic takes into consideration the benefit of Max-min as well as Min-min and selects among the two in each of the iterations. At first, the tasks are arranged in increasing order and completion time of each task for a machine is found out. As in the next step, the minimum of the expected completion time is found out. The standard deviation gives an idea about the number of longer or shorter tasks present in the meta-task. If the number of shorter task is more, Max-min is implemented; else Min-min is implemented.

To minimize the shortcoming and improve the efficiency of Max-min heuristic, a better form of Max-min is stated in [15]. In improved Max-min heuristic, the process starts with the summation of the execution time and ready time of all tasks in meta-task. In next step instead of extracting the task having maximum completion time, improved Max-min finds the task with maximum execution time. Finally, the selected task is assigned to the machine which gives the minimum completion time.

Enhanced Max-min [16] gives a better way of selecting a particular task for execution. This is an enhanced version of improved Max-min heuristic. Improved Max-min selects a task having maximum execution time. In doing so, it selects the largest task first. But in enhanced Max-min, this selection process of the task is altered. In enhanced Max-min, average of the execution time is calculated. In next step, task having average or nearest greater than average execution time is chosen for execution. The selected task is then submitted to the machine having the minimum completion time. This reduces the makespan, and load balancing is also improved.

Sufferage heuristic [10, 17] is built up on the concept that better mapping can be triggered by allocating a task to a resource that would 'suffer' maximum in terms of expected completion time. Sufferage for a task is calculated by finding the difference between second minimum completion time and first minimum completion time for the particular task. First of all the Sufferage value for all the tasks is calculated. Task having maximum Sufferage amount is allocated to the machine giving first minimum completion time.

As effective load balancing is not done in Sufferage heuristic, an improved method is proposed. In advanced MaxSufferage heuristic [18], there are three phases. In the first phase, the Sufferage value is calculated for each task called as the SV calculation phase. The second phase is the MSV calculation phase, and the third is the task dispatching phase.

In deadline and Sufferage-aware algorithm [19], all the tasks are arranged according to the deadline time. The task having the minimum deadline time is kept at the start of the list. In the next step, the Sufferage value is calculated and the task having maximum Sufferage is selected for execution.

Like QoS-guided Min-min algorithm, QoS Sufferage heuristic [17] also takes into account network bandwidth for QoS parameter. Here, the task demands for a high or low QoS and accordingly the tasks are assigned to machine having high QoS value or machine having low QoS value, respectively. After the completion time of each task is calculated, the tasks requesting for high QoS are executed first.

RASA [20] is a hybrid heuristic which is a combination of two other, i.e., Min-min and Max-min. RASA is named as *resource-aware scheduling algorithm*. Here, the number of resources in the list is taken into consideration for working. It starts by counting the number of machines available. If the number of resources is even, then the first round is executed using Max-min heuristic else Min-min heuristic. Then, the rest of the tasks are executed applying Min-min and Max-min alternately. In other words, let us say there are five resources.

Like RASA, TASA [21] too combines two heuristics which are Min-min and Sufferage. As TASA is *task-aware scheduling algorithm*, here the algorithm takes into account the numeric count of tasks in the task list. When task count is odd, then

Min-min heuristic is executed; else Sufferage method is executed for assignment of task to machines. In each iterating loop, the algorithm starts with the calculation of the task number. As the number of tasks goes on decreasing by one, the number of tasks is alternately odd or even.

The optimized task scheduling heuristic [22] tries to remove the disadvantage of Max-Min, Min-Min and RASA. It selects one of these algorithms according to the values calculated.

4 Experimental Setup

To experiment and analyze the different batch mode heuristics used in cloud computing, MATLAB simulation environment is used. MATLAB version 2011 is chosen to carry out the evaluation of the methods. The testing is done for the cases of task and resources as follows.

- Case I—All the algorithms are tested for a set of 10 tasks, i.e., {T1, T2, T3… T10}, and a set of 3 resources, i.e., {R1, R2, R3}.
- Case II—All the heuristics are tested for a set of three resources along with an increasing number of tasks. The task numbers are increased as 20, 40, 60, 80 and 100.
- Case III—Testing is done for a constant 10 tasks but with an increasing number of resources. The resources are increased as 5, 8 and 10.

5 Analysis of Different Heuristics

For the analysis of case I, the description about the group of task and resource is mentioned in Tables 3 and 4, respectively.

The expected execution time (EET_{ij}) is calculated using the formula given below, and the calculated values are mentioned in Table 5.

Table 3 Task description

Task	Instruction volume (MI)	Data volume (Mb)	Task	Instruction volume (MI)	Data volume (Mb)
T1	1315	1406	T6	657	1471
T2	627	1414	T7	1458	985
T3	1132	597	T8	1301	642
T4	778	1047	T9	922	1416
T5	1458	1465	T10	1292	1460

Table 4 Resource description

Resource	Instruction volume (MIPS)	Data volume (MbPS)
R1	166	103
R2	185	194
R3	168	176

Table 5 Expected execution time of each task on each resource

Task	EET on R1	EET on R2	EET on R3	Task	EET on R1	EET on R2	EET on R3
T1	21.57	14.35	15.82	T6	18.24	11.13	12.27
T2	17.50	10.68	11.77	T7	18.35	12.96	14.27
T3	12.62	9.19	10.13	T8	14.07	10.34	11.39
T4	14.85	9.60	10.58	T9	19.30	12.28	13.53
T5	23.01	15.43	17.00	T10	21.96	14.51	15.98

Table 6 Makespan of each algorithm

Algorithm	Makespan
Min-min	55.42
Max-min	51.04
Sufferage	50.04
RASA	54.94
TASA	49.92

$$\text{Expected Execution Time} = (\text{MI}/\text{MIPS}) + (\text{Mb}/\text{MbPS})$$

The above data is used to find out the makespan for all algorithms studied and considered for analysis. The respected makespan values are shown in Table 6. From table, it is observed that Min-min has given a maximum makespan. The lowest makespan is given by TASA. As TASA combines the advantages of Min-min along with Sufferage, it gives the most efficient result. The different heuristics are compared in terms of makespan. The comparison of makespan of different heuristics is shown in graphical representation.

In Fig. 2a, Min-min, Max-min and Sufferage heuristics are analyzed. Among these three, Sufferage is showing a better result. Sufferage executes the task which will be suffering most in the set. It minimizes the disadvantage of Min-min and Max-min giving better output. Figure 2a shows makespan of Min-min, Max-min and Sufferage. In Fig. 2b, the algorithms compared are Min-min, Max-min and RASA. In this set of tasks and resources, RASA is having a better makespan than Min-min but a worse makespan than Max-min. As mentioned by the author in [20], RASA can give an equivalent result to Min-min or Max-min when the task is of almost same size. Hence in this case, RASA is giving a result nearly equal to Min-min. There can

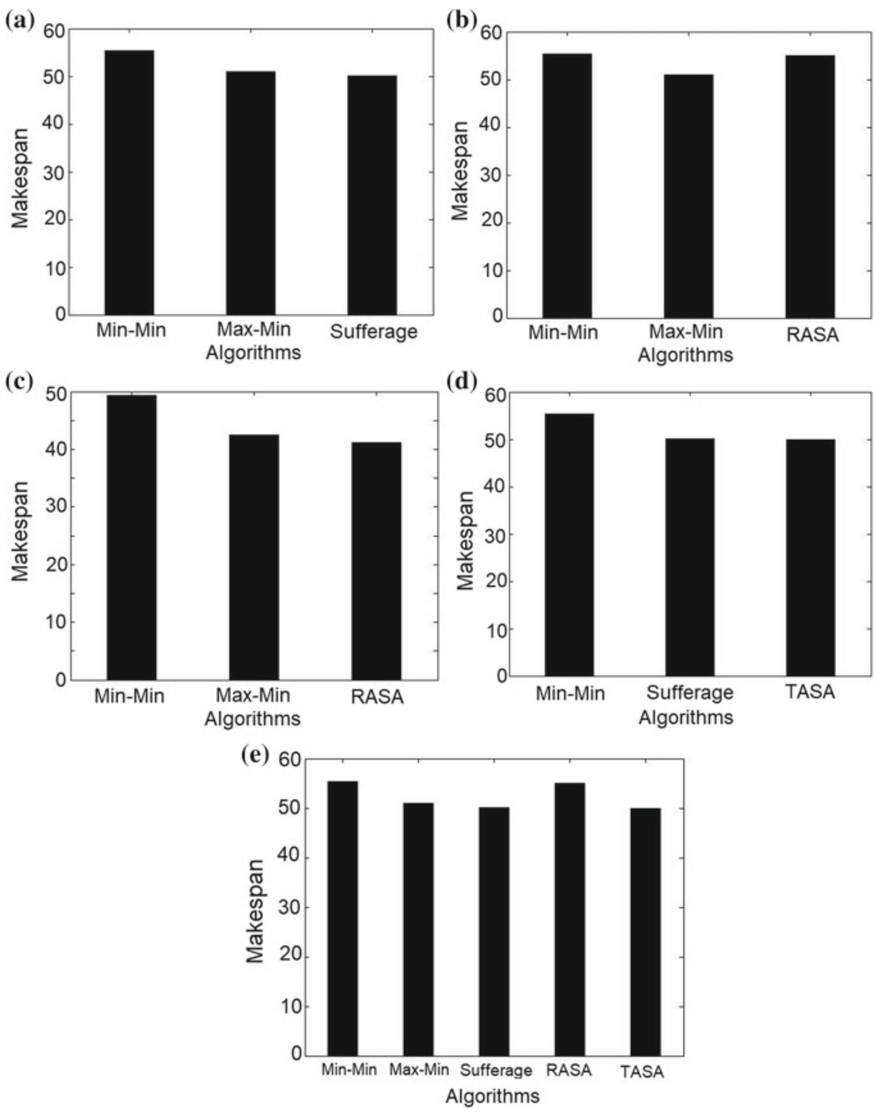


Fig. 2 Comparison of makespan versus algorithm. **a** Makespan of Min-min, Max-min, Sufferage. **b** Makespan of Min-min, Max-min, RASA. **c** Makespan of Min-min, Max-min, RASA (best). **d** Makespan of Min-min, Sufferage, TASA. **e** Makespan of the entire batch mode heuristic

also be situation where RASA outperforms than the other two methods. This can be shown by taking another set of task but with the same resource specification.

Tables 7 and 8 give the details of task and resource specification, respectively. Table 9 gives the expected execution time of tasks for which RASA gives a better makespan than Min-min and Max-min. Figure 2c shows the makespan comparison.

Table 7 Task description

Task	Instruction volume (MI)	Data volume (Mb)	Task	Instruction volume (MI)	Data volume (Mb)
	1155	662	T6	1006	1199
T2	619	998	T7	1391	1460
T3	1460	840	T8	1047	638
T4	1085	724	T9	649	757
T5	1252	755	T10	1341	754

Table 8 Resource description

Resource	Instruction volume (MIPS)	Data volume (MbPS)
R1	166	103
R2	185	194
R3	168	176

Table 9 Expected execution time of each task on each resource for better result of RASA

Task	EET on R1	EET on R2	EET on R3	Task	EET on R1	EET on R2	EET on R3
T1	13.38	9.65	10.64	T6	17.70	11.62	12.80
T2	13.42	8.49	9.35	T7	22.55	15.04	16.57
T3	16.95	12.22	13.46	T8	12.50	8.95	9.86
T4	13.57	9.59	10.57	T9	11.26	7.41	8.16
T5	14.87	10.66	11.74	T10	15.40	11.13	12.27

Figure 2d shows a result comparison of Min-min, Sufferage and TASA. For the taken test sample, TASA is giving better makespan than the other two methods. TASA combines the benefit of both Min-min and Sufferage to give a minimum makespan. Figure 2e gives a comparative plot of all the five heuristics. Here, it is seen that TASA is giving the best makespan among all the schedules.

In case II, the comparison is made for a different number of tasks for a constant number of resources. The heuristics have been compared for 20, 40, 60, 80, 100 tasks. As in a cloud environment, the number of tasks will go on increasing time to time having the number of resources constant, and it is required to check if there will be any major effect in relation to the heuristics.

Table 10 shows the makespan values for the changing task count, and Fig. 3a gives a pictorial representation of case II. For this scenario, it is observed that in most of the cases TASA is giving the best result. The comparison is also carried out for varying number of resources. The test is done for 3, 5, 8 and 10 numbers of resources. As the demand goes on increasing, the number of resources also has to increase. Hence, it is also necessary to test the efficiency of the algorithms for a constant number of

Table 10 Makespan for varying number of tasks

Algorithm	Makespan for # tasks (20)	Makespan for # tasks (40)	Makespan for # tasks (60)	Makespan for # tasks (80)	Makespan for # tasks (100)
Min-min	78.29	160.23	248.19	317.7	417.4
Max-min	78.27	160.66	252.54	329.2	420.1
Sufferage	76.64	158.38	246.98	322.6	412.4
RASA	76.40	161.00	245.94	321.2	416.8
TASA	76.64	156.39	245.38	322.1	418.2

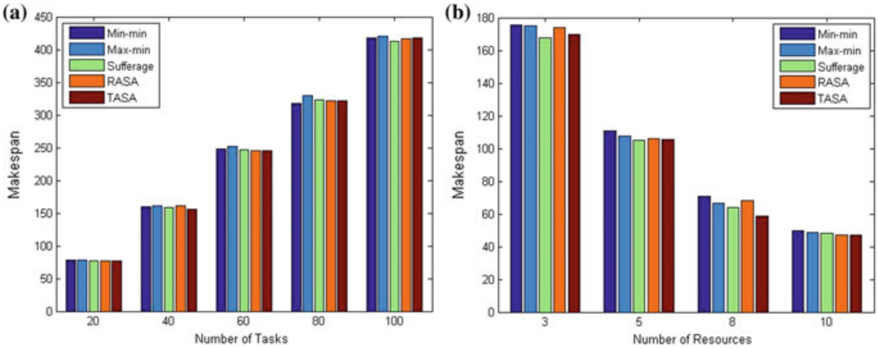


Fig. 3 Comparison of (**a**) makespan versus number of resources and (**b**) makespan versus number of tasks

tasks and increasing number of resources. Table 11 shows the makespan value for the changing count of resources, and Fig. 3b gives a graphical representation of case III. For this scenario, also it is observed that in most of the cases TASA is giving the best result.

Table 11 Makespan for varying number of resources

Algorithm	Makespan for # resources (3)	Makespan for # resources (5)	Makespan for # resources (8)	Makespan for # resources (10)
Min-min	175.78	110.94	70.92	49.55
Max-min	174.96	107.38	66.46	48.77
Sufferage	167.87	105.20	63.89	48.18
RASA	173.77	106.12	68.27	47.22
TASA	169.65	105.42	58.85	47.11

6 Conclusions

For the varying test done to check the efficiency of the batch mode algorithms, most of the tests show the best result for TASA. In the increasing client demands, it is highly necessary to reduce the makespan as much as possible. The mentioned algorithms are proposed keeping the problem in view. In some cases, RASA is able to perform better than Min-min and Max-min. But, at times when the task size is almost same RASA also has a worse case timing. As future work, we will try to propose an algorithm to get a reduced makespan, considering the problems of existing heuristics.

References

1. Durao, F., Carvalho, J.F.S., Fonseka, A., Garcia, V.C.: A systematic review on cloud computing. J. Supercomput. **68**, 1321–1346 (2014)
2. Beloglazov, A., Abawajy, J., Buyya, R.: Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Futur. Gener. Comput. Syst. **28**(5), 755–768 (2012)
3. Singh, R.M., Paul, S., Kumar, A.: Task scheduling in cloud computing: review. Int. J. Comput. Sci. Inf. Technol. **5**(6), 7940–7944 (2014)
4. Jain, A., Kumar, R.: A comparative analysis of task scheduling approaches for cloud environment. In: IEEE International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1787–1792 (2016)
5. Buyya, R., Vecchiola, C., Selvi, S.T.: Mastering cloud computing: foundations and applications programming. Newnes (2013)
6. Mittal, R., Soni, K.: Analysis of cloud computing architectures. Int. J. Adv. Res. Comput. Commun. Eng. **2**(5), 2088–2091 (2013)
7. Braun, T.D., Siegel, H.J., Beck, N., Boloni, L., Maheswaran, M., Reuther, A.I., Robertson, J.P., Theys, M.D., Yao, B.: A taxonomy for describing matching and scheduling heuristics for mixed-machine heterogeneous computing systems. In: Proceedings of IEEE Symposium on Reliable Distributed Systems, pp. 330–335 (1998)
8. Yonggui, W., Ruilian, H.: Study on cloud computing task schedule strategy based on MACO algorithm. Comput. Meas. Control **19**(5), 1203–1204 (2011)
9. Wadhonkar, A., Theng, D.: A survey on different scheduling algorithms in cloud computing. In: IEEE International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB16) (2016)
10. Maheswaran, M., Ali, S., Siegel, H.J., Hensgen, D., Freund, R.F.: Dynamic mapping of a class of independent tasks onto heterogeneous computing systems. J. Parallel Distrib. Comput. **59**, 107–131 (1999)
11. Sobhanayak, S., Turuk, A.K., Sahoo, B.: Task scheduling for cloud computing using multi-objective hybrid bacteria foraging algorithm. Futur. Comput. Inform. J. 1–21 (2018)
12. Nagadevi1, S., Satyapriya, K., Malathy, D.: A survey on economic cloud schedulers for optimized task scheduling. Int. J. Adv. Eng. Technol. **4**(I), 58–62 (2013)
13. Freund, R.F., Gherrity, M., Ambrosius, S., Campbell, M., Halderman, M., Hensgen, D., Keith, E., Kidd, T., Kussow, M., Lima, J.D., Mirabile, F.: Scheduling resources in multiuser, heterogeneous, computing environments with SmartNet. In: 7th IEEE Heterogeneous Computing Workshop (HCW '98), pp. 184–199 (1998)
14. He, X., Sun, X., Von Laszewski, G.: QoS guided min-min heuristic for grid task scheduling. J. Comput. Sci. & Technol. **18**(4), 442–451 (2003)

15. Elzeki, O.M., Reshad, M.Z., Elsoud, M.A.: Improved max-min algorithm in cloud computing. Int. J. Comput. Appl. **50**(12), 22–27 (2012)
16. Bhoi, U., Ramanuj, P.N.: Enhanced Max-min Task Scheduling Algorithm in Cloud Computing. Int. J. Appl. Innov. Eng. & Manag. **2**(4), 2319–4847 (2013)
17. Munir, E.U., Li, J., Shi, S.: QoS sufferage heuristic for independent task scheduling in grid. Inf. Technol. J. **6**(8), 1166–1170 (2007)
18. Hiang, M.L., Hsieh, H.C., Tsai, W.C., Ke, M.C.: An improved task scheduling and load balancing algorithm under the heterogeneous cloud computing network. In: IEEE 8th International Conference on Awareness Science and Technology (iCAST), pp. 290–295 (2017)
19. Chaudhary, P., Varsha,: Deadline and suffrage aware task scheduling approach for cloud environment. Int. Res. J. Eng. Technol. (IRJET) **4**(8), 972–977
20. Parsa, S., Entezari-Maleki, R.: RASA: a new task scheduling algorithm ingrid environment. World Appl. Sci. J. (Special Issue of Computer & IT) **7**, 152–160 (2009)
21. Taherian Dehkordi, S., Khatibi Bardsiri, V.: TASA: a new task scheduling algorithm in cloud computing. J. Adv. Comput. Eng. Technol. **1**(4), 25–32 (2015)
22. Mittal, S., Katal, A.: An optimized task scheduling algorithm in cloud computing. In: 6th IEEE International Conference on Advanced Computing, pp. 197–202 (2016)
23. Etminani, K., Naghibzadeh, M.: A min-min max-min selective algorithm for grid task scheduling. In: Internet, ICI 2007. 3rd IEEE/IFIP International Conference in Central Asia on IEEE (2007)

Offer Based Auction Mechanism for Virtual Machine Allocation in Cloud Environment

Sasmita Parida, Bibudhendu Pati, Suvendu Chandan Nayak and Chhabi Rani Panigrahi

Abstract Nowadays, the cloud service providers provide different resources for computation such as CPU power, memory to a wide range of users by virtual machines (VM). The VMs vary according to their capacity of resources, where each has a limited number of resources with a different price. For this, it is very much required to allocate the available resources efficiently among the requested users, So that there should be the minimization of the cost of resources and maximization of profit from the resources. Therefore, resource allocation is one of the most challenging issues in cloud computing and fog computing. Presently, in cloud computing VMs are allocated with a fixed price mechanism which is not minimizing the cost. In this paper, to solve such problems, we purposed a hybrid combinatorial auction-based approach for allocation of VMs. Here, we calculate the cost with maximum resource utilization which is beneficiary to users on demand.

Keywords Auction theory · Virtual machine (VM) · Combinatorial auction mechanism · Cloud computing · Resource allocation

---

S. Parida (✉) · B. Pati · C. R. Panigrahi
Department of Computer Science, Rama Devi Women's University, Bhubaneswar, India
e-mail: sasmitamohanty5@gmail.com

B. Pati
e-mail: patibibudhendu@gmail.com

C. R. Panigrahi
e-mail: panigrahichhabi@gmail.com

B. Pati · C. R. Panigrahi
Rama Devi Women's University, Bhubaneswar, India

S. C. Nayak
Department of Information Technology, I-Nurture Education Solutions Pvt. Ltd., Bengaluru, India
e-mail: suvendu2006@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_29

1 Introduction

Cloud computing is merged technology in between traditional computing, distributed computing and virtualized computing, which helps to satisfy business strategies by using pay per use model [1]. Cloud also provides a flexible elastic approach to acquire the resources with optimized and cost-effectiveness. The cloud resource pool provides resources such as CPU, storage, platform, and other services to the clients. Out of these VM is a vital resource in cloud computing [2, 3]. Moreover, resource scheduling and allocation not only deal with on-demand computational resources but, also how these resources should be available to users anytime anywhere with short-term basis payment [4, 5]. Nowadays basically in the market, both commercial (e.g. Amazon EC2 and Microsoft Azure) and different open source clouds are available [6].

Mostly the cloud providers provide computational resources via VMs and charge price according to the size of VM instances as well as time period of demand by the customers [7]. Customers have to pay in ways: long-term contract and low term contract according to pay as you go [8]. Here the user has to pay on the basis of a fixed price model. Due course, there arises a problem of decreased revenue at cloud providers. A solution to this, auction-based resource allocation mechanism came into the market [9]. It provides economical efficiency along with efficient resource utilization in a cloud environment. There are various kinds of auction mechanisms where cloud providers offer and users acquire the instances at the proposed price. Here, the bidders specify their demand list along with the maximum price they agree to pay. The cloud provider refreshes periodically and auctions among the current price of active bidders. Then every user has to pay the same price per instance as per the auction. A user having allocation may be disconnected later if the auction decided price raised beyond his bid [10]. Here the issue is dynamically changing of pricing of bidders. To address such types of resource allocation problems in clouds is solved by combinatorial auctions. It solves problems like VM pricing, allocation, and provisioning.

In this paper, we presented a systematic overview of auction-based VM allocation mechanisms in cloud computing. This paper is organized as follows: Sect. 2 discusses a framework of auction-based VM allocation and focuses on VM allocation problem in cloud environment. Section 3 we present the mechanisms to solve the resource allocation problem. In Sect. 4, we explain the results. In Sect. 5, we conclude the paper and represent other possible directions for future research in this problem domain.

2 Related Work

In this section, we have represented different type of auction algorithms related to cloud computing. An auction mechanism was introduced by Sutherland in 1968 to allocate the computing capacity of a computer [11]. The application of the auction algorithm is not limited in the field of computers. Bertsekas [12] discussed the auction algorithm to solve linear transportation problems and claimed that the auction algorithm is very effective for certain types of transportation problems. Bogdanowicz and Coleman [13] proposed and studied the auction algorithm to optimize the performance of weapons/targets pairing. Zavlanos et al. [14] proposed a distributed auction mechanism to solve the assignment problem in a global information system. Feng et al. [15] proposed double auction for spectrum allocation in wireless communication. Many researchers considered resource allocation as an assignment problem and solved it by auction algorithm. Schnizler et al. [16] proposed a combinatorial auction of double-sided to allocate resources in the grid. Li et al. [17] proposed a second-price sealed-bid auction mechanism for reasonable profit and efficient resource allocation in cloud computing.

Cramton et al. [18] investigated, studied and presented combinatorial double auction in detail. Zaman and Grosu [19] proposed combinatorial auctions for VM allocation in cloud computing. Prasad et al. [20] discussed a combinatorial auction for multiple resource procurement in hybrid cloud computing. Samimi et al. [21] proposed a new market-based CDARA (Combinatorial Double Auction Resource Allocation) algorithm for cloud computing environments. A multi-attribute combinatorial auction is discussed by Baranwal and Vidyarthi [22].

3 Proposed Auction Mechanism for VM Allocation

3.1 Traditional VM Allocation Policy

In the cloud, the cloud providers can use efficient resource allocation techniques to allocate VM instances with maximum revenue. On the other hand, users demand resources with minimum pay along with enhanced QoS. Thus resource Basically VM allocation is crucial and vital for the cloud environment. However, allocation of the limited number of VM to users to obtain maximum utilization is complex max-min optimization problem [22]. The resources are provided in form of VMs as per on-demand from a pool of available resources. The broker plays an important role to allocate on-demand resources to different users from physical machines. Different allocation policies and mechanisms: QOS based, pricing, deadline based, VM based, etc. are implemented in it rather than market-based.

VM allocation is nothing the economical sharing of VMs to obtain maximum revenue to a cloud provider as it provides a different configuration of VMs [23]. This allocation mechanism should enhance the performance of the cloud with limited resources and the users pay according to the time of use. There must be a concrete objective to allocate VMS with a reduced time limit and maximum utilization [24–26]. Assume that m different types of virtual machines $\mathrm{VM}_1, \mathrm{VM}_2, \ldots, \mathrm{VM}_m$ provided by a cloud provider. A weight vector $W = (w_1, w_2, \ldots, w_m)$ represents the relative computing power of the VMs, where $w_1 \leq \ldots w_m$ and $w_1 = 1$. Means, the capacity of VMs in a number of processors in GHz, the main memory in GB and storage. Assume, there is n number of users $u_1, u_2, \ldots, u_n$ requesting for VMs in the cloud. The traditional VM allocation policy focused on profit and resource utilization in such a way that with maximum profit utilization where each request's weight should not exceed the maximum resource capacity. The user requests must be compatible in nature. Let $R \langle S, F, C, \text{PU} \rangle$, Q is set of user request $Q = \{q_i \text{I} i = 1, 2, \ldots, n\}$ where s_i= start time, f_i= finish time and w_i = weight of the user request. Let assume the total profit ($P_{\max}$) and weight of the solution (W) must represent:

$$P\text{max} = \sum_{i=1}^{n} x_i P_i \text{ such that}$$

$$W = \sum_{i=1}^{n} x_i P_i \leq C$$

where, $P_i = w_i(f_i - s_i) \times \text{PU}, s_i \geq S, f_\text{i} \leq F$ and $x_i = \begin{cases} 1 \; s_i \cap f_i = \emptyset \text{ and } w_i < C \\ 0 \qquad \text{otherwise} \end{cases}$

```
VM_Allocation_algorithm(user_request,cost)

Step1: Collect the requests  from all users
Step2: (Selection of VM )
         Compute the allocation mechanism according to best fit(Freely available VM)
          and  payment
Step3: User pays to the cloud provider
Step4: User gets access to requested resources
```

Rather than profit and resource utilization, traditional allocation policy should consider various parameters like Quality of Service (QoS). It also focuses on parameters such as response time, performance, availability, reliability, security, and throughput. Moreover, maintaining QoS dynamically is a big challenge in today's business computing world. The computing parameters are not limited to QoS, it is changed due to the requirement of user and cloud service provider. So far, different allocation policies have been proposed by many researchers, meanwhile, the traditional allocation policy has been adopted with many mechanisms and techniques, rather than these mechanisms, the auction-based mechanism is one the preferred mechanism among the user and the service provider. The different types of auction mechanisms are briefly discussed in the next section.

3.2 Auction-Based VM Allocation Policy

Auctions can be classified in different ways depending on the knowledge of bidding information either "open-cry or sealed-bid". Contrarily, auctions may be forward or reverse. In the case of forwarding, auction users bid for buying whereas in reverse auction providers bid for selling resources. An auction-based VM allocation framework in the cloud is shown in Fig. 1. Users submit their needs for VM along with bids and the corresponding winner will be allocated VM by considering a certain auction policy. Let the user set $U = \{u_1, u_2, \ldots, u_m\}$ and the winner set $W = \{w_1, w_2, \ldots, w_p\}$ where, $W \subseteq U$. The price user u_j pays to the cloud provider is represented by c_j, $j = 1, 2, \ldots, m$. The allocation problem can be presented as:

$$\sum_{j:u_{j\in W}} r_i^j \leq k_i = 1, 2, \ldots, m \tag{1}$$

$$0 \leq p_j \leq v_j \text{ if } u_j \in W \tag{2}$$

$$p_j = 0 \text{ if } u_j \notin W \tag{3}$$

The auction-based VM allocation mechanism is fixed price mechanism means, the VM cost is fixed by the cloud provider before the auction starts. The basic problem is the additional offered benefits are not available to the user. Similarly, during peak hour the cost of the VM is static as it was in non-peak hour. So the benefit should not available to the cloud service provider. The cost of VM in peak hours should be dynamic depending upon the QoS. Moreover, there is no specific objective function either in case of simple auction or fixed price VM allocation mechanism. In most of the auction algorithms consider the objective function as the goal to maximize cloud provider's revenue.

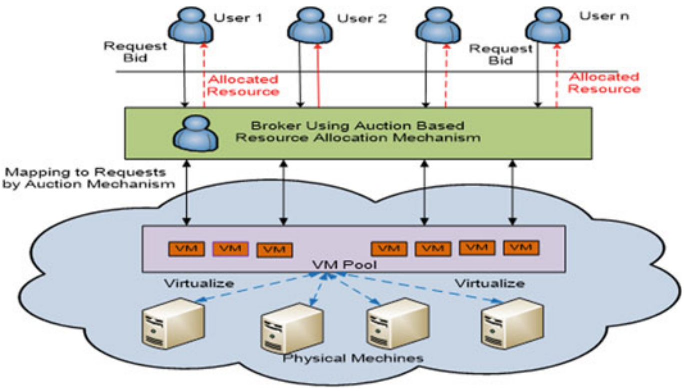


Fig. 1 VM allocation auction model

```
VM_Allocation_Auction_algorithm(user_request,user_bid,cost)

Step1: Collect the requests and bids from all users
Step2: If(submission==finish)
            Goto Step3.
        Else
            Repeat step1 and 2.
Step 3:  Sort all user bids.
Step 4: Find the winner based upon auction policy and match the
          VM configurations to potential winners.
Step 5: If (auction termination condition satisfied)
            goto Step 6.
          Else
            Repeat Step 4 & 5.
Step 6: The winner user pays for VM instances.
Step7: Bind the VM with winning user.
```

3.3 Combinatorial Auction VM Allocation Policy

The bidders are allowed to submit their bids for auctioning along with the combinations of different items, are usually called combinatorial auctions. In a combinatorial auction, "multiple, heterogeneous goods are auctioned simultaneously and bidders may bid for arbitrary combinations of goods". An auction mechanism can be single-item or multi-item based. Conversely, in multi-item auction users bid for bundles of items instead of single. So, the bidders can express their requirements in a more meaningful way. Hence combinatorial auctions are most applicable to solve VM allocation and pricing problems in a cloud environment. There are two approaches used for Combinatorial Auction, CA-LP ("Combinatorial Auction—Linear Programming") and CA-GREEDY ("Combinatorial Auction—Greedy").

3.4 CA-LP VM Allocation Policy

Archer et al. [27] proposed CA-LP to solve the VM allocation problem. The problem can be expressed as:

$$\max \sum_{i=1}^{n} x_i v_i \tag{4}$$

$$\sum_{i=1}^{n} x_i r_j^i < k_j', \quad j = 1, \ldots, m \tag{5}$$

where $0 \leq x_i \leq 1,\ i = 1, \ldots, n$.

Let k'_j the total number of available VM_j instances used in the constraint in Eq. 5. The constraint k_i reduces the allocation probability for VMs and it randomly rounds up in CA-LP. It is presented as:

$k'_j \leftarrow (1 - \varepsilon)k_j$ where, $0 < \varepsilon < 1,\ \ j = 1, \ldots, n$

The user u_i is selected as a winner in the auction with probability x_i, if u_i does not violate any constraint in Eq. 5. The payment p_i for each u_i is computed using the following steps:

Payment ( )
Step 1: for each user $\text{u}_\text{i} \in \text{W}$
do
Step 2: Compute binary search for v'_i in the range $[0, \text{v}_\text{i}]$
(a) Set valuation of ui as v'_i in Eq.4
(b) Solve the LP, let x'_i is the fractional allocation computed for ui
(c) Until v'_i is found such that, setting $\text{u}_\text{i} < \text{v}'_\text{i}$
Compute $\text{x}'_\text{i} < \text{y}_\text{i}$ and find the greater value than v'_i
Evaluate $\text{x}'_\text{i} > \text{y}_\text{i}$ where, the critical value is v'_i
Step 3: $\text{p}_\text{i} \leftarrow \text{v}'_\text{i}$
Step 4: end for
Step 5: for each user $\text{u}_\text{i} \notin \text{W}$
do
Step 6: $\text{p}_\text{i} \leftarrow 0$
Step 7: end for

3.5 Modified CA-GREEDY VM Allocation Policy

Lehmann et al. [28] proposed "CA-Greedy mechanism based on $\sqrt{M}$-approximation with single-minded bidders where the total number of items that need to be allocated is M". It can be computed as $M = \sum_{i=1}^{n} k_i w_i$. The size s_j of the bundle in bid B_j requested by user u_j is as:

$$s_j = \sum_{i=1}^{n} w_i r_i^j \tag{6}$$

In CA-GREEDY winners are selected by "first-ranking", the ranks are decreasing order of their "bid density". We formulate the bid density as: $v_i/\sqrt{s_i}$. The computing mechanism for winner and payment is below.

Let assume B_j is the collection of bids for each user, $B_i = \left\{r_1^i, \ldots, r_n^i, v_i\right\}$ where, $i = 1, \ldots, n$. The winner is selected based upon the bid density. In existing work bid density does not consider the number of VMs; however, the VMs are also associated with the start time and finish time. The auction cost may vary with the time factor for which we considered the start time and end time to evaluate the bid density. For evaluating bid density, we computed resource factor (r_f) for each bidder.

$$r_{f_i} = \frac{\sum_{i=1}^{n} \mathrm{VM}_i(\mathrm{ST}_i - \mathrm{FT}_i) \big/ D_i}{n} \tag{7}$$

where

VM_i is VM instances of user U_i,
ST_i start time of VM_i,
FT_i finish time of VM_i,
D_i duration of VM_i, and
n number of VM instances participated by user U_i in the auction

Similarly, we also expressed weight in terms of a total number of VM instances, the number of bidders and the auction time. The weight is a constant factor during the auction. In existing work, the weight is constant for all bidders, whereas in the proposed work it varies from bidder to bidder to achieve better resource utilization. For each auction, the weight is computed for each bidder based upon the on-demand VM instances. The weight is expressed as:

$$w_i = \frac{\sum_{i=1}^{n} (D_i \times \mathrm{VM}_i)}{M} \tag{8}$$

where M is the number of users participated in the auction.

So the modified bid density S_j can be computed using Eqs. 7 and 8 as:

$$S_i = \sum_{i=1}^{n} w_i r_{f_i} \tag{9}$$

The proposed work computes two mechanisms: Winner () and Payment (). The winner () computes the winner in the auction based on the modified bid density using Eq. 9. The resource factor is computed using Eq. 7 and the dynamic weight is evaluated using Eq. 8. The impact of r_{f_i} provides a better winner selection with resource utilization, whereas, the modified weight provided better revenue and payment for the user. In this work, we formulate the problem of VM instance allocation in cloud computing as a combinatorial auction problem. The main goal is to provide VM instance allocation efficiently among the bidders demanding a variety of VM instances. In this work, we discussed "different auction-based VM allocation mechanism and extended the combinatorial auction mechanism with single-minded bidders where we address the presence of multiple identical copies of different items".

Winner ( )
Step 1: set $W \leftarrow \emptyset$
Step 2: for i=1, . . . , n do
Step 3: compute s_i using Eq.10
Step 4: end for
Step 5: Find the bid density and sort it
Step 6: for i=1, . . . , n do
Step 7: for all j=1, . . . , m
Step 8: $r_j^i + \sum_{u_{j'} \in W} r_j^{i'} < k_j$ then
Step 9: $W \leftarrow W \cup v_i$
Step 10: end for
Step 11: end for

Payment ( )
Step 1: for all $u_i \in W$ do
Step 2: $W_i' \leftarrow \{u_a : u_a \notin W => u_a \in W\}$
(a is the minimum index in W_i')
Step 3: if $W_i' \leftarrow \left({}^{v_a}/_{\sqrt{s_a}}\right)\sqrt{s_i}$
else
Step 4: $p_i \leftarrow 0$
Step 5: end if else
Step 6: end for
Step 7: for all $u_i \in W$ do
Step 8: $p_i \leftarrow 0$
Step 9: end for

Role of Offers for Revenue

In this proposed work, we also considered QoS, for which the VM cost is changed dynamically as the VM instances increased for users. The offer provides additional benefits to the user and revenue for service providers. In the proposed mechanism, we consider the "total VM cost and the target revenue". The target revenue is computed by considering the average cost of maximum VM cost and minimum VM cost. The target revenue Rev_T is expressed as:

$$\mathrm{Rev}_T = \pi r^2$$

4 Simulation Results and Discussion

The proposed work is simulated in Cloud Analysts with different data center and user configuration including different cost parameters. The Cloud Analyst is a JAVA-based programming tool to analyze the research work in the field of cloud computing. It provides a detailed analysis of both users and service providers. We simulated and compared the performance results in terms of response time. The "response time" is evaluated in the "average case", "worst case", and "best case". Moreover, we computed the price and the payment as discussed in CA-LP and proposed CA-Greedy.

Figure 2 shows the comparison of response time, it is quite less in CA-GREEDY than CA-LP. Similarly, we computed the total VM cost ($\mathrm{VM}_{\mathrm{cost}}$) in the experiments and observed initially the $\mathrm{VM}_{\mathrm{cost}}$ is more in CL-GREEDY than CA-LP. It is occurring based upon the user's request. Whenever the on-demand VMs have increased the cost of VMs goes decreases whereas in CA-LP the cost of VMs is constant as shown in Fig. 3. The comparative performance is evaluated average overall response time, the response time of different data centers to process user's requests in ms, the total

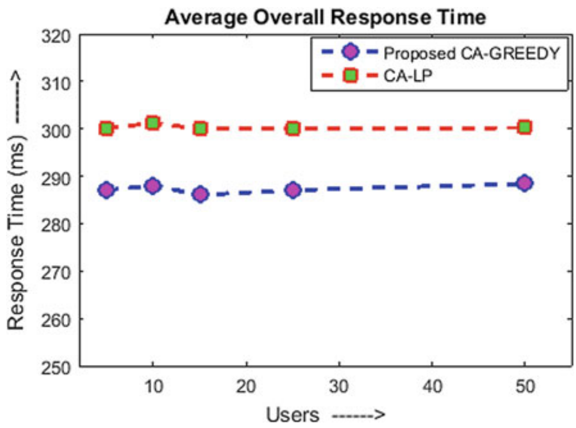


Fig. 2 Average overall response time for 50 users in 5 data centers

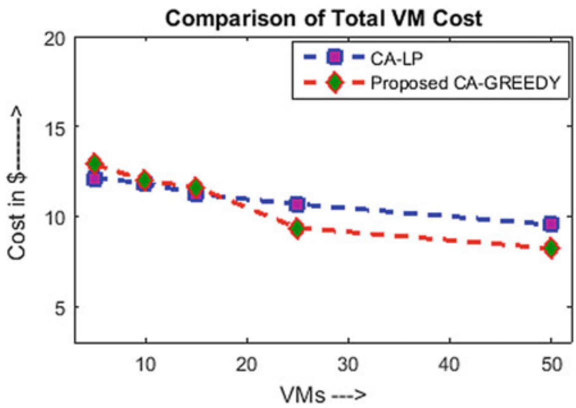


Fig. 3 Total VM cost in data centers

revenue collected in an auction for data centers and the total cost for the VM in \$. The total VM cost is computed as:

$$\text{Total VM cost } (VM_{total}) = \text{VM cost } (VM_{cost}) + \text{Data Transfer cost } (DT_{cost})$$

The total revenue is shown in Fig. 4 which are close in both the mechanism, but the GA-GREEDY is dynamic in nature for which the cost of VMs varies. So the mechanism is more flexible than CA-LP to adopt additional offers to the users during the auction. Finally, we evaluated the average time required to process the user requests in ms and shown in Figs. 5 and 6. Figures 5 and 6 show the five data center processing time for a different number of users. In both the mechanism the average processing time is close to each other. This signifies both the mechanisms have nearly equal processing time.

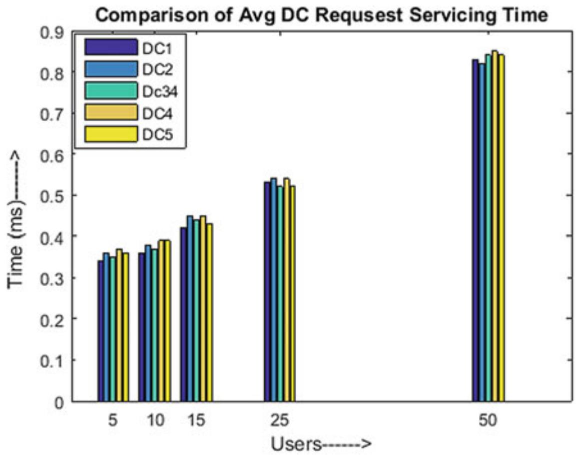


Fig. 4 Average data center request servicing time in CA-LP

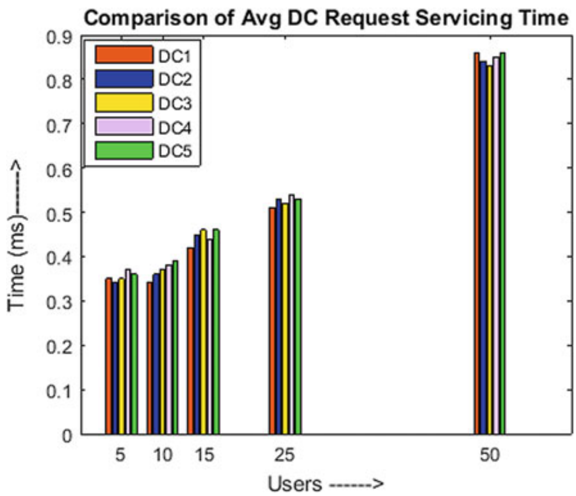


Fig. 5 Average data center request servicing time in proposed CA-GREEDY

5 Conclusion

In this work, we studied different types of auction mechanism used for VMs allocation in cloud computing. We compared the performance of CA-LP and CA-GREEDY auction mechanism in terms of "average response time", "request processing time", "total VM cost", and "revenue" collected by cloud service providers. The work adopts dynamic mechanism in an auction using CA-GREEDY mechanism. The mechanism provides flexibility to change the VM cost dynamically during the auction that is depending upon the number of VMs requested by the user. Similarly, the cloud service provider can provide additional offers to attract users and a better QoS can

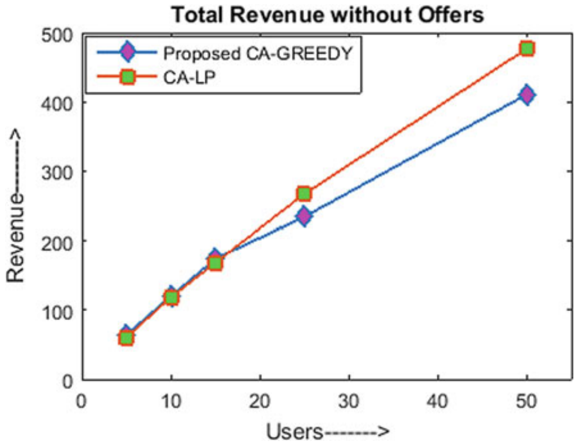


Fig. 6 Comparison of total revenue without offers

be achieved. Moreover, the work does not consider the deadline in the auction and the conflicts among the similar users in the auction that is considered as the future scope of the proposed work.

References

1. Rodriguez, M.A., Buyya, R.: Deadline based resource provisioning and scheduling algorithm for scientific workflows on clouds. IEEE Trans. Cloud Comput. **2**(2), 222–235 (2014)
2. Mustafa, S., Nazir, B., Hayat, A., Khan, R., Madani, S.A.: Resource management in cloud computing: taxonomy, prospects, and challenges. Comput. Electr. Eng. **47**, 186–203 (2015)
3. Parida, S., Nayak, S.C., Tripathy, C.: Truthful resource allocation detection mechanism for cloud computing. In: ACM International Conference Proceeding Series (2015)
4. Nayak, S.C., Tripathy, C.: Deadline sensitive lease scheduling in cloud computing environment using AHP. J. King Saud Univ.-Comput. Inf. Sci. (2016)
5. Parida, S., Nayak, S.C.: Study of deadline sensitive resource allocation scheduling policy in cloud computing. Int. J. Comput. Sci. Mob. Comput. **3**(12), 521–528 (2014)
6. Zaman, S.: Combinatorial auction-based virtual machine provisioning and allocation in clouds. Wayne State Univ. Diss. Pap. **720** (2013)
7. Nayak, S.C., Tripathy, C.: Deadline based task scheduling using multi-criteria decision-making in cloud environment. Ain Shams Eng. J. (2018)
8. Einav, L., Farronato, C., Levin, J.: Auctions versus posted prices in online markets. J. Polit. Econ. Forthcom. (2016)
9. Huang, J., Han, Z., Chiang, M., Poor, H.V.: Auction-based resource allocation for cooperative communications. IEEE J. Sel. Areas Commun. **26**(7), 1226–1237 (2008)
10. Fraser, M.D., Schaefer, M.E.: Auction allocation of computing resources. Commun. ACM **38**(6), 88–102
11. Report, S.: A futures market iu. Commun. ACM **11**(6), 449–451 (1968)
12. Bertsekas, D.P.: The auction algorithm for the transportation problem. Ann. Oper. Res. **20**, 67–96 (1989)
13. Bogdanowicz, Z., Coleman, N.: Auction algorithm for weapons/targets pairing application. In: Conf. Numer. Math. FIRE Control BOMBING Syst. pp. 2–4 (1992)

14. Zavlanos, M.M., Spesivtsev, L., Pappas, G.J.: A distributed auction algorithm for the assignment problem. In: Proceedings of the IEEE Conference on Decision and Control, pp. 1212–1217 (2018)
15. Feng, X., Chen, Y., Zhang, J., Zhang, Q., Li, B.: TAHES : a truthful double auction mechanism for heterogeneous spectrums. IEEE Trans. Wirel. Commun. **11**(11), 4038–4047 (2012)
16. Schnizler, B., Neumann, D., Veit, D., Weinhardt, C.: Trading grid services—a multi-attribute combinatorial approach. Eur. J. Oper. Res. **187**(3), 943–961 (2008)
17. Li, X., Ding, R., Liu, X., Liu, X., Zhu, E., Zhong, Y.: A dynamic pricing reverse auction-based resource allocation mechanism in cloud workflow systems. Sci. Program. **2016**, 17–28 (2016) Article ID 7609460
18. Cramton, P., Shoham, Y., Steinberg, R.: Combinatorial auctions. In: Cramton, P., Shoham, Y., Steinberg, R. (eds.) pp. 29–55. MIT Press (2016)
19. Zaman, S., Grosu, D.: Combinatorial auction-based allocation of virtual machine instances in clouds. J. Parallel Distrib. Comput. **73**(4), 495–508 (2013)
20. Prasad, V.G., Rao, S.A., Prasad, S.: A combinatorial auction mechanism for multiple resource procurement in cloud computing. Int. Conf. Intell. Syst. Des. Appl. ISDA **7161**(c), 337–344 (2012)
21. Samimi, P., Teimouri, Y., Mukhtar, M.: A combinatorial double auction resource allocation model in cloud computing. Inf. Sci. **357**, 201–216 (2016)
22. Baranwal, G., Vidyarthi, D.P.: A fair multi-attribute combinatorial double auction model for resource allocation in cloud computing. J. Syst. Softw. **108**, 60–76 (2015)
23. Kumar, D., Baranwal, G., Raza, Z., Vidyarthi, D.P.: A systematic study of double auction mechanisms in cloud computing. J. Syst. Softw. **125**, 234–255 (2017)
24. Shirvastava, S.: Best fit based VM allocation for cloud resource allocation. **158**(9), 25–27 (2017)
25. Chandan, S., Parida, S., Tripathy, C., Kumar, P.: An enhanced deadline constraint based task scheduling mechanism for cloud environment. J. King Saud Univ.-Comput. Inf. Sci. (2018)
26. Nayak S.C.: Multicriteria decision - making techniques for avoiding similar task scheduling conflict in cloud computing. Int. J. Commun. Syst. 1–31 (2019)
27. Archer, A., Papadimitriou, C., Talwar, K., Tardos, É.: An approximate truthful mechanism for combinatorial auctions with single parameter agents. Internet Math. **1**(2), 129–150 (2003)
28. Lehmann, D., O'Callaghan, L.I., Shoham, Y.: Truth revelation in approximately efficient combinatorial auctions **49**(5) (2002)

IoT-Based Cardiac Arrest Prediction Through Heart Variability Analysis

Santosh Kumar, Usha Manasi Mohapatra, Debabrata Singh and Dilip Kumar Choubey

Abstract Current machine learning methods for sudden cardiac arrest have not been tested against physically active heart rates. Developments in wearable technology and advancements in non-intrusive heart rate monitors may allow for a future where people can stream their heart rate readings, with the readings automatically analyzed by robust machine learning algorithms which will alert cardiac arrest risk. This paper presents a new sudden cardiac arrest prediction technique, a random forest classifier implementation, a prospective physical activity heart rate dataset, and an Internet of things solution toward heart rate monitoring and sudden cardiac arrest warning. In this paper, five minutes advance warning is provided with 97.03% accuracy and a 0.9485 F-score for the classification of sudden cardiac arrest prediction. The result shows the efficiency of our method compared to other existing methods.

Keywords Heart rate variability (HRV) · Random forest · Heart rate monitoring band (HRM) · IoT

S. Kumar (✉) · U. M. Mohapatra · D. Singh
Department of CSIT, Institute of Technical Education and Research S‘O’A Deemed to be University, Bhubaneswar, Odisha, India
e-mail: santoshkumar@soa.ac.in

U. M. Mohapatra
e-mail: ushamohapatra@soa.ac.in

D. Singh
e-mail: debabratasingh@soa.ac.in

D. K. Choubey
Department of Computer Science & Engineering National Institute of Technology, Patna, Bihar, India
e-mail: dilipchoubey_1988@yahoo.in

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_30

1 Introduction

Researchers believe that cardiac arrest may be predicted in advance [1]. Specifically, through heart rate variability (HRV) analysis algorithms, machine learning, the Internet of things (IoT), and big data, we may be able to monitor at-risk individuals and give them advance warning (1–4 h) to get to a hospital. There is a divide between papers focused on electrocardiogram (ECG) features and HRV features. Murukesan et al. [2] achieve 96.36% accuracy through the use of a support vector machine (SVM) and HRV features. This paper uses a five minute advance warning and two minutes of sample data. This is the baseline that this paper uses. HRV is the preferred feature set due to allow for a broader spectrum of potential commercial wearable devices. A significant benefit of such advance warning is corroborating evidence of the need to go to the hospital rather than ignoring symptoms. Developments in wearable technology and advancements in non-intrusive heart rate monitors may allow for a future where people can stream their heart rate readings, with the readings automatically analyzed by robust machine learning algorithms which will alert cardiac arrest risk. This field is new and there is room for progress through additional studies and the development of HRV analysis algorithms. This is supported by the increasing number of medical records stored electronically.

Current machine learning methods for sudden cardiac arrest have not been tested against physically active heart rates. All tested samples come from in-hospital data where the individuals are at a resting state. This paper hopes for better cardiac arrest prediction for the wearable world through HRV analysis and develop a solution that can be reliable with high accuracy and F-score.

1.1 Heart Rate Variability

Heart rate is a feature which can be derived from the ECG signal [3]. This is done by comparing the placement between the RR intervals. Each peak on the R wave represents a node in the heart rate array. The time between each R wave peak is the instantaneous heart rate. It is potentially inconvenient to spend the whole day with 12 electrodes attached to the body, and there is technology to track the heart rate without use of an ECG signal. This creates a preference to have a system which relies on more convenient sensors (wearable devices) in comparison to 12 ECG electrodes, despite the electrodes having the potential to provide more information.

1.2 Arrhythmias

A regular ECG signal is called a regular sinus rhythm. There are specific classes of ECG signals representative of what is going on inside the heart. There are a couple key vocabulary words that are used to identify these different classes such

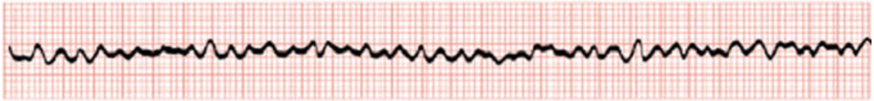


Fig. 1 ECG signal for ventricular tachyarrhythmia

as arrhythmia, tachycardia, bradycardia, atrium, and ventricles [3]. Based on these terms, we can determine an arrhythmia such as ventricular tachyarrhythmia (VT) is a rapid heartbeat that starts in the bottom chambers of the heart [3]. VT is one of the main arrhythmias associated with sudden cardiac arrest. These arrhythmias have certain classifying features. For example, VT is easily identified by the fast oscillatory waves, which represent the rapid twitching of the heart. This is shown in Fig. 1.

1.3 Cardiac Arrest

Cardiac arrest occurs when the beating of the heart and all electrical activity stops [3]. This means that blood stops pumping to the body, and it is especially important because brain damage can occur within ten minutes from blood loss to the brain. This is different from a heart attack, which is a physical failure in comparison to an electrical failure. Heart attacks are caused by the blockage of blood flow to the heart. Sudden cardiac arrest (SCA) is cardiac arrest that occurs unexpectedly and can result in death, called sudden cardiac death (SCD). There are three main arrhythmia classes that correlate with cardiac arrest: ventricular tachyarrhythmia (VT), ventricular asystole (VA), and pulse-less electrical activity (PEA). VT is described in the previous section and is the most common predecessor to *SCA* [3]. VA is represented by a flat line. This is shown in Fig. 2.

1.4 Machine Learning

In the history of *SCA* prediction, various machine learning algorithms are in use. This section highlights the four machine learning classification algorithms explored in this paper. In this case, datasets are broken up into a healthy (non-SCA) class and a *SCA*

Fig. 2 ECG signal for ventricular asystole

warning class (five minutes until cardiac arrest) to create a two-class classification problem. Kotsiantis et al. [4] provide a strong review of many classification algorithms. Much of the early work in ECG classification is done with neural networks [5].

Classification: Classification is a machine learning problem where a data sample can belong to one of two or more classes [4]. While classification algorithms can deal with more than two classes, this paper focuses on a binary classification problem where the two possible classes are either normal sinus rhythms or the onset of sudden cardiac arrest. Machine learning classifiers use the feature vectors derived from the data samples to learn how these features help identify what class future data samples belong to.

Support Vector Machine: A *SVM* is a machine learning algorithm for classification proposed by Cortes and Vapnik in 1995 [6]. The main idea of *SVM*s is to learn a nonlinear function by a combination of linear mappings in high-dimensional feature space. The desired function would be able to map all training data within a certain margin of error. Support vectors are especially good in high-dimensionality spaces because *SVM*s do not depend upon the dimensionality of the input space.

Decision Tree: Decision trees build a ruleset to classify input data [7]. Each node in the tree that is built represents a decision. As the input data traverses through this tree, it will eventually arrive to a leaf node which dictates the class of the input data. The trees built have logarithmic run times due to the structure. Unfortunately, the ruleset is prone to overfitting for which there are a number of techniques to avoid.

Naive Bayes: Naive Bayes is a classification algorithm based on the Bayes theorem. This classifier uses the theorem to derive the probability that a given feature vector is associated with a specific class. The algorithm naively assumes that there is independence between every pair of features [4, 7]. This assumption creates a weakness in the algorithm, because there will almost never been an independence between every pair of features given enough features. Regardless, the classification algorithm is proven to be strong in smaller training sets.

Random Forest: This paper utilizes the random forest algorithm demonstrated by Breiman [8]. The random forests used as classification algorithms are common in machine learning applications, because of providing good results, the ease of use, and the scalability. In many implementations, the single parameter which requires tuning is the number of trees used to build the forest, which usually denotes the trade-off between computational cost and the performance. The depth of individual trees is of less importance, as with big forests the variance of the model imposed by the deep trees is reduced, therefore, the trees do not require pruning.

2 Proposed Method

With the increased interest in wearable technology, there will be a market for *HRV* monitoring [9]. With the large amount of cardiac arrests that occur outside of the hospital, constant *HRV* monitoring may be able to prevent a large amount of deaths.

This constant *HRV* monitoring will have less data than typical in-hospital data and will have to be mostly reliant on heart rate data as well as some patient-level data. There has been success in predicting cardiac arrest with in-hospital data, but the potential for success may exist with limited data with physical activity as well. Success, in this case, is defined as a system that is more accurate, in terms of accuracy and F-score.

2.1 Proposed System

Figure 3 details an outline of the proposed system, similar to the system that builds the physical activity dataset. The wearable device collects heart rate information using its heart rate monitor. This example uses the specific heart rate monitoring band (HRM), but the system may be built for any wearable device that has a heart rate monitor or electrodes for *ECG* recording. That data is sent to a smartphone via Bluetooth. An application on the phone holds this data in two-minute intervals. Data may be sent to the phone on each beat or all at once. At the end of each two minute interval, the sample is sent via *HTTP* request to a server hosting the Python machine learning code. Also, on this server is the dataset of labeled heart rate samples, which has the opportunity to grow with each new input sample. The running code will classify the sample and send back a response. If the sample is classified as the onset of sudden cardiac arrest, then the phone and wearable device's notification system are used to notify the user. Otherwise, the phone sends no warning notification.

The *HRM* is a commercial product available to the public. It features a heart rate monitor, skin temperature sensor, *GPS* tracker, and other sensors. It is visible in Fig. 4. The band communicates via Bluetooth to a synchronized smartphone.

2.2 Datasets

The datasets are a prospective physical activity dataset collected via the *HRM*. Prior to this work, there is only one dataset that provides the heart rates of individuals

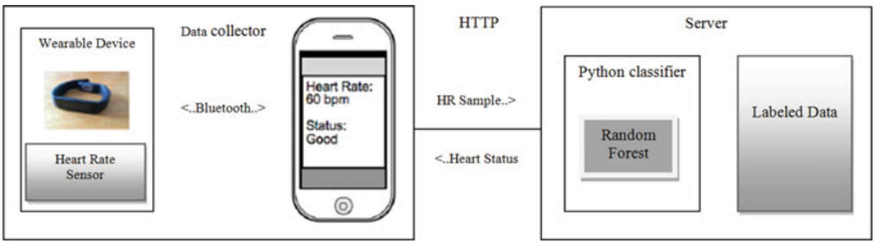


Fig. 3 Proposed system

Fig. 4 Heart rate monitor band

performing physical activities called ‘PAMAP2’ dataset [10]. The dataset is sampled using an *HR* monitor with a frequency at 9 Hz collected in 2012. The dataset is comprised of the heart rates of nine individuals performing 18 different physical activities. The dataset also holds other biometric values such as body mass index, acceleration data, temperature, gyroscopic data, and orientation.

This paper introduces the use of a prospective study that may be representative of future heart rates of people using a commercially available wearable device. The prospective physical activity dataset in this paper is collected from a *HRM*. This dataset features the two-minute samples of the heart rates of five individuals performing ten unique activities. A table is provided with the counts as well in Table 1. There are a total of 105 samples in this dataset.

Table 1 Physical heart rate activity dataset distribution

Activity	Sample count
Sitting	4
Walking	10
Running	18
Computer work	13
Stretching	3
Car driving	9
Video games	22
Lying	6
Eating	8
Post workout	12
Total	105

2.3 Features

The independent variables for this paper are based on common features that act as indicators for cardiac arrest and arrhythmia classes [11, 12]. This includes common *HRV* features such as mean *HR*, minimum *HR*, maximum *HR*, standard deviation of the *HR*s, standard deviation of the beats per minute, median of the *HR*, root mean square of the standard deviation of the *HR*, and the number of outliers. These features are shown in Table 2.

2.4 Labels

The dependent variable in this experiment is the class of arrhythmia. This experiment tests using binary classification. The person may have a heart rate that is normal or a heart rate that is exhibiting an onset of sudden cardiac arrest. Being able to classify the rhythm is important so that individuals may take certain precautions or be able to detect and treat specific symptoms.

2.5 Accuracy

All data comes labeled regarding the class of the heart rates as well as the time until cardiac arrest [3, 11]. The output of the machine learning algorithms is directly compared to these classes and is judged as wrong or right on a binary scale. Accuracy is used to determine the proficiency of the algorithm. The accuracy of these classifiers is compared against baselines such as the one provided by Murukesan et al. [2]. In all cases, at least five runs were performed to collect an average accuracy.

Precision, Recall, and *F*-score Oftentimes, it has been determined that unbalanced datasets such as this one cannot have reliable classification results with accuracy

Table 2 Features extracted by the system

Number	Fatures
1	Mean HR
2	Minimum instantaneous HR
3	Maximum instantaneous HR
4	Standard deviation of the HRs
5	Standard deviation of the beats per minute
6	Median of the HRs
7	Root mean square of the standard deviation of the HRs
8	The number of outliers (greater than 50 ms difference)

alone [7]. In this case, a classifier has the potential to nearly always guess that the sample is a normal sinus rhythm, because the number of normal sinus rhythms is twice that of the sudden cardiac arrest onsets. Because of this, the random forest algorithm is also evaluated in terms of precision, recall, and F-score.

$$\text{Precision} = \frac{\text{No. of true positives}}{\text{No. of true positives} + \text{No. of false positives}} \tag{1}$$

$$\text{Recall} = \frac{\text{No. of true positives}}{\text{No. of true positives} + \text{No. of false negatives}} \tag{2}$$

$$F\text{-Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precesion} + \text{Recall}} \tag{3}$$

2.6 Experimental Protocol

The empirical study is broken down into multiple parts. First, data is acquired from the *HRM*. Second, multiple machine learning classifiers are randomly trained upon a 70–30 % of the total dataset. Third, the rest of the dataset is used to test against these classifiers and provide an accuracy to demonstrate the proficiency of the algorithm. The first step is the collection and normalization of all data. This creates 105 samples that are provided from the physical database collected via *HRM*. The machine learning algorithms that are compared include a support vector machine with a linear kernel, decision trees, naive Bayes, and random forest. Comparisons are done in terms of accuracy.

3 Results and Discussion

3.1 Accuracy of Random Forest Versus SVM

The first tables of this section compare the random forest classifier accuracy to the SVM classifier accuracy. Table 3 compare the random forest classifier accuracy to the SVM classifier accuracy based on the data collected by *HRM* band. The result shows the efficiency of random forest classifier in terms of accuracy over SVM.

The rest of the section is devoted to the individual runs of each classifier at 70–30 training-testing percentages. Table 4 shows the accuracies with the physical activity dataset. This table illustrates the high accuracy of random forest over the three other algorithms.

Table 3 SVM versus random forest classification results

Run 70–30%	SVM accuracy	Random forest accuracy
0	0.9275	0.9824
1	0.9384	0.9602
2	0.9513	0.9527
3	0.9625	0.9824
4	0.9554	0.9756
Average	0.9470	0.9703

Table 4 Classification results with physical activities

70–30%	SVM	Naive Bayes	Decision tree	Random forest
0	0.8624	0.8139	0.8755	0.9521
1	0.8217	0.8356	0.8692	0.9478
2	0.8448	0.8297	0.8867	0.9502
3	0.8768	0.8215	0.8893	0.9461
4	0.8309	0.8344	0.8972	0.9521
Average	0.8473	0.8270	0.8835	0.9496

3.2 Precision, Recall, and F-Score Evaluation

Table 5 details the precision, recall, and F-score of the random forest classifier on training-testing dataset splits. These values are the average of five runs each.

The values of F-score are detailed in Table 6 and Fig. 5. Similar to the accuracies, the results for random forest are higher than any other classifier at training-testing split.

Table 5 Random forest accuracy, precision, recall, and F-score results

Algorithms	Training-testing split
Accuracy	0.9484
Precision	0.9702
Recall	0.8674
F-score	0.9485

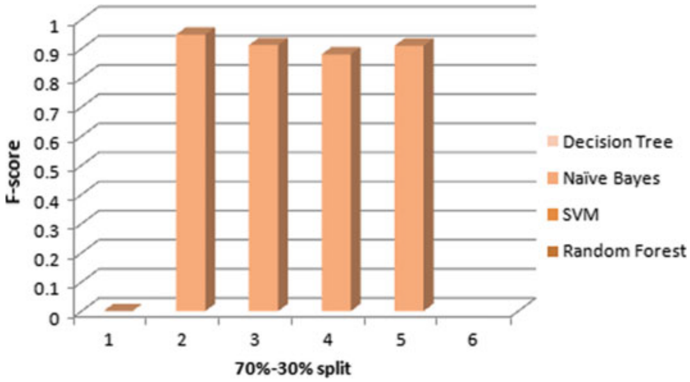


Fig. 5 F-score of each classifier

Table 6 F-score results

Algorithms	Training-testing split
Random forest	0.9485
SVM	0.9124
Naive Bayes	0.8809
Decision tree	0.9102

4 Conclusion

This paper presents a new sudden cardiac arrest prediction technique, a random forest classifier implementation for multiple weak learners, a prospective physical activity heart rate dataset, an *IoT* solution toward heart rate monitoring, and sudden cardiac arrest warning. Using a 70–30% training-testing split, the work in this paper is able to achieve 97.03% accuracy with a 0.9485 F-score for the classification of sudden cardiac arrest prediction. Comparably, Murukesan et al. [2] achieve 96.36% accuracy, but their approach uses a smaller dataset and does not include physical activity heart rates. This paper's approach uses *HRV* derived features on two-minute samples of heart rate data. This paper proves that it is possible to classify resting as well as active heart rates against the heart rates of people about to go into sudden cardiac arrest. The heart rate data from a wearable device proves that there is a future for sudden cardiac arrest prediction through wearable devices.

References

1. Ebrahimzadeh, E., Pooyan, M., Bijar, A.: A novel approach to predict sudden cardiac death (SCD) using nonlinear and time-frequency analyses from HRV signals. PLoS ONE **9**(2) (2014)

2. L. Murukesan, Murugappan, M., Iqbal, M.: Sudden cardiac death prediction using ECG signal derivative (heart rate variability): a review. In: 2013 IEEE 9th International Colloquium on Signal Processing and its Applications (CSPA). IEEE, New York (2013)
3. Goldberger, A.L., Goldberger, Z.D., Shvilkin, A.: Clinical Electrocardiography: A Simplified Approach E-Book. Elsevier Health Sciences (2017)
4. Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E.: Machine learning: a review of classification and combining techniques. Artif. Intell. Rev. **26**(3), 159–190 (2006)
5. Dokur, Z., lmez, T.: ECG beat classification by a novel hybrid neural network. Comput. Methods Progr. Biomed. **66**(2–3), 167–181 (2001)
6. Cortes, C., Vapnik, V.: Support-Vector networks. Mach. Learn. **20**(3), 273–297 (1995)
7. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res., 2825–2830 (2011 Oct 12)
8. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
9. Jin, Z. Sun, Y., Cheng, A.C.: Predicting cardiovascular disease from real-time electrocardiographic monitoring: an adaptive machine learning approach on a cell phone. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009. EMBC 2009
10. Reiss, A., Stricker, D.: Introducing a new benchmarked dataset for activity monitoring. In: 2012 16th International Symposium on Wearable Computers (ISWC). IEEE, New York (2012)
11. Lichman, M.: UCI Machine Learning Repository (2013)
12. Joo, S., Choi, K.-J., Huh, S.-J.: Prediction of ventricular tachycardia by a neural network using parameters of heart rate variability. Comput. Cardiol. IEEE, New York (2010)

Advanced Computer Networks

Application of Group Mobility Model for Ad Hoc Network

Kailash P. Dewangan, Padma Bonde and Rohit Raja

Abstract The mobility model represents the movement pattern of mobile nodes, position, velocity and acceleration with respect to time. Mobility management is a perilous issue in the wireless ad hoc network. Mobility model can be organized into two sorts: Entity Mobility Model (EMM) and Group Mobility Model (GMM). This paper presents a classification of GMM for wireless ad hoc network and their appropriate application with respect to GMM. Group motion happens repeatedly in an ad hoc network and further change in group motion parameters, the behavior of mobility model changes. Different GMM behaviors are observed and compared these models on different parameters. Further, it is reviewed to see that GMM can be easily adapted for existing applications.

Keywords RPGM · RRGM · SGM · VTBGM · Ad hoc network

1 Introduction

In recent times ad hoc networks have been fascinating raising and consciousness from the industrial and research perspective. Ad hoc networks are categorized convenient to their usage such as mobile ad hoc network, vehicular ad hoc network, and military ad hoc network so on. Mobility model is used to describe the behavior of mobile nodes with respect to some parameter such as velocity, acceleration, position and their movements. It has been observed that changing this parameter results in different behavior which emerges a reason to develop a new mobility model. Mobility model is two sorts such as Entity mobility model and Group mobility model [1]. In EMM

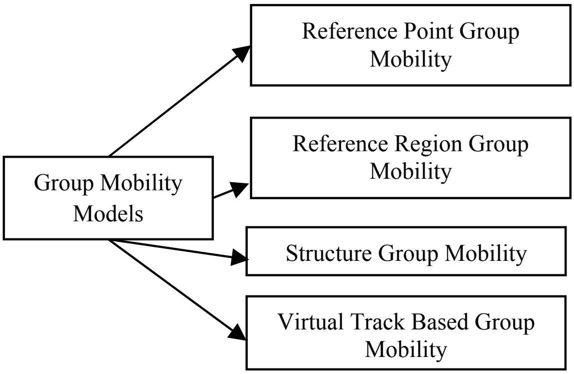
K. P. Dewangan (✉) · P. Bonde
Department of Computer Science and Engineering, SSTC, SSGI, Bhilai, Chhattisgarh, India
e-mail: kaishapt@yahoo.com

P. Bonde
e-mail: drpadmabonde@sstc.ac.in

R. Raja
Department of Computer Science and Engineering, SIET, Hyderabad, India
e-mail: drrohitraja1982@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_31

Fig. 1 Classification of group mobility models



there is no correlation between mobile nodes, every node is move independently. On the other hand in GMM mobile nodes work in a group to complete a usual task. GMM have multiple applications which are applicable in real life scenarios such as war zone, crowd control, search and rescue operation, social networking and so on.

There is an enormous and active literature associated with mobility models. In [2] mobility models are divided into four sets as random models, temporal dependence, spatial dependence and geographic restriction. In [3] different mobility models are compared with respect to certain parameters. In [4] GMM is classified into two classes as point base group mobility model and region-based group mobility model and compared seven GMM with different parameters. In [5], comparison of GMM and EMM is carried out by following different parameters and concluded that GMMs provides better metrics than EMMs in terms of packets loss and packet delivery ratio while they give the lowest value in throughput.

After reviewing the available literature, this paper presents a brief discussion of four GMMs and its applications for encouragement of a real life scenario. This assists the researchers to utilize these applications in their future work. Figure 1 illustrates the classification of GMM.

2 Reference Point Group Mobility Model

Reference Point Group Mobility (RPGM) Model [7] describes the behavior of mobile nodes. The mobile node movements are identical and these shape certain groups of mobile nodes. This is a circular model. Every group has a middle point which is known as leader node. This leader node determines the movement behavior of all nodes such as location, speed, direction, and acceleration. Each node has a reference point which traces the leader node. The reference point method permits a node to have independent random movement and preserve a fixed distance and direction from leader node.

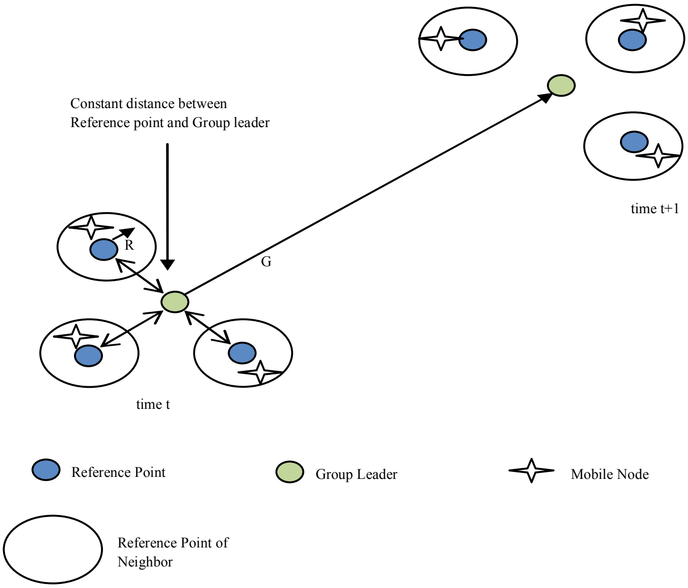


Fig. 2 Movement of mobile node using RPGM model

RPGM model has two kinds of motion as Group motion and Random motion. Group motion represents the motion of leader node which particularizes the group motion. Random motion defines the motion of mobile nodes according to the reference point. The current location of the mobile node is calculated by the adding these two motions. Calculation of the movement of a jth node in the ith group as follows:

$$\text{Node movement: } V_{j,i}(t) = G_i(t) + R_{j,i}(t) \tag{1}$$

where $G_i(t)$ is Group motion and $R_{j,i}(t)$ is Random motion.

Figure 2 describes an example of RPGM model. When the group leader shifts from time t to $t + 1$ by group motion, the reference point of every mobile node also shifts from old place to new place. Through this random motion the mobile node's position is updated to a new place.

In [8] three other mobility models are cited which are infers from RPGM model; they are Column mobility model, Nomadic community model and Purse mobility model.

2.1 Advantage RPGM Model

This model defines checkpoints. A checkpoint is a motion path that is followed by a group. By using checkpoints many realistic applications can be designed. In [9] it is found that RPGM model has a huge effect on throughput, delay and its execution is better than individual node's movement. For limited networks, RPGM provides the finest execution and is not impotence by amplifying the speed of mobile nodes [10]. RPGM generates lesser PDR and extreme end-to-end delay, however it delivers minimal overhead.

2.2 Disadvantage of RPGM Model

- It is impracticable for run time partition prediction [11].
- It is challenging to conceive the group motion pattern of nodes and alters the fashion in the network topology.

2.3 Applications of RPGM Model

The RPGM model has following applications:

- *In-place mobility model*: An area can be divided into various sub-area, and miscellaneous groups are located in this sub-area. It is pertinent in a war zone, where different arm forces are performing same assignments on different areas to achieve a common task.
- *Overlap mobility model*: Diverse sets are dispensed for dissimilar assignments in the even zone. This is pragmatic for the scenario of disaster recovery, search and rescue operation and etcetera.
- *Convention mobility model*: This model is pragmatic for the scenario when demonstrators and students are framed. Various demonstrators give presentations about their investigation work in a distinct room and a lot of students can stir from one room to another room.

3 Reference Region Group Mobility Model

The Reference Region Group Mobility (RRGM) model [12] is utilized to define the behavior of group movement as well as individual motion. Mobile nodes will move towards the region or move inside the region then expect for other mobile nodes of the similar group to nexus that region. The destination is pre-decided here. Inside the RRGM every node uses random way point mobility model for movement. When all mobile nodes arrive into the reference region then they move towards the goal. Before reaching the goal, the mobile nodes continuously search for new region and nexus that region until they reach the goal. Figure 3 demonstrates the moving pattern of RRGM model.

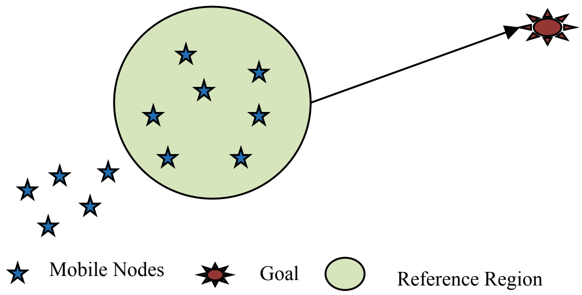


Fig. 3 Moving pattern of RRGM model

There are two types of groups in this model they are the Active group and Stand by group. In Active group, goals are pre-decided, and nodes can move towards their reference region that is node movement inside the region is defined. In the standby group, there is no pre-decided goal, here nodes can move only inside the reference region to nexus the groups for accomplishing the prevail task.

3.1 Advantage of RRGM Model

- It is based on the principles of split and merges. The group gets divided into different sub-groups for completing different tasks. After completing the sub-dividing task, these sub-groups merge into the main group.
- The velocity of mobile nodes is constant in this model.

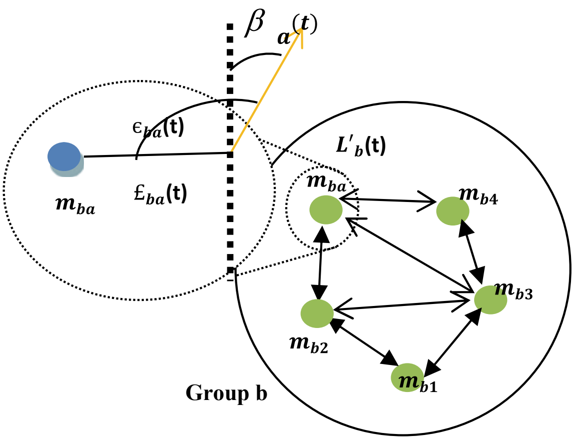
3.2 Application of RRGM Model

This model is applicable to the army, hunt and saving, demonstration hall visiting, house hunt, and so on.

4 Structured Group Mobility Model

The structured group mobility model (SGMM) [13] frames the set structure and it establishes the motion sequence that will be utilized in the simulation. The earlier consciousness of group structure spawns the SGMM model extra precise mobility example of group mobility. Individual mobile nodes are placed here to form a group structure and the group has to move to accomplish a common goal. It is a scalable model that is also known as an advance group. In this model, the member of the main group is authorized to be a structured group itself. Every set has a reference point

Fig. 4 Placement of node $m_{j,I}$ of group j in structure group mobility model



which is assumed as the middle of the set or group leader. Figure 4 describes the motion of SGMM.

The location of node a, of group b can be calculated by:

$$\text{Loc}_{b,a}(t) = f(L'_b(t), \beta_a(t), C_{b,t}(\pounds), \alpha_{b,a}(\epsilon)) \tag{2}$$

where t is time, b is a given set, $\text{Loc}_{b,a}(t)$ is the position of node a of set b at time t, $L'_b(t)$ is the position of the selected reference point of set b at time t, $\beta_a(t)$ is $L'_b(t)$'s steering alignment 0° on a universal coordinate structure, $C_{b,a}(\pounds)$ is the endless circulation from which $\pounds_{b,a}(t)$ is selected, $\alpha_{b,a}(\epsilon)$ is the constant supply from which $\epsilon_{b,a}(t)$ is selected. The individual node is not associated with velocity vector. The motion of the group may be done by its own mobility model as an entity mobility model for $R_j(t)$.

4.1 Advantage of SGM Model

In this model, user can accurately represent the actual behavior of groups with natural structure. Since SGMM allows rigid scripting of simulation, so the performance of some sets can be coordinated by related velocities. This model is self-comparative and in advance itself for arranges their structure.

4.2 Disadvantage of SGM Model

The SGMM is more complex than other GMM.

4.3 Applications of SGM Model

Its application is seen at the places where the prior information of group structure is located, or it is tightly controlled. Some of its application is explained below:

- *Firefighting Team in a Building*: There are a few members in firefighting team which work together in other small groups for assault the fire. Group structure and control are two basic parameters in this team.
- *Military Units on the Battlefield*: SGMM is mainly used at the military region where every vehicle moves in discipline and their movement or structure is pre-decided. Figure 5 is an illustration that can be utilized to portray a tank brigade moving in the creation of the SGMM. The tank battalion obeys the ordered disciplines that are from the controller.

5 Virtual Track-Based Group Mobility Model

This model is also applicable in diverse GMM. The virtual track-based group mobility (VTBGM) model [14] works in the principle of splits and merge the model in real-life scenario. In this model some switch stations are placed randomly then these stations are interconnected by virtual paths. At the switch station, a group might be divided into sub-groups or different sub-groups can be merged into a larger group.

Nodes in the similar sets are moved in the same path. Each node has its own random mobility in the life of a set. The velocity of this group is randomly chosen between a maximum and minimum velocity. Figure 6 outlines a fundamental thought of this VTBGM model. In this illustration, five switch stations are arbitrarily put in the ground associated by means of seven virtual paths with equivalent path width.

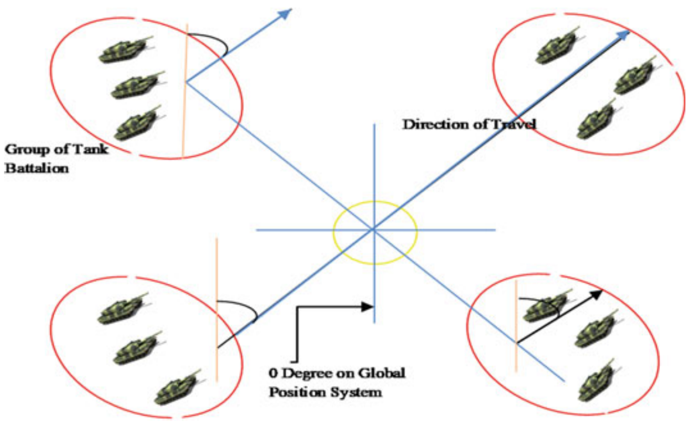
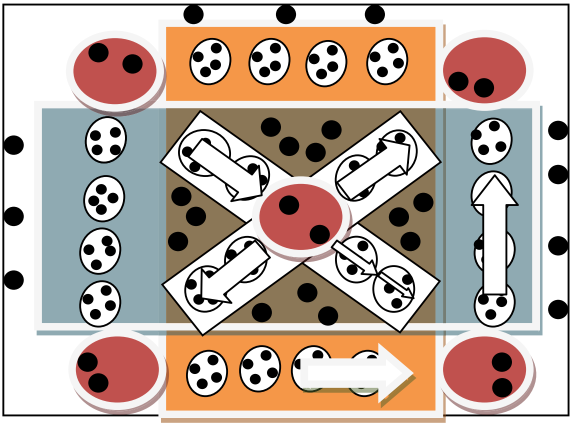


Fig. 5 Military units including sub-groups on the move

Fig. 6 Virtual track-based group mobility model



A group of nodes moves toward switch stations along the paths. They split and merge at switch stations as appeared in the figure. The dark nodes in Fig. 6 speak to the exclusively progressing nodes and stationary nodes. They are set and travel autonomously on paths and switch stations.

5.1 Advantage of VTGM

It can represent the heterogeneous mobility performance with the group and distinct movement. The virtual path can control the motion of group nodes. This model can depict randomly and entirely moving nodes and also static nodes.

5.2 Application of VTGM Model

This model is appropriate for both military and rural situations.

- *Switch Station and Virtual Tracks*: Fig. 7 demonstrates the war zone, the “switch stations” be hot spots of army powers. The virtual paths are streets which connect the switch stations. The battalions used this track for their movement.
- *Tour Guide*: The VTGM model is applicable for tourism. Here a tour guide can visit the different historic places which are located in an area. Each group member must follow the tour guide. At some moment any member can leave that group and join another group according to their interest.

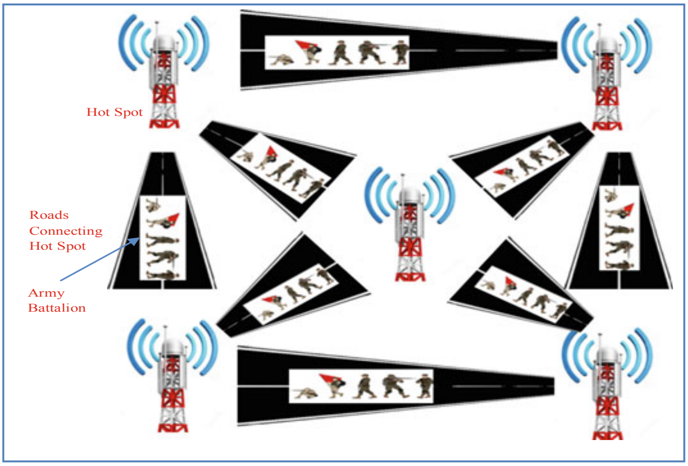


Fig. 7 Army battalion with hot spot and roads

6 Discussion Analysis

To the best lore, there are not many delves in the terrain of the survey on GMM. This paper is an outcome of exhaustive and efficient work and spans the lore of practitioners in this arena. Having inspected GMMs, it was observed that the mobility models can have different possessions and illustrates unique mobility attributes. As a result, it is perceived that the mobility models vary with the tweak in parameters and impact the performance in numerous manners. In this way, the GMM can impose the performance of ad hoc network perfectly. Therefore, it is usually preferred to utilize a GMM rather than EMM. The comparison of different GMM models has done with respect to different performance metrics which is illustrated in Table 1.

7 Conclusion and Future Work

After reviewing the available literature on GMMs, this paper compares different GMM. Every GMM has its own excellent and mobility characteristics. It is concluded that different GMM demonstrate different behavior according to performance metrics. Therefore, it provides the researchers to select a suitable GMM for their research work. The RPGM Model offers the best execution for small scale networks. The RRGM model tries to deliver a well replication of the group motion design with group separator and merge. The SGM Model is more complex than past Mobility Models. The VTBGMM is used in different GMM.

Although the outcome of this paper was quite inspiring, there were several possible hitches that deserve the concentration of offing researchers. Foremost, one must

Table 1 Comparison of GMM

S. No.	Performance metrics	GMM				References
		RPGM	RRGM	SGM	VTBGM	
1	Link up/down	High	Low	*	*	[7, 15–18]
2	Link duration	High	*	*	*	[4, 7, 15, 17]
3	Cluster head change	Low	*	*	*	[7]
4	Throughput	High	*	*	Low	[4, 6, 7, 16, 17]
5	Routing overhead	Average	*	Low	*	[4, 13, 16]
6	Packet delivery ratio	Low	High	High	Low	[16]
7	Heterogeneous group	*	*	*	Yes	[14]
8	Split and merge	*	Yes	*	Yes	[11, 12, 14]
9	Delay	Average	*	*	High	[16]
10	Complexity	Low		High	*	[13]
11	Reference point	Yes	*	Yes	*	[1–19]
12	Average speed	Low	High	*	*	[16, 17]
13	Average hop count	High	*	*	*	[4, 7]
14	Average neighbor count	Low	*	*	*	[4, 7]

*These parameters listed are not addressed by the respective reference

concentrate that evaluations of the GMMs were generally accomplished in the different simulation environments. Second, the performance metrics were different for various GMM. It is fascinating to further explore the behavior of GMMs in a unique simulation and observed these behaviors according to similar performance metrics. We will also try to observe the different behavior of GMM in IOT [20]. Thus, it would be immersing to explore strategies to describe these properties and ripen a model that captures it.

References

1. Camp, T., Boleng, J., Davies, V.A.: A survey of mobility models for ad-hoc network research. Wirel. Commun. Mob. Comput. (WCMC): Spec. Issue Mob. Ad Hoc Netw.: Res. Trends Appl. **2**(5), 483–502 (2002). http://doi.org/10.1177/0897190011431143
2. Bai, F., Helmy, A.: A survey of mobility models in wireless ad-hoc networks. In: Wireless Ad Hoc and Sensor Networks, pp. 1–30 (2004)
3. Zheng, Q., Hong, X., Ray, S.: Recent advances in mobility modeling for mobile ad hoc network research, pp. 70–75 (2004)
4. Bai, F., Sadagopan, N., Helmy, A.: Important: A Framework to Systematically Analyze the Impact of Mobility on Performance of Routing Protocols for Ad hoc Networks. In: INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 2, pp. 825–835 (2003). IEEE. http://dx.doi.org/10.1109/INFCOM.2003.1208920

5. Aung, C.Y., Seet, B.C., Zhang, M., Xie, L.F., Han, P., Chong, J.: A review of group mobility models for mobile ad hoc. Wirel. Pers. Commun. (2015)
6. Lamaazi, H., Benamar, N., Imaduddin, M.I., Habbal, A., Jara, A.J.: Mobility support for the routing protocol in low power and lossy networks. In: The 30th IEEE International Conference on Advanced Information Networking and Applications (AINA-2016), pp. 809–814 (2016). https://doi.org/10.1109/WAINA.2016.122
7. Hong, X., Gerla, M., Pei, G., Chiang, C.-C.: A group mobility model for ad hoc wireless networks, pp. 53–60 (1999). https://doi.org/10.1145/313237.313248
8. Roy, R.R.: Handbook of Mobile Ad-hoc Networks for Mobility Model. Springer. E-ISBN 978-1-4419-6050-4. https://doi.org/10.1007/978-1-4419-6050-4
9. Ciullo, D., Martina, V., Garetto, M., Leonardi, E., Member, S.: Impact of correlated mobility on delay—throughput performance in mobile ad hoc networks, vol. 19, no. 6, pp. 1745–1758 (2011)
10. Rahman, M.U., Alam, A., Abbas, S.: Investigating the impacts of entity and group mobility models in MANETs (n.d.). https://doi.org/10.1109/ICECUBE.2016.7495219
11. Sacko, D., Software, C.: Overview of reference region group mobility model for ad-hoc networks. **3**(1), 1–14 (n.d.)
12. Ng, J.M., Zhang, Y.: Reference region group mobility model for ad hoc networks, pp. 1–5
13. Blakely, K., Lowekamp, B.: A structured group mobility model for the simulation of mobile ad hoc networks. In: International Workshop on Mobility Management & Wireless Access Protocols, pp. 111–118 (2004). http://doi.acm.org/10.1145/1023783.1023805
14. Zhou, B.Z.B., Xu, K.X.K., Gerla, M.: Group and swarm mobility models for ad hoc network scenarios using virtual tracks. In: IEEE MILCOM 2004. Military Communications Conference, vol. 1, pp. 1–6 (2004). https://doi.org/10.1109/MILCOM.2004.1493283
15. Williams, S.A., Huang, D.: A group force mobility model. In: Communications and Networking Simulation Symposium (2006)
16. Gowrishankar, S., Sarkar, S.K., Basavaraju, T.G.: Simulation based performance comparison of community model, GFMM, RPGM, manhattan model and RWP-SS mobility models in MANET. In: 1st International Conference on Networks and Communications, NetCoM 2009, pp. 408–413 (2009). https://doi.org/10.1109/NetCoM.2009.31
17. Williams, S.A., Huang, D.: Group force mobility model and its obstacle avoidance capability. Acta Astronaut. **65**(7–8), 949–957 (2009). https://doi.org/10.1016/j.actaastro.2009.01.030
18. Wu, K., Yu, Q.: A multi-group coordination mobility model for ad hoc networks. In: Proceedings of IEEE Military Communications Conference MILCOM, pp. 1–5 (2007)
19. Borrel, V., de Amorim, M.D., Fdida, S.: A preferential attachment gathering mobility model. IEEE Commun. Lett. **9**(10), 900–902 (2005). https://doi.org/10.1109/LCOMM.2005.10023
20. Lenka, R.K., et al.: Building scalable cyber-physical-social networking infrastructure using IoT and low power sensors. IEEE Access **6**, 30162–30173 (2018)

A PSO-Based Approach for Improvement in AODV Routing for Ad Hoc Networks

Shruti Dixit and Rakesh Singhai

Abstract In MANET, the routing issue is solved by the nodes themselves, thus reducing computational and resource costs. The particle swarm optimization algorithm (PSO) is utilized in the research work, to choose the propitious value of parameters for ad hoc on-demand distance vector routing protocol to improve the quality of service (QoS) in MANET. The routing problem is solved where PSO uses agents like entities from insect communities as a metaphor. Swarm agents based on routing explain a collection of rules for the participating nodes to pursue. Swarm agents interchange information about their behavior adaptively and efficiently for the successful completion of their assigned tasks. PSO algorithm uses the maximum flow objective to prefer the best locations of the swarm agents during each step of network operation. MATLAB language is used for implementation of PSO, and the results of it are used for simulation of routing protocol AODV in QUALNET software. PSO is used for majoring the performance of AODV with the help of QoS parameters: jitter, throughput, and average delay.

Keywords MANET · Particle swarm optimization · Quality of service · AODV · Swarm agents

1 Introduction

A cluster of mobile nodes in mobile ad hoc network (MANET) [1] performs self-configuration in a dynamic way. The mobile network is formed with no fixed infrastructure or centralized supervision. Routing protocols proposed for MANETs have discovered that the reactive protocols are more efficient than proactive routing algorithm in which mobile devices or nodes communicate through a wireless medium. The routes are determined by the routing protocol. This routing protocol provides communication through intermediate nodes between source node and destination node where nodes act as router or end host. AODV [2–5] is an illustrious on-demand

S. Dixit (✉) · R. Singhai
Department of Electronics and Communication, UIT RGPV, Bhopal, India
e-mail: shrutikdixit@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_32

protocol for routing in MANETs. Route request (RREQ) is sent first when a message from a source node is broadcasted in the network. RREQ received by neighboring node with the addresses of source node and destination node compares with the target node's address.

The research papers [6–8] address the challenges intricate in facilitating QoS in MANET. The methods of QoS provided at different levels including those at the levels of routing, medium access control (MAC), and cross-layer have been considered. It is difficult to convey desired QoS such as delay, jitter, and packet loss. Translation of end user inputs to technical variables at MAC layer, routing, etc. and is recommended. The QoS at routing layer is considered in this research work.

Following are some of the schemes used for improving the routing in MANETs found in the literature: Heuristic methods, for example, genetic algorithm, fuzzy logic, artificial neural networks, and ant algorithm [9–12].

Genetic algorithm [9] is multicast and multi-routing method. The appropriate parameters are selected such as crossover and population size, by flooding limited use of available resources. In dynamic conditions, less computing time is required, resulting in improved and optimized routes. Trusted ant routing protocol based on fuzzy logic [10], considers the shortest path, and has faith in the trustworthiness of intermediate nodes. The trust values in terms of evaluation scheme for trust and time-ratio variables are calculated to distinguish between the healthy and malicious nodes. Artificial neural network [11] technique is used for exact forecast of end-to-end delay of packets. The mean number of neighbors between source and target nodes and the path length are used as input variables for computation of end-to-end delay.

Ant colony optimization [12] takes inspiration from ants' behavior. Collectively but decentralized way shortest routes are discovered to the food resource. The new QoS variables such as stability of links and existing buffer size are improved for establishing favorable routes for end-to-end interaction. Architecture and design of multi-agent system are the primary criteria of artificial intelligence based on swarm intelligence (SI) [13]. SI is inspired from the combined activities of sociable insects like bees, ants, wasps, termites, etc., and similar behavior is witnessed in other organisms, for example, schools of fish or flocks of birds. The concept of swarm behavior is used in PSO [14] which is a heuristic technique having capability for searching the best possible solutions of a problem. A fitness sharing in PSO helps in enhancing the performance of routing protocol. A candidate's problem with regard to a given measure of quality is improved by iteratively using this computational method. The existing approaches accepted the challenge of QoS in mobile environment [15–20]. The proposed technique helps to develop an energy-aware multiple-path routing scheme based on PSO which discovers the best route to reduce the routing overhead which enhances the reliability of the network in terms of transmission cost, energy, and traffic ratio [21]. The problems of mobility and decaying energy efficiency are overcome by using PSO based on social natural activities of insects as well as birds. The algorithm based on clustering enhances lifetime of network, energy utilization of mobile nodes, packet delivery ratio, connectivity between nodes, etc. in comparison with other existing algorithms [22–24]. The research paper focuses on the extra energy consumption by cluster heads in wireless sensor network (WSN).

PSO-based energy-efficient algorithm for selection of cluster head for WSN is proposed for proper consumption of energy and increases lifetime of network and better performance in terms of PDR.

The proposed approach takes advantage of PSO algorithm for improving the QoS of AODV routing protocol. In a given search space, swarm agents having specific position and velocity are moving randomly. Using mathematical formulae, it evaluates the problem and provides the best possible solution. Movement of each swarm agent is effected by its local best well-known position and is directed in the search space which is reformed as better positions. The proposed methodology evaluates the accomplishment of AODV QoS parameters like jitter, average delay, and throughput and accordingly improves the overall performance of a MANET.

The remainder of the paper contains five sections. Section 2 gives overview of AODV routing protocol, PSO, and QoS. Section 3 describes the proposed methodology. Section 4 presents simulation parameters and a relevant performance analysis. Results of proposed methodology are discussed in Sect. 5. Finally in Sect. 6, brief conclusion and future direction are given.

2 Overview

2.1 Introduction of AODV Protocol

Source-driven type routing protocol AODV uses on-demand approach for finding routes to establish communication between source node and destination node. Data packets are communicated via intermediate routes. RREQ packet is flooded by the source node in case of non-availability of route for the destination node. Intermediate nodes forward RREQ packet or check the routings in the routing table that could reach the destination node and then send RREP to the source node or continue to flood RREQ in the network. Local connectivity is maintained by AODV protocol by broadcasting hello messages in the network. It sends routing ERROR message to transmitting nodes in case of link breakage; meanwhile, broken records are deleted or repair the routing.

2.2 Particle Swarm Optimization

The routes are dynamically discovered from source node to destination node in PSO. Initially, nodes are identified by sending a route request signal RREQ. After receiving route request, any node equivalent to swarm particle initialized position and velocity uses the concept of PSO. In PSO, swarms wander through a three-dimensional space, in search of the best possible route. Particles are treated as individuals, and the whole population is recognized as a swarm. The primary swarms are formed first, and the population of the particles is assigned arbitrarily over the search space. The process is repeated again and again; each particle is reorganized by two "best" values, called pbest (individual best previous location) and gbest (swarm best previous location). Algorithm steps are summarized as given below:

1. Initialization of individuals by allocating arbitrary position.
2. Approximation of the best value for individual.
3. The individual's best value is compared with its pbest for each individual particle. This pbest is compared with the present particle's position. Updation of pbest is possible if the present particle's position is superior to the pbest value, and then pbest is set with the present particle's position, xi as pi.
4. The individual having the best fitness value will be identified as gbest fitness function.
5. Modification of pbest and gbest of all the individuals using (1) and (2).
6. Repeating steps 2{5 until an acceptable value is achieved}.

The progress of the individual is directed by renewing its velocity and position parameters according to following equation:

$$V_1^{t+1} = WV_i^t + C_1 r_1 \left(x_{\text{pbest}}\, X_i^t\right) + C_2 r_2 \left(x_{\text{gbest}} X_i^t\right) \tag{1}$$

$$X_I^{t+1} = X_i^t + V_i^{t+1} \tag{2}$$

c—acceleration coefficient, w—inertia weight, r—the random values between 0 and 1, X—position of particle, and V—velocity of particle. The flow chart of PSO is shown in Fig. 1.

2.3 QoS

QoS can be outlined as "The collaborative effect of service performance which calculates the degree of satisfaction of a user of the service." It means that the objective of QoS is to attain a more acceptable network behavior in terms of data delivery and resource utilization. The adornment of QoS in a MANET is a challenging assignment. Based on the specific application, the network should be capable of providing approved grade of service. The QoS parameter to be measured is selected depending on the necessities of an application. Examples of QoS parameters include jitter, availability of bandwidth, packet loss probability, average delay, etc. The number of challenges in the refinement of QoS is due to the unique characteristics of MANET. Few challenges are given: (1) The network topology changes rapidly. Hence, difficult to frame a scheme which provides assurance about the QoS required by the particular application. (2) The light-weight scheme should be provided because of limited use of device resources. PSO is used for improving the QoS of the network scenario and

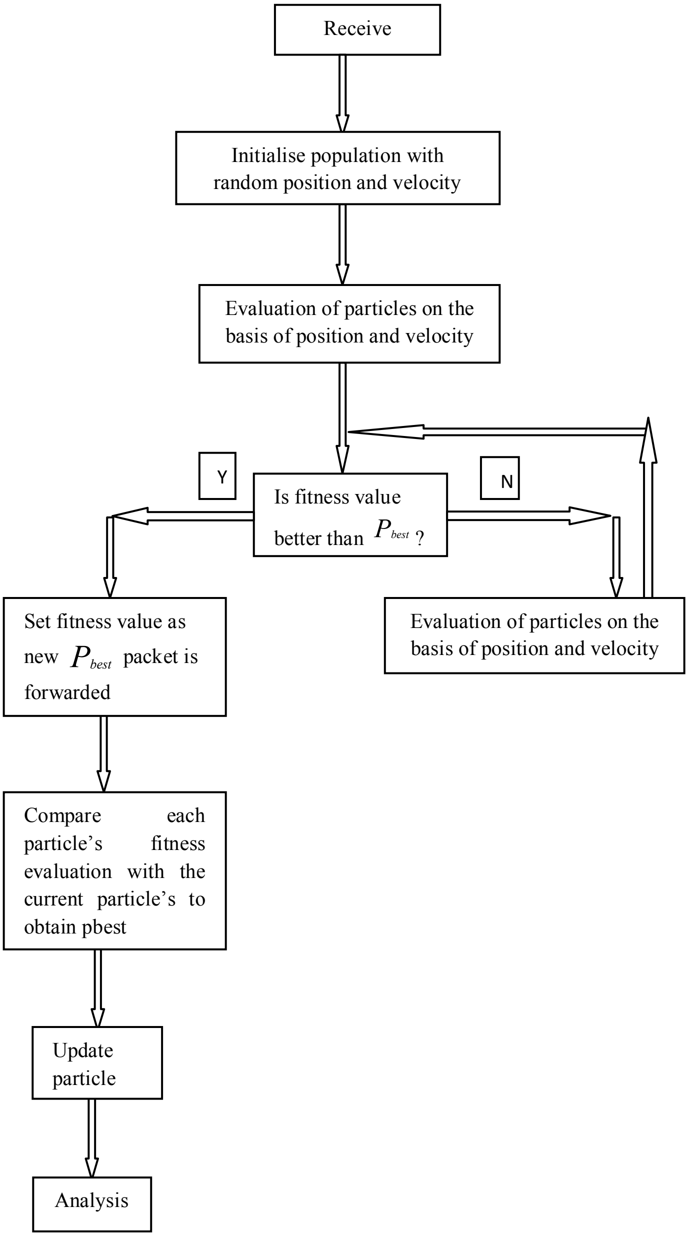


Fig. 1 PSO flowchart

for minimizing the number of control messages in the network. It means if the optimal route in terms of performance metrics is obtained, then QoS can be improved. The basic theory behind optimizing network route is to manage the data flows in the network. The research is focused on QoS challenge in the MANET. The proposed research works on MANET routing protocol and improves the QoS by modifying the optimization algorithm. The performance of MANET is evaluated according to the following QoS metrics:

1. Average Delay: The average amount of time required for a data to be transmitted from the source node to the target node.
2. Throughput: It is the ratio of the total amount of data which is received by the target node from the source node upon the time it takes for the target node to get the final resulting packet. The throughput is the number of bits transmitted per second.
3. Jitter: It is the variations of average delay between the two packets. The delay between packets must be lower than the required threshold value.

Here, the objective is to design a network with maximum throughput, reduced average delay, and reduced jitter with respect to velocity. The utility of PSO algorithm is to enhance the attainment of AODV.

3 Proposed Methodology

The routing is the major issue in MANET. Traditional heuristic techniques are used to find a solution to the routing problem. This technique is suitable for any routing protocol. The shortest route is established during route establishment in AODV routing. The method is proposed to enhance the QoS of AODV in MANET. Analysis of AODV without PSO with default values and AODV with PSO is measured. The QoS metrics like average delay, throughput, and jitter of AODV with PSO are improved. There are two categories of research work. First category and second category are used for the purpose of optimization and solution evaluation, respectively. The method used for optimization is PSO, and QUALNET simulator is used for second category. It is used to obtain favorable parameters in terms of position and velocity in search space. Then velocity is used for solution evaluation of AODV and then testing the performance of routing protocol with optimized value using the QUALNET. The performance of protocol is evaluated in QUALNET by setting up MANET and configuring this with AODV. The overall performance of network is improved in terms of the performance majoring parameter like average delay, throughput, and jitter of AODV. The performance of AODV is characterized by using fitness function which is delineating as follows:

$$\text{Fitness} = W_1\,\text{throughput} - (W_2\,\text{delay} + W_3\,\text{jitter}) \tag{3}$$

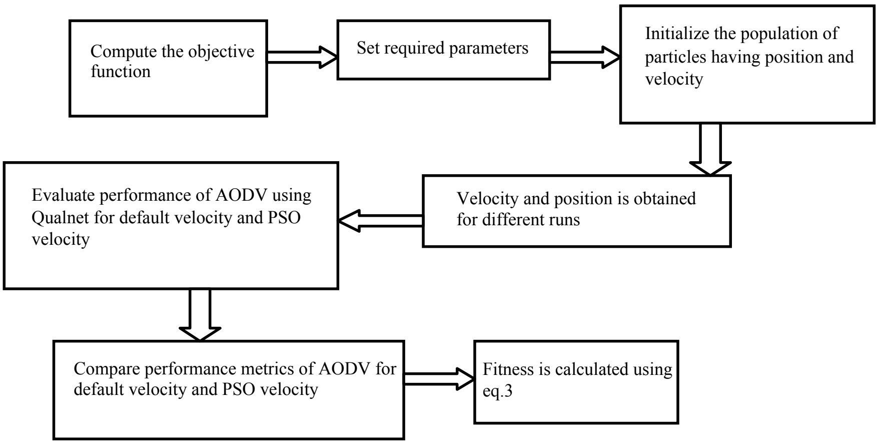


Fig. 2 Working model

The main aim of fitness function is to increase throughput and to reduce average delay and jitter. Equation 3 is the maximizing function because positive sign is for throughput and negative signs for average delay and jitter. The variables W_1, W_2, and W_3 are the weight variables that define the impact of performance metrics on the resultant fitness value. For the simulation purpose W_1, W_2, and W_3 are 0.5, 0.25, and 0.2, respectively. Figure 2 shows the working model.

4 Simulation Parameters

PSO programming codes have been explained successfully in MATLAB environment which establish the competency of the algorithm. To write code of PSO algorithm in MATLAB, two script files (.* m) are needed. In the first, the objective function is defined, whereas in second file the main PSO program is developed. The maximum iterations and maximum runs are changed evenly as 10, 20, 30, 40, 50, 60, and 70 and 5, 10, and 15, respectively. Swarm agent's velocity and position are measured for the PSO algorithm. AODV routing protocol is simulated in QUALNET 5.0.2 software which is used for simulation of AODV routing protocol with default velocity and swarm agent's velocities. During simulation, all nodes are moving randomly. The QoS metrics are determined for the default velocity and swarm agent's velocity. PSO is implemented on MATLAB software, and then output velocity of it is utilized as input variables into the network simulation tool QUALNET. In this way, the performance of AODV is measured. After getting results from simulation, the fitness function is calculated using Eq. 3 and represented graphically.

5 Results

The simulation parameters are summarized below:

- In the scenario, the packet size taken is 512 bytes. The transmission rate is 2 Mbps, and simulation time taken is 300 s.
- The maximum node speed is 10 mps, and pause time is 30 s.
- AODV protocol is used for analysis.

Figure shows the graphical representation of AODV protocol is based on PSO algorithm.

Figure 3 shows the end-to-end delay for AODV when the velocities are increased from 10 to 15 m/s. AODV routing protocol with and without PSO is analyzed. For the constant number of nodes and pause time, AODV with PSO delay is less as compared with AODV without PSO.

In Fig. 4, the average throughput of AODV routing protocol is constant which 3684 for the default velocity is, but the throughput goes on increasing as the velocity increases.

The graph in Fig. 5 displays the jitter of AODV for fixed number of wireless mobile nodes, varying velocities, and pause time. The lowest value of jitter is for velocity 12 m/s and highest for velocity 14 m/s.

Fig. 3 Average delay

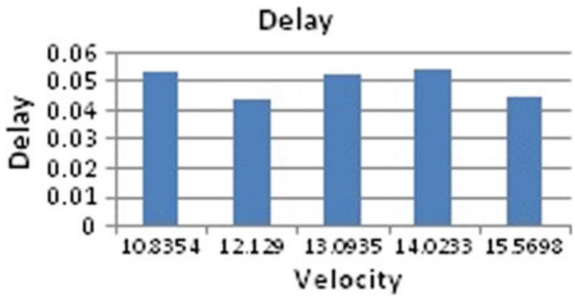


Fig. 4 Throughput

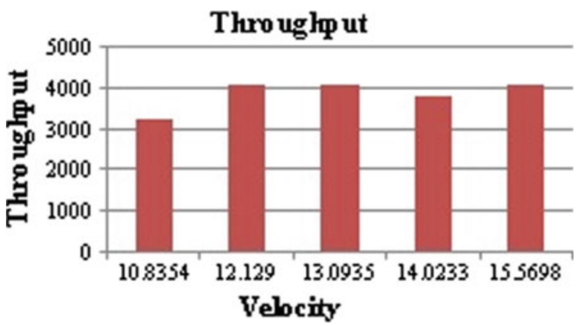


Fig. 5 Jitter

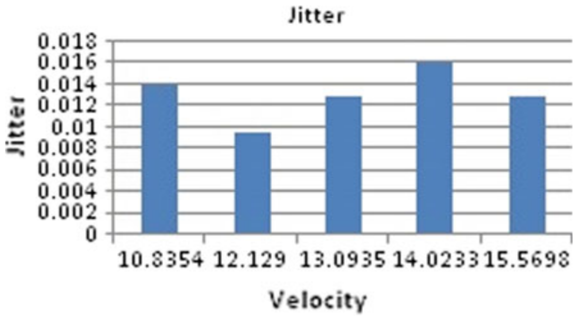
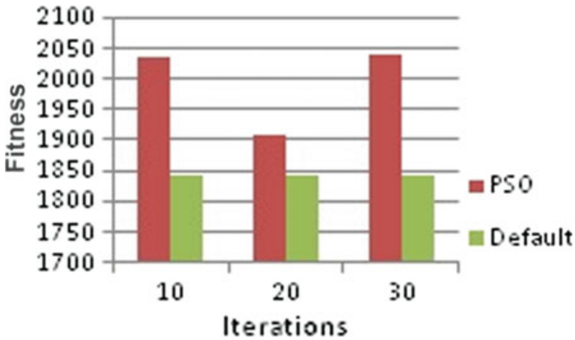


Fig. 6 Fitness function



In Fig. 6, fitness function is calculated for default velocity and PSO velocity using jitter, throughput, and average delay. The default value of parameters for AODV gives inferior results for parameter fitness function. The fitness function for AODV with PSO gives better result.

6 Conclusion

In this work, a PSO-based AODV routing algorithm is simulated and the results are obtained. The value of performance QoS metrics plays a very important role in MANET because the routing protocol performance depends on metrics. The PSO-based algorithm is used to choose the optimized value of variables, so that the output result is improved. Here, AODV performance is compared along with PSO-based AODV on the basis of three performance metrics, namely throughput, average delay, and jitter. After the analysis of the simulation results, the performance of PSO-based AODV outclasses the existing routing algorithm AODV. Routing algorithm based on PSO is more favorable for particular quality of ad hoc networks. It outperforms in practical environmental conditions and gets a useful as well as proficient routing protocol for ad hoc network. Future research can be extended by applying PSO algorithm for other proactive and reactive routing protocols, and few more performance metrics can be analyzed.

References

1. Siva, R., Manoj, B.: Ad-Hoc Wireless Networks: Architecture and Protocols. Pearson, USA (2004)
2. Liu, S., Yang, Y., Wang, W.: Research of AODV routing protocol for ad-hoc networks. In: Conference on Parallel and Distributed Computing and Systems, pp. 21–31. Elsevier (2013)
3. Bai, R., Singhal, M.: DOA: DSR over AODV routing for mobile ad hoc networks. IEEE Trans. Mob. Comput. **5**(10), 1403–1416 (2006)
4. Rath, M., Pattanayak, B., Pati, B.: Comparative analysis of AODV routing protocols based on network performance parameters in mobile ad hoc networks. In: Foundations and Frontiers in Computer, Communication and Electrical Engineering, pp. 461–466 (2016)
5. Rath, M., Pattanayak, B., Pati, B., Panigrahi, C., Sarkar, J.: Load balanced routing scheme for MANETs with power and delay optimization. Int. J. Commun. Netw. Distrib. Syst. Indersci. **19**(4), 394–405 (2017)
6. Abbas, A., Kure, O.: Quality of service in mobile ad hoc networks: a survey. Int. J. Ad Hoc Ubiquitous Comput. Indersci. **6**(2), 1–24 (2008)
7. Mieghem, P., Kuipers, F.: Concepts of exact QoS routing algorithms. IEEE/ACM Trans. Netw. **12**(5), 851–864 (2004)
8. Wang, W., Chatterjee, M., Kwiat, K.: User input driven QoS management in ad hoc networks. J. Comput. Commun. **32**(11), 1306–1315 (2009)
9. Chang, W., Ramakrishna, R.: A genetic algorithm for shortest path routing problem and the sizing of populations. IEEE Trans. Evol. Comput. **6**(6), 566–579 (2002)
10. Sethi, S., Udgata, S.: Fuzzy-based trusted ant routing (FTAR) protocol in mobile ad hoc networks. In: International Workshop on Multi-disciplinary Trends in Artificial Intelligence MIWAI 2011: Multi-disciplinary Trends in Artificial Intelligence, pp. 112–123. Springer (2011)
11. Singh, J., Dutta, P., Pal, A.: Delay prediction in mobile ad hoc network using artificial neural network. Procedia Technol. **4**, 201–206 (2012)
12. Mohan, B., Baskaran, R.: Survey on recent research and implementation of ant colony optimization in various engineering applications. Int. J. Comput. Intell. Syst. **4**(4), 566–582 (2011)
13. Caro, G., Ducatelle, F., Gambardella, L.: Swarm intelligence for routing in mobile ad hoc networks. In: Proceedings of the 2005 IEEE Swarm Intelligence Symposium, SIS 2005, pp. 76–83 (2005)
14. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of the IEEE International Joint Conference on Neural Networks, pp. 1942–1948 (1995)
15. Lobiyala, D., Kattia, C., Giria, A.: Parameter value optimization of ad-hoc on demand multipath distance vector routing using particle swarm optimization. In: International Conference on Information and Communication Technologies (ICICT 2014), Elsevier, pp. 151–158 (2015)
16. Shu-Kai, S., Changa, J.: A parallel particle swarm optimization algorithm for multi-objective optimization problems. Eng. Optim. **41**(7), 673–697 (2009)
17. Trivedi, M., Sharma, A.K.: QoS improvement in MANET using particle swarm optimization algorithm. In: Proceedings of the International Congress on Information and Communication Technology, vol. 439, pp. 181–189. Springer, Berlin (2016)
18. Manickavelu, D., Vaidyanathan, R.: Particle swarm optimization (PSO)-based node and link lifetime prediction algorithm for route recovery in MANET. EURASIP J. Wirel. Commun. Netw. **107**, 1–10 (2014)
19. Saleem, M., DiCaro, G., Farooq, M.: Swarm intelligence based routing protocol for wireless sensor networks: survey and future directions. Inf. Sci. **181**(20), 4597–4624 (2011)
20. Lalwani, S., Singhal, S., Kumar, R., Gupta, N.: A comprehensive survey: applications of multi-objective particle swarm optimization (Mopso) algorithm. Trans. Comb. **2**(1), 39–101 (2013)
21. Robinson, Y., Rajaram, M.: Energy-aware multipath routing scheme based on particle swarm optimization in mobile ad hoc networks. Sci. World J. **2015**, 1–9 (2015). (Article ID 284276)
22. Akay, B., Karaboga, D.: A modified artificial bee colony algorithm for real-parameter optimization. Inf. Sci. **192**(1), 120–142 (2012)

23. Rao, S., Jana, P., Banka, H.: A particle swarm optimization based energy efficient cluster head selection algorithm for wireless sensor networks. Wirel. Netw. **23**(7), 2005–2020 (2017)
24. Singh, S., Sharma, S.: Implementation of a PSO based improved localization algorithm for wireless sensor networks. IETE J. Res. **64**(5), 1–14 (2018)

Health Monitoring Planning for On-Board Ships Through Flying Ad Hoc Network

Sudesh Kumar, Abhishek Bansal and Ram Shringar Raw

Abstract Flying ad hoc networks (FANETs) have developed as an emerging technology with a number of applications that may have a significant impact on the society. FANET-based health monitoring, which is used for transferring data using small unmanned aerial vehicles (UAVs) as drones in 3D space, is one of the promising application under information and communication technologies for improving communication between medical professionals and patients. This planning can be improved to monitor patients using wireless body area network (WBAN) by sending the personal health information (PHI) to the healthcare center with the help of FANET. In this paper, we have proposed a framework for a health monitoring plan with the help of FANET by using WBAN technology for on-board ships, which is used to provide immediate response to on-board patients in emergency situations that are vulnerable to communication difficulties like sea and extend this planning for the on-board ships monitoring with wireless sensor network (WSN) through FANET.

Keywords FANET · WBAN · WSN · PHI · UAVs

1 Introduction

The shipment of goods (cargo), food, raw material, cars, fuels, and people by sea and other waterways through ships or boats is one of the oldest transportation mechanisms and still plays a very important role in development for developing countries. Fishing

S. Kumar (✉) · A. Bansal
Department of Computer Science, Faculty of Computronics, Indira Gandhi National Tribal University, Amarkantak, MP, India
e-mail: sudesh.kumar@igntu.ac.in

A. Bansal
e-mail: abhishek.bansal@igntu.ac.in

R. S. Raw
Department of Computer Science and Engineering, Ambedkar Institute of Advanced Communication Technologies and Research, New Delhi, Delhi, India
e-mail: rsrao08@yahoo.in

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_33

boats are too utilized by millions of fishermen throughout the world for business point of view. Military forces operate vessels for combat, transport and support forces ashore. This is also one of the most cost-effective ways of goods transportation through long distances and also more reliable and less polluting than transporting goods by street or rail. According to recent news, over 55,000 cargo ships are active in international trade. The fleet is represented in over 150 nations, crewed with over 1.5 million mariners working around the world. In this transportation industry, thousands of people are daily traveling through these heavy on-board ships as worker and tourist. Therefore, ships manufacturers' prime aim to use advanced tool and technology for on-board ships as much as possible in order to provide comfort, safety, and infotainment.

In other word, the most of the medical emergencies like heart attack, high/low blood pressure, and accident recovery depends upon the how quickly patient gets medical care. The lack of medical attention at the time of trauma is leading to nearly 27% deaths in India annually. Therefore, our main concern is to provide medically related services to all humans as well as patients who travel through these transportation types. Thus, we require to improvement in existing communication technology for better emergency services to the patient for the greatest chance of survival.

1.1 Flying Ad Hoc Network

Recently, the technological advancements, wireless technologies have addressed the demands of portable, flexible, and wide range communications. Wireless ad hoc network (WANET) allows communication between nodes without the assistance of any centralized structures. A new grooving research field of WANET is called FANETs, which establish wireless communication between small, portable, and flexible devices like unmanned UAVs system, which can fly in the sky (far away from the ground) and controlled remotely without carrying any human being. In the last few years, FANETs have become more popular and its use for different application is appreciated in military purposes as well as civilian purposes [1]. In FANET, every UAV called smart vehicle (node) because these are equipped with high-resolution camera, wireless device, computing device, sensors, digital map, GPS, etc.

In FANET, multiple UAVs communicate to each other by forming an ad hoc network in following ways as shown in Fig. 1.

1. UAV-to-UAV communication (U2U)
2. UAV-to-infrastructure communication (U2I)

All UAVs in the network can communicate to each other through U2U communication, in which UAVs directly communicates to each other or through other UAVs. UAVs can use short- as well as long-range communication according to required data and information transfer rate. Other communication is called UAV-to-infrastructure (U2I). At this, either singly or more UAVs connect to the ground station (base station) to transmit or receive data and information about performing various operations. If

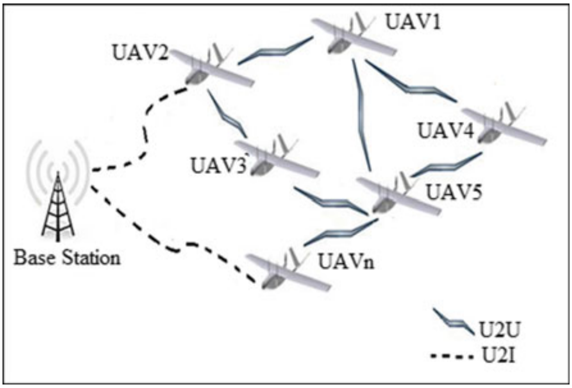


Fig. 1 Architecture of FANET

more number of UAVs takes part in FANET, tasks can be parallelized which in turn reduce the completion time of mission [2]. This network is first time used during Hurricane Katrina deadly cyclone in 2005 for search and rescue operation, it was the worst hurricane in Louisiana and then after in the March 11, 2011 Fukushima Daiichi nuclear disaster and also in April 24, 2015 used in Nepal earthquake. Many of the human lives were saved by using this concept and technology, which demonstrated the validity FANETs.

Wireless ad hoc network is categorized according to their objective and utilization. By the definition of FANET, it is created by multi-UAVs communication system. On the other hand, it is also the specialized form of previously existing mobile ad hoc network (MANET) as well as vehicular ad hoc network (VANET). There are many common popular characteristics as well as certain important differences between FANET and previously existing ad hoc networks. In this section, we discuss some common characteristics in detail manner.

Node Mobility: Mobility degree of flying nodes (UAVs) is much higher than the VANET nodes (like cars) and MANET nodes (like laptops). The UAV has a speed from 30 to 450 km/h; therefore, in FANET, different types of issues occur to frame communication protocol [3].

Mobility model: In most of the cases of the FANETs, the mobility model has been always regular. But, due to mission updates, the flight plan is not predetermined in case of autonomous multi-UAV systems. Therefore, the flight plan is required recalculation [1].

Node density: In a MANET, node density means the average number of nodes present in a particular unit area. In FANET, the distance between all UAVs (scattered in the sky) can be several kilometers. Therefore, FANET node density is very low other than MANET and VANET.

Network topology change: The topology changes very frequently in FANET due to the higher node's mobility. In such a case, the communication link between UAVs is also broken. Therefore, it is required to design and develop some new routing methods for effective communication between UAVs without any breakage [1].

Radio propagation model: The MANET and VANET nodes are very near to the ground or may be on the ground, but FANET nodes most of the time are very far away from the ground. Therefore, there is a need for developing some new efficient radio propagation model FANET [1].
Power consumption and network lifetime: In FANET, communication hardware may not be power sensitive because, it is powered by the energy source of the UAV, unlike in MANET applications. However, it can be a challenging issue for mini-UAVs.
Computational power: In FANET, all UVAs nodes behave like routers; therefore, all nodes have sufficient energy as well as space to support high computational operation and power. VANET also includes high computational power. However, in MANET, because of the small size of nodes as well battery constraint, the nodes have limited or very small computational power.
Localization: Most of the FANETs' application required highly efficient and accurate localization data with smaller time intervals. Therefore, each special UAV must be equipped with a special device GPS as well IMU, which is providing the exact location of UAV to the other UAVs at any time, although GPS is sufficient for MANET and AGPS & DGPS used by VANET [1].

Now, based on the common characteristic between all types of ad hoc networks, given Table 1 shown the comparison among MANETs, VANETs, and FANETs.

Basically, FANETs overcome the limitation where the previous traditional network is not functional, such as disaster areas and military fields. In a MANET, it is very difficult to install movable nodes, on the surface of disaster and critical areas. FANETs able to provide an effective solution by using flying nodes like drones, also this is very useful for critical applications like search, monitoring, and rescue operation.

1.2 Wireless Body Area Network

Recent technological advances in wireless communications have resulted in the design of small, low-cost, lightweight, low-power, and intelligent physiological sensor nodes. These devices can form a wireless body area network (WBAN), which is used for monitoring behavior related to human physiological responses in medical emergencies through a wireless link. In other word, WBAN is a special type of advanced wireless networking technology, which is based on radio frequency that consists of smart sensor nodes which do not hamper the daily life activities of humans and are useful in detecting chronic diseases like heart attack, asthma, blood pressure, oxygen level, diabetes, etc., also, environmental parameters like location, temperature, humidity, light, etc., and to warn the patients in case of emergency conditions [4]. According to their different usages, these special types of sensor nodes operate in close vicinity to on or few cm inside a human body to gather information and share with other device. Figure 2 describes the basic architecture of WBAN, in

Table 1 Comparison between ad hoc networks

FANET	VANET	MANET
Very high node mobility	High node mobility	Very low node mobility
Regular mobility model, but sometime different models required for autonomous multi-UAVs	Regular mobility model required	Random mobility model required
Very low node density	High node density	Low node density
Very frequently topology changed	Fast topology changed	Slow topology changed
UAVs nodes very far away from ground base node. Therefore, most of the cases line of sight (LoS) between the nodes	All vehicles moving on the ground therefore, LoS is not required	All nodes are close to ground therefore, LoS is not required
Small UAVs are not power sensitive, but challenging issues for mini-UAVs	Not power sensitive	Energy efficient protocol required
Computational power of nodes (UAVs) are very high	Computational power of nodes (vehicles) are average	Computational power of nodes are limited
GPS, AGPS, DGPS and specially inertial measurement unit (IMU) used to provide correct location of UAV to other UAVs	GPS, AGPS, DGPS used to provide accurate location	GPS is sufficient for location information

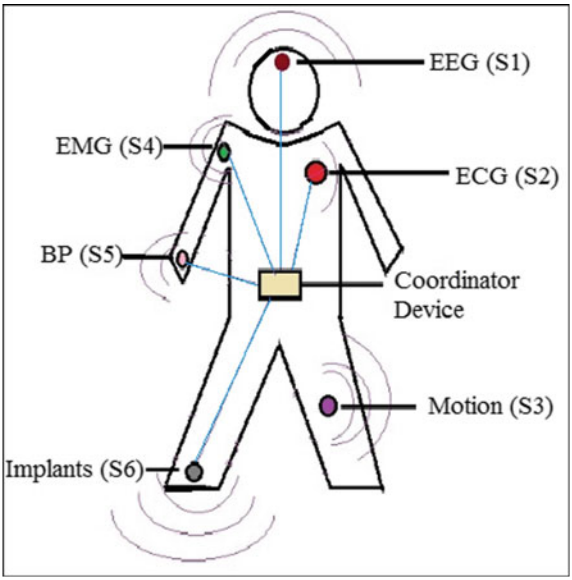


Fig. 2 Wireless body area network

which different technologies like Zigbee, Bluetooth, or Wi-Fi are used by WBAN for monitoring the PHI of the patient's body. All the different types of sensor nodes like EEG, ECG, EMG, BP, etc., represented by S_1, S_2, …, S_n are attached with human body for monitoring all PHI and after that these important data forwards from on-body terminal (coordinator device) to the other different node or gateway through Wi-Fi. This technology helps to cure the patient in the best possible way.

In Sect. 2, literature work is reviewed. In Sect. 3, the proposed framework is introduced with the help of detailed flowchart. In Sect. 4, the case study is presented, the proposed framework also extended in Sect. 5, and the study is concluded in Sect. 6.

2 Literature Survey

A flying ad hoc network (FANET) is special type of wireless network that consists of a grouping of small or mini-UAVs, which are connected to each other in an ad hoc manner. The applications of UAVs promise new ways for military as well as civilian, such as search and rescue operations [5], border surveillance [6], managing wildfire [7], relay for ad hoc networks [8, 9], disaster monitoring [10], and traffic monitoring [11]. In [2], basics of FANETs, various challenges, and several applications are explained in detail and motivated to research to find various solutions and new analysis for the open challenges faced in FANETs.

On other side, WBAN is a wireless network technology, which is formed by attaching a variety of sensors to the human body to transmit and receive health information like body temperature, blood pressure, breathing, heartbeats count, etc. [12]. Also, it is the most popular application in remote medical diagnosis [13], sports training as well as military training [14]. Kamal [15] proposed a model for a health monitoring system with the use of WBSN, mobile communication, and VANET. This model provides high-quality health care and immediate response to patient in emergency situations in urban and disaster area. [16] designed a basic framework for aerial data collection scattered on the ground by expanding existing wireless sensor network utilizing aircraft. Kaćunić [17] discusses the possibility of using drones for evaluating the condition of the transport infrastructure network, with a high emphasis on their different types of advantages and limitations, is explored. Recent studies [18] also designed UAV based health monitoring system for prevention of human life disaster in an area like mountains which are vulnerable to communication difficulties. However, a study of the health monitoring plan for on-board ships through FANET with the help of WBAN to collect data from patient body for life-saving purpose and also monitoring on-board ship for disaster prevention has not been conducted until now.

Therefore, the proposed framework works basically in those outdoor activity areas where communication technology is limited. With the combined effort of FANET and WBAN, the data is collected and transferred wirelessly through FANET from the patient and get transferred to the ground station (medical center) on doctor's system in a nearby city for the real-time diagnosis. This technology helps to cure the patients in the best possible way by delivering air medical services (air ambulance flight) and takes the patient to the medical professional at the nearest hospital. Therefore, the study of this exiting and welfare work may be a novel work for the society.

3 Proposed System Frameworks

In this paper, we propose a system framework for health monitoring to ensure safety at an early stage by collecting health information for people who are active with on-board ship in sea area and transmit it through FANETs to the medical emergencies. Here, we have used FANET and WBAN to enhance our work in the more elaborated manner. In this exertion, we aim to deploy the UAVs to observed on-board ship. So, whenever on-board patient feels unhealthy, his\her physiological information will be automatically send to the hospital server in the healthcare monitoring system. When the patient is in emergency condition has to send him\her to the hospital, the proposed system immediately passes the message to air ambulance for an on-board patient in a timely fashion.

The flowchart of the proposed framework for the monitoring system is shown in Fig. 3. The main working concept of the proposed system can be summarized as follows.

In the first phase, gather health information from an active human body by attaching different body sensors (ECG, PPG, temperature, BP, position, motion, etc.) node as well as emergency alarm is attached to a human body. These special kinds of

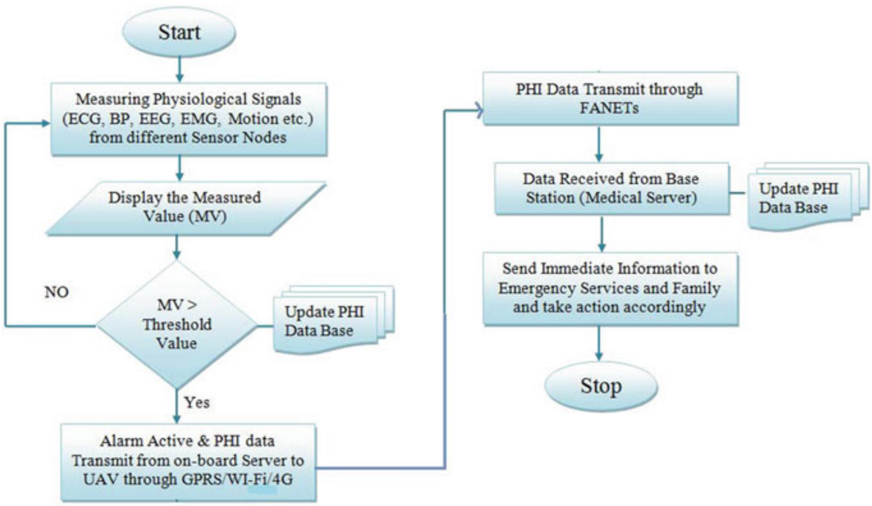


Fig. 3 Flowchart for health monitoring planning

sensor nodes have wireless transmission capability and sense biological information from the human body with a central coordinator device worn on the body or placed in an accessible location [4]. In this, we can use different network such as Bluetooth (IEEE 802.15.1), Wi-Fi (IEEE 80211b/g), Zigbee (IEEE 802.15.4) with parameters like range, bandwidth data rate, frequency band, and multiple access for wireless transmission. Moreover, all collected information is then transferred to the gateway, such as a mobile, PDA, or local computer system.

Secondly, we transfer all collected PHI data with the help of FANETs environment, in which dedicated UAV collect the PHI data through the data receiver device from the WBAN working with on-board ship at sea through WAN, that are vulnerable to communication with ground for emergency services. In which there are different types of path selection methods for sending PHI data to the nearest UAV. Every UAV is treated as a smart UAV, which is highly configure with camera, wireless sensor and computing devices, digital map, GPS, etc. After collecting data from WBAN, data may move through FANET. In which, UAVs directly communicates to each other or through other UAVs by U2U communication. UAVs can use short as well long-range communication according to required data and information transfer rate. A variety of efficient routing techniques [19, 20] is used for transmission of data and information. The communication between the UAVs is controlled by the base station at ground.

In the third level, we establish a base station in the nearest urban area or smart city to collect the transferred data from FANET. Now, the base station is either medical server or if not, then data is transmitted to the nearest medical server at the smart hospital by using infrastructure to infrastructure (I2I) communication. The key work of medical server is, it keeps electronic medical records of registered users and provides various kinds of services to the users. In this exertion, our aim is that data collected through the UAV is reported to the medical staff, family, and emergency personnel like air ambulance services, allowing them to take appropriate measures.

4 Case Study

In this part, we discuss a hypothetical case study to explain the significance of our proposed framework. We also consider an issue in which patient facing heart attack problem and describe how our proposed framework can be used to firstly address the problem and then provide the immediate solutions [21]. Mr. Rajan is recovering from a heart attack. After the one month from the recovery, he decides to visit some place through boat or ship transportation system via sea. Our proposed health monitoring framework offers a solution for Ranjan. Equipped with a WBAN, tiny sensors provide constant observation of vital statistics and estimate induced energy expenditure [18]. Tiny electronic inertial sensors measure movement while electrodes on the chest can measure Ranjan heart activity. Think about on-board Ranjan suddenly feel pain in heart and goes into a serious condition where the heart does not pump blood around the body efficiently. Now, coordinator device continuously captures and transmits the ECG signals and all heart-related activities to the on-board patient's server. If

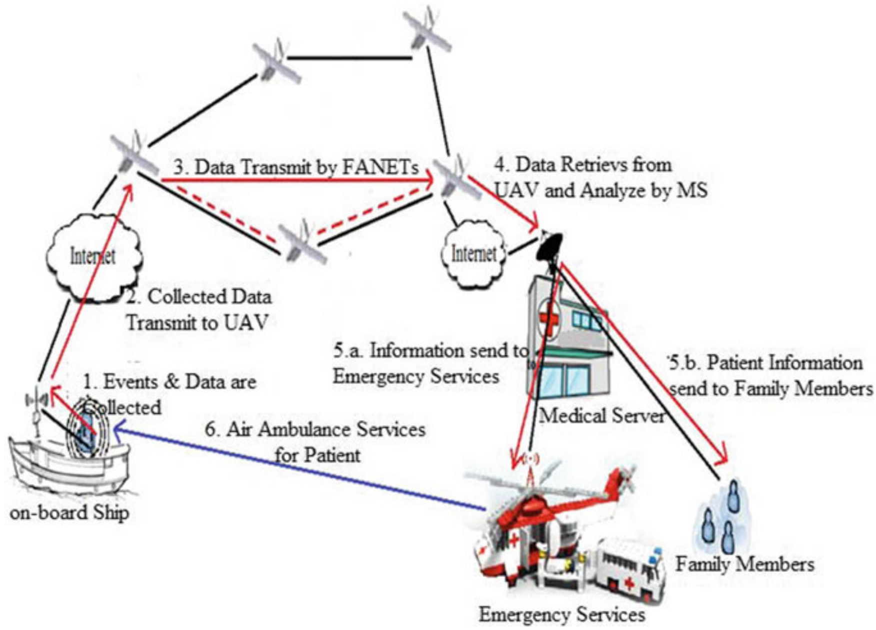


Fig. 4 Data flow in the proposed framework

all the data is higher than the predetermined threshold values, the fastest alert will be issued to on-board staff. Now, all PHI data is received by nearest dedicated UAV through the Internet and further transmitted to the ground base station with the help of FANETs environment. After that transferred PHI data collected by the ground base station (Hospital Medical Server) is outside the allowable range or greater than a predetermined threshold value then alarm are generating. Meanwhile the data collected by UAV also reported to the family members and also the ground medical server can directly contact emergency medical services (EMS) like air ambulance. When the patient is in emergency conditions and has to send him/her to the hospital, EMS immediately dispatches an air ambulance with doctor to the patient's location in a timely fashion (Fig. 4).

5 Application Scenario for Monitoring On-Board Ships with WSN and FANETs

Proposed work also extends to monitoring the on-board ships with the help of wireless sensor network (WSN), which is composed large number of different types of sensor nodes through the FANET. The actual on-board ships contain several of sensors which are connected to a central processing unit (on-board control room) via Wi-Fi/ZigBee. When a disaster occurs, the most important issue that needs to be solved is to preserve on-board human lives on the ships. Because, the first 24 h, after the disaster are the most critical situation in human life. With the advancements in sensor communication technology, the concept of smart on-board ship aims to enhance ship performance. Smart on-board ship configures with different types of sensor nodes like temperature monitoring, GPS, environment monitoring, water-level monitoring, weight monitoring, etc. Whenever, monitoring node reaches and exceeds the any set value, it is intimated by alarm/siren to the ship operator and this information is also transmitted to observe UAV, and through FANETs, this information will reach to a base station for emergency services if the condition is out of control. Following are some key points of importance of FANET framework [16].

- This framework application is to search for and to rescue the unfortunate on-board ships that happen to be lost, which is trapped by the UAVs during the disaster.
- Observed the damage by using UAV with high-resolution video inspection.
- If the weather condition is not suitable for on-board ship, then it is intimated by siren with information moving through FANETs to emergency services.
- Compute the exact location detection of the illegal ships.
- Sea traffic monitoring in dense areas.
- Usefulness for monitoring sea boarder areas.
- Maintain relay network between ships and ground base station, because sometimes they cannot communicate directly due to long distance and some physical obstacles like mountains.
- Health monitoring system.

6 Conclusion

In this study, we proposed a framework for on-board ships or boats through FANETs. This method is monitoring on-board human life disasters in area like the sea that is vulnerable to communication difficulties. The working of FANET and WBAN has been explained together to elaborate the concept of both technologies for health monitoring. This concept explains a flowchart and hypothetical case study included an important key object that provides high-quality health care at an early stage and provides an immediate response to patients in emergency situations. Moreover, the system can be utilized to monitor on-board ships also with WSN.

We believe that in future, we have a large scope to improve the health monitoring planning if additional sensors are mounted on the body for the betterment of the patient in any crucial situations.

References

1. Bekmezci, I., Sahingoz, O.K., Temel, S.: Flying ad-hoc networks (FANETs): a survey. Ad Hoc Netw. 1244–1270 (2013)
2. Kumar, S., Raw, R.S.: Flying ad-hoc networks (FANETs): current state, challenges and potentials. In: INDIACom-2018, 5th International Conference on "Computing for Sustainable Global Development", Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi, India, IEEE Conference ID: 42835, pp. 4233–4238, 14–16 Mar 2018
3. Han, Z., Swindlehurst, A.L., Liu, K.J.R.: Optimization of MANET connectivity via smart deployment/movement of unmanned air vehicle. IEEE Trans. Veh. Technol. **58**, 3533–3546 (2009)
4. Khan, J.Y., Yuce, M.R.: Wireless body area network (WBAN) for medical applications. In: Campolo, D. (ed.) New Developments in Biomedical Engineering (2010). ISBN: 978-953-7619-57-2
5. George, J., Sujit, P.B., Sousa, J.: Search strategies for multiple UAV search and destroy missions. J. Intell. Robot. Syst. **61**, 355–367 (2011)
6. Sun, Z., Wang, P., Vuran, M.C., Al-Rodhaan, M., Al-Dhelaan, A., Akyildiz, I.F.: BorderSense: border patrol through advanced wireless sensor networks. Ad Hoc Netw. **9**(3), 468–477 (2011)
7. Barrado, C., Messeguer, R., Lopez, J., Pastor, E., Santamaria, E., Royo, P.: Wildfire monitoring using a mixed air-ground mobile network. IEEE Pervasive Comput. **9**(4), 24–32 (2010)
8. de Freitas, E.P., Heimfarth, T., Netto, I.F., Lino, C.E., Pereira, C.E., Ferreira, A.M., Wagner, F.R., Larsson, T.: UAV relay network to support WSN connectivity. In: ICUMT, pp. 309–314. IEEE (2010)
9. Jiang, F., Swindlehurst, A.L.: Dynamic UAV relay positioning for the ground-to-air uplink. In: IEEE Globecom Workshops (2010)
10. Maza, I., Caballero, F., Capitan, J., Martinez-De-Dios, J.R., Ollero, A.: Experimental results in multi-UAV coordination for disaster management and civil security applications. J. Intell. Robot. Syst. **61**(1–4), 563–585 (2011)
11. Semsch, E., Jakob, M., Pavlicek, D., Pechoucek, M.: Autonomous UAV surveillance in complex urban environments. In: Web Intelligence, pp. 82–85 (2009)
12. Ullah, S., Higgins, H., Braem, B.: A comprehensive survey of wireless body area networks on PHY, MAC, and network layers solutions. J. Med. Syst. **36**(3), 1065–1094 (2012)
13. Gao, T., Massey, T., Selavo, L.: The advanced health and disaster aid network: a light-weight wireless medical system for triage. IEEE Trans. Biomed. Circuits Syst. **1**(3), 203–216 (2007)
14. Smeaton, A.F., Diamond, D.: Aggregating multiple body sensors for analysis in sports. In: Proceedings of International Workshop on Wearable Micro and Nanosystems for Personalised Health-pHealth, Valencia, Spain (2008)
15. Kamal, P., Raw, R.S., Singh, N., Kumar, A.: VANET based health monitoring through wireless body sensor networks. In: 3rd International Conference on Computing for Sustainable Global Development (INDIACom) 16–18 Mar 2016
16. Wang, C., Ma, F., Yan, J., De, D., Das, S.K.: Efficient aerial data collection with UAV in large-scale wireless sensor networks. Int. J. Distrib. Sens. Netw. **2015**, 1–19 (2015). (Article ID 286080)
17. Kaćunić, D.J., Librić, L., Car, M.: Application of unmanned aerial vehicles on transport infrastructure network. Građevinar **68**(4), 287–300 (2016)
18. Ha, I., Cho, Y.-Z.: Unmanned aerial vehicles health monitoring system for prevention of disaster in activities of the mountain. Int. J. Control Autom. **9**(9), 353–362 (2016)
19. Tareque, M., Hossain, M., Atiquzzaman, M.: On the routing in flying ad hoc networks. In: Proceedings of the Federated Conference on Computer Science and Information Systems, vol. 5, pp. 1–9 (2015)
20. Fosati, S., Kruzelecki, K., Heitz, G., Floreano, D., Rimoldi, B.: Dynamic routing for flying ad hoc networks. IEEE Trans. Veh. Technol. **65**(3), 1690–1700 (2016)

21. Otto, C., Milenković, A., Sanders, C., Jovanov, E.: System architecture of a wireless body area sensor network for ubiquitous health monitoring. J. Mob. Multimed. **1**(4), 307–326 (2006)
22. Erdelj, M., Król, M., Natalizio, E.: Wireless sensor networks and multi-UAV systems for natural disaster management. Comput. Netw. (2017)

Performance Evaluation of MAC Protocols with Static and Dynamic Duty Cycle in WSNs

Gulshan Soni and Kandasamy Selvaradjou

Abstract Prolonging the lifetime of sensor nodes is still a challenging area of research in wireless sensor networks (WSNs) and other related wireless communication technologies. WSNs have numerous extensive research works, which are dedicated to design energy-efficient MAC protocols. WSNs have wide range of applications that cover various fields such as surveillance, defense, and healthcare. The WSNs can become an invaluable resource for realizing the vision of the Internet of thing (IoT). The integration of WSNs and IoT can easily solve countless difficult problems and make the life easy. This paper proposes a dynamic duty cycle MAC (D^2CMAC) protocol, which emphasizes mainly on accomplishing less energy consumption with acceptable packet loss and latency. The D^2CMAC protocol attempts to increase the sleep interval, while data traffic is very low that leads to consuming less energy as compared to SMAC and tunableMAC (CSMA/CA with duty cycle) protocols. The simulation results illustrate that D^2CMAC consumes less power of battery up to 57% and 9%, when compared with SMAC and tunableMAC protocols, respectively.

Keywords MAC layer · SMAC · Duty cycle · WSNs · Sleep interval

1 Introduction

The reducing consumption of power endures a critical challenge in WSNs and other wireless communication network technologies such as IoT. In these types of networks, often the nodes are battery powered and periodically changing or replacing these batteries is not realistic. In the WSNs, deterministic placement of sensor nodes

G. Soni (✉) · K. Selvaradjou
Computer Science and Engineering Department, Pondicherry Engineering College, Puducherry 605014, India
e-mail: gsoni@pec.edu

K. Selvaradjou
e-mail: selvaraj@pec.edu

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_34

is not realistic because of mainly two reasons [1]. The first reason is that the target areas for sensor network placement (deployment) are generally inhospitable or remote location, which avoid planned placement of sensor nodes. Second reason is that the quantity of sensor nodes is generally huge so individual placement of sensor nodes will increase the cost and complexity of deployment. The WSNs have a huge range of applications, which include observation (monitoring) of environmental conditions (i.e., temperature, pollution, humidity, pressure, noise, etc.) in specific areas, smart traffic management, smart home system, smart grid management, and many more. These variety of applications make the WSNs significant technology for human. The WSNs consist of many sensor nodes.

Figure 1 illustrates the power consumption of MICA2 mote during different radio modes such as transmit, receive, idle listening, and sleep [2]. Through observation of Fig. 1, sleep mode consumed very negligible amount of energy (0.003 mW) as compared to other modes (transmit, receive, and idle listening) so that keeping sensor node in sleep mode is good choice to save energy. The sleep mode of sensor node is a good contender to reduce the consumption of energy, but it also has many constraints such as increased packet drop, latency, overhead of control packet, etc.

The foremost source for waste of energy is idle listening, collision, overhearing, overemitting, and control packet overhead [3]. Figure 2 illustrates major causes of energy waste in WSNs.

The protocol stack used in wireless communication consists of physical layer, data link layer, network layer, transport layer, and application layer. This paper emphasizes on data link layer or more precisely MAC sublayer.

The main contribution of this paper is adopting a dynamic duty cycle approach, so nodes get more time for sleeping that will lead to less consumption of energy by

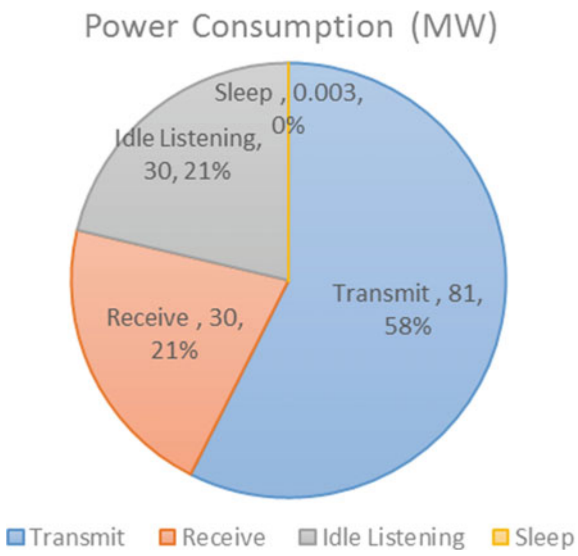


Fig. 1 Radio power consumption (MICA2 mote)

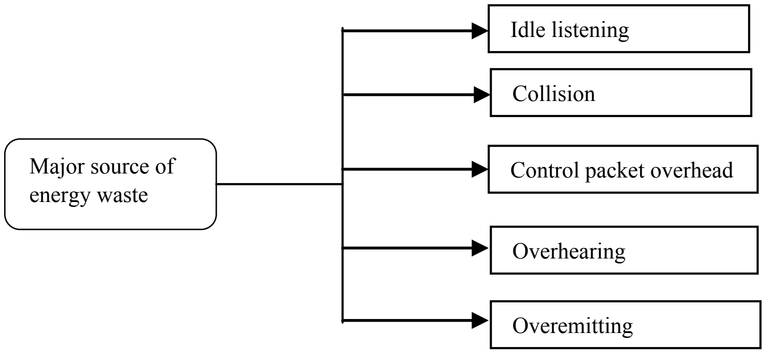


Fig. 2 Major source of energy waste in WSNs

sensor nodes. The proposed D^{2}CMAC protocol has dynamic sleep interval, which is efficient under low data traffic conditions.

The rest of the paper is structured as follows. Section 2 outlines exiting energy-efficient MAC protocols along their pros and cons. Section 3 gives an overview of the problem statement and D^{2}CMAC protocol. Section 4 describes our simulation experiments and results analysis, and finally, Sect. 5 concludes the paper with a brief summary of observations.

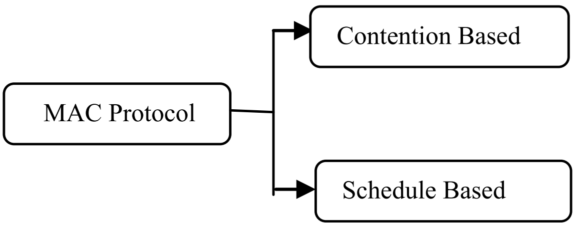
2 Related Work

In the literature, many researchers from academia and industries proposed several protocols or algorithms those provide new solutions or enhancement of existing solutions for MAC protocol in WSNs and other related wireless technologies.

Generally, MAC protocols can be classifying in two broad categories based on the methodology used to manage medium access: contention-based and schedule-based [4]. Figure 3 indicates about classification of MAC protocols.

The contention-based MAC protocols have distributed access on medium and therefore no requirement for central coordinator such as sensor MAC (SMAC) [3]. The second approach to access medium is based on a schedule (or order) for every

Fig. 3 Classification of MAC protocols



node to transmit, receive, or be inactive that is called schedule-based MAC protocol. In more specific manner, every individual node has specific time slot(s) for active and sleep mode. The examples of schedule-based MAC protocol are low-energy adaptive clustering hierarchy (LEACH) [5].

In conventional MAC protocol such as IEEE 802.11 (CSMA/CA) for WSNs, the sensor nodes expend a lot of time on detecting whether they need to transmit or not, but actual time required for transmission is generally very tiny so that it causes huge waste of energy. This process is called idle listening. A lot of works are focused on designing energy-efficient MAC protocols, which suggest how to avoid factors such as idle listening, collision, overhearing, control packet overhead, and overemitting, which lead to evade unnecessary waste of energy in sensor nodes.

Singh and Raghavendra [6] proposed PAMAS, which tried to avoid overhearing among neighboring nodes, but this did not address problem of idle listening. This protocol was based on the multiple access with collision avoidance (MACA) protocol and works on the concept of turning off radio under certain conditions.

The SMAC [3] protocol is first and foremost MAC protocol, which introduced a scheme of periodic listen and sleep for a predefined interval for nodes to solve the problem of idle listening. The SMAC has significant advantages over other classical MAC protocols such as consumption of less energy, overhearing avoidance, long message is fragmented into many small fragments, and then they are transmitted in a burst. In SMAC protocol, the packet latency was increased because senders always had to wait for their receivers to wake up before sending out data packet.

To overcome the constraints of SMAC, Van Dam and Langendoen [7] introduced TMAC protocol, which had an adaptive duty cycle in the place of fixed duty cycle. In this approach, one node transmits all the messages to another node in bursts of variable length and sleep between the bursts. The TMAC suffered from early sleep problem.

In last few year, various researchers proposed numerous enhancements on SMAC protocol for encounter existing limitations. Some of the important enhancements on SMAC protocol are as follows:

To improve latency in SMAC, Abu-El Humos and Alhalabi [8] proposed FASMAC protocol. The FASMAC protocol utilized the mechanism of adaptive listening and FRTS packets (TMAC protocol) to migrate packet latency and maintained energy efficiency.

Roy and Sarma [9] introduced an adaptive energy efficiency MAC (AEEMAC) protocol to optimize energy efficiency of nodes. The AEEMAC protocol uses some optimization techniques; those are adaptive sleeping, reusing of channel, and use of combined control packets (ACK-RTS, SYNC-RTS) in bidirectional and multihop data transmission.

To address the trade-off between energy utilization and latency. Hu et al. [10] proposed Adaptive Duty Cycle SMAC (ADC-SMAC) protocol. In this protocol, the adjustment of duty cycle is done dynamically on the basis of utilization, average sleeping delay, and upper and lower bounds of duty cycle.

Wang et al. [11] proposed a dynamic duty cycle SMAC (DC-SMAC) protocol to replace the fixed duty cycle concept in SMAC. The DC-MAC protocol uses adaptive duty cycle mechanism and priority discriminant function. The objective of priority discriminant function is providing priority to node for access channel based on number of packets on nodes.

Xu and Wang [12] proposed node correlation based on flow adaptive efficient adaptive MAC (EA-MAC) protocol. The EA-MAC protocol uses node correlation analysis algorithm and traffic adaptive duty cycle mechanism. The access of channel was allocated based on available remaining energy of nodes.

To reduce the idle listening problem, Kuo and Liu [13] proposed a new mechanism for SMAC. This work introduced proactive exchanging the status between sensors nodes and utilized information from physical layer. The authors also developed a bug-fixed module for SMAC protocol in ns2 simulator.

For mission critical scenarios, Sakya and Sharma [14] proposed a MC-MAC protocol. The MC-MAC protocol uses novel regression based on adaptive duty cycle approach and adaptive to the traffic scenarios while considering the residual energy of nodes. The proposed MC-MAC protocol model is not implemented in any simulation/real test bed.

Jieying et al. [15] proposed adaptive energy saving mechanism (AES-SMAC). The AES-SMAC protocol utilized two mechanisms, adaptive synchronization period (ASP) and dual-element adaptive contention window (DEACW).

In this extensive literature survey, we observed various approaches (solutions) are available that are dedicated to reducing consumption of battery power and the turning off the radio of node is one of the existing solutions. The mechanism of turning on and off the radio of sensor node is called duty cycle. The duty cycle decides how much time the sensor node will be in sleep mode and active mode (transmission or receiving mode).

3 Problem Statement and D^{2}CMAC Protocol

In our D^{2}CMAC protocol, we tried to resolve the problem of fixed duty cycle. The fixed duty cycle approach uses periodic listen and sleep interval during entire lifetime of sensor nodes. Figure 4 illustrates the concept of periodic listen and sleep approach. The periodic listen and sleep concept is not suitable for several cases such as when the data traffic is very low but those time the sensor nodes is on for fixed listening interval. In this scenario, the listen interval is large, but node only needs very tiny interval to send its packet due to low data traffic. Hence, major portion of this listen interval is wasted due to the fixed nature of the duty cycle.

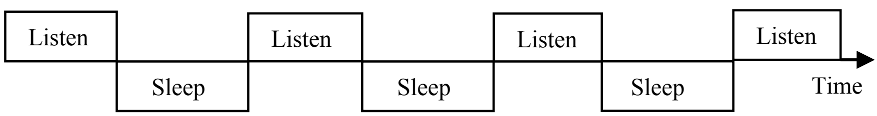


Fig. 4 Periodic listen and sleep

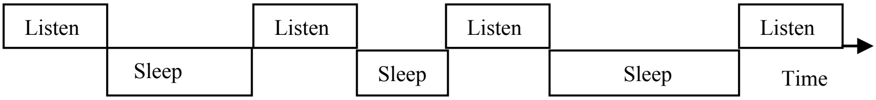


Fig. 5 Dynamic sleep interval

We considered the scenario, where the data rate is very low. This situation fixed duty cycle approach will not work appropriately. In our D^2CMAC protocol, we tried to resolve fixed sleep interval problem. The D^2CMAC increases the sleep interval whereas packet rate is very low, so this will lead to consume less battery power as compared to SMAC and tunableMAC [16] protocols.

Figure 5 illustrates the concept of dynamic sleep interval. Our D^2CMAC algorithm implements the concept of dynamic sleep interval, so the sleep interval is not fixed.

Most of the features of D^2CMAC are like the tunableMAC [16] protocol, which is implemented in Castalia Simulator [16]. This protocol supports for broadcast communication, so there is no provision for acknowledgments and control packets (RTS and CTS). This protocol adopted CSMA/CA (IEEE 802.11) mechanism for transmitting data packets.

The major difference between tunableMAC and IEEE 802.11 is the duty cycle approach for sensor nodes. The tunableMAC has duty cycle mechanism, but simple CSMA/CA has no such type of concept. With the help of duty cycle approach, sensor nodes can turn off or on their radio, so this way those can reduce idle listening and save battery power. These nodes are not synchronized to their sleeping schedules; therefore, each node uses to transmit train of beacons to wake up likely receivers before each data transmission.

The relation between sleep interval and consumed energy is expressed in Eq. 1.

$$\text{Sleep Interval}\,\alpha\,\frac{1}{\text{Consumed Energy}} \tag{1}$$

$$\text{Number of Transmission}\,\alpha\,\text{Packet Reception Rate (PRR)} \tag{2}$$

$$\text{Number of Transmission}\,\alpha\,\frac{1}{\text{Packet loss rate}} \tag{3}$$

Algorithm 1 Pseudocode for the D^{2}CMAC protocol

```
------------------------------------------------------------------------
Input: numTx, numTxTries
Output: Packets will transmit or drop.
1  Begin
2    IF Packet can buffer THEN
3      IF Packet is first THEN
4        Cancel listen and sleep timer
5        numTxTries<--numTx
6          IF numTxTries<0 THEN
7            delete front packet
8            IF buffer has packet THEN
9              numTxTries = numTx
10             go to step 4
11           ELSE
12             start sleeping for random amount of time
13           END IF
14         ELSE
15           start carrier sensing
16           IF channel is clear THEN
17             transmit the packet
18             numTxTries<--numTxTries-1
19               IF all packet has transmitted THEN
20                 IF numTxTries<=0 THEN
21                   delete front packet
22                     IF any packet left THEN
23                       numTxTries<--numTx
24                       go to step 15
25                     ELSE
26                       start sleeping for random amount of time
27                     END IF
28                 ELSE
29                   numTxTries<--numTx
30                   go to step 15
31                 END IF
32               ELSE
33                 perform carrier sense for each packet
34                 go to step 4
35               END IF
36           ELSE
37             Backoff
38             go to step 15
39           END IF
40         END IF
41     ELSE
42       already in transmission loop
43     END IF
44   ELSE
45     MAC buffer overflows
46   END IF
47 END
------------------------------------------------------------------------
```

Equation 1 states that if sleep interval will be high, then consumption of energy will be less. The sleep interval should not be too large; otherwise, packet loss will be increased. The number of transmission and packet reception rate (PRR) are proportion to each other, which is illustrated in Eq. 2. If number of transmissions will increase, then PRR value will also increase under certain conditions. If number of transmissions increases, then packet loss rate will decrease up to some limit, which is shown in Eq. 3.

Algorithm 1 describes the working of D^2CMAC protocol. As we mentioned earlier D^2CMAC and tunableMAC both having same working approach, nonetheless both behave diverge for sleeping interval. Therefore, D^2CMAC is modified version of tunableMAC protocol. The nature of this algorithm is distributed. This algorithm first tries to buffer the coming packet, but if buffer is already full, then it indicates MAC buffer overflow. The sleep and listen timer of sensor nodes will cancel, if nodes have any packet to transmit. The number of transmissions (numTx) plays an important role in entire procedure of MAC protocol. The attempt of transmitting a packet repeats until it reaches to maximum number of transmissions. The node starts to sense the channel; if channel is found free, then it initiates the transmission of packet, otherwise backoff. The backoff is a random amount of time, which node must wait when channel is not found free to avoid the conflicts of concurrent transmissions. The algorithm has option to transmit all the packets in single time sense of channel, else sense the channel for every packet. After all packets' successful transmission, then node can go for sleep mode.

4 Simulation and Results Analysis

The main objective of our simulation experiments is to measure average consumption of energy, PRR, and latency of packet for our D^2CMAC protocol and to do compare with SMAC and tunableMAC protocols.

4.1 Simulation Environment and Topology Setting

The simulation environment is Castalia 3.3 [16], and simulation platform is Ubuntu 16.04. We used star topology, and Fig. 6 shows a specific scenario of our simulation experiment. This is single-hop network with four sources and one sink. Packets from all the sources nodes (1, 2, 3, and 4) flow toward sink node 0.

In simulation experiments, we considered two different scenarios to evaluate performance of our D^2CMAC protocol. The first scenario is retaining the topology (Fig. 6) constant and varying the rate of packets. In second scenario, we are keeping rate of packets constant and vary number of source nodes. We calculated average consumed energy (mJ), average PRR, and average application level latency (ms) for D^2CMAC, SMAC, and tunableMAC, respectively.

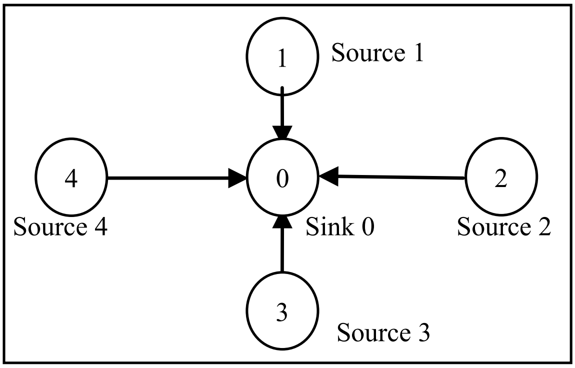


Fig. 6 Typical star topology

Table 1 Parameter for tunableMAC and D^{2}CMAC protocols

Simulation parameters	Parameter values
dutyCycle	0.1
beaconIntervalFraction	1
numTx	5
listenInterval	10 ms

Table 2 Network simulation parameter

Simulation parameters	Parameter values
Simulation time	30 s
Sensor node field size	200 * 200 m^2
Sensor node application name	"ThroughputTest"
TxOutput power	"−5 dBm"
Application packet rate	0.2
Application constant data payload	200 bytes
Node initial energy	18,720 J
Idle listening consumption	62 mW
Transmit power	62 mW per packet
Receive power	62 mW per packet
Sleep power consumption	1.4 mW
Listen timeout (SMAC)	100 ms

4.2 Parameter Configuration

Table 1 consists of parameters for tumableMAC and D^{2}CMAC protocols. The main experimental parameters are shown in Table 2. Table 2 consists of all essential parameters for all set of experiments, which is required to perform in Castalia Simulation Tool.

4.3 Analysis of Experimental Results

The length of packet is fixed during entire simulation experiments. We measured three metrics (consumed energy, PRR, and latency) in two different scenarios. The first scenario is that the number of nodes is fixed and packet rate (traffic) is changing. Second case is that packet rate (traffic) is persistent and number of nodes is variable.

i. **Consumed energy analysis**

Figure 7 illustrates the average consumption of energy during various data traffic. Figure 7 depicts that D^2CMAC is consuming less energy as compared to SMAC and tunableMAC protocols. This proves that when data traffic is very low, then D^2CMAC protocol consumes less energy that will lead to prolong life of sensor network. The D^2CMAC protocol consumes a lesser amount of energy up to 57% and 9% when compared to SMAC and tunableMAC, respectively.

Figure 8 demonstrates the average consumed energy while number of nodes is dynamic. Figure 8 shows significant improvement of consumed energy of D^2CMAC when compared to SMAC and tunableMAC protocols.

Our experiments show that D^2CMAC consumes very less energy when compared to SMAC and there is noteworthy difference between D^2CMAC and tunableMAC. The reason behind the consumption of less power of D^2CMAC protocol is increase of sleep interval of nodes when compared to SMAC (fixed duty cycle) and tunableMAC (CSMA protocol with duty cycle).

ii. **Analysis of Packet Reception Rate**

The PRR indicates that how many packets delivered to sink node successfully from various source nodes. The PRR is just opposite of packet lost rate. Figure 9 illustrates

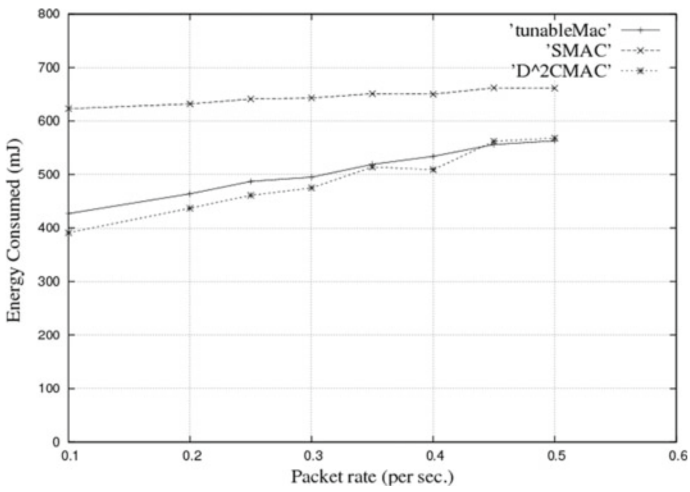


Fig. 7 Average energy consumption (first scenario)

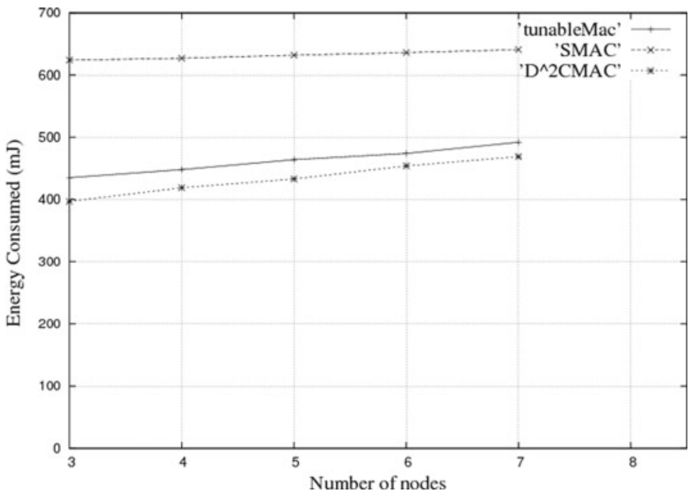


Fig. 8 Average energy consumption (second scenario)

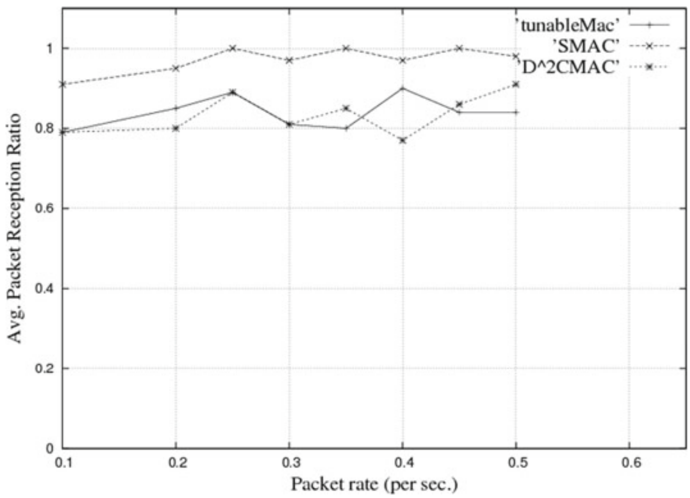


Fig. 9 Average packet reception rate (first scenario)

that D^{2}CMAC has PRR among 80–90%, which is in quite acceptable limit under the condition of changing packet rate.

Figure 10 verifies that D^{2}CMAC shows quite acceptable PRR when compared with SMAC and tunableMAC. It is specified form Fig. 10 that D^{2}CMAC is preserving PRR approximately 80%.

The simulation experiments display proposed D^{2}CMAC protocol preserves PRR at least 80% and this is quite acceptable. The D^{2}CMAC protocol performs equal or

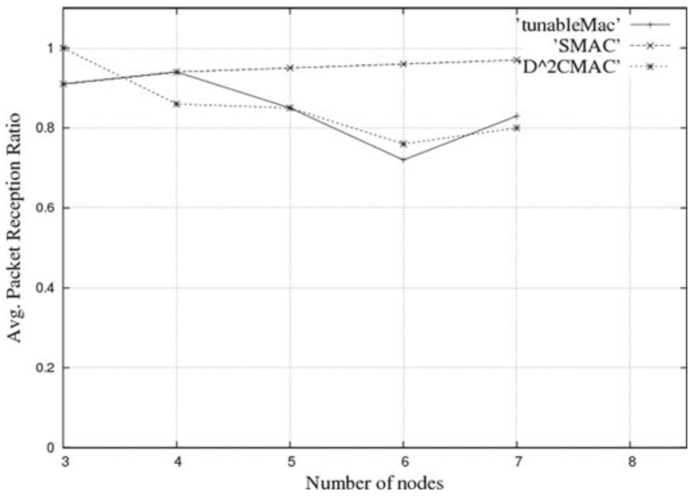


Fig. 10 Average packet reception rate (second scenario)

lesser when compared to SMAC and tunableMAC protocols in terms of PRR because nodes are not synchronized to their sleeping schedules in D^2CMAC.

iii. Analysis of packets latency

Figure 11 shows that D^2CMAC protocol consists of packet latency between SMAC and tunableMAC under the condition of vary data packet. When compared to SMAC, D^2CMAC has less latency of data packet.

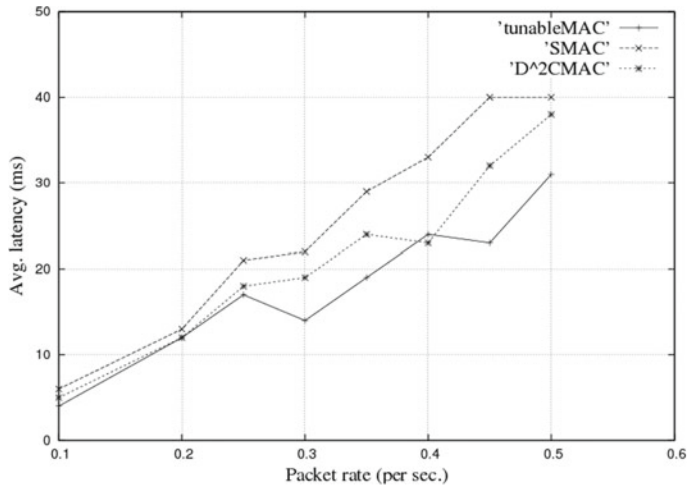


Fig. 11 Average latency (first scenario)

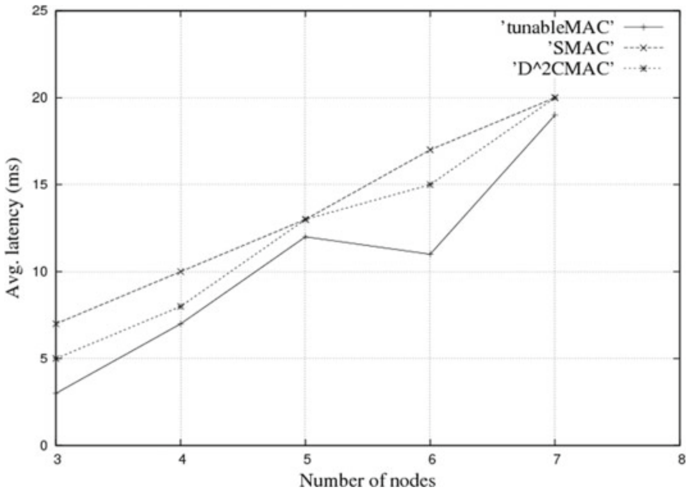


Fig. 12 Average latency (second scenario)

Figure 12 depicts average latency during variation of number of nodes. Figure 12 shows D^{2}CMAC which has less latency when compared to SMAC protocol.

In terms of packet latency, the D^{2}CMAC protocol performs better when compared to SMAC in both scenarios. The D^{2}CMAC protocol performs better because it is a very simple protocol (CSMA/CA with dynamic duty cycle) when compared to SMAC.

5 Conclusion

This research emphasizes on designing the MAC protocol for WSNs, which will reduce the amount of consumed energy of sensor nodes. To prolong the survival time of the sensor nodes, we proposed a D^{2}CMAC protocol. Generally, idle listening is major cause for energy depletion in WSNs, so D^{2}CMAC chief objective is to reduce the idle listening problem. We considered sensor network having very less data traffic, and our experiment results show that our D^{2}CMAC protocol exhibits superior performance in energy consumption up to 57% and 9%, when compared to SMAC and tunableMAC, respectively. For future work, we would like to implement our D^{2}CMAC protocol on the real test bed and analyze the performance metrics. We will also want to test our D^{2}CMAC protocol in various other topologies such as two-hop networks.

References

1. Slijepcevic, S., Potkonjak, M.: Power efficient organization of wireless sensor networks. In: IEEE International Conference on Communications (ICC 2001), vol. 2, pp. 472–476. IEEE, Helsinki (2001)
2. Miller, M.J., Vaidya, N.H.: Power save mechanisms for multi-hop wireless networks. In: First International Conference on in Broadband Networks, pp. 518–526. IEEE, San Jose (2004)
3. Ye, W., Heidemann, J., Estrin, D.: An energy-efficient MAC protocol for wireless sensor networks. In: Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2002), vol. 3, pp. 1567–1576. IEEE, New York (2002)
4. Rajendran, V., Obraczka, K., Garcia-Luna-Aceves, J.J.: Energy-efficient, collision-free medium access control for wireless sensor networks. Wirel. Netw. **12**(1), 63–78 (2006)
5. Heinzelman, W.B., Chandrakasan, A.P., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. IEEE Trans. Wirel. Commun. **1**(4), 660–670 (2002)
6. Singh, S., Raghavendra, C.S.: PAMAS—power aware multi-access protocol with signalling for ad hoc networks. ACM SIGCOMM Comput. Commun. Rev. **28**(3), 5–26 (1998)
7. Van Dam, T., Langendoen, K.: An adaptive energy-efficient MAC protocol for wireless sensor networks. In: 1st International Conference on Embedded Networked Sensor Systems, pp. 171–180. ACM (2003)
8. Abu-El Humos, A., Alhalabi, B.: FASMAC: a low latency and energy efficient MAC protocol for wireless sensor networks. In: Advances in Computer, Information, and Systems Sciences, and Engineering, pp. 179–184. Springer, Berlin (2007)
9. Roy, A., Sarma, N.: AEEMAC: adaptive energy efficient mac protocol for wireless sensor networks. In: Annual IEEE Conference (INDICON), pp. 1–6. IEEE (2011)
10. Hu, H., Min, J., Wang, X., Zhou, Y.: The improvement of S-MAC based on dynamic duty cycle in wireless sensor network. In: International Conference on Computer Science and Automation Engineering (CSAE), vol. 1, pp. 341–345. IEEE (2011)
11. Wang, C., Chen, Y., Hou, Y.: The analysis and improvement of SMAC protocol for Wireless sensor networks. In: 9th International Conference on Mobile Ad-hoc and Sensor Networks, pp. 437–441. IEEE (2013)
12. Xu, D., Wang, K.: An adaptive traffic MAC protocol based on correlation of nodes. EURASIP J. Wirel. Commun. Netw. (1), 258 (2015)
13. Kuo, Y.-W., Liu, K.-J.: Enhanced sensor medium access control protocol for wireless sensor networks in the ns-2 simulator. IEEE Syst. J. **9**(1), 1311–1321 (2015)
14. Sakya, G., Sharma, V.: MAC protocol with regression based dynamic duty cycle feature for mission critical applications in WSN. Int. J. Adv. Comput. Sci. Appl. **8**(6), 198–206 (2017)
15. Jieying, Z., Shi, P., Yinglin, L., Shaopeng, H.: Adaptive energy-saving mechanism for SMAC protocol in wireless sensor network. In: International Conference on Communications and Networking, pp. 47–57. Springer (2016)
16. Boulis, A.: Castalia—a simulator for wireless sensor networks and body area networks. NICTA, User's Manual Version 3.2 (2011)

Bandwidth Enhancement of Microstrip Array in Wireless Communication with Optimized Geometry

Sarmistha Satrusallya and Mihir Narayan Mohanty

Abstract With improved technology for communication, antenna is an essential component. A single element antenna cannot satisfy the desire of the user. The antenna array is an alternative to this solution. In this paper, an antenna array has been designed for application in wireless communication. A triangular patch element is considered for 3×3 array antenna. These elements are placed around the central patch that is of rectangular and received the feeding. The work is emphasized on location geometry of the patches. The geometry of the patch is optimized using Genetic Algorithm techniques. It is found that the optimized antenna provides better result as compared to Unoptimized one in terms of bandwidth and return loss. Also, the radiation pattern is clearly observable. The bandwidth resulted as 2.5 GHz at the 15.5 GHz frequency with a return loss of 26.5 dB.

Keywords Array antenna · Bandwidth · Gap coupled · Optimization

1 Introduction

Wireless communication develops rapidly for the transmission of data in the form of voice, image, and video where Microstrip antenna an important role. Its different application is radar, military, and satellite communication due to its weight, size, and good bandwidth. The challenges for the researchers are the limitations which include low impedance gain, low efficiency. Different shapes of the patches and slots are used to design the Microstrip antenna such as rectangular, circular, and triangular to enhance the performance of the antenna. Authors also studied the use of GA for broad band application [1, 2]. In [3], authors have designed a U shaped parasitic patch. They have analysed the coupling by changing the patches in horizontal and vertical

S. Satrusallya · M. N. Mohanty (✉)
Department of Electronics and Communication Engineering, ITER, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India
e-mail: mihir.n.mohanty@gmail.com

S. Satrusallya
e-mail: sarmisthasatrusalya@soa.ac.in

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_35

orientation. The bandwidth of the proposed antenna was calculated to be 1.5 GHz at 5.5 GHz centre frequency, which was slightly greater than the conventional antenna. Other designs as a parasitic structure defined by a set of microstrip element and microstrip line-fed U slot patch with 2×2 array provided good bandwidth, reliability as claimed by authors [4, 5]. Antenna array for MIMO communication system was designed to improve the link capacity. In this case, series-fed array was used with pattern synthesis [6]. An antenna and the EBG structure based array has been proposed by the authors in [7, 8]. Also, four types of EBG structures were proposed to be used in the design of a patch antenna array to improve the bandwidth reducing the overall size. Another design of dual polarization microstrip patch antenna with quasi-cross slots having high isolation was proposed over a wide bandwidth [8–10]. A F shaped antenna with compact size was designed for practical application. The size of the antenna is useful for generating an antenna array. Similarly, a Sierpinski fractal antenna with different patches was proposed for generating a 2×2 antenna array to be operated in 28 GHz [11, 12].

Based on optimization technique PSO, complementary PSO, Evolutionary Algorithm has been applied by authors [11–14]. Complementary PSO was applied to the partially coupled array in [11, 12], for optimized radiation pattern PSO was utilized [13], and EA has been applied to T slot antenna to enhance the bandwidth [14].

In this paper, an array is designed and is described in Sect. 2. A further element of the array is optimized using Genetic Algorithm and is explained in Sect. 3. The optimized dimensions are utilized for bandwidth enhancement.

2 Antenna Array Design

The shape of the radiating component of the antenna can be used for the expansion of the antenna without increasing the size. In this proposed method to shape the antenna array are used. Basically, an array is the combination of multiple components where the addition of the individual element determines the array field. The array may have a different shape such as rectangular, circular, triangular etc. Eight identical triangular parasitic patches are used to form the antenna array. Parasitic patches are useful for the increment of gain, bandwidth, and directivity of the antenna by changing the shape of the patch.

Figures 1 and 2 represent the proposed antenna array. The ground plane is of square structure. The patch antenna is present at the center with a probe fed and coupled (gap) with the other parasitic patches. The two corners of the rectangular patch was cut diagonally and the feed point is provided at the upper patch cut. The parasitic patches are of triangular shape. A 3×3 antenna array is designed with a different dimension of the substrate. The material of the substrate is chosen to be FR4. The dimension of the center patch is of dimension 11×11(mm). The design is realized with HFSS.

The antenna parameters are calculated as per the following equations [14].

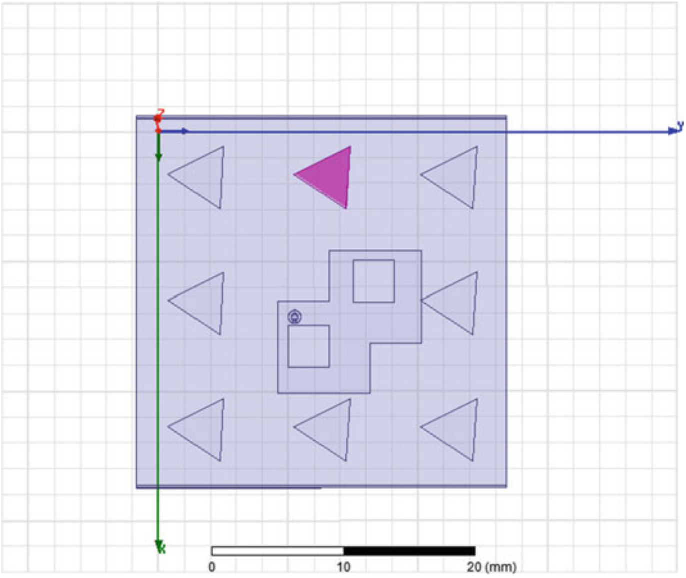


Fig. 1 The geometry of the unoptimized 3 × 3 array

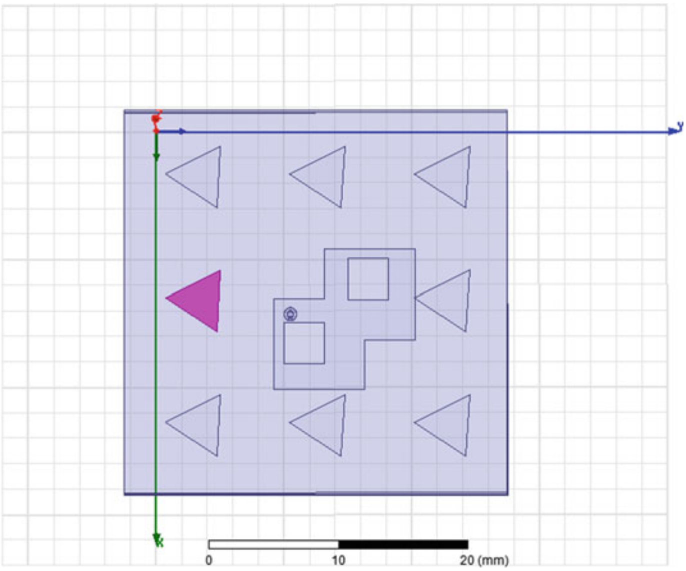


Fig. 2 The geometry of optimized 3 × 3 array

For rectangular patch antenna, the patch length is L, the dielectric constant is $\in_r$, Lightspeed is c, the Resonant frequency is f_r Wavelength is λ, The rectangular patch width W_0 and the dielectric thickness H_0

Then,

$$W_0 = \frac{c}{2f_1} = \left(\frac{\in_r + 1}{2}\right)^{-1/2} \tag{1}$$

$$L = \frac{C}{2\sqrt{\in_r} + f_r} \tag{2}$$

The effect of dielectric is given by

When $\left(\frac{W_0}{H_0}\right) \leq 1$

$$\in_r = \frac{\in_r + 1}{2} + \frac{\in_r - 1}{2}\left\lfloor\left(1 + 12\left(\frac{H_0}{W_0}\right)\right)^{-1/2} + 0.04\left(1 - \left(\frac{W_0}{H_0}\right)\right)^2\right\rfloor \tag{3}$$

When $\left(\frac{W_0}{H_0}\right) \geq 1$

$$\in_e = \frac{\in_r + 1}{2} + \frac{\in_r - 1}{2}\left(1 + 12\left(\frac{H_0}{W_0}\right)\right)^{-1/2} \tag{4}$$

The triangular patch is derived from the rectangular patch with dimension 5 mm. The derived figure is given in Fig. 3. The resultant triangle is an Equilateral triangle where the base arm is of 5 mm and other two arms are of 4.71 mm.

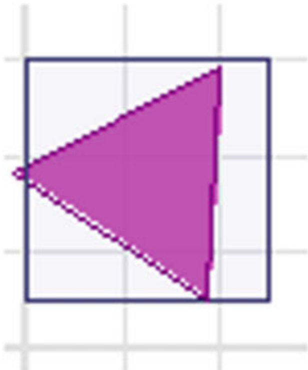


Fig. 3 Geometry of the triangular patch

3 Optimization Using GA

The dimension of the substrate has been optimized using EA. An antenna array of 3 × 3 has been designed using the triangular patch. The patch is repeated on a substrate of length 28.5 mm. The substrate is optimized using an evolutionary algorithm. The EA is very accurate and based on the theory of survival of the strongest and extinction of the weak species for the next generation.

The genetic algorithm contains the solution to a given optimization problem for a population of the encoded candidate. The optimization of the parameter is according to [14] is as follows:

(i) The design is initialised with specific f_c, ε_r, and h as 15.5 GHz, 4.3 and 1.6 mm respectively.
(ii) The patch dimension is considered with a random value and studied the return loss.
(iii) When the return loss achieves more than 20 dB the process will stop else repeat from step (ii).

4 Result and Discussion

The VSWR of the antennas are represented in Figs. 4 and 7. The impedance bandwidth of the Unoptimized and optimized antenna is presented in Figs. 5 and 8. The gain is calculated from Figs. 6 and 9.

VSWR of the antenna must lie in the range of 1–2. The proposed antenna has a VSWR of less than 1. The optimized antenna design converges at 11.5 and 15.5 GHz

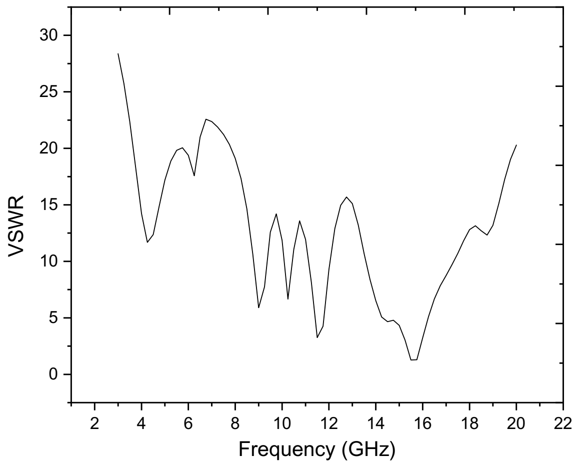
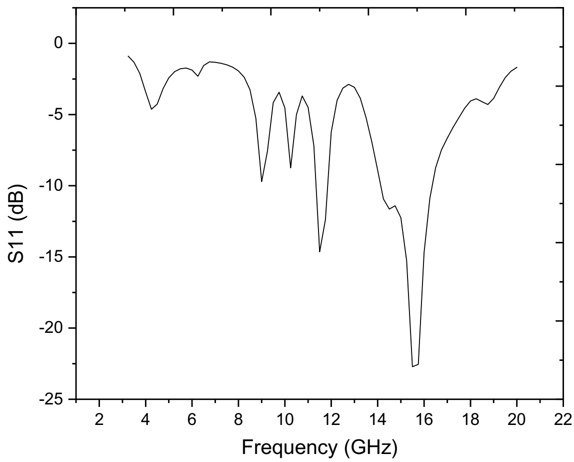


Fig. 4 VSWR of the unoptimized antenna

Fig. 5 Impedance bandwidth of unoptimized antenna array



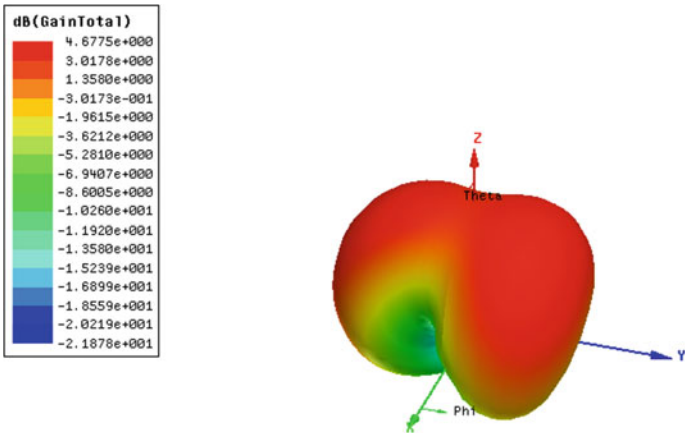


Fig. 6 Polar plot of the unoptimized antenna array

with a bandwidth of 2.5 GHz as compared to the unoptimized design having a bandwidth of 2GHZ. Also, the gain of the proposed antenna is measured to be 4.81 dB better than unoptimized design having a gain of 4.67 dB (Tables 1, 2 and 3).

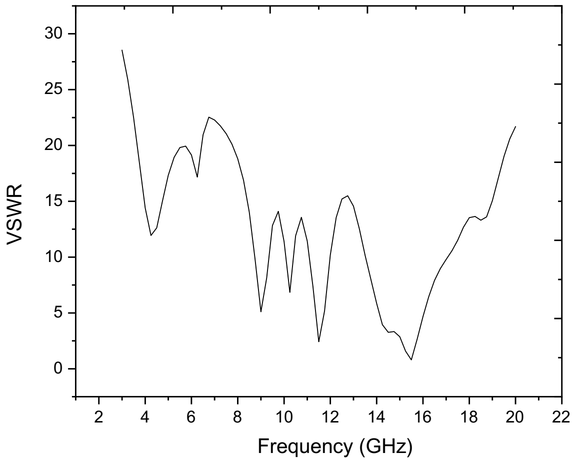


Fig. 7 VSWR of the optimized antenna array

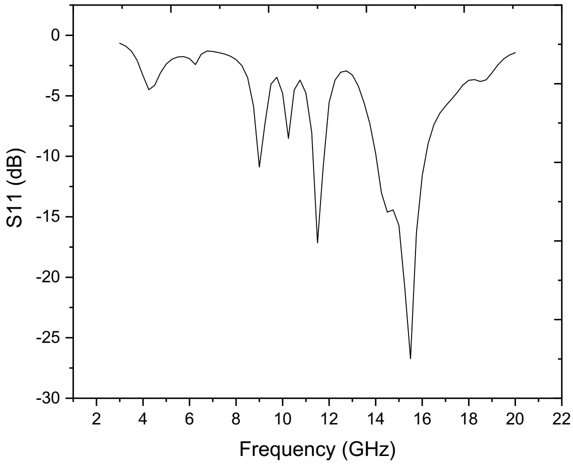


Fig. 8 Impedance bandwidth of optimized antenna array

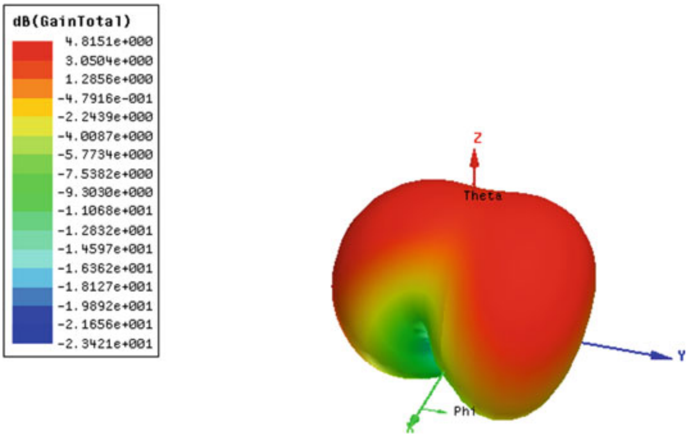


Fig. 9 Polar plot of the optimized antenna array

Table 1 Optimized dimension of parasitic patch parameter

Substrate material	FR4 epoxy (dielectric constant = 4.4)	FR4 epoxy (dielectric constant = 4.4)
Loss tangent	0.02	0.02
Substrate length	28.5 mm	30 mm
Substrate width	28.5 mm	30 mm
Substrate thickness	1.6 mm	1.6 mm
Length of centre patch	11 mm	11 mm
Width of patch	11 mm	11 mm
Length of parasitic patch	5 mm (base arm)	5 mm (base arm)
	4.71 mm (other)	4.71 mm (other)
Feed type	Coaxial feed	Coaxial feed

Table 2 Performance analysis

	Return loss (dB)	Bandwidth (GHz)
Un optimized	−22.8	2
Optimized	−26.5	2.5

Table 3 Comparison of earlier method with proposed method

Author	Return loss (dB)	Bandwidth (GHz)
Choo et al. [2]	−17	1.5
Wi et al. [3]	−26	1.5
Lu et al. [10]	−35	2
Proposed method	−26.5	2.5

5 Conclusion

The proposed antenna arrays are analyzed considering the parasitic patches to be triangular. The performance of the array is compared by changing the substrate length and width to get a better performance. The comparison results in delivering a return loss of −26.5 dB and impedance bandwidth of 2.5 GHz for the optimized design as compared to the unoptimized design. The proposed antenna will perform better in the specify frequency range as compared to the others.

References

1. Garg, R., Bhartia, P., Bahl, I.J., Ittipiboon, A.: Microstrip antenna design handbook. Artech House (2001)
2. Choo, H., Hutani, A., Trintinalia, L.C., Ling, H.: Shape optimisation of broadband microstrip antennas using genetic algorithm. Electron. Lett. **36**(25), 2057–2058 (2000)
3. Wi, S.H., Lee, Y.S., Yook, J.G.: Wideband microstrip patch antenna with U-shaped parasitic elements. IEEE Trans. Antennas Propag. **55**(4), 1196–1199 (2007)
4. Donelli, M., Azaro, R., Fimognari, L., Massa, A.: A planar electronically reconfigurable Wi-Fi band antenna based on a parasitic microstrip structure. IEEE Antennas Wirel. Propag. Lett. **6**, 623–626 (2007)
5. Wang, H., Huang, X.B., Fang, D.G.: A single layer wideband U-slot microstrip patch antenna array. IEEE Antennas Wirel. Propag. Lett. **7**, 9–12 (2008)
6. Piazza, D., Kirsch, N.J., Forenza, A., Heath, R.W., Dandekar, K.R.: Design and evaluation of a reconfigurable antenna array for MIMO systems. IEEE Trans. Antennas Propag. **56**(3), 869–881 (2008)
7. Chen, Z., Otto, S.: A taper optimization for pattern synthesis of microstrip series-fed patch array antennas. In: IEEE Wireless Technology Conference, EuWIT. European, pp. 160–163 (2009)
8. Chen, X.P., Wu, K., Han, L., He, F.: Low-cost high gain planar antenna array for 60-GHz band applications. IEEE Trans. Antennas Propag. **58**(6), 2126–2129 (2010)
9. Nashaat, D., Elsadek, H.A., Abdallah, E.A., Iskander, M.F., Elhenawy, H.M.: Ultrawide bandwidth microstrip patch array antenna using electromagnetic band-gap structure (EBG). IEEE Trans. Antennas Propag. **59**(5), 1528–1534 (2011)
10. Lu, J., Kuai, Z., Zhu, X., Zhang, N.: A high-isolation dual-polarization microstrip patch antenna with quasi-cross-shaped coupling slot. IEEE Trans. Antennas Propag. **59**(7), 2713–2717 (2011)
11. Xu, J., Zhao, M., Zhang, R., Lei, M., Gao, X., Huang, S., Bi, K.: A wideband F-shaped microstrip antenna. IEEE Antennas Wirel. Propag. Lett. **16**, 829–832 (2017)
12. Mishra, G.P., Maharana, M.S., Modak, S., Mangaraj, B.B.: Study of Sierpinski fractal antenna and its array with different patch geometries for short wave Ka band wireless applications. Procedia Comput. Sci. **115**, 123–134 (2017)
13. Minasian, A.A., Bird, T.S.: Complementary particle swarm antennas for next generation wireless communication systems. In: IEEE Wireless Communication Systems (ISWCS), pp. 895–898 (2012)
14. Minasian, A.A., Bird, T.S.: Particle swarm optimization of microstrip antennas for wireless communication systems. IEEE Trans. Antennas Propag. **61**(12), 6214–6217 (2013)
15. Schlosser, E.R., Tolfo, S.M., Heckler, M.V.: Particle swarm optimization for antenna arrays synthesis. In: IEEE Microwave and Optoelectronics Conference (IMOC), pp. 1–6 (2015)
16. Das, A., Mohanty, M.N., Mishra, L.P.: EA based optimized T-slot patch antenna for bandwidth improvement. IJET (Eng. J.) **7**(4), 1357–1360 (2015)

QoS Optimization of Wireless Sensor Network for Large Foodgrain Warehouse Monitoring Using NS-2

Dipak Shelar, Arvind Shaligram and Damayanti Gharpure

Abstract Warehousing of foodgrains, proper storage and protection from bad weather effects are important issues for assuring the food security for the people. With the use of wireless sensor network, the environmental parameters inside foodgrain warehouse can be monitored and corrective action can be taken. The deployment of wireless sensor network inside foodgrain storage is practically difficult because infrastructure inside the storage is different compared to other monitoring applications. Energy consumption is a crucial QoS parameter that affects sensor network lifetime. The present work carried out a performance evaluation of QoS parameters, viz. packet delivery ratio, throughput, average end-to-end delay and energy consumption using NS-2 simulator for foodgrain warehouse monitoring.

Keywords Wireless sensor network (WSN) · User datagram protocol (UDP) · NS-2 · Energy consumption · Packet delivery ratio

1 Introduction

In wireless sensor networks, sensor nodes are distributed in an application area to monitor as well as control different parameters cooperatively. They are useful in various applications such as environmental monitoring, home monitoring and control, medical, precision agriculture and industry. Each application of WSN consists of several sensor nodes and one base station called sink node. All sensor nodes sense

D. Shelar (✉)
Department of Electronic Science, New Arts, Commerce and Science College, Ahmednagar, Maharashtra, India
e-mail: dipakshelar27@gmail.com

A. Shaligram · D. Gharpure
Department of Electronic Science, Savitribai Phule Pune University, Pune, Maharashtra, India
e-mail: adshaligram@gmail.com

D. Gharpure
e-mail: dcgharpure@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_36

the parameters as per application requirement and send it to the sink. Sensor node data reaches the destination in one hop or multi-hops. While considering the data transfer from source to sink, we have to consider the parameters like delay, reliability, congestion status and node lifetime. All these parameters are important for maintaining the quality of service (QoS) of WSN [1].

Grains are the biggest source of foods in most of the countries. In many countries, grains are harvested once a year or seasonally. Therefore, to provide food to the population, produced foodgrains are mostly stocked in foodgrain warehouses. Foodgrains such as rice, maize, wheat, sorghum and millets are stored for few months to years, and this storage plays a crucial role in the economic system of developed and developing countries [2]. Foodgrain warehouses are intended for the storage and physical protection of bagged grain. In India, excess foodgrains are stored by State Warehousing Corporation (SWC), Food Corporation of India (FCI) and Central Warehousing Corporation (CWC) [3].

The wireless sensor network can be deployed inside the foodgrain warehouse for monitoring temperature, humidity, light intensity and CO_2 gas. This paper presents a simulation study of WSN to understand the effect of various parameters on the QoS of the network. In [4], we evaluate the performance of ZigBee-based wireless sensor network foodgrain storage monitoring. We have used WSN planer tool software to decide suitable arrangement of sensor nodes in the network, and NS2 simulator is used with IEEE 802.15.4 standard to study performance characteristics of the WSN. The result showed that the coordinator position at the center gives a good packet delivery ratio and throughput compared to other positions. Also, simulations related to the effect of path loss exponent and shadowing deviation showed that by increasing the value of the path loss exponent, end-to-end delay increases and packet delivery ratio decreases. In this paper, we have performed the simulation study of QoS parameters of WSN such as a packet delivery ratio, throughput, average end-to-end delay and power consumption by increasing the number of hops and distance between two nodes. Also, we have optimized the lifetime of wireless sensor network.

2 Related Work

This section gives the details of NS-2-based simulation studies carried out for WSN. Zhang et al. [5] used simulation software NS-2 for simulation study of wireless sensor network with 50 nodes. Packet delivery ratio, node density and average end-to-end delay were used for sensor network performance evaluation using MAC layer protocol IEEE 802.15.4. Gautam et al. [6] implemented wireless sensor network in NS2 and presented simulation details for creating a wireless sensor network. The packet delay and average energy consumption of the sensor network are evaluated using MAC protocol 802.11.

Panse et al.'s [7] work was focused on performance evaluation of VANET. The work carried out using SUMO and NS2 was based on data delivery rate, average end-to-end delay, residual energy and routing overhead. Xu et al. [8] designed wireless network model using four nodes, throughput and average end-to-end delay were measured, and results analysis was performed using NAM and gnuplot tools.

Charoo et al. [9] introduced multi-region pre-routing (MRPR) scheme for large-scale wireless ad hoc network. The pre-routing scheme used to have a good improvement in average energy consumption as compared to AODV protocol. In addition, the average routing time also improved. Bhavsar et al. [10] presented a model based on IEEE 802.15.4 standard for performance investigation of WSN in animal health monitoring application. Their simulation results reported better packet delivery ratio and less delay.

Zhu et al. [11] described two fundamental issues related to WSNs such as coverage and connectivity. These two parameters have a great impact on QoS of wireless sensor networks. They also discussed existing problems, challenges and summarized typical issues of coverage and connectivity in WSNs. Said [12] constructed a simulation environment for WSN using the NS-2 simulator and proposed system for guaranteeing WSN QoS. Wireless sensor network was divided into different groups where every group comprises of a number of nodes. Performance parameters, viz. packet loss, throughput, latency and sensor power consumption, were measured. Asif et al. [13] reviewed QoS routing protocols used in wireless sensor network. The work was focused on QoS satisfaction, challenges and requirements at each layer. QoS aware protocols for WSN are discussed comprehensively, and computational intelligence techniques for QoS management are described.

3 WSN Infrastructure for Foodgrain Warehouse

Foodgrain warehouses are used to store different types of foodgrains for longer duration. Temperature, humidity, light intensity and concentration of CO_2 gas are important physical parameters inside foodgrain storages. A wireless sensor network can be placed inside the foodgrain warehouse for monitoring these physical parameters. A wireless sensor network equipped with sensor, microcontroller, transceiver and battery can provide the information related to the environmental conditions inside foodgrain warehouse. With the help of information provided by wireless sensor network, a corrective action taken by the concerned authority avoids damage of foodgrains and various kinds of losses occur due to microorganisms, insects, mites, rodents and birds. The WSN-based foodgrain warehouse monitoring system will be practically implemented in Maharashtra State Warehousing Corporation Warehouse in Kedgaon, Ahmednagar. Figure 1 shows the proposed architecture of WSN and foodgrain stacks inside the foodgrain warehouse.

As foodgrain warehouse is of large size, it requires several numbers of wireless sensor nodes for covering entire warehouse. Maharashtra State Warehousing Corporation godown at Kedgaon, Ahmednagar (M.S.), has dimensions of 135 × 70 feet and total capacity of godown is about 1580 MT. It consists of 12 stacks, and each stack consists of 2500 foodgrain bags. The deployment of sensor nodes is also important because the foodgrain bags or sacks are arranged in a stack. In the present work, a

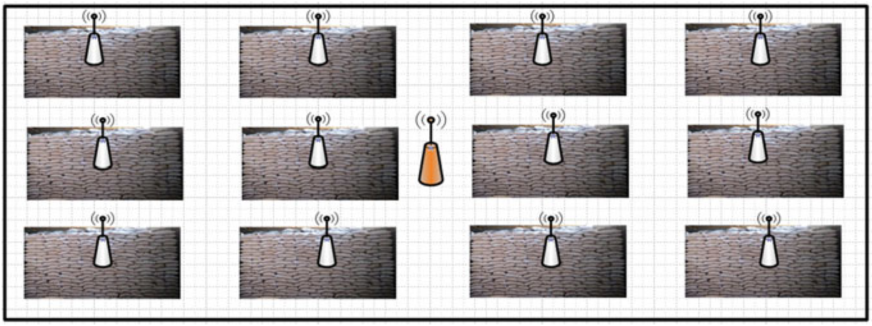


Fig. 1 Proposed architecture of WSN in foodgrain warehouse

simulation study of wireless sensor network inside foodgrain warehouse using NS-2 performed successfully. For this base station (N12) positioned at the center and 12 sensor nodes (N0–N11) are placed as one node in every stack. All source nodes sent the data to the base station in a single or multiple hops.

As all sensor nodes are placed inside foodgrain warehouse, simulation tests are performed using NS-2 with the shadowing propagation model, because receiving signal power fluctuates due to foodgrain bags and the warehouse walls cause dispersion and intersymbol interference.

4 Simulation Scenario

To perform simulation study of WSN inside foodgrain warehouse, we have used the NS-2 simulator, which is a discrete event-driven simulation tool and is open source. It is useful to study the dynamic nature of wire as well as wireless communication networks.

4.1 Performance Metrics

Definitions of important QoS parameters are stated below.

4.1.1 Packet Delivery Ratio (PDR)

PDR is the ratio of successfully received packets at sink to the number of packets sent from the source. A higher packet delivery ratio value indicates that the network performance is good and packet loss is less.

4.1.2 Throughput

The throughput is the maximum rate at which the information is transferred from source to sink. It is measured as the number of packets arriving at the sink in bits per second (bps).

$$\text{Throughput} = \frac{(\text{Number of bytes received} \times 8)}{(\text{Time} \times 1000)}$$

4.1.3 Average End-to-End Delay

The measuring of the average time taken by each packet to transfer the data from the source to sink is called as average end-to-end delay. If it goes higher, the network suffers congestion.

4.1.4 Energy Consumption

Amount of energy used by sensor node is referred as energy consumption. Energy consumption is important parameter that decides the era of sensor network. Energy consumption can be minimized by putting the sensor nodes in low power modes for most of the time and active for short duration.

4.2 Basic Simulation Experiments

Basic experiments are performed using two sensor nodes with a shadowing propagation model to study the effect of distance variation and packet size. The range for communication of sensor node is fixed to 40 m. Node N0 is a source node and node N1 is a sink node. Distance between two nodes is varied from 10 to 60 m. The experiment is repeated for packet size of 20, 40, 60, 80 and 100. Performance parameters such as packet delivery ratio (PDR), throughput and average end-to-end delay are studied.

The PDR is 100% up to 30 m distance and more than 90% for the distance between 30 and 40 m. Figure 2 shows that packet size does not affect the packet delivery ratio.

Figure 3 shows that as we increase the packet size, throughput is increased. Throughput is not much affected by an increased distance between nodes. As the communication range is 40 m, throughput reduces sharply for the distance greater than 40 m.

Average end-to-end delay increases with distance, is independent of packet size and increases sharply for distance greater than 40 m, as shown in Fig. 4.

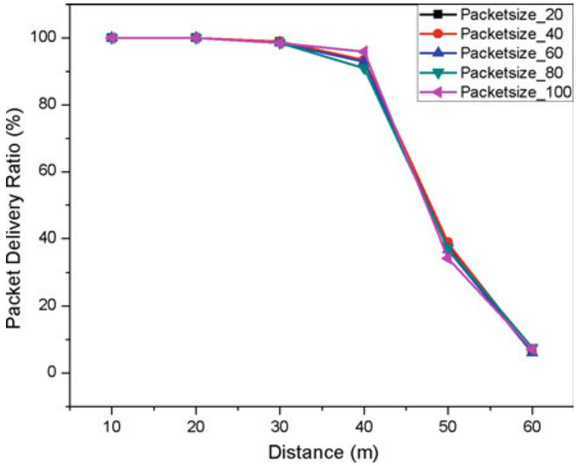


Fig. 2 PDR versus distance

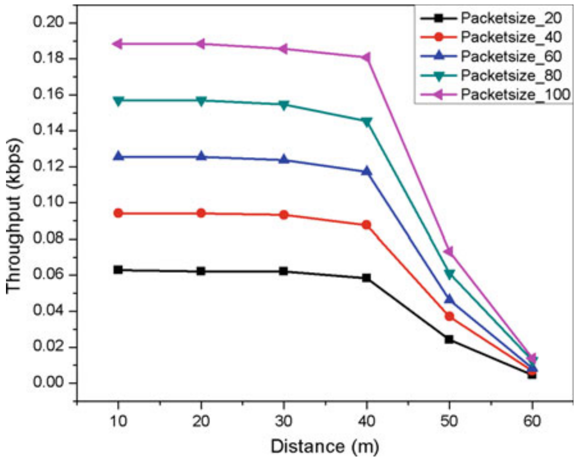


Fig. 3 Throughput versus distance

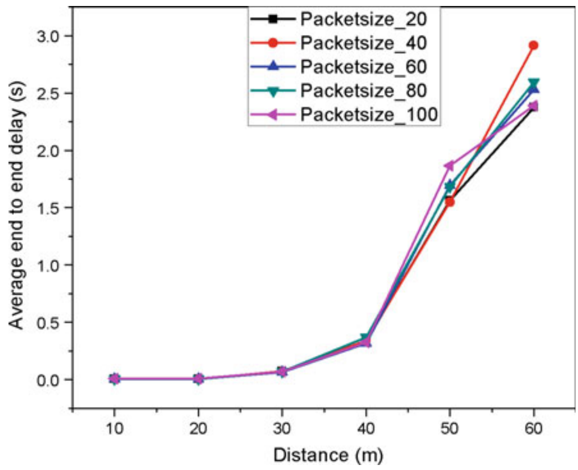
4.3 Study of Energy Consumption

The study of energy consumption is performed by varying the distance between two nodes and increasing the number of hops.

NS2 simulations are run for 1000 s and graphs are plotted for energy used against distance. The energy model is as follows:

```
$ns node-config  -energyModel "EnergyModel"\
                 -initialEnergy 100\
                 -rxPower 0.3\
                 -txPower 0.3\
                 -transitionTime 0.0015\
```

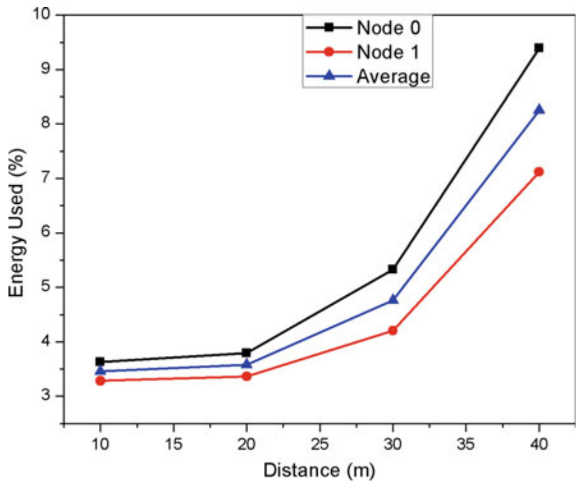
Fig. 4 Average end-to-end delay versus distance



All nodes are powered with 100% energy. The first experiment was conducted to study energy consumption with increasing distance between two nodes by 10 m step.

Figure 5 shows that as we increase the distance between two nodes, the energy consumption also increases and lifetime decreases. As the number of hops was increased, the amount of energy consumption was reduced compared to single-hop communication as shown in Fig. 6.

Fig. 5 Energy consumption versus distance



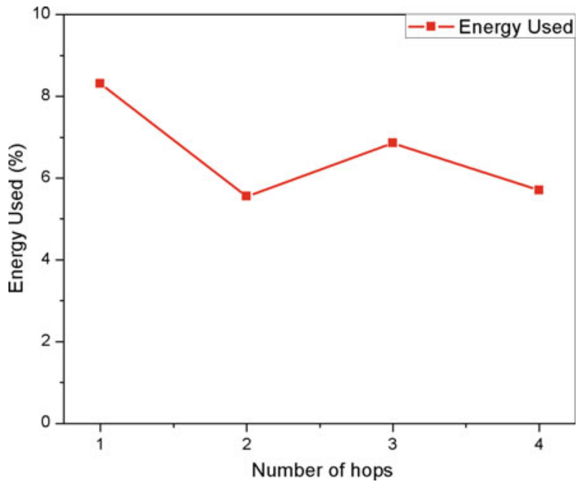


Fig. 6 Energy used versus number of hops

5 Network Scenario for Foodgrain Warehouse Monitoring

For foodgrain warehouse monitoring application, 12 sensor nodes are used as source nodes and one base station is used as a sink node or base station in the NS-2 simulator. One sensor node is assigned to each foodgrain stack, and base station is placed at the center as shown in Fig. 1, with the parameters set as shown in Table 1.

Sensor nodes are deployed using Cartesian arrangement. Figure 7 shows NAM window of the NS-2 simulator, where sensor nodes send data toward base station.

For foodgrain warehouse simulation, we have performed two experiments using NS-2 simulator. Perl script was used for analysis of trace file. In experiment 1, all nodes send the data toward the base station simultaneously, which gives the packet delivery ratio 66.57%, a throughput of 1.24 kbps and average end-to-end delay of 9.5 ms, which is shown in Fig. 8.

Table 1 Simulation parameters

Sr. No.	Parameters	Details
1.	Number of nodes	13
2.	Radio propagation model	Shadowing
3.	Antenna model	Antenna/omni antenna
4.	MAC type	802.15.4
5.	Routing protocol	AODV
6.	Internet protocol	UDP
7.	Traffic type	cbr
8.	Packet size	512
9.	Simulation area	50 × 30 m
10.	Simulation time	1000 s

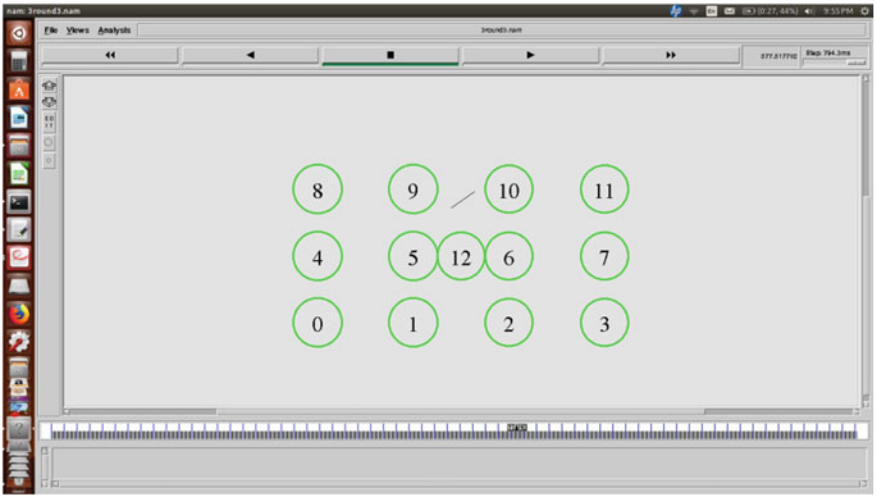


Fig. 7 NAM window of NS-2 simulator

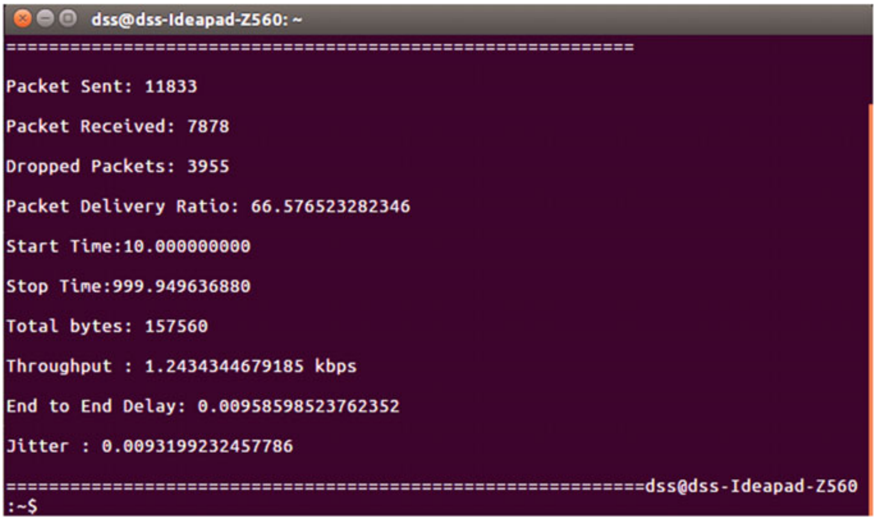


Fig. 8 Network performance while all nodes send data simultaneously

To improve the packet delivery ratio, we have performed experiment 2 in which each node sends data one after another in round-robin technique. This method improves the packet delivery ratio up to 75% and energy consumption is also reduced.

Figure 9 shows that packet delivery ratio is improved, throughput is reduced and average end-to-end delay is increased. But as the foodgrain warehouse monitoring does not require real-time monitoring, average end-to-end delay can be acceptable.

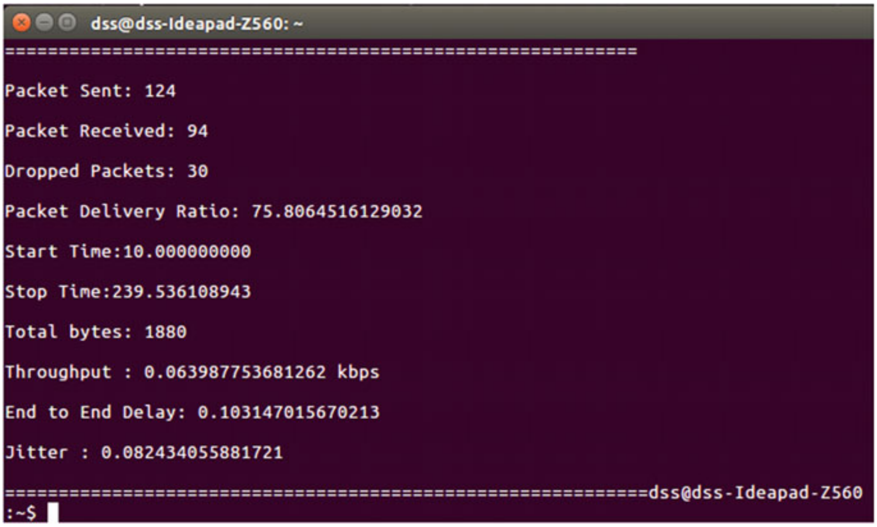


Fig. 9 Network performance while all nodes send data at specific time interval

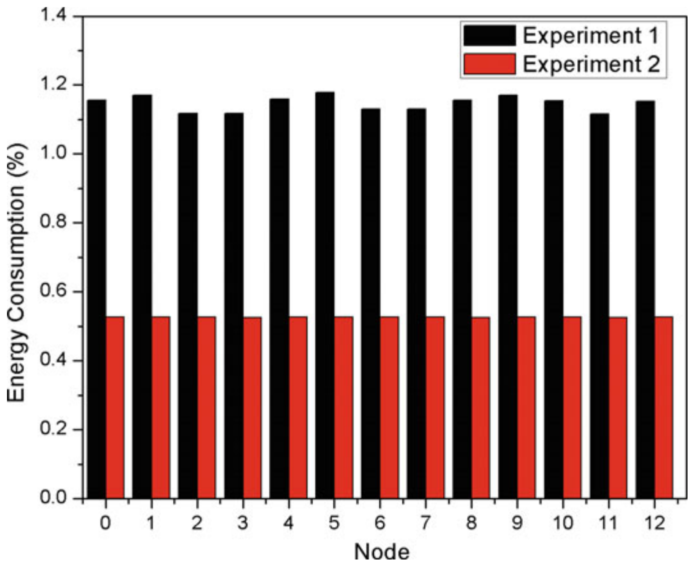


Fig. 10 Energy consumption versus node

Also, there is no need to send sensor data continuously to base station, and throughput is acceptable.

Energy consumption of each node is calculated and graph is plotted. Figure 10 shows energy consumption of each node for both experiments. Energy consumption of sensor nodes in experiment 1 is more than experiment 2 where sensor nodes send the data at specific interval which saves the energy.

6 Conclusion

The wireless sensor network is useful in different types of monitoring applications. In this work, simulation study of wireless sensor network for foodgrain warehouse monitoring is performed using the NS-2 simulator with IEEE 802.15.4 standard. Some basic simulation experiments are performed with two nodes. The experimental results show that the packet size does not affect the packet delivery ratio, throughput escalates with an increase in packet size and average end-to-end delay remains constant for different packet sizes. The delay increases with an increase in distance between two nodes as expected. Energy consumption increases with the distance and decreases as we increase the number of hops. In a simulation study of foodgrain warehouse monitoring application, packet delivery ratio increases as the duty cycle is reduced. Energy consumption is reduced with periodic data transmission. The lifetime of wireless sensor network can be improved by reducing the duty cycle and use of multi-hop sensor network.

References

1. Deshpande, V.S.: Handling the energy in wireless sensor networks. In: Wireless Sensor Networks from Theory to Applications, pp 369–387. CRC Press (2014)
2. Proctor, D.L.: Grain storage techniques: evolution and trends in developing countries. In: Food and Agriculture Organization of the United Nations (FAO) Rome (1994)
3. Esther Magdalene Sharon, M., Kavitha Abirami, C.V., Alagusundaram, K.: Grain storage management in India. J. Postharvest Technol. 12–24 (2014)
4. Shelar, D.S., Gharpure, D.C., Shaligram, A.D.: Performance analysis of ZigBee based wireless sensor network for grain storage monitoring. Int. J. Comput. Sci. Inf. Technol. **6**(6), (2017)
5. Zhang, F., Zhou, H., Zhou, X., Liao, Q.: Simulation of Temperature Measurement System for Grain Storage Based on Zigbee Technology. IEEE (2009)
6. Gautam, G., Sen, B.: Design and simulation of wireless sensor network in NS2. Int. J. Comput. Appl. (2015)
7. Panse, P., Shrimali, T., Dave, M.: Performance evaluation of vehicular ad hoc network using SUMO and NS2. In: Proceedings of the International Congress on Information and Communication Technology, Advances in Intelligent Systems and Computing, pp. 127–135
8. Xu, S., Yang, Y.: Protocols Simulation and Performance Analysis in Wireless Network Based on NS2. IEEE (2011)
9. Charoo, M.A., Mishra, D.K.: Multi-region pre-routing in large scale mobile ad hoc networks (MRPR). In: Proceedings of the International Congress on Information and Communication Technology, Advances in Intelligent Systems and Computing, pp. 637–645
10. Gehlaut, S., Koti, J., Sakhardande, K.: To Improve the Lifetime of Wireless Sensor Network. IEEE (2016)
11. Zhu, C., Zheng C., Shu, L., Han, G.: A survey on coverage and connectivity issues in wireless sensor networks. J. Netw. Comput. Appl. (Elsevier)

12. Said, O.: Performance evaluation of WSN management system for QoS guarantee. EURASIP J. Wireless Commun. Netw. (2015)
13. Asif, M., Khan, S., Ahmad, R., Sohail, M., Singh, D.: Quality of service of routing protocols in wireless sensor networks: a review. IEEE Access. pp. 1846–1871 (2017)

Monitoring and Analysis of Power Quality Issues in Distribution Network—A Case Study

V. S. Jape, D. S. Bankar and Tejaswini Sarwade

Abstract Tracking of system's overall performance in phrases of Power Quality disturbances and its ill effects on distribution network is growing attention of application toward tracking of Power Quality indices like voltage sag, voltage swell, and harmonics. This paper introduces new index as Power Quality Distortion Index (DI) which offers the contribution of every load on the total distortion of the power system. The presented system set the basis of tracking and analysis of Power Quality indices for distribution network [1]. The characterization of Power Quality problems is found through non-widespread currents, voltages, and frequencies. The Power Quality is related to variations of supply voltage in the form of sags, swells, harmonics, transients, and so on. These issues result in deterioration of energy delivered to end customers, also various disturbances are generated. The loads used by consumers also account for deterioration of Power Quality. Penalties are charged for low power factor loads but at the same time negligence to sizable existence of Power Quality parameter issues like voltage sags and harmonic distortions. Introduction of Custom Power Devices (CPD) is an effective solution over Power Quality problems in distribution network [2]. This paper presents diverse Power Quality indices like Distortion Index as the computational parameter of Power Quality and design of Dynamic Voltage Restorer (DVR) to compensate voltage sag in the system. Effect of DVR on Distortion Index (DI) is also observed, and outcomes are analyzed with the assist of MATLAB/Simulink.

Keywords Power quality · Harmonics · Custom power devices (CPD) · Dynamic voltage restorer (DVR) · Distortion index (DI)

V. S. Jape (✉) · D. S. Bankar
BVDUCOE, Pune, India
e-mail: jape_swati@yahoo.co.in

D. S. Bankar
e-mail: dsbankar@bvucoep.edu.in

T. Sarwade (✉)
PES's MCOE, Pune, India
e-mail: sarwadet@gmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_37

1 Introduction

The power system layout has become extra complicated every day. It comprises numerous generating stations and loads whose interconnection is through numerous transmission and distribution strains. Also, increased use of power electronics based circuits leads to growth in nonlinearity. These sorts of loads are sensitive to Power Quality parameters including voltage sags, swells, harmonics, sparkles, fluctuations, etc.

Existence of harmonic distortion is due to deviation in voltage, current, or fundamental frequency. The voltage sag is a drop off in root mean square value of voltage or current usually between 0.1 and 0.9 per unit at power frequency lasting for half cycle to 60 s. Fault clearing time refers to the variety of 3 to 30 cycles [2, 3]. Voltage swell is rise in root mean square value of voltage at power frequency between 1.1 and 1.8 per unit lasting for half cycle to 60 s. Transients are the part of change within the variable disappeared in the course of alteration from one consistent state to every other.

2 Power Quality Improvement

Enhancement in efficiency of power system desires continuous working which in addition attributes the importance of monitoring for any form of disturbances that is to be taken into consideration as Power Quality issues and also offers corrective measures on such problems to restrict the occurrences of those events. Mitigation of Power Quality problems is difficult with the aid of the use of traditional equipments including tap changing transformers, lightning arresters, surge arresters, capacitor banks, and many others. Also existence of power electronics devices performs a vital role to decide performance of PQ issues. Power electronics based solutions are categorized as as FACTS controllers for transmission systems and Custom Power Devices (CPDs) which contribute fundamental function in Power Quality improvement of distribution network. Various CPDs are available consisting of Distribution Static Compensators (D-STATCOM), Active Power Filters (APF), Dynamic Voltage Restorer (DVR), Battery Energy Storage Systems (BESS), Static VAr Compensator (SVC), and so on. Dynamic Voltage Restorer (DVR) is identified as more efficient device among all devices

2.1 Control Methods of DVR

Control of DVR circuit topology is critical element for design and modeling factor of view, which includes voltage disturbances identification with proper recognition strategies. Voltage supply converter at once impacts DVR's overall performance

Fig. 1 DVR controls

A)Linear	B)Nonlinear
1. Feedback Control	1. Fuzzy logic Control
2. Feedforward Control	2. ANN Control
3. Multiloop Control	3. SVPWM Control

because it satisfies reactive power requirement. Therefore, it is taken into consideration as important part of DVR [4]. The inverter control strategies are usually categorized as follows (Fig. 1).

3 Monitoring and Analysis of Power Quality Parameters: A Case Study

To analyze the impact of Power Quality parameter, a 11 kV/440 V, 200 KVA substation is considered which supplies power to an educational institute. The effect on voltage sag, voltage harmonic distortion, and current harmonic distortion is observed by implementing the layout of DVR system in MATLAB/Simulink as shown in Fig. 2.

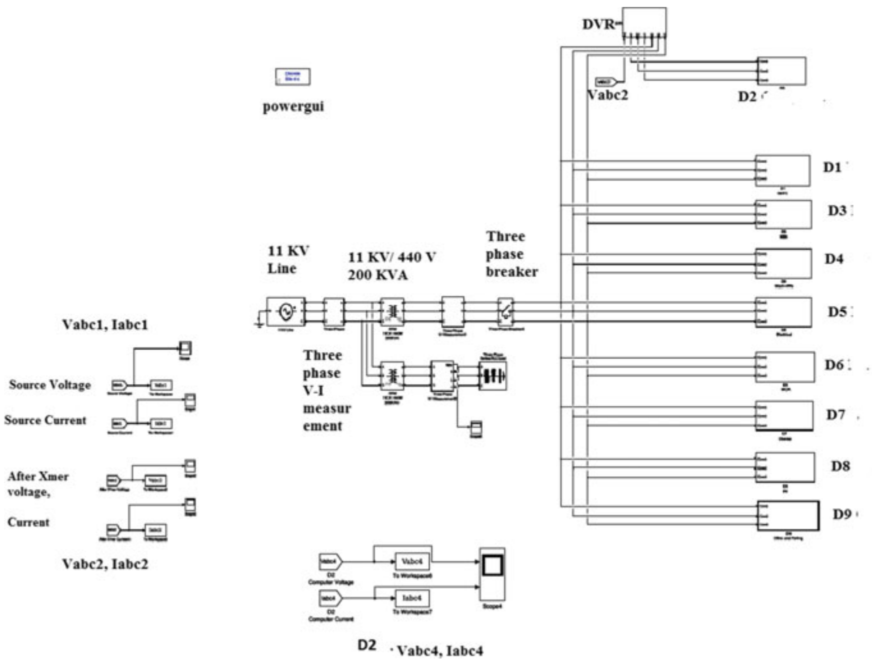


Fig. 2 Simulink diagram of institution with DVR

Table 1 System parameters

S. No.	Parameter	Value
1	Transmission line voltage	11 kV
2	Transformer	11 kV/440 V, 200 KVA
3	Line frequency	50 Hz
4	Nominal voltage	440 V
5	Transition time	[0 1]

DVR is a sequence-connected Custom Power Device that is injected in between distribution network and the load. Basic characteristic of DVR is to inject a required compensation voltage to mitigate Power Quality issues.

Nine distinct departments (D1-D9) of the institute are taken into consideration. The overall load linked to the system is 600.68 kW. DVR is connected between supply side and department D2 considering different load conditions.

Table 1 shows system parameters.

The DVR subsystem is shown in Fig. 3.

3.1 Results

Figure 4 indicates source voltage waveform, wherein voltage sag is observed in among 0–1 s. The voltage and current waveforms of department in which nonlinear loads are substantial with inclusion of DVR are shown in Fig. 5.

Also, Figs. 6 and 7 indicate voltage THD values earlier than and after DVR connection for a linear and nonlinear load (Figs. 8 and 9).

THD readings are determined for nonlinear load. Also, current THD readings are observed for linear and nonlinear loads before and after DVR implementation inside the system, and effects are represented in Table 2.

Table 2 indicates the decreased values of THD because of inclusion of DVR in the system. Same results can be observed after connecting DVR in other departments of device with distinct loads.

3.2 Distortion Index Calculations

Along with THD analysis, another parameter known as Distortion Index (DI) [5] is calculated for the system and is analyzed through simulation. Following formulae are used to calculate DI.

Following formulae are considered to calculate Distortion Index.

Let us consider V_1 and I_1 be the fundamental voltage and current. V_3, V_5, $V_7 \ldots$ V_n and I_3, I_5, $I_7 \ldots$ I_n are orders of harmonic voltages and currents, respectively.

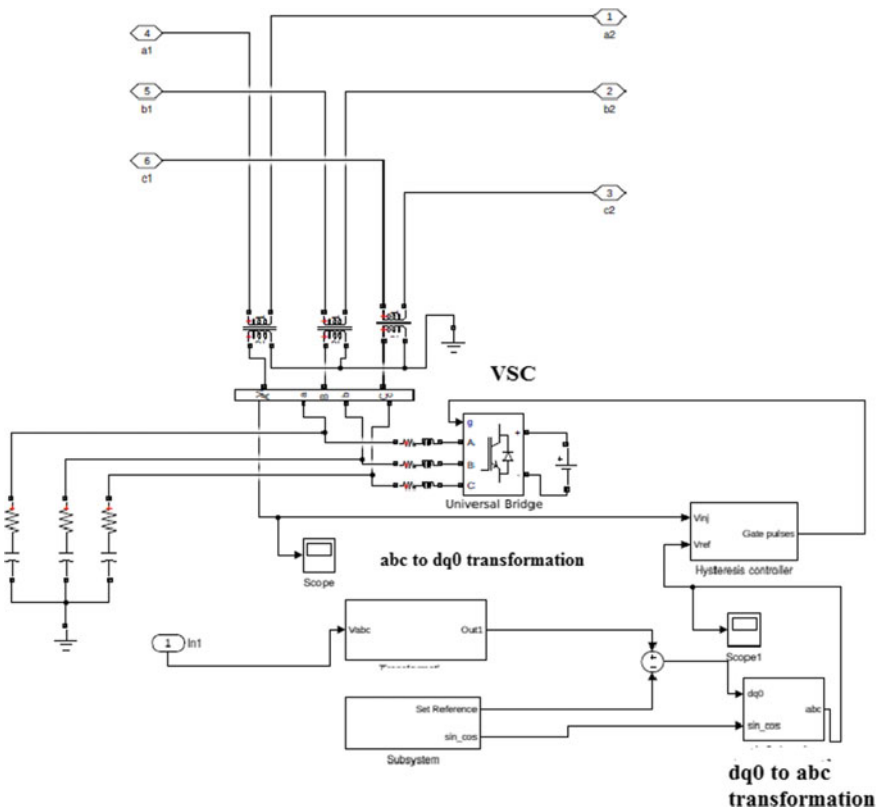


Fig. 3 DVR subsystem

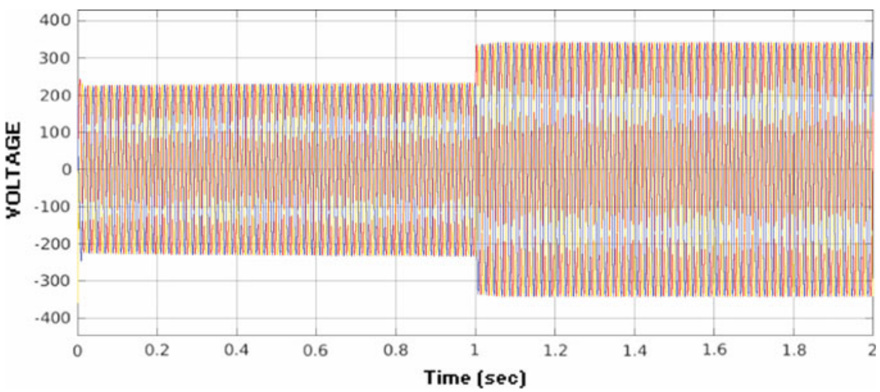


Fig. 4 Source voltage waveform

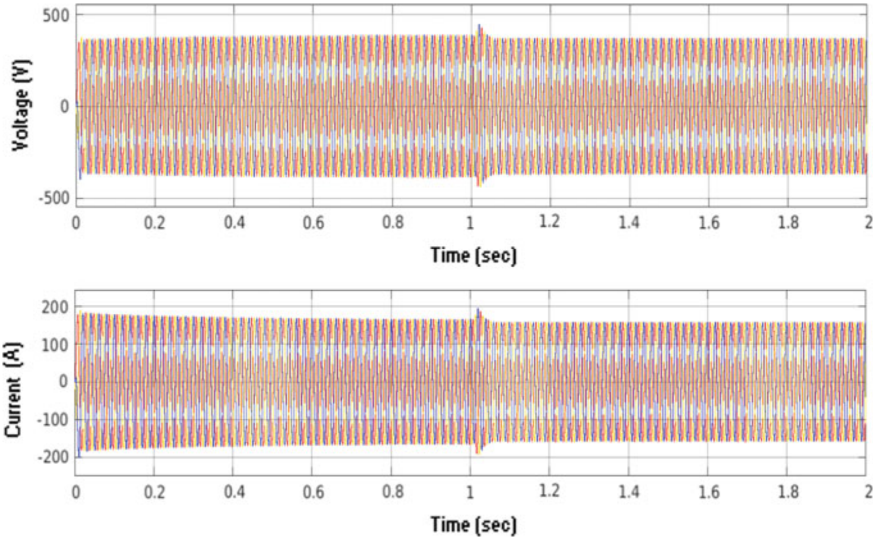


Fig. 5 Voltage and current waveforms of D2

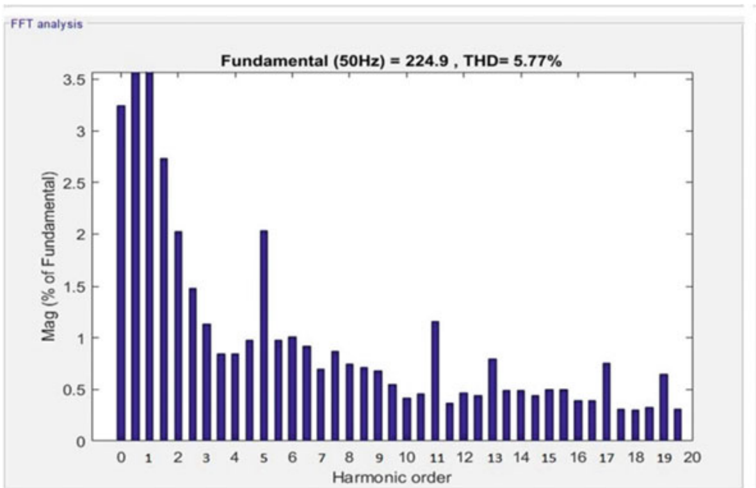


Fig. 6 Voltage THD for nonlinear load (before DVR connection)

$$I_H = \sqrt{I_3^2 + I_5^2 + I_7^2} \tag{1}$$

$$V_H = \sqrt{V_3^2 + V_5^2 + V_7^2} \tag{2}$$

$$\text{Total Harmonic Distortion Voltage } V_{\text{THD}} = V_H / V_1 \tag{3}$$

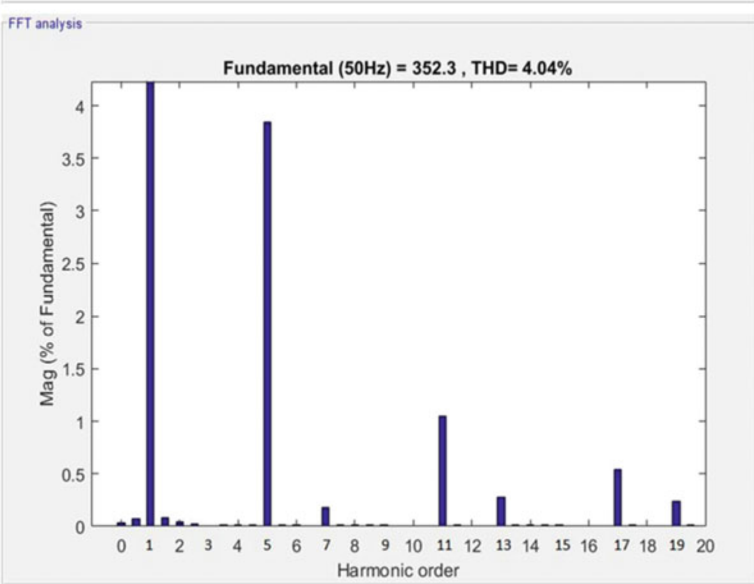


Fig. 7 Voltage THD for nonlinear load (after DVR connection)

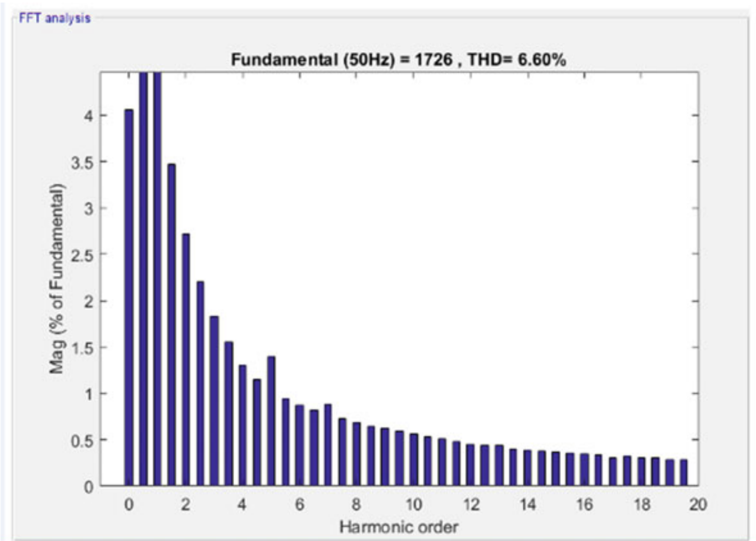


Fig. 8 Current THD for nonlinear load (before DVR connection)

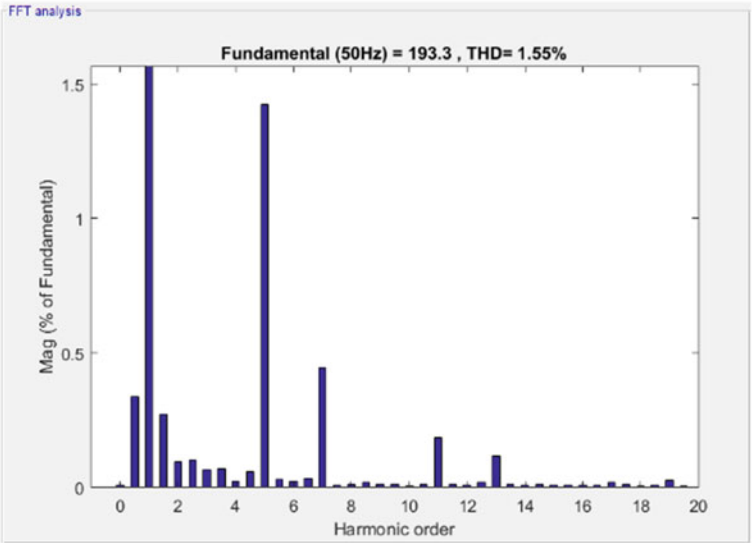


Fig. 9 Current THD for nonlinear load (after DVR connection)

Table 2 System results

Load type	Voltage THD (%)		Current THD (%)	
	Before DVR	After DVR	Before DVR	After DVR
Linear	4.92	1.90	4.77	0.66
Nonlinear	5.77	4.04	6.6	1.55

$$\text{Total Harmonic Distortion Current } I_{\text{THD}} = I_H / I_1 \tag{4}$$

$$\text{Fundamental Apparent Power FAP} = V_1 I_1 \tag{5}$$

$$\text{Current Distortion Power CDP} = V_1 I_H \tag{6}$$

$$\text{Voltage Distortion Power VDP} = V_H I_1 \tag{7}$$

$$\text{Harmonic Distortion Power HDP} = V_H I_H \tag{8}$$

$$\text{Non-linear Apparent Power NAP} = \sqrt{\text{CDP}^2 + \text{VDP}^2 + \text{HDP}^2} \tag{9}$$

$$\text{Total Apparent Power TAP} = \sqrt{\text{FAP}^2 + \text{NAP}^2} \tag{10}$$

$$\text{Distortion Index DI} = \frac{\text{NAP}}{\text{FAP}} * 100 \tag{11}$$

Table 3 shows DI values for departments D1–D9 in the system before DVR connection and after DVR connection. Fig. 10 depicts graphical representation of DI variation with various departments.

4 Conclusions

From various observations, the outcomes are compared. It has been highlighted that the harmonic contents and Distortion Index are reduced significantly with the inclusion of DVR in system. In this regard, a new procedure is presented which evaluates the Distortion Index with DVR and without DVR for nonlinear loads connected to the power distribution network.

Table 3 Distortion Index

Departments	Distortion index (%)	
	Before DVR	After DVR
Source voltage	2.6507	1.9964
After transformer voltage	34.9758	31.8338
D1	34.995	25.4469
D2	41.9388	37.2259
D3	48.5182	25.3123
D4	35.0075	25.3128
D5	34.9939	25.312
D6	48.4725	25.3124
D7	34.992	25.3089
D8	34.9906	25.3069
D9	34.9951	25.3094

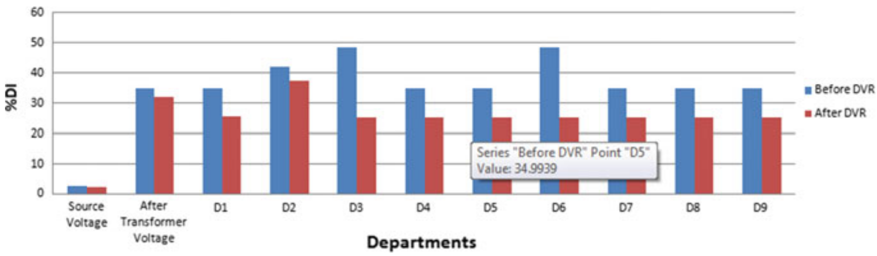


Fig. 10 Graphical representation of DI variation with departments

Contribution of Power Quality indices is the most important subject to reveal Power Quality stages often. From observations, Total Harmonic Distortion displays most effective voltage or the current distortion, while Distortion Index (DI) pertains to distortion in distribution power. Hence, DI can be introduced as main Power Quality index for identification and evaluation of Power Quality levels present inside the distribution network. The regulatory authority can consider this index as “Quality factor” or “Penalty Factor” [6, 7].

Acknowledgements The research work is being carried out in Bharati Vidyapeeth Deemed University COE, Pune, in consent with MSEDCL, Pune. The authors desire to thank authorities of BVDUCOE, Pune, for granting permission to publish the work.

References

1. Jape, V.S., Bharadwaj, D.G.: Substation monitoring system (SMS) for remote power quality monitoring and analysis of power distribution network using object oriented web enabled method. IJCTA **10**(9), 533–539 (2017) International Science Press. ISSN: 0974–5572
2. Sarwade, T., Jape, V.S., Bharadwaj, D.G.: Power quality problems mitigation using dynamic voltage restorer (DVR) using pi controller and fuzzy logic controller. IJET **2**(12), 214–218 (2018)
3. Dugan, R.C., McGranghan, M.F.: Electrical Power System Quality, 2nd edn. Tata McGraw Hill
4. Raval, C.H., Doorwar, A.: An overview of various control methods of DVR. In: ICCPCT, pp. 53–57 (2014)
5. Vlahinic, S., Brnobic, D., Stojkovic, N.: Indices for harmonic distortion monitoring of power distribution systems. Instrumentation. 12–15 (2008)
6. IEEE Std 519–1992: IEEE recommended practices and requirements for harmonic control in electrical power systems
7. IEEE STD 1159–1995: IEEE recommended practice for monitoring electric power quality

Vulnerability Analysis of Real-Time Operating Systems for Wireless Sensor Networks

M. Rajesh and B. Sreevidya

Abstract The faultless operation of systems which are safety-critical like the anti-lock braking system (ABS), fuel control system (especially during cruise control), traction control system, etc., is equally dependent on both the logical as well as the temporal correctness of the output generated, and hence, systems like these can be called as the real-time systems (RTS), specifically, hard real-time systems. In this paper, various security issues relevant to such automotive systems have been studied and tested. Based on the observations made, a feasibility study is done on using OSEK/VDX for wireless sensor network applications.

Keywords Automotive systems · Connected car · Security · Wireless sensor network

1 Introduction

The correctness of any real-time system (RTS) depends equally on both, the logical as well as the temporal correctness of the output generated [1]. Some examples of real-time systems are—aircraft control, temperature monitoring in nuclear power plants, mobile phones, online ticket reservation systems, etc. There are two kinds of real-time systems—hard RTS and soft RTS. A hard real-time system is a system for which the output must be produced within the given deadline, missing which, would result in catastrophic consequences such as loss of life. The aircraft control, temperature monitoring in nuclear power plants, etc., can be categorized as hard RTS. A soft real-time system is a system that should deliver the output within the deadline, but the penalty for missing the deadline is not as severe as that of the hard

M. Rajesh (✉) · B. Sreevidya (✉)
Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, India
e-mail: m_rajesh@blr.amrita.edu

B. Sreevidya
e-mail: b_sreevidya@blr.amrita.edu

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_38

RTS. The systems such as digital cameras, mobile phones, audio-visual systems, and online ticket reservation systems can be categorized as soft RTS.

A real-time operating system (RTOS) is the heart of any real-time system. The operating system that will serve a real-time application, process data, and produce the output without much of buffering delays and latencies is referred as real-time operating system. The processing time requirement for an RTOS (including all sorts of delays and latencies) is typically shorter than or equal to a few tenths of seconds. Any general RTOS will usually focus on central time-critical services like task scheduling, time management, inter-process communication, and management of the device driver layer [2]. Some of the examples of RTOS are—TinyOS, OSEK/VDX, VxWorks, QNX, Windows CE, etc.

In this paper, OSEK/VDX is chosen as the RTOS to be analyzed for vulnerability analysis. OSEK/VDX is a set of standards for distributed, real-time architecture that was developed by a consortium of European automobile manufacturers and suppliers in conjunction with the University of Karlsruhe, Germany [3, 4]. OSEK/VDX is scalable, portable, configurable, and supports time-triggered architectures. It has a statically defined OS model that helps maintain low time and space operating costs by overcoming the need for bookkeeping, and also in the detection of dynamic allocation bugs. Considering all these features and specifications, the work focuses on analyzing the feasibility and compatibility of using OSEK/VDX which is originally designed for automotive applications to be used in wireless sensor network applications. To achieve this, a study is carried out on the security aspects of OSEK/VDX on an in-vehicular network application.

2 Literature Review

Ever since the automotive industry has introduced a new communication schemes, such as vehicular network and intra-vehicle communication technologies have gained prominent public usage, it has been pointed out that the safety of the vehicle is under threat due to its connection to the external network. As a result, the security of automotive control systems has gained more importance. An electronic control unit (ECU) does most of the functions in a vehicle. There are several number of ECUs in a vehicle. The data exchange between these ECUs is done using a communication protocol for an in-vehicular network known as the controller area network (CAN). The data used for the controlling of the vehicle, such as the velocity of the vehicle, the state of the brake are all handled by the CAN. This kind of data can be considered as safety-critical because the safety of the vehicle lies entirely dependent on the logically and temporally correct transfer of this data to and from the ECUs to the appropriate actuators for an appropriate and an accurate action at critical times.

To protect this safety-critical data in the in-vehicular communication network, there are certain standards laid and followed by certain countries and companies [5]. Connected cars enable many functions, such as remote preconditioning, keyless entry, wireless tire pressure monitoring, downloading of route maps, live traffic information, uploading of vehicle diagnostic information for predictive maintenance, and over-the-air (OTA) update of the ECU programs. With the enabling of these features, the connected car opens an unintended wireless access to the gateway to the vehicle called the body control module (BCM). Thus, the BCM now becomes an easy point of access for attackers. By unauthorized access to the vehicle, simple exploits such as information loss, vehicle theft, to complex incidents such as causing accidents remotely and invisibly [6].

2.1 Security Threats in Connected Cars

There can be three areas which are vulnerable as per the safety aspects of a connected car [6]:

- The group of under-the-hood elements like ECUs, the in-vehicle network, and the gateway
- Use of mobile devices
- Use of vehicular cloud computing.

On-board diagnostics (OBD-II) port is a gateway which is used for reading and writing control information from and to the network inside the vehicle. The same is used for software update of ECUs. The information required to access the OBD-II port is to be made available to all maintenance garages [7]. Sagstetter et al. [7], also states that the OBD-II is easiest attack points, as the hacker can get complete access to all ECUs; and since the vehicle network is fully connected, the in-vehicle network and all the ECUs connected to it can be manipulated by any hijacked ECU. Hence, any ECU accessible from outside the vehicle (especially the wirelessly accessible ones) becomes a potential intrusion point. Weak security measures in the wireless access points to the car, as in the case of keyless entry systems allowed the unlocking of the car without the car key, and tire pressure monitoring systems (TPMS) might allow data readout [7].

According to Bécsi et al. [6], with the connection of mobile devices, threats can come from various apps, SMS, Web browsers, and other holes that malicious hackers can exploit; and hence, mobile devices will always remain as the most vulnerable points in the connected cars, as these devices are ideal for spoofing due to their role in the network. Even though the code can be protected in several ways, protection of the device's actions can be ensured by a carefully designed authentication.

Vehicular cloud computing is a recent hybrid technology that is a combination of various networks, such as vehicular ad hoc networks (VANETs), wireless sensor networks (WSNs), mobile ad hoc networks (MANETs), and cloud computing [8]. It has a noteworthy impact on managing traffic and ensuring safety on road by using resources present on the vehicle, such as computing, storage, and Internet to make decisions.

Vehicular cloud computing (VCC) security is a very important aspect of safeguarding the privacy and maintaining the trust of its users. Some of the regularly faced security issues in this area are: Denial of services (DoS), data tampering, information disclosure, and other network attacks such as identity spoofing, and Sybil attack [8].

2.2 RTOS Security

As most of the automotive control systems are considered as hard real-time systems, their safety and security lie with the security of the RTOS used. The connection of an RTS to a network means that it not only is susceptible to the security issues caused by its RTOS's design, but also to the security issues caused due to its interfacing to external networks [1]. Denial of determinism (DoD) is a term introduced and analyzed in [2], which describes the attacks against hard real-time systems, which aim to deny the logically and temporally accurate services.

(a) Security within the RTOS: The performance of an RTOS can slow down, causing the deadline miss leading to the failure of the RTS, due to the security issues such as

    (1) Modification of arbitrary memory
    (2) Manipulation of system entities
    (3) Forging of inter-process communication (IPC) objects
    (4) Manipulation of the system clock
    (5) Priority inheritance (PI) chain attack
    (6) Buffer overflow.

(b) Security of RTOS in a network: RTOS usually controls certain automated machinery and is often linked to a monitoring system that connects to other RTOS for management of an entire facility. The intruders can reprogram the RTOS in such a network, to either crash or to harm the one that controls the physical machinery. These control networks are usually called as supervisory control and data acquisition (SCADA) or distributed control systems (DCS). The security of both intra-RTOS communication and SCADA is similar to that of non-RTOS and computer networks [1]. The vulnerabilities in a network could be due to the factors like broadcasting behavior, deficiency in infrastructure, and environment being hostile [9, 10]. Most of the network attacks fall into any one of the following categories [10–12]

    (1) In-vehicular network attacks
    (2) Sensor-based attacks
    (3) Spoofing/impersonation attacks
    (4) Blackhole/sinkhole attacks
    (5) The Sybil attack
    (6) Wormholes
    (7) Man-in-the-middle attack.

Most of the automotive systems are hard real-time systems. For the faultless operation of any real-time system (RTS), two kinds of security issues are to be considered—security within the RTOS and security of the RTOS in a network. Through this literature survey, it was found that most of the research and study has been done in the aspect of security of RTOS in a network and not much study has been done in the aspect of security within the RTOS. Hence, the scope of this paper is channeled toward studying and testing some of the various security issues that might exist within the chosen RTOS.

3 Design and Implementation

3.1 The OSEK/VDX Architecture

OSEK/VDX is an industry standard to define operating systems that interface automotive electronic systems. It provides the environment for the continuous development of the ECU software and helps building compatible applications. OSEK is a specification for a single-core processor. The advantages of OSEK/VDX are—portability and reliability [13].

There are four major parts of the OSEK/VDX specification. They are:

1. The operating system (OS)—specifies the base services of the real-time kernel
2. The communication (COM)—specifies the communication services
3. The OSEK/VDX network management (NM)—specifies the network management services
4. The OSEK implementation language (OIL)—specifies the configuration of an implementation (Fig. 1).

Since the scope of this paper is channeled toward the study and testing of security issues within the chosen RTOS, the study of the architecture is limited to the

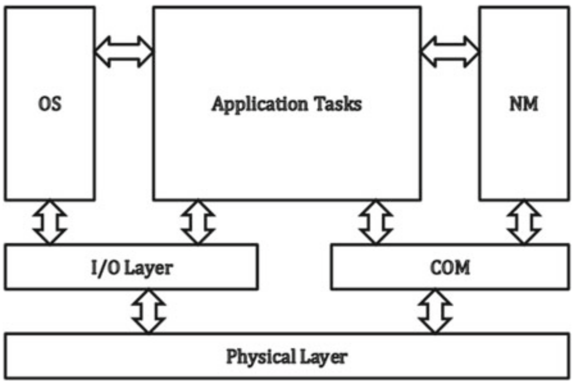


Fig. 1 Block diagram of the OSEK/VDX architecture

OSEK/VDX OS specification. The OSEK/VDX-based operating system is meant for distributed embedded control units. It provides services like task management, scheduling policy, synchronization, resource management, interrupts, inter-process communication, etc.

3.2 Implementation

The first step in taking control of a program flow is to exploit vulnerability. In this project, an attempt to exploit the buffer overflow, stack overflow, and memory leak issues have been made.

a. **About the ERIKA Enterprise Virtual Machine**: ERIKA Enterprise is an OSEK/VDX certified open-source RTOS for selected architectures. The ERIKA Enterprise virtual machine was done in collaboration with Lauterbach [14]. It includes the demo license of the Lauterbach TRACE32 software that simulates the CPU and the basic peripherals like the LEDs, buttons, and timers, of the Axiom MPC5674evbfxmb evaluation board hosting a Freescale MPC5674F Mamba IC. The virtual machine also contains RT-Druid, which is based on the eclipse framework and generates scripts and OSEK ORTI files for interfacing with Lauterbach TRACE32.
b. **System Assumption**: An attacker having access to the on-board diagnostics (OBD-II) port has re-flashed and replaced the properly functioning code in the ECU with his malicious code.
c. **The Base Application**: A basic cruise control application has been considered, where the speed of the vehicle, which is to be maintained constant is set by the driver by pressing a button and can be terminated by pressing either the clutch or the brake pedal. The controller is a proportional-integral (PI) controller and is as shown in Fig. 2.

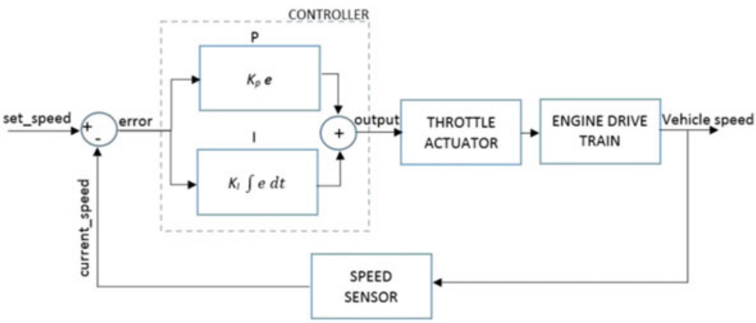


Fig. 2 Generic block diagram of a cruise control system

The previous nine error values along with the current error value are stored in a buffer called error_buffer of length 10 elements in order to calculate the 'integral (I)' part of the PI controller.

d. **The Buffer Overflow Issue**: The application considered in (c) is modified to write the error data into and beyond the size of the error_buffer. A macro MAX_BUFFER was defined such that it is greater than the size of error_buffer. The error data is to be written into the error_buffer, MAX_BUFFER number of times in a loop in order to check if this data could be used to overwrite the data in the adjacent memory locations of the error_buffer.
e. **The Stack Overflow Issue**: For the purpose of context switching, the stack stores the contexts of the tasks in the order of execution. If the stack overflows, the data in the stack might overwrite the data in the adjacent memory locations beyond the stack.
f. **The Memory Leak Issue**: Memory leak occurs when memory allocations are incorrectly managed in a computer program. If the memory assigned dynamically is not released, it leads to filling up of the heap that eventually leads to memory exhaustion. As the chosen RTOS is a statically allocated OS, the heap where the dynamically allocated memory can be stored does not exist, and the memory leak issue could not be implemented.
g. **The Timer Storm Attack**: A timer storm attack is usually done in order to impose the undesirable and unacceptable amount of delays on either the target task, a group of tasks, or on the system as a whole. The attack pattern used in this project is that a periodic timer generates an interrupt that activates a task that induces a delay in execution. Alternatively, multiple timers could also be used in order to generate multiple interrupts. The study of this attack should be of greater importance, especially when the safety-critical hard real-time systems such as the anti-lock braking system (ABS) or the traction control system of the vehicle are at stake, whose disruption might prove fatal.
h. **The Interrupt Storm Attack**: The idea of this attack has been evolved from the timer storm attack. The main aim of this attack is to disrupt the OS by storming it with a large number of interrupts. The frequency of these interrupts should be such that the OS is unable to handle the context switch and crashes. The nature of the interrupts could be either internal or external. The internal-interrupts for this attack could be generated with the help of a timer with a very low periodicity. The high-frequency external-interrupts can arrive from a spoofed ECU controlled by the attacker. The criticality of this attack is also very high as the failure of safety-critical hard real-time systems could lead to fatal accidents.

4 Results and Analysis

The simulation can be viewed in the Lauterbach TRACE32 window.

Fig. 3 Base application—variable watch window (set_speed = 80 kmph)

4.1 The Base Application

The value of the variable set_speed is chosen from the speed_value buffer, which is the replacement for real-time speed sensor data. With each click of the set speed button, the value of set_speed is updated to the next value in the speed_value buffer. In Fig. 3, the value of set_speed is set to be 80 kmph as the set speed button was pressed once which is denoted by the variable *j*. The variable *i*2 denotes the position of the current_speed buffer (which is a replacement of the real-time speed sensor data) using which the current speed of the vehicle at that particular time is taken. The value of error is calculated by subtracting set_speed from the current_speed. These error values are written into the buffer error_buffer of length 10. The output is the PI controller output value that will be fed to the throttle actuator as shown in Fig. 2.

4.2 Buffer Overflow

The buffer overflow issue was tested positive. The data in the adjacent memory locations beyond the buffer length could be accessed and overwritten. The error data is written MAX_BUFFER number of times into the buffer error_buffer, which is of length 10 elements. The parameter MAX_BUFFER is defined such that MAX_BUFFER > length (error_buffer). Here, the value of MAX_BUFFER is defined as 12. Hence, the two adjacent memory locations of the buffer error_buffer have been overwritten by the older error_buffer elements.

4.3 Stack Overflow

The stack overflow issue was tested negative. Despite various attempts to completely fill the stack by trying to declare and activate maximum number of tasks, the system allows only 70% of the stack to be filled and the maximum number of tasks that could be only 16. Below is the supporting screenshot for the implementation (Fig. 4).

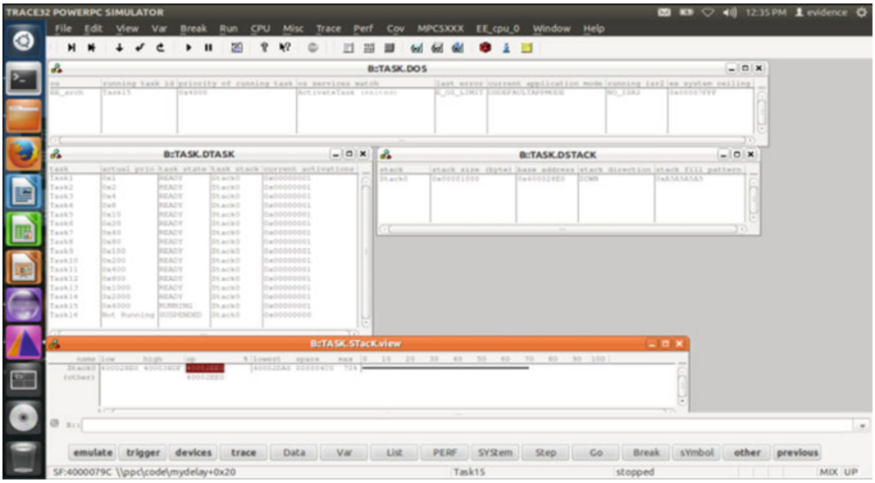


Fig. 4 Stack overflow—maximum limit for the stack (70%)

The maximum number of tasks that can be declared and activated is 16. Upon declaring and activating task number 17, the system throws an error message saying, "'Task17' undeclared (first use in this function)" as shown in Fig. 7. By this, it can be understood that, despite the proper syntactical declaration and activation of Task17, the system ignores the declaration part and throws an error message for the activation part of the task. Hence, in this OS, the stack can never be filled to its 100%; and the maximum limit is only 70% of the stack with just 16 tasks (Fig. 5).

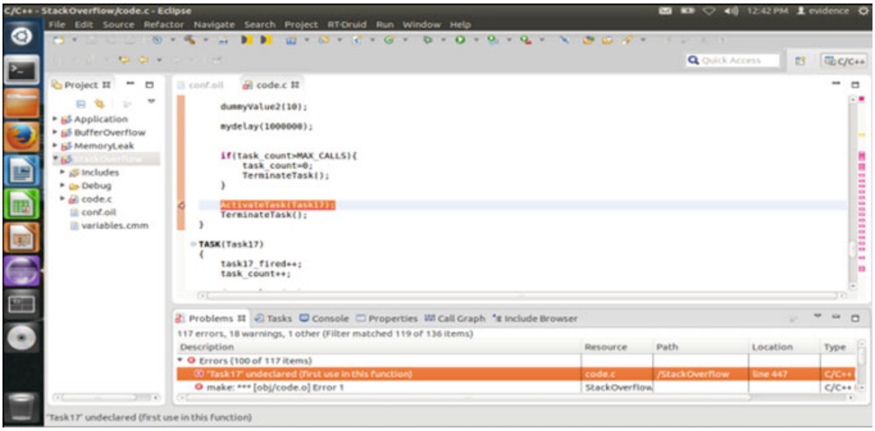


Fig. 5 Stack overflow—error message upon declaring and activating task number 17

4.4 The Timer Storm Attack

The timer storm attack was tested positive. The periodic interrupt generated by the timer introduces a delay in the execution of the subsequent tasks. The simulator used does not display any clock signal in the output window, in order to measure the delay in execution. It is only possible that this result (the delay in execution of tasks) can only be observed and perceived by vision, and hence, the screenshots could not be taken in this case.

4.5 The Interrupt Storm Attack

The interrupt storm attack was tested positive. The high-frequency periodic interrupt generated by the timer requires a high-frequency context switch that is beyond the OS capability making it to crash. It has been observed that, if the context switch frequency requirement is more than 10 clock-cycles, the system works fine. But if it is equal to or lesser than 10 clock-cycles, the system crashes. Figures 6 and 7 are the supporting screenshots for the implementation.

The system runs properly until the execution reaches Task8, in which the frequency at which the interrupt is triggered is set to 10 clock-cycles. Hence, after running Task8, an interrupt is triggered for every 10 clock-cycles and so, the OS being incapable of handling this high-frequency context switch crashes and stops functioning.

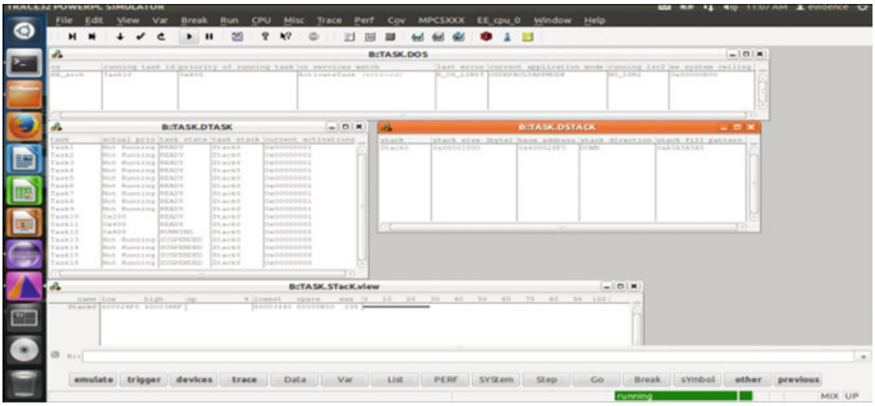


Fig. 6 Interrupt storm attack—on-going simulation (Task12)

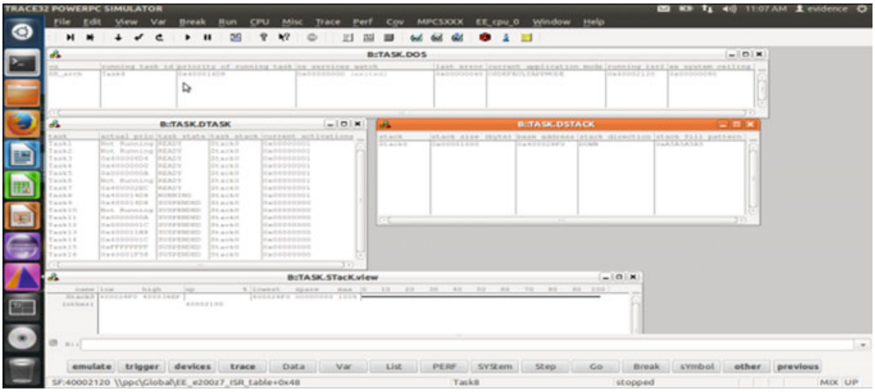


Fig. 7 Interrupt storm attack—the final status (after system crash)

5 Conclusion and Future Scope

For the faultless operation of any real-time system (RTS), two kinds of security issues are to be considered—security within the RTOS and security of the RTOS in a network. In this paper, a base application, which is a basic cruise control application, had been written and the vulnerability check for the chosen RTOS (OSEK/VDX-based ERIKA Enterprise) was performed using this base application. Security issues such as buffer overflow, stack overflow, memory leak, timer storm attack, and interrupt storm attack were tested for the chosen RTOS after making certain system assumptions.

From the results obtained, it can be observed that the chosen RTOS (OSEK/VDX-based ERIKA Enterprise) was tested positive for buffer overflow, timer storm attack, and also the interrupt storm attack. It is immune to stack overflow and memory leak issues. There is possibility that the issues that were tested positive could be exploited by the attackers to either overwrite the important and mandatory data using buffer overflow, or even cause accidents remotely by either inducing undesirable and unacceptable amounts of delay using the timer storm attack, or by simply crashing the system by the interrupt storm attack. Considering these results, a wireless sensor network application, which is time critical and data critical can use OSEK/VDX platform its real-time operating system.

The future scope for this paper could be exploring the possibility of exploiting the buffer overflow issue for malicious code injection; and framing new standards and OS policies for vehicular ad hoc networks (VANETs) in order to have safer connected cars. Also, apply this knowledge to build a data secured wireless sensor network application.

References

1. Yu, W.D., Baheti, D., Wai, J.: Real-Time Operating System Security. Computer Engineering Department, San Jose State University, San Jose, 95192-0180
2. Kiszka, J., Wagner, B.: Modelling security risks in real-time operating systems. In: 5th IEEE International Conference on Industrial Informatics (2007)
3. Lemieux, J.: Programming in the OSEK/VDX Environment. CMP Books, Lawrence, Kansas (2001)
4. Peter, H.F.: Real-time application development with OSEK: a review of the OSEK standards (2003)
5. Yoshikawa, M., Sugioka, K., Nozaki, Y., Asahi, K.: Secure in-vehicle systems against Trojan attacks. In: 14th International Conference on Computer and Information Science, pp. 29–33. IEEE/ACIS (2015)
6. Bécsi, T., Aradi, S., Gáspár, P.: Security issues and vulnerabilities in connected car systems. In: International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), pp. 477–482. IEEE (2015)
7. Sagstetter, F., Lukasiewycz, M., Steinhorst, S., Wolf, M., Bouard, A., Harris, W.R., Jha, S., Peyrin, T., Poschmann, A., Chakraborty, S.: Security challenges in automotive hardware/software architecture design. In: Design, Automation and Test in Europe Conference and Exhibition (DATE), pp. 458–463. IEEE (2013)
8. Whaiduzzaman, M., Sookhak, M., Gani, A., Buyya, R.: A survey on vehicular cloud computing. J. Netw. Comput. Appl. **40**, 325–344 (2014)
9. Pillai, A.S., Isha, T.B.: ERTSim: an embedded real-time task simulator for scheduling. In: IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai, India, pp. 724–727 (2013)
10. Pillai, A.S., Punnekkat, S., Isha, T.B.: Energy minimization in real time embedded systems by control of context switches. In: Proceedings of Second International Conference on Smart Technologies for Materials, Communication, Controls, Computing and Energy (ICS), Chennai, India, pp. 428–432 (2013)
11. Shantilal Patel, J., Chavda, V.M.: Security vulnerability and robust security requirements using key management in sensor network. Int. J. Grid Distrib. Comput. **7**(3), 23–28 (2014)
12. Chockalingam, S., Lallie, H.S.: Alarming! security aspects of the wireless vehicle: review. Intl J. Cyber-Secur. Digit. Forensics (IJCSDF) 200–208 (2014)
13. Han, K., Weimerskirch, A., Shin, K.G.: Automotive cyber security for in-vehicle communication. IOT Q. **6**(1), 22–25 (2014)
14. Adib, F.M., Hajji, H.M.: VSpyware: spyware in VANETs. In: 35th Conference on Local Computer Networks (LCN), pp. 621–624. IEEE (2010)

Performance Evaluation of Delay-Tolerant Routing Protocols on Bangladesh Map

Md. Shahedul Islam, Sadia Islam Thaky and Md. Sharif Hossen

Abstract This paper evaluates the performance of delay-tolerant routing techniques, namely prophet, maxprop, binary spray and wait (B-SNW), epidemic, and spray and focus (SNF) on Bangladesh map scenario. The investigation is performed to change the number of the message copies in the network and the number of nodes, respectively. Hence, the results show that maxprop exhibits better delivery for having message generation rate less than or equal to 5, and the number of nodes greater than or equal to 40 while SNF shows a good delivery tendency for having message copies greater than 5 and the number of nodes less than 40. Maxprop shows a higher average latency, and lower overhead than epidemic and prophet while higher overhead than SNF and B-SNW. That means, SNF shows good performance with lower delay, lower overhead, and higher delivery compared to all except the case of delivery where message copies are less than or equal to 5 and the number of nodes greater than or equal to 40. We see a greater number of contacts during the times: 2, 4, 10, 11, and 12 h for varying message generation rates as specified in Table 1, while approximately a constant result for varying the number of nodes, i.e., the same result per hours except the last hour.

Keywords Delay-tolerant network · Opportunistic network environment simulator · Openjump · Geographic information system · Routing protocols

Md. S. Islam
Department of Computer Science & Engineering,
Stamford University Bangladesh, Dhaka, Bangladesh
e-mail: shahed.polimi@gmail.com

S. I. Thaky
Department of Computer Science and Engineering,
University of Science & Technology, Chittagong, Bangladesh
e-mail: sadiataki11@gmail.com

Md. S. Hossen (✉)
Department of Information & Communication Technology,
Comilla University, Comilla 3506, Bangladesh
e-mail: mshossen@cou.ac.bd
URL: http://cou.ac.bd/ict/mshossen

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_39

1 Introduction

Due to the speedy increase of wireless devices such as tablets, laptops, smartphones, etc. educational and commercial wireless applications have become available in our daily lives [1, 2]. These applications make new means for user's interaction and easier communication between different parties establishing a temporary network, i.e., ad-hoc network (AN) with the increase of working and storage capability of the mobile or other wireless movable devices [3, 4]. AN [5] is considered as a delay-tolerant network (DTN) [6], which is qualified with big latency, long-duration, partitioning, and shortage of coeval routes among the nodes [7, 8]. In such a network, when a node with no delay (means very lower delay where user cannot feel that delay) cannot find another node in the simulation network [9–11], delay-tolerant routings apply store-and-forward principle where movable devices, e.g., buses [12] and cars [13] are used to transfer data via a proper route [14]. This principle is to forward and store the message copies from one node to the next node available instead of waiting to send the message copies to the destination [15–17]. In this paper, we see the performance of several DTN routings, namely epidemic [18], binary spray and wait (B-SNW) [19], prophet [20], maxprop [15], as well as, spray and focus (SNF) [21] on a highly populated country (Bangladesh) map, for changing the number of message copies and the number of nodes. These routings are evaluated based on three performance metrics, namely delivery ratio, overhead ratio, and average delay. Rest of the paper is prepared as following: in Sect. 2, we summarize investigated routing protocols in DTN. Section 3 describes openjump, an event-driven simulator, and simulation setting. The investigated results have been discussed in Sect. 4. And in Sect. 5, we have mentioned the concluding remarks about this research work.

2 Brief Discussion of DTN Routing Protocols

This section gives essential information of the considered DTN routing protocols.

2.1 Epidemic

Epidemic means something that is flooded or widespread in nature. In the network, a node can repeat a message at any instant without any consideration but only by taking into account that the encountered nodes do not have the duplicate copy [18].

2.2 Prophet

It forwards, a copy of message to another node having a greater delivery tendency according to the following three rules [20]:

When a node X encounters Y, delivery ratio of Y is raised.

$$P_{(x,y)} = P_{(x,y)old} + (1 - P_{(x,y)old}) * P_{in} \tag{1}$$

where P_{in} is a fixed value during first meeting.

When X and Y do not encounter themselves within a fixed time, the delivery ratio seems older and will also loss the possibility of forwarding messages

$$P_{(x,y)} = P_{(x,y)old} * \gamma^{l} \tag{2}$$

where γ^{l} is a fixed value to indicate aging.

According to transitivity property, when a client X often meets another client Y and client Y meets another client Z, then client Z becomes potential to send the message intended for client X.

$$P_{(x,z)} = P_{(x,z)old} + (1 - P_{(x,z)old}) * P_{(x,y)} * P_{(y,z)} * \beta \tag{3}$$

where β is a fixed value indicating the impact of the transitivity among nodes.

2.3 MaxProp

MaxProp routing uses Dijkstra shortest path algorithm [22] to determine which message copies need to be transmitted first and dropped first. It gives more precedency to those copies, which have the lower hop counts, taking into account no duplicate packets or message copies.

2.4 SNW

Spray and wait (SNW) [19] obtains good delivery by reducing the number of the message copies in the network. There are two steps. In the first step (spray), if the starting relay caries L message copies, it forwards them first to L or L/2 distinct relays. In wait phase, nodes directly meet and forward message copies to the destination node. It has vanilla as well as binary variations. In vanilla, starting relay sends L copies to first L relays. However, in the second variation (binary), the starting node forwards half of the copies from L copies to half of the distinct relays. The L/2 relays again forward to the next relays until having only one message. Only one message means that the relay is in the waiting stage. Waiting stage is same for both binary and vanilla as discussed above.

2.5 SNF

Compared to SNW with the focus phase of spray-and-focus (SNF) [21] routing, having only one message of a node can be forwarded to another relay instead of the destination node using the following criteria: A source X can successfully send a message to the destination Z with forwarding it through intermediate node Y iff Y has the higher utility value to meet Z compared to the utility of X to meet Z plus the threshold value, U_{th}.

$$U_{Y(Z)} > U_{X(Z)} + U_{th} \tag{4}$$

3 Simulation Tool and Environmental Setting

This section explains ONE simulator and the environment modeling parameters.

3.1 OpenJump and WKT Format

We have used a Java unified mapping platform called openjump, which is a geographic information system (GIS), to read and edit Bangladesh map as well as make the map into a well-known text (WKT) format. In this map, some disconnected areas are either connected or removed by using openjump [23] for simulation purposes. During the WKT format, some unwanted characters are generated, which are removed by developing an application made by Java as shown in Fig. 1.

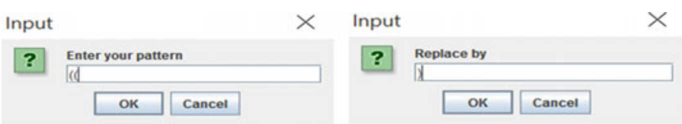


Fig. 1 String removal application

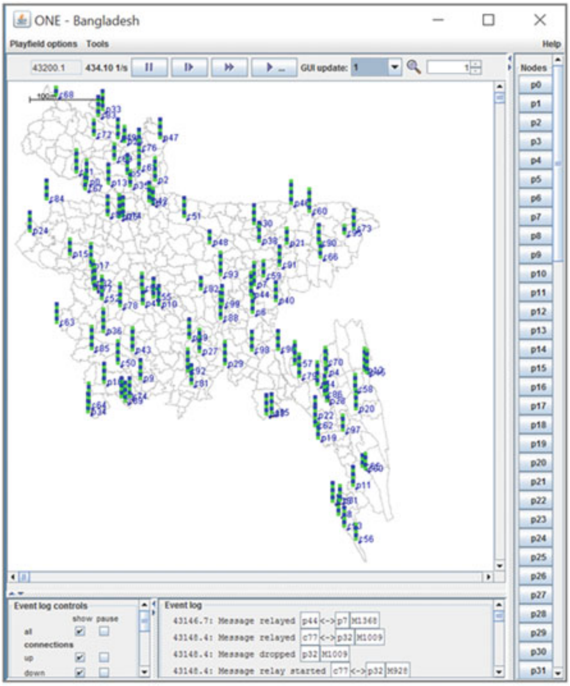


Fig. 2 Bangladesh simulation area

3.2 ONE Simulator

Opportunistic network environment (ONE) simulator is a Java-based simulator developed by Kernenan et al. used for analyzing the effects of various events like tracing the movements of nodes and predicting their behaviors. Besides, we can easily investigate the performance of DTN routing protocols due to having all the events combined in a single package [24, 25] (Fig. 2).

3.3 Simulation Environment Setting

Parameters of simulation setup for the abovementioned routings are specified here. We simulate the considered scenario about half of the day (43,200 s). Two categories of host, namely pedestrians and cars have been regarded with the respective speeds from 0.5 to 1.5 m/s and from 2.7 to 13.9 m/s. We have used the communicating interface as Bluetooth with the speed of 250 kilobytes/s and coverage of 10 m. We have changed the message generation rates as 2, 3, 4, 5, 6, 10 (msg/min) and the number of nodes as 20, 50, 80, 100 where for varying the first, second is kept fixed at 100 nodes while for varying the second, the first is kept fixed at two copies. In both cases, we have applied shortest path map-based movement with 5MB buffer, 5 h TTL (time-to-live), 500 kilobytes to 1 Megabyte as the size of message, and the simulation area of 147,570 * 147,570 m^2 (width * height).

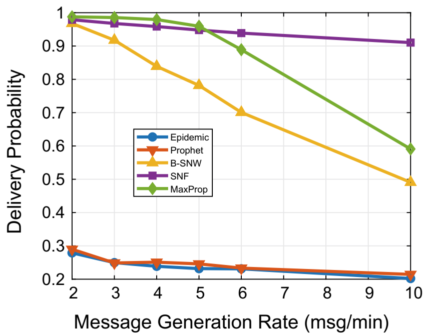
4 Result and Discussion

This section includes the discussion of the analysis of simulated results with the performance metrics of delivery, latency, and overhead.

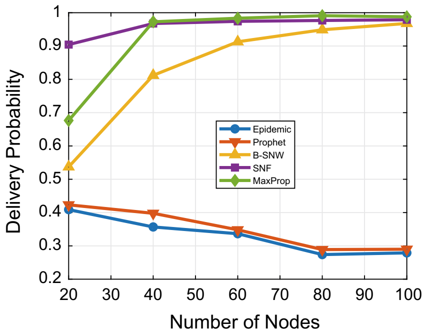
4.1 Evaluation on Delivery Probability

Delivery probability is a value that is obtained by dividing the total number of message copies received to the destination by the total number sent from source.

In Fig. 3a, we see that maxprop shows a better delivery compared to other routings when message generation rate is less than or equal to 5. SNF exhibits better performance when the number of message copies is greater than 5. In case of increasing the number of nodes as shown in Fig. 3b, maxprop shows better delivery when the number of nodes is equal to or greater than 40, While, SNF shows higher delivery when nodes are not greater than 40. That means, although maxprop and SNF have higher delivery compared to other routings, they compete with each other in the abovementioned terms.

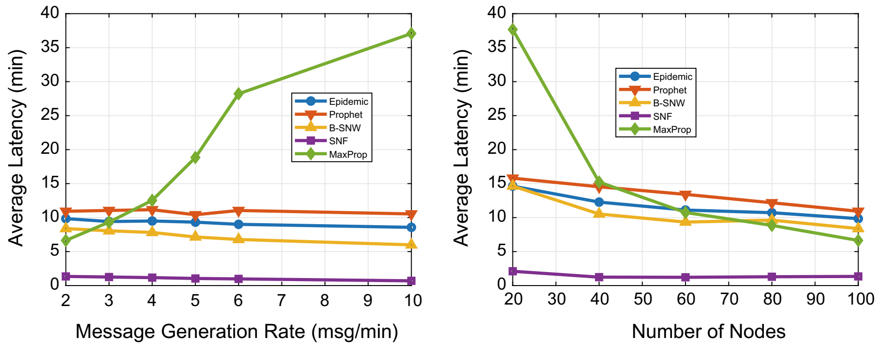


(a) Delivery ratio with message copies.



(b) Delivery ratio with mobile nodes.

Fig. 3 Variation of delivery compared to the number of message copies per minute and number of mobile nodes



(a) Average delay with message copies. (b) Average delay with mobile nodes.

Fig. 4 Variation of average delay compared to the number of message copies per minute and number of mobile nodes

4.2 Evaluation on Average Delay

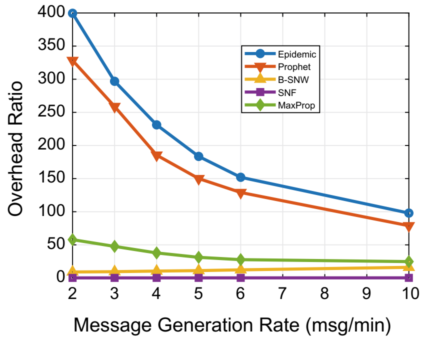
The average time that messages take to be transferred from the starting node to the final is called average delay. From Fig. 4a, we see that maxprop has the growing latency while other routings have the decreasing latency. On the other hand, in the case of varying the number of nodes as shown in Fig. 4b, maxprop has a higher latency. In both cases, as shown in Fig. 4a and b, SNF has a lower latency compared to all.

4.3 Evaluation on Overhead Ratio

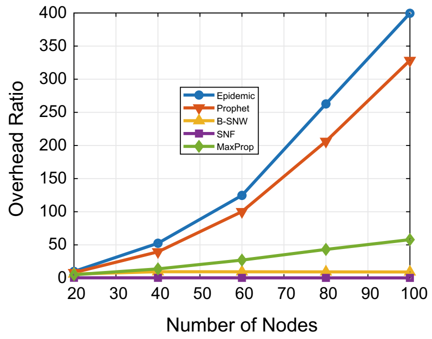
To send the messages successfully from a source to a destination, the source node needs to resend copies repeatedly due to the loss of message transmission. The total number of repetitions for successfully sending a single message is called overhead ratio. Hence, we see that maxprop has lower overhead than epidemic and prophet. On the other hand, SNF and B-SNW routings have lower overhead than others where lowest is found at SNF as shown in Fig. 5a and b.

4.4 Analysis on Number of Contacts for Varying Message Copies

Hence, we see from Fig. 6 that the numbers of contacts is maximum (considering above 14,000) for using the message generation rates during the time mentioned in Table 1.



(a) Overhead ratio with message copies.



(b) Overhead ratio with mobile nodes.

Fig. 5 Variation of overhead ratio compared to the number of message copies per minute and number of mobile nodes

Table 1 Better number of contacts during the time for specified message generation rates

Time (h)	Message generation rates (msg/min)
2	2, 5, 6, 10
4	10
10	5
11	3
12	4

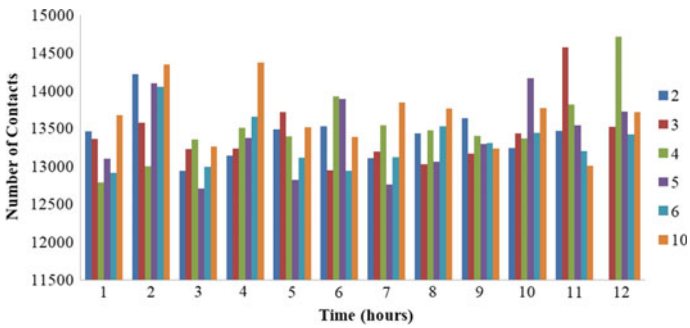


Fig. 6 Number of contacts with time for varying the message copies

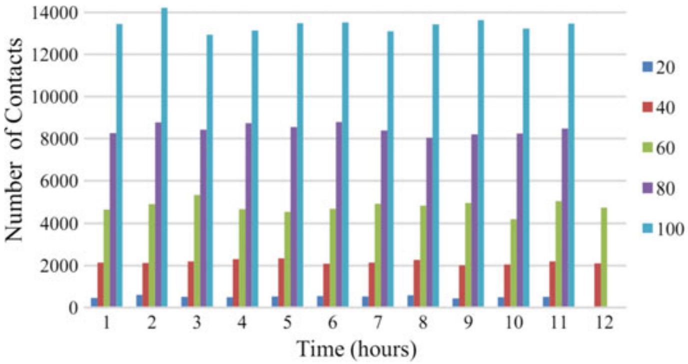


Fig. 7 Number of contacts with time for varying the number of nodes

4.5 Analysis on Number of Contacts for Varying Mobile Nodes

Figure 7 shows the number of contacts with time which is approximately same at any hour with the increase of the mobile nodes except the time equal to 12 h where we only get the number of contacts for using 40 and 60 nodes.

5 Conclusion

In this research investigation, the analysis of several delay-tolerant routing protocols has been evaluated on Bangladesh map scenario. We have converted the Bangladesh map into wkt format using an openjump tool and implemented it in ONE simulator. Hence, the investigation is done for changing the number of message copies and mobile nodes for several ad-hoc routing protocols, namely epidemic, spray and wait, prophet, maxprop, and spray and focus (SNF). In the investigation, we see that maxprop deserves better delivery if message copies are less than or equal to 5 and nodes are greater than or equal to 40 while for SNF these are greater than 5 copies and less than 40 nodes. Although we see good message delivery for both maxprop and SNF with the certain conditions, the chance of the delay of sending message using maxprop is very higher than other routings while SNF shows lower latency. Again, SNF shows lower overhead compared to others where epidemic shows the higher. Therefore, SNF routing shows good performance with the lower delay, lower overhead, and the above terms of delivery. Besides, we have observed the duration of sending the message copies per minutes for obtaining a better number of contacts mentioned in Table 1.

Acknowledgements Authors would like to thank Muhammad Sajjadur Rahim, Associate Professor, Dept. of ICE, University of Rajshahi, Bangladesh. Authors also would like to thank Farhana Rea Deeba, Lecturer, Dept. of English Language and Literature, Jatiya Kabi Kazi Nazrul Islam University, Bangladesh.

References

1. Massri, K., Vitaletti, A., Vernata, A., Chatzigiannakis, I.: Routing protocols for delay tolerant networks: a reference architecture and a thorough quantitative evaluation. J. Sens. Actuators Netw. (2016)
2. Alenazi, M.J.F., Cheng, Y., Zhang, D., Sterbenz, J.P.G.: Epidemic routing protocol implementation in ns-3. In: ACM WNS3, Barcelona, Spain (2015)
3. Moreira, W., Mendes, P.: Social-aware forwarding in opportunistic wireless networks: content awareness or obliviousness? In: IEEE AOC (2014)
4. Ciobanu, R.I., Dobre, C., Cristea, V.: SPRINT: social prediction-based opportunistic routing. In: IEEE WoWMoM (2013)
5. Pelusi, L., Passarella, A., Conti, M.: Opportunistic networking: data forwarding in disconnected mobile ad hoc networks. IEEE Commun. Mag. **44**, 134–141 (2006)
6. Fall, K.: A delay-tolerant network architecture for challenged internets. In: ACM SIGCOM, Germany, pp. 27–34 (2003)
7. Jain, S., Fall, K., Patra, R.: Routing in a delay-tolerant network. In: ACM SIGCOMM, USA, pp. 145–157 (2004)
8. Feng, Z.: A unified study of epidemic routing protocols and their enhancements. In: IEEE IPDPSW, pp. 1484–1493 (2012)
9. Jindal, A., Psounis, K.: Fundamental mobility properties for realistic performance analysis of intermittently connected mobile networks. In: IEEE PerComW (2007)
10. Ghosh, J., Ngo, H.Q., Qiao, C.: Mobility profile based routing within intermittently connected mobile ad hoc networks. In: ACM IWCMC, Vancouver, British Columbia, Canada (2006)
11. Zhang, Z.: Routing in intermittently connected mobile ad-hoc networks and delay tolerant networks: overview and challenges. IEEE Commun. Surv. Tutor. (2006)
12. Komnios, I., Tsaoussidis, V.: CARPOOL: extending free internet access over DTN in urban environment. In: ACM MobiCom LCDNet, Miami, USA (2013)
13. Cheng, P.C., Lee, K., Gerla, M., Harri, J.: GeoDTN+Nav: geographic DTN routing with navigator prediction for urban vehicular environments. Mob. Netw. Appl. **15**, 61–82 (2010)
14. Moreira, W., Mendes, P.: Impact of human behavior on social opportunistic forwarding. Ad Hoc Netw. **25**, 293–302 (2015)
15. Burgess, J., Gallagher, B., Jensen, D., Levine, B.N.: MaxProp: routing for vehicle-based disruption-tolerant networks. In: IEEE INFOCOM, Spain (2006)
16. Juang, P., Oki, H., Wang, Y., Martonosi, M., Peh, L.S., Rubenstein, D.: Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences with ZebraNet. In: ASPLOS, USA, pp. 96–107 (2002)
17. Chaintreau, A., Hui, P., Crowcroft, J., Diot, C., Gass, R., Scott, J.: Impact of human mobility on opportunistic forwarding algorithms. IEEE Trans. Mob. Commun. (2007)
18. Vahdat, A., Becker, D.: Epidemic routing for partially connected ad hoc networks. Tech. Rep. (2000)
19. Spyropoulos, T., Psounis, K., Raghavendra, C.S.: Spray and wait: an efficient routing scheme for intermittently connected mobile networks. In: ACM WDTN, USA (2005)
20. Lindgren, A., Doria, A., Scheln, D.: Probabilistic routing in intermittently connected networks. ACM Mob. Commun. Commun. Rev. **7**, 19–20 (2003)
21. Spyropoulos, T., Psounis, K., Raghavendra, C.S.: Spray and focus: efficient mobility-assisted routing for heterogeneous and correlated mobility. In: IEEE PerCom, USA (2007)

22. Thakur, G.S., Kumar, U., Helmy, A., Hsu, W.: Analysis of spatio-temporal preferences and encounter statistics for DTN performance. arXiv preprint arXiv:1007.0960 (2010)
23. Project page of the OpenJump. http://www.openjump.org/
24. Kernen, A., Ott, J., Krkkinen, T.: The ONE simulator for DTN protocol evaluation. In: ICSTT, Italy (2009)
25. Project page of the ONE simulator. Tkk/Comnet (online) (2017)

Author Biographies

Md. Sharif Hossen is currently working as an Assistant Professor in Dept. of Information and Communication Technology (ICT) at Comilla University, Bangladesh. He was formerly Lecturer in Dept. of ICT at Comilla University and in Dept. of CSE at Southeast University, and a research assistant in Dept. of Information and Communication Engineering (ICE) at the University of Rajshahi (RU). He completed his B.Sc. and M.Sc. from Dept. of ICE at the University of Rajshahi, Bangladesh. He achieved faculty first position. His research interests include delay-tolerant networking, wireless ad hoc, sensor, underwater and vehicular networks, as well as considerations of privacy in the Internet of Things. He is also interested in applying big data techniques. He has published eight international journal papers, thirteen international conference papers, and two book chapters. He is the reviewer of several international journals (viz. Springer, IEEE Access, Wiley, Scopus index). He received several scholarships like ICT research fellowship (for M.Sc. thesis), UGC scholarship (for B. Sc. highest result in faculty), and Merit scholarship for outstanding academic results at the University of Rajshahi. He received Best Paper Award at the IEEE ICISET 2016 conference at IIUC, Bangladesh.

Md. Shahedul Islam is working as an Assistant Professor in Dept. of Computer Science and Engineering at Stamford University Bangladesh, Dhaka.

Sadia Islam Thaky completed her B.Sc. and M.Sc. degree in Dept. of Computer Science and Engineering at University of Science & Technology, Chittagong, Bangladesh.

Upgraded Proportional Integral (UPI): An Active Queue Management Technique for Enhancing MANET's Performance

Naveen Ranjan and B. Nithya

Abstract A Mobile Ad hoc Network (MANET) is a group of many autonomous and free wireless nodes in which network does not have any fixed infrastructure and centralized administration. In such conditions, reliable conveyance of control packets and management of large unused buffer are the main issues. Many queue management algorithms allow queues to maintain a status of full or almost full queue for the maximum duration of time, leading to more packet loss thereby results in full buffer and bufferbloat problems. To resolve this problem, a novel algorithm called Upgraded Proportional Integral (UPI) algorithm for active queue management is proposed. The proposed UPI algorithm makes a wise decision for every incoming packet to decide whether the packet needs to be enqueued or dropped. Instead of random dropping, the run time parameters such as arrival rate, service delay and drop probability are utilized to make this judgment to enhance QoS of MANET. The performance results of the proposed UPI algorithm are measured and compared with Conventional PI and Blue using NS2 simulator over different metrics. The result shows that the proposed algorithm minimizes average end-to-end delay and routing overhead without compromising throughput.

Keywords MANET · Blue · PI · Bufferbloat problems · Packet overhead · NS2

1 Introduction

A Mobile Ad hoc Network (MANET) is a group of many independent, autonomous, self-operative, self-adaptive and free wireless nodes. This network does not have any fixed infrastructure and centralized administration, i.e., communication is possible without any centralized coordinator [1, 2]. In this scenario, the condition reliable

N. Ranjan (✉) · B. Nithya
Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, Tamil Nadu 620015, India
e-mail: naveenranjan50@gmail.com

B. Nithya
e-mail: nithya@nitt.edu

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_40

conveyance of control packets and management of large unused buffer are the main issues. The queue management techniques, which are applied to a router, play an important role in providing quality of service (QoS). There are several queue management techniques proposed in the literature [3–6]; they are broadly classified into two categories: Passive [6] and Active [7] Queue Management.

In Passive Queue Management (PQM), the source knows about queue status only after packet is dropped. It has two states: in the first state, if packet is dropped, then the intermediate node does not send any early warning regarding congestion to the sender. In the second state, when there is 100% chance of packet drop, then all sender nodes go to back off state. Drop tail [8] and Drop From Front [9] are two such examples of PQM. In Drop tail [8], the working principle is based on First-In, First-Out (FIFO). If the queue capacity is reached to its maximum size, then incoming packets are rejected until the buffer has enough space to hold the packet. It is a very simple technique to deploy and computational overhead is also very low. In Drop From Front [9], packet is dropped from front rather than tail or rear of the queue. The packet loss probability with Drop From Front is lesser than Drop tail because it rejects the packet which has been for a longer time inside the queue. The drawback of both PQM techniques is that it does not handle the buffer space using preventive actions and it maintains a status of full queue or almost full queue for maximum duration of time. Hence, it leads to lockout, full queue and bufferbloat problems [10].

In Active Queue Management (AQM) [4–7], the packet is dropped before the capacity of queue is full. It operates by probabilistically dropping, random dropping and marking probabilities of the incoming packet. Dropping probability for every packet is based on a total number of hops that the packet has traversed. It assigns higher drop probability for those packets which have traversed less number of hops; similarly, it assigns smaller drop probability for the packets which have traversed more number of hops. Some of the active queue management techniques are RED [11], Blue [12], PI [13] and Codel [14]. In RED [11], packet is dropped based on drop probability and it accepts the entire packet when the buffer is empty. As the buffer size grows, probability for packet drop also grows. If queue is full or almost full, then the drop probability reaches at most to one and after that, every incoming packet is rejected. Here, link utilization is high but it is very difficult to configure and space requirement is more. Unlike RED, Blue [12] maintains a packet drop probability P and constant value i; if the queue is full, then P is incremented by a constant i, and whenever queue is empty, it is decremented by i. It is very easy to configure and space requirement is also very less. But it performs better only when the queue capacity is small.

In PI [13], it maintains a packet drop probability and a random generator function for every packet. It drops the packet randomly based on the comparison between the packet drop probability and random generator function. Even though it performs well for larger queue size, fairness is not good. In Codel [14], to avoid the problem of unfair queuing, it uses packet scheduling scheme with a probabilistic approach. It recognizes the behavior of queue as two types: Bad and good queue based on whether it exhibits bufferbloat problem or not, respectively. This technique experiences minimum packet delay and good link utilization. The drawback of this method is that it distinguishes the queue frequently which is an overhead and incorrect recognizing of queue leads to bufferbloat problem.

Overall for reliable conveyance of control packets and for the management of large unused buffer, there should be an effective solution for queue management. To achieve this, an efficient algorithm called Upgraded Proportional Integral (UPI) algorithm for Active Queue Management is proposed. It manages the large unused queue using current network status to enhance the network performance.

The remaining section is described as follows: in Sect. 2, related works with their advantages and disadvantages are explained. The proposed UPI algorithm is described in Sect. 3. Section 4 concise about the simulation and performance analysis using NS2 and results are compared with Blue and PI. In the end, the paper is concluded in Sect. 5.

2 Related Work

Many techniques [3–6] have been proposed for managing the large unused buffer space to reduce congestion and packet loss. A few proposals that are suggested and related to the proposed algorithm are given below.

Kulkarni et al. [15] proposed a technique in which buffer size is based on traffic variation and it predicts average queue size for the next finite intervals with respect to previous interval value. So, the packet drop probability lies in that interval; hence, variation in input traffic is captured accurately. The simulation performance of the technique is compared with random early detection (RED) algorithm and results show that it minimizes the packet loss ratio and increases the link utilization. But the scalability is an issue because it does not maintain flow state, and in different network conditions, it is difficult to evaluate traffic variation.

Friderikos and Aghvami [16] suggested the technique in which dropping probability for every packet is based on a total number of hops that the packet has traversed. It assigns higher drop probability for those packets which have traversed less number of hops; similarly, it assigns smaller drop probability for the packets which have traversed more number of hops. The advantage of this technique is that it reduces end-to-end delay and avoids the problem of buffer overflow. But in this queue management scheme, if the number of hop is more, then connection timeout happens frequently.

Niharika [17] proposed a technique in which the maintenance of queue is based on round-robin fashion rather than FIFO. At every node, it maintains a table to record the flow status and packet number. Due to round-robin scheme, if the packet number is greater than quantum value, then it does not accept the packet. The simulation result shows that it avoids a starvation problem and reduces packet loss. But to avoid starvation, it needs to perform two more tasks such as packet scheduling and queue monitoring function which increases end-to-end delay and energy consumption.

Chandra and Kavitha [18] gave the idea in which the concept of virtual queue is used to avoid the packet loss. The size of virtual queue is less than the original queue. Whenever the packet arrives, a dummy packet will be sent to the virtual queue, and if virtual buffer space is full, then information about congestion is intimated to the sender. The simulation result is compared with REM and RED algorithms, and its performance is better than these schemes. The drawback is that it requires extra buffer space for virtual queue which is an additional overhead.

Chen and Bensaou [19] proposed the technique to avoid the problem of disparity in high-bandwidth network using simple synchronized method. It reduces its transmission rate in one interval until flow does not respond to congestion indication. The process continues for the next interval and so on, and the effect is reflected immediately on the congestion condition of a link. It has higher link utilization and lesser packet loss. Its performance degraded when network speed is not high.

Dimitriou et al. [20] proposed a method of service differentiation in which dropping policies is based on packet size, i.e., dropping probability is prioritized accordingly. It measures the queue size in packet, not in bytes and it finds the packet size dynamically based on current buffer space occupancy. It increases goodput and bandwidth utilization. But if many small packets of same size arrive, then its computational overhead increases because it does not have any scheduling mechanism.

Agrawal et al. [21] suggested a technique that dynamically adjusts the value for queue length in such a way, so that it can utilize the buffer space to its maximum extent. The drop probability value is set to one whenever average queue length lies between maximum and current queue length; otherwise, drop probability value is reset to zero. The simulation result shows that it has achieved better network throughput and goodput than RED but when the uncertain load increases, network become unstable.

2.1 Overall Inference and Motivation

The inference from the abovementioned queue management algorithms is that they set the maximum queue length for handling the buffer space for every node, and when the queue size is reached to maximum buffer length, then it simply rejects the packet. Some of the algorithms face connection timeout problem whenever the number of active nodes is increased. Few algorithms specifically [18] require extra buffer space for virtual queue which is an additional overhead. Moreover, one queue becomes unstable when uncertain load on the network is increased. To mitigate these problems, there should be an effective solution for queue management. This motivates us to propose a novel algorithm called Upgraded Proportional Integral (UPI) algorithm for Active Queue Management. It manages the large unused buffer space using service delay, arrival rate of packets, service probability and drop probability. These four run time parameters depict the dynamic nature of MANET. So, the proposed algorithm enhances the network performance by making a wise judgment about whether the packet needs to be dropped or enqueued for the possible transmission.

3 Proposed Algorithm

The proposed Upgraded Proportional Integral (UPI) algorithm is an Active Queue Management algorithm that efficiently manages the buffer space to mitigate the lockout, full queue and bufferbloat problems. Instead of using random approach, the proposed UPI algorithm adopts probabilistic approach that appropriately mimics the current network condition. It uses the concept of proportional integral to manage the queue size at networking devices such as routers, switches with higher stability. The classic Proportional Integral (PI) algorithm drops the packet randomly based on the value of drop probability but the proposed UPI algorithm avoids the random dropping of incoming packet and calculates the service probability using optimum network parameters, and it makes the decision about dropping packet based on service probability and drop probability. Moreover, the traditional queue management approaches set the maximum buffer length for every node, and when the queue size is reached to maximum buffer length, it rejects the packet. The drawback of the traditional approach is that it allows buffers to have full or almost full queue status for the maximum duration of time. The abovementioned two situations result in high packet drop probability leading to performance degradation.

The proposed UPI algorithm ensures the availability of buffer for the packet and it does not allow queue to maintain a status of full queue or almost full queue. It reduces the packet delay by assigning the proper service probability for every incoming packet as shown in Eq. (1).

$$S_p = r/s \tag{1}$$

where r is the arrival rate and s is service delay of the packet. If the arrival rate (r) is high, then the service probability of the corresponding packet is also high as it is directly proportional to the arrival rate (r). Due to this, full queue problem is minimized thereby significant reduction in packet loss at every node.

If this service probability S_p is higher than drop probability p, then the corresponding packet is enqueued; otherwise, it is dropped. The proposed UPI algorithm determines the drop probability (P) as shown in Eq. (2).

$$P = \alpha(q_{\text{len}} - D_{\text{q}}) + \beta(q_{\text{curr}} - D_{\text{q}}) + P_{\text{old}} \tag{2}$$

It is inferred from the Eq. (2), drop probability P depends on the current queue length (q_{curr}), total queue length (q_{len}), desired queue size (D_{q}), previous drop probability (P_{old}) and the scaling factors α and β. The desired queue size (D_{q}) is calculated in the following equation.

$$D_{\text{q}} = (\text{payload}/\text{packet size}) \tag{3}$$

Moreover, the scaling factors α, β and w (sampling rate) are determined as in [12, 13] and it is selected based on the dynamic network condition. The functionality of proposed UPI algorithm to manage the queue size and sustain adequate path or link utilization is described in Algorithm 1. The enqueue() function stores the return value of drop_early() method in the Boolean variable X, and if the value of X is true, then the packet is dropped otherwise enqueue the packet. After that, for the next incoming packet, calculate_p() function is called from enqueue() method. In drop_early method, the return value of service_p() is stored in the variable Y; similarly, the return value of calculate_p() is stored in the variable Z. If the value of Z is greater than Y, then drop_eraly() function return true otherwise return false. The calculate_p() function calculates drop probability for every incoming packet at a sampling rate represented by W. The service_p() method finds the service probability for every incoming packet using optimum network parameters. Using above four methods, the proposed algorithm makes the decision whether the incoming packet should enqueue or dequeue the queue. The proposed UPI algorithm does not make decision of dropping packet randomly rather than it depends upon the current network environment measured in terms of service probability, arrival rate of packet and service delay of the packet.

Algorithm 1 Proposed UPI Algorithm

Parameters used:

α= // Scaling factor
β= // Scaling factor
P = // drop probability value
i = 0 // integer variable
W = sampling frequency
$P_{old=}$ old probaility

Formulas:

$P = \alpha(q_{len} - D_q) + \beta(q_{curr} - D_q) + P_{old}$

Dq= (payload/packet size)

where,

Dq = queue desired

q_{curr}= current queue length

qlen= maximum queue size

```
Enqueue()
  {
     bool x;             // x is Boolean variable
     X= drop_early();  // it return true or false value
     If (x)
          Drop packet;
     else
          Enqueue Packet;
     Dequeue();       // it invokes Dequeue() method in queue.h
     Calculate_p();  // it invokes Calculate_p() from pi.h
  }
```

```
Drop_early()                              service_P()
  {                                           {
     float y,z;                                  float p,r,s;
     Y= service_p();                       // r is arrival rate of packet;
     Z= Calculate_p();                    //  s is service delay of packet;
     If(z>y)                                     Sp= r/s;
          return true;                         return Sp;
    else                                        }
          return false;
  }
Calculate_p()
  {
     float p;
     P = α(q_len - D_q)+ β(q_curr - D_q ) + P_old; // finding Drop probability
     int i=w;       // w is rate sampling frequency
   While(i!=w)
          i++;
   Drop_early();
  }
```

4 Simulation and Performance Analysis

The proposed UPI algorithm is simulated using NS2 and its performance is analyzed in the following subsections. The metrics PDR, throughput, delay and routing overhead are measured and compared with Blue and PI algorithms

4.1 Simulation Parameter

The proposed UPI algorithm is simulated with the following performance. The running duration of simulation is 120 s with varying nodes from 10 to 200, and nodes are scattered randomly in the range of $1000 \times 1000\ \text{m}^2$. The packet size of 1024 Bytes is sent with bottleneck bandwidth of 10 Mbps. The proposed UPI algorithm executes four important functions such as enqueue(), drop_early(), service_p() and calculate_p() methods to manage the large unused buffer and calculate drop probability and service probability of packet being served using optimum network parameters. It makes the decision about dropping packet based on comparison result of service

Table 1 Simulation parameters

Parameters	Values
Simulator	Ns 2.35
RTT	100 ms
Antenna type	Omni-directional antenna
Queue size	100 Packets
Scale factor (α)	0.0017×10^{-2}
Scale factor (α)	0.0018×10^{-2}
Sampling rate (w)	170
Routing algorithm	AODV
MAC layer	IEEE 802.11
Transmission range	100 m
Number of nodes	10, 30, …, 200
Packet size	1024 Bytes
Simulation area	1000×1000
Simulation time	120 s
Bandwidth	10 Mbps

probability and drop probability for every incoming packet. And, this action is continued by every node till the end of the simulation. The values of other simulation parameters are given in Table 1.

The packet delivery ratio (PDR) attained by the proposed UPI algorithm, Blue and PI are plotted and shown in Fig. 1. It is inferred from Fig. 1, that the proposed UPI algorithm has gained better result than Blue and PI for the varying simulation time. This is mainly due to the ability of the proposed UPI algorithm that determines the desired queue size value dynamically based on current network conditions. Based on this, it takes the decision about dropping of packets. Whereas in Blue algorithm, if the queue size reaches its maximum value, it decrements packet drop probability by a small constant value. In PI algorithm also, desired queue size is not determined

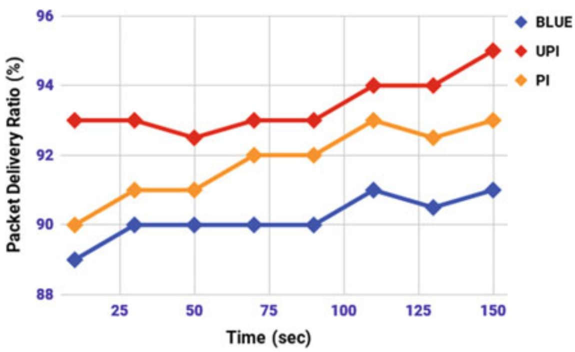
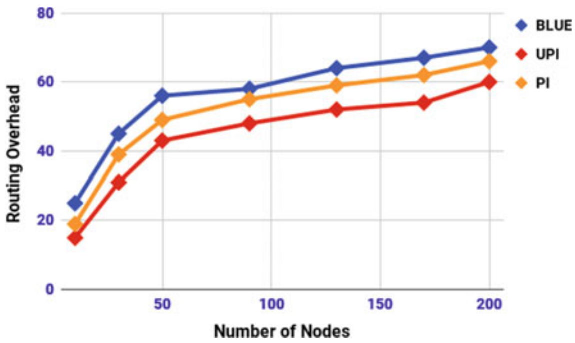


Fig. 1 Packet delivery ratio

Fig. 2 Routing overhead



dynamically. As a consequence, these two algorithms attained lesser PDR compared with the proposed UPI algorithm.

The graph in Fig. 2 shows that the routing overhead is high in Blue and PI compared with the proposed UPI algorithm. Because Blue and PI do not manage the buffer space dynamically thereby many control packets are exchanged. These transmissions consume more resources like bandwidth, energy and queue space. In addition, these transmissions cause congestion and collisions in the network. But UPI algorithm manages the buffer space dynamically using a probabilistic approach rather than random approach. It increases the possibility of successful transmission of control packets without wasting scarce resources thereby minimizing routing overhead.

Figure 3 shows the end-to-end delay that occurred in the proposed UPI, Blue and PI algorithms. It is inferred from Fig. 3 that In Pi algorithm, Calculate_p() method takes more time to calculate drop probability because it does not only calculate drop probability rather it updates the old queue length and current queue length at regular intervals of time. In Blue algorithm, incorrect setting of Pmark values can complicate the whole execution process and can lead to wrong results also. These two algorithms apply the same procedure for all intermediate nodes from source to destination. As a consequence, the end-to-end delay of the packet is increased, whereas the proposed UPI algorithm does not waste time in setting or updating the queue values or other parameters frequently. It does dynamically for every incoming packet during their

Fig. 3 End-to-end delay

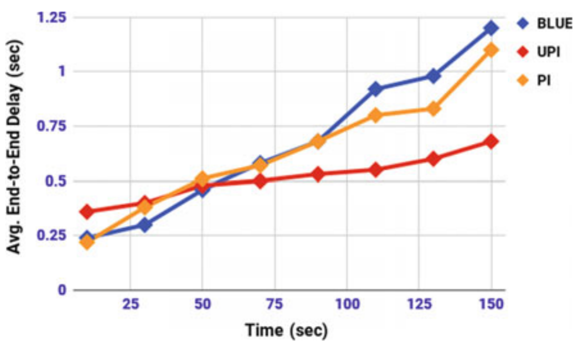
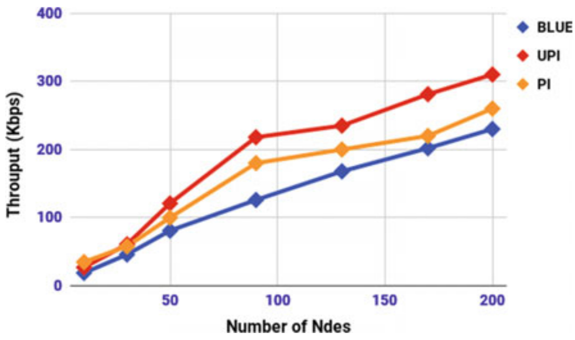


Fig. 4 Throughput



arrival at queue, so that end-to-end delay incurred by the proposed algorithm is minimized compared with other algorithms.

The graph for throughput gained by Blue, Pi and proposed UPI algorithm is depicted in Fig. 4 with varying nodes, since the improvement in result of the three parameters helps to gain better throughput of UPI than Blue and Pi algorithms. In dense network, the value of service probability is more due to higher arrival rate of packet because service probability is directly proportional to arrival rate. If the service probability value is high, it is possible to insert more number of packets in the queue without dropping the packets. It is guaranteed that all these packets are transmitted, and the proposed UPI algorithm achieves better throughput than Blue and PI algorithms.

5 Conclusion

The effective management of buffer space guarantees better network performance by reducing packet loss. But most of the traditional approaches simply drop the packets whenever the queue is full or almost full. In the case of high traffic, most of the time queue is in saturated condition leading to packet drops. Consequently, it triggers several problems like lockout, full queue and bufferbloat problems. To overcome these problems, a novel method called Upgraded Proportional Integral (UPI) is proposed for Active Queue Management in MANET. Based on the current network condition and service probability, the proposed UPI algorithm decides whether to drop the packet or not. Due to run time decision parameters, the random dropping of packets is completely avoided by the proposed UPI algorithm. The NS2 simulation results of proposed UPI algorithm are compared with PI and Blue algorithm over different performance metrics. The results show that the proposed UPI algorithm minimizes end-to-end delay and packet loss ratio without compromising throughput.

References

1. Toh, C.K.: Ad hoc Mobile Wireless Networks: Protocols and Systems. Pearson Prentice Hall (2005)
2. Gupta, P.: A literature survey of MANET. Int. Res. J. Eng. Technol. (IRJET) **3**(2), 95–99 (2016)
3. Hassan, M. Jain, R.: High Performance TCP/IP Networking. Pearson Prentice Hall (2003)
4. Mahida, P.T., Patel, K.: A comparative analysis of queue management techniques using NS-2 simulator. Int. J. Comput. Appl. **65**(6) (2013)
5. Ali Ahammed, G.F., Banu, R.: Analyzing the performance of active queue management. Int. J. Comput. Netw. Wirel. Commun. **2**(2) (2010)
6. Pujeri, U.R., Palaniswamy, V.: Comparative analysis and comparison of various AQM algorithm for high speed. Indian J. Sci. Technol. **8**(3) (2015)
7. Singh, P.K., Sahani, S.K., et al.: Performance evaluation of queue management techniques. Int. J. Comput. Netw. Wirel. Commun. **3**(4) Oct (2013)
8. Mohamed, M.H., Woodward, M.E.: Adaptive drop tail: an algorithm for maintaining delay bounds in a buffer with time-varying arrival rate. In: 3rd International Conference on Information and Communication Technologies, pp. 1–6 (2008)
9. Yin, N., Hluchyj, M.G.: Implication of dropping packets from the front of a queue. IEEE Trans. Commun. **41**(6), 846–851 (1993)
10. Cerf, V., Jacobson, V.: Buffer bloat problem with the internet. ACM Commun. J. **14**(2) (2012)
11. Siregar, B., Manik, M.S.: Implementation of network monitoring using (RED) method. In: International Conference on Communication, Networks and Satellite, pp. 42–47 (2017)
12. Feng, W.-C., Kanlur, D.D., Debanjan Saha, D., Shin K.: BLUE: a new class of active queue management algorithms. J. ACM Trans. Netw. **10**(4) (2002)
13. Abolmasoumi, A et al.: Congestion control of tcp-west wood using AQM-PI controller. Int. J. Comput. Netw. Commun. (IJCNC) **3**(2) (2011)
14. Raghuvanshi, D.M., Annappa, B. Tahiliani, Mohit.: On the effectiveness of CoDel for active queue management. In: 3rd International Conference on Advanced Computing and Communication Technologies (ACCT), pp. 107–114 (2013)
15. Kulkarni, P.G., McClean, S., Perr, G.P., Black, M.: Proactive predictive queue management for improved QoS in IP networks. In: International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies (ICNICONSMCL'06), pp. 7–7 (2006)
16. Friderikos, V. Aghvami, A.H.: Hop based queueing (HBQ): an active queue management technique for multi hop wireless networks. In: Fourth International Conference on Information, Communications and Signal Processing, vol. **2**, pp. 1081–1085 (2003)
17. Niharika, P.T.: A novel method for better queue management. Int. J. Eng. Res. Gen. Sci. **3**(3) (2015)
18. Chandra, A., Kavitha, T.: Adaptive virtual queue with choke packets for congestion control in MANETs. Int. J. Comput. Netw. Wirel. Commun. **4**(2), 2250–3501 (2014)
19. Chen, S., Bensaou, B.: Can high-speed networks survive with DropTail queues management. Int. J. Comput. Netw. **51**(7), 1763–1779 (2007)
20. Dimitriou, S., Tsioliaridou, A., Tsaoussidis, V.: Introducing size-oriented dropping policies as QoS-supportive functions. IEEE Trans. Netw. Serv. Manag. **7**(1), 14–27 (2010)
21. Agrawal, R., Sharma, P., Malviya, V.: A novel method for queue management using RED technique in mobile ad hoc network. In: 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 175–180 (2017)

An Efficient Selection Procedure with an Enhanced Load-Balancing Scheme for GridSim

Deepak Kumar Patel and Chitaranjan Tripathy

Abstract A computational Grid is a distributed network of heterogeneous resources dedicated to execute the user-defined jobs in a more efficient manner. Therefore, a load-balancing scheme is always required in a Grid computing environment for better utilization of the resources and time-efficient execution of the jobs. In this paper, the proposed scheme LBDC provides an efficient selection procedure for better scheduling of jobs among the computational resources so that every resource will remain balanced, and every job will be executed in a time-efficient manner. The proposed scheme LBDC delivers better results in terms of response time as compared to previous load-balancing schemes on GridSim.

Keywords GridSim · Load balancing · Gridlet · Deadline · Response time

1 Introduction

The traditional computing environment has now shifted to more powerful form of distributed computing environment with the inclusion of more powerful efficient computational resources, distributed worldwide connected through a wide area network [1]. A computational Grid has the infrastructure to connect all these resources so that any high-performance computing activities can be performed with respect to various user's necessity [2]. The role of a computational resource in an environment of Grid computing means everything. Because of this, a resource should be always balanced for time-efficient execution of the jobs. A resource should never be in the state of heavilyloaded or lightlyloaded. This status of resource states may harm the overall

D. K. Patel (✉)
Department of Computer Science and Information Technology, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Khordha, Odisha 751030, India
e-mail: deepakpatel@soa.ac.in

C. Tripathy
Department of Computer Science and Engineering, Veer Surendra Sai University of Technology, Burla, Sambalpur, Odisha 768018, India
e-mail: crt.vssut@yahoo.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_41

functionality of the Grid activities. Therefore, an effective load-balancing scheme is necessary for all heterogeneous resources to maintain the complete balance and also to improve the Grid system's performance [3]. A proper resource management and an efficient scheduling are guaranteed by an effective load-balancing scheme which leads to the execution of more number of jobs and also the reduction in the execution time of all jobs [4].

In this paper, we have developed a scheme for load balancing known as "LBDC" which always provides an efficient selection procedure for better scheduling of the many difficult jobs among the best computational resources so that every resource will remain balanced, and every job will be executed in a time-efficient manner. Therefore, a selection method for both jobs and resources is proposed to perform load balancing. Our scheme decreases the response time of the jobs in comparison with previously proposed schemes for GridSim.

The rest portion of the paper is arranged as follows. Section 2 introduces the GridSim simulator and presents the related work. Section 3 describes a load-balancing scheme LBDC. In the next Section, results of the simulation are presented. Finally, Sect. 5 concludes the paper.

2 GridSim

The GridSim simulator is built to simulate a hierarchical Grid environment [5]. In this hierarchy, the top most position is taken by the Gridbroker. It manages the complete hierarchy so that all the Gridresources can schedule all the Gridlets in a load-balanced manner. In the context of GridSim, the job is also recognized as a Gridlet, and a resource is recognized as a Gridresource. A Gridlet contains million of instructions which need to be executed by a Gridresource according to the user's requirement. A Gridresource, which comes after the Gridbroker in this hierarchy, is accountable for the execution of all the Gridlets submitted by the Gridbroker. The Gridlets are assigned to one of the machines present under a Gridresource, and then, the machine executes it using one of its processing elements which come next in the hierarchy [6, 7] (Fig. 1).

2.1 Grid Topology

A topological view of a hierarchical Grid environment on GridSim is shown in the figure as mentioned below. The Grid is structured as a collection of many clusters. Every cluster maintains a number of PEs which are connected together by a local network or LAN. Each cluster is interconnected together by a global network or WAN [5] (Fig. 2).

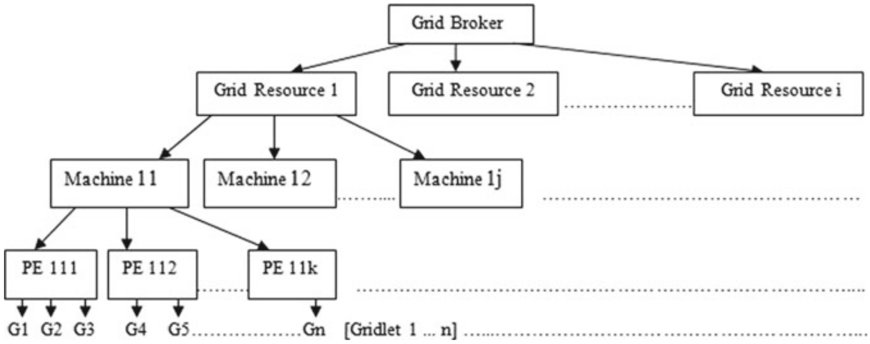
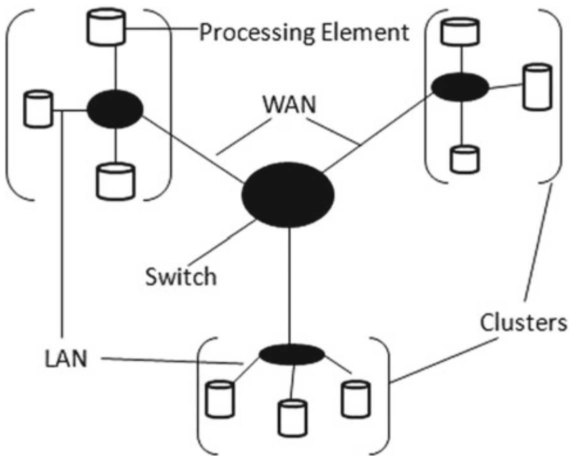


Fig. 1 Hierarchical grid architecture of GridSim

Fig. 2 Topological view of GridSim



2.2 *Related Work*

Some distributed load-balancing schemes have been proposed to maintain the load among all Gridresources present in a hierarchical Grid environment of GridSim. The most mentioned load-balancing schemes are WLB [8, 9], LBGS [10–12], LBEGS [13] and EGDC [14]. The WLB scheme simulates only the scheduling procedure of jobs to resources. But there is no mention of any procedure to do load balancing for the resources. To eliminate this problem, both the LBGS and LBEGS schemes are proposed where the algorithm to maintain the load balance among the resources is simulated successfully. Then, more improvised scheme EGDC is proposed in which job deadline is considered as a main factor in calculation of resource load. Also, this scheme acknowledges that every job must execute within its mentioned deadline with the help of better healthy resources. This scheme has the better results in terms of finished jobs, unfinished jobs and the execution time required to finish that set of jobs than the others. But all these schemes do not have any selection procedure for both jobs and resources to do an efficient scheduling.

3 The Proposed Scheme: Load Balancing on Enhanced GridSim with Deadline Control

Here, a complete stepwise discussion of our proposed scheme (LBDC) is presented (Table 1).

Step 1: Before assigning any jobs to any resources, the most important thing is that not any job should be assigned to any heavilyloaded resource. Therefore, the Gridbroker calculates the current state of all the resources to know the status of every resource present under its influence using the following equation.

$$/ * \text{ Calculating the current load of all resources } * /$$
$$\mathrm{CRL} = \mathrm{JL}/(\mathrm{JD} * \mathrm{PEC} * \mathrm{PE}) \tag{1}$$

Step 2: The Gridbroker predicts the current status of all the resources after calculating every resource's load using the following procedure. In any Grid environment, a resource can be in any three states: *heavilyloaded, balanced* and *lightlyloaded.* Here, the prediction procedure is as follows: a heavilyloaded resource's load is more than REL, a lightlyloaded resource's load is less than MEL, and a balanced resource's load is in between MEL and REL. Due to this procedure, the Gridbroker can list the resources in the proper lists such as HLRL, LLRL and BRL to do load balancing.

Table 1 Notations

Notations	Description
PE	Number of PEs under a resource
CRL	Calculated resource's load
PEC	Calculated PE's capacity
JL	Job Length
LLRL	Lightlyloaded resource's list
HLRL	Heavilyloaded resource's list
BRL	Balanced resource's list
UAJL	Job's list going to be scheduled
JD	Job deadline
CET	Calculated execution time
REL	Resource estimate load
MEL	Machine estimate load

/*Checking every resource's state*/
for every resource does
Estimating CRL;
if (CRL < MEL)
Assigning the lightlyloaded resources into LLRL;
else if (CRL > REL)
Assigning the heavilyloaded resources into HLRL;
else
Assigning the balanced resources into BRL;
end for

Step 3: Now, the Gridbroker transfers some jobs from the heavilyloaded resources to make them balanced and assign those jobs to lighlyloaded resources to make them more useful until all the resources become balanced. These jobs are stored in the list UAJL waiting to be scheduled on the lightlyloaded resources.

Step 4: Now, the priority is to assign the many difficult jobs coming from the UAJL to one of the best lightlyloaded resources present in the LLRL. Therefore, we provide a selection method for both jobs and resources to find out the difficult jobs and also the best underloaded resources to do an efficient scheduling.

In this selection method, there is one arrangement for jobs. Every job of the UAJL must be completed on any lightlyloaded resources within the Gridbroker's assigned deadline. Therefore, from the beginning of scheduling, the aim always must be to execute the very difficult jobs means the jobs with the lowest deadline on the best lightlyloaded resources. Therefore, the Gridbroker arranges all the jobs of UAJL according to the lowest deadline assigned to those jobs.

For resources, this selection method has one calculation and one arrangement. First, the Gridbroker predicts the approximated execution time of the job with the lowest deadline of the UAJL on all the lightlyloaded resources of the LLRL using the following equation.

$$/ * \text{ Calculating the execution time on the lightlyloaded resources } * /$$
$$\text{CET} = \text{GL}/\text{PEC} \tag{2}$$

Second, the Gridbroker arranges the lightlyloaded resources of the LLRL according to the lowest execution time which is calculated using Eq. 2. So, the Gridbroker schedules jobs with the lowest deadline on the most lightlyloaded resources having the possibility of least execution time. Due to this, more number of jobs can be executed within less amount of time.

Step 5: Here, the Gridbroker can assign the jobs of the UAGL to the lightlyloaded resources of the LLRL using the following scheduling procedure.

/*Executing jobs of UAJL to the lightlyloaded resources of LLRL*/
while (LLRL and UAJL are not empty)
for all resource r of LLRL
for all job j of UAJL

factor = RL + ((JL)/(JD *PEC*PE));
if (factor <= REL)
job j is assigned to resource r;
else
job j is assigned to LLRL's next lightlyloaded resource;
end if
end for
end for
end while

4 Simulated Results

The proposed scheme was simulated by the simulator GridSim5.0 [10]. Here, each resource provider has two machines, and each machine has five PEs. So, ten PEs are present in every resource. Here, each PE has a capacity to execute 1–5 MIPS. Also, every job length is expected in between 1 and 5 million instructions. Each job is provided with a deadline of 1–6 s. Now, the proposed scheme is simulated to evaluate the response time of all the jobs using two cases as given below.

4.1 Parameters

See Table 2.

Table 2 Parameter and values

Parameter	Values
Job deadline	1–6 s
Job length	1–5 million instructions
Each resource's machine no.	2
Network bandwidth	0.5–10 Mbps
Rating of PE	1–5 million instructions per second
Each machine's PEs no.	5
Resource estimate load	0.8
Machine estimate load	0.75

4.2 Performance Evaluation

1st Case: With the consideration of fixed resources
2nd Case: With the consideration of fixed Jobs

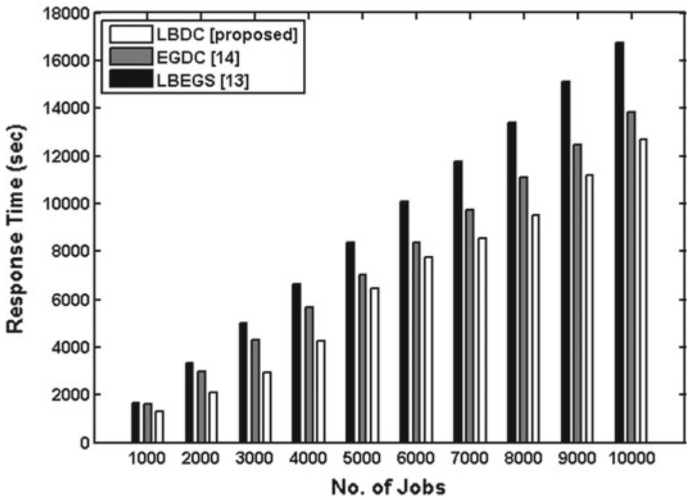


Fig. 3 Response time versus number of jobs

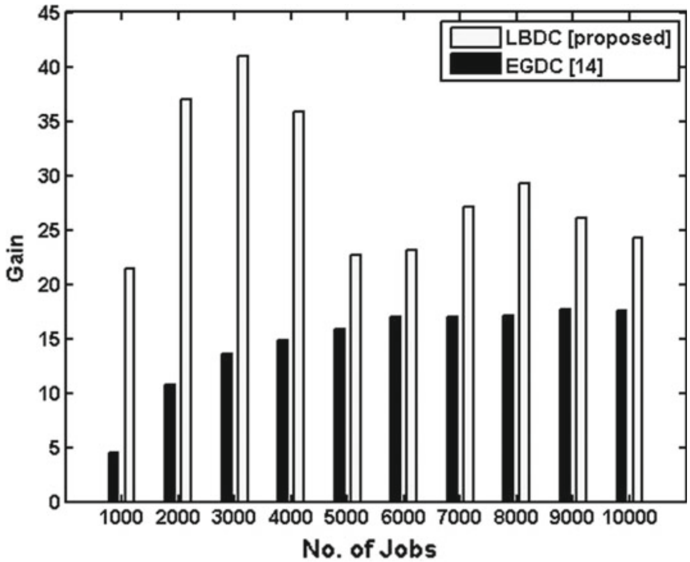


Fig. 4 % Gain in response time versus number of jobs

In both cases, this proposed scheme LBDC performs better than the previously discussed schemes [13, 14] in comparison with response time. The LBDC takes less time to execute all the submitted jobs than the existing ones due to the proposed selection method as shown in Figs. 3, 4, 5, 6 and Table 3.

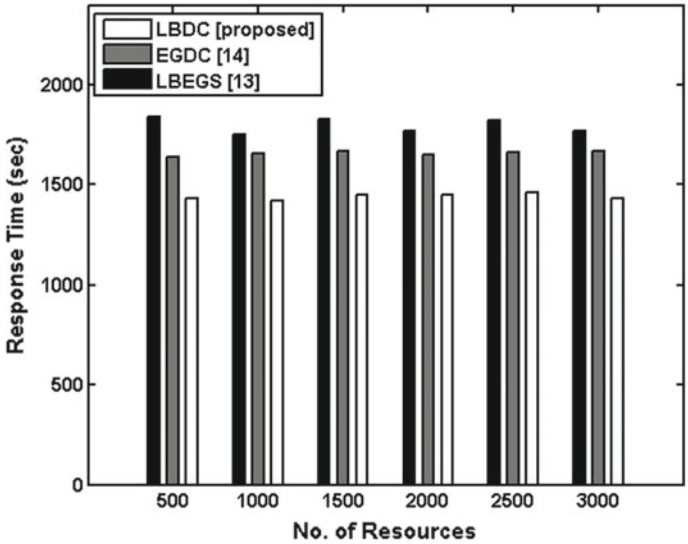


Fig. 5 Response time versus number of resources

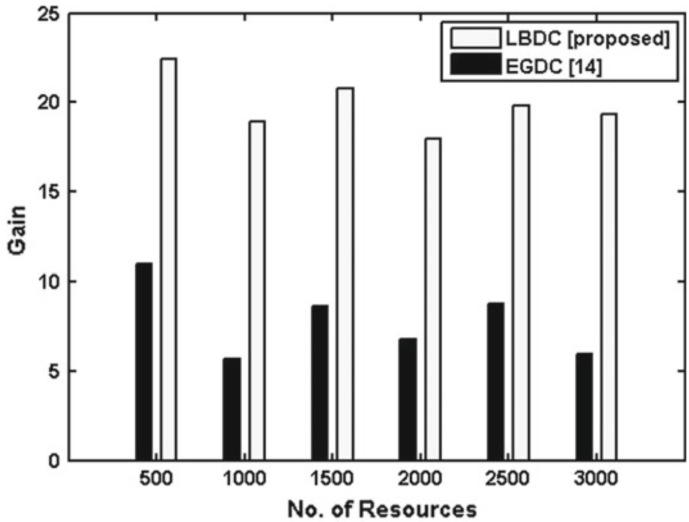


Fig. 6 % Gain in response time versus number of resources

Table 3 Results

Cases	Figure No.	Performance metric	Performance evaluation
1st case	Figure 3	Response time (RT)	Decrement
	Figure 4	% Gain in RT	Increment
2nd case	Figure 5	Response time (RT)	Decrement
	Figure 6	% Gain in RT	Increment

5 Conclusion

Here, a new effective load-balancing approach (LBDC) is presented according to the hierarchical Grid computing environment. This proposed scheme provides an efficient selection procedure for better scheduling of jobs among the computational resources. The proposed scheme LBDC minimizes the execution time of all the jobs as compared to previous load-balancing schemes on GridSim due to the proposed selection procedure. The proposed scheme was simulated by the GridSim5.0 simulator. The proposed scheme can be enhanced further with the inclusion of an efficient state prediction procedure on GridSim.

References

1. Patel, D.K., Tripathy, D., Tripathy, C.R.: Survey of load balancing techniques for Grid. J. Netw. Comput. Appl. **65**, 103–119 (2016)
2. Patel, D.K., Tripathy, C.R.: An Improved approach for load balancing among heterogeneous resources in computational Grids. Eng. Comput. **31**, 825–839 (2015)
3. Patel, D.K., Tripathy, C.R.: An efficient load balancing mechanism with cost estimation on GridSim. In: Proceedings of the 15th IEEE International Conference on Information Technology (ICIT'16), Bhubaneswar, India, pp. 75–80 (2016)
4. Patel, D.K., Tripathy, D., Tripathy, C.R.: An improved load balancing mechanism based on deadline failure recovery on GridSim. Eng. Comput. **32**, 173–188 (2016)
5. http://www.buyya.com/GridSim/
6. Sulistio, A., Poduval, G., Buyya, R., Tham, C.K.: On incorporating differentiated levels of network service into GridSim. Futur. Gener. Comput. Syst. **23**, 606–615 (2007)
7. Howell, F., McNab, R.: SimJava: a discrete event simulation package for Java with applications in computer systems modeling. In: Proceedings of the 1st International Conference on Web-based Modelling and Simulation, Society for Computer Simulation, San Diego, CA (1998)
8. Buyya, R., Murshed, M.: GridSim: a toolkit for the modeling and simulation of distributed management and scheduling for Grid computing. J. Concurr. Comput. Pract. Exp. **14**, 13–15 (2002)
9. Murshed, M., Buyya, R., Abramson, D.: GridSim: A toolkit for the modeling and simulation of global grids. Technical Report, CSSE, Monash (2001)
10. Yagoubi, B., Slimani, Y.: Task Load balancing strategy in grid environment. J. Comput. Sci. **3**(3), 186–194 (2007)
11. Yagoubi, B., Slimani, Y.: Load balancing strategy in Grid environment. J. Inf. Technol. Appl. **4**, 285–296 (2007)

12. Yagoubi, B., Slimani, Y.: Dynamic load balancing strategy for grid computing. World Academy of Science, Engineering and Technology, pp. 90–95 (2006)
13. Qureshi, K., Rehman, A., Manuel, P.: Enhanced GridSim architecture with load balancing. J. Supercomput. 1–11 (2010)
14. Hao, Y., Liu, G., Wen, N.: An enhanced load balancing mechanism based on deadline control on GridSim. Futur. Gener. Comput. Syst. **28**, 657–665 (2012)

Advanced Algorithms and Software Engineering

Designing a BYOD Supported Formative Learning Interface and Analysis of Classroom Interaction

Suman Deb, Nisha Saha and Paritosh Bhattacharya

Abstract Increase in the use of portable and personalized devices like smartphone, tablet, etc., that has changed the way of traditional interaction in face-to-face communication. This work is to explore the popularity of these devices in an affirmative way to enhance the classroom interaction. In the present classroom, even with the use of multimedia demonstration of educational content, it is difficult to set up an interactive one-to-one relation between a teacher and students. With the affordable self-paced learning platforms students are already equipped with the content to be discussed in the classroom; thus, making it difficult to maintain the integrity and interest level of a classroom beyond a certain time. To maintain pace with the latest technology that pre-feeds and provided on-demand knowledge into the student population, we are proposing a BYOD-based platform to increase the interaction of students in classroom by a one-to-one interaction relation with teacher. This can be summarized as a computer-mediated predictive analysis of students interest that may allow the teacher to steer the classroom interest for catering to every students individual needs. As most of higher classes students move with their devices, we propose here a framework to utilize the advent of BYOD to increasing the formative learning and assessment of student interest to avert the boredom of monotonous learning in classroom.

Keywords BYOD interaction · Formative learning · Predictive analysis of performance · Classroom interaction

S. Deb (✉) · N. Saha · P. Bhattacharya
Computer Science and Engineering Department,
National Institute of Technology Agartala, Jirania, India
e-mail: sumandeb.cse@nita.ac.in

N. Saha
e-mail: nisha.kanta93@gmail.com

P. Bhattacharya
e-mail: pari76@rediffmail.com

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_42

1 Introduction

BYOD—Bring Your Own Device, a very popularized concept to bring ones interactive platform to your workplace, schools and colleges has been enabling employees to work more freely and interactively and same goes for students getting access to the huge knowledge database of the Internet to enhance their classroom learning experience. But the access to Internet is often debatable during classroom lectures whether beneficial or distracting. But if tailor-made systems are developed [10, 19] for cross-platform devices with proper monitoring, the same concept can be revolutionary in increasing the interaction level of the teacher and students. The mundane experience of the production line system of current education can be converted [2, 15] to a one-to-one interaction between a mentor and an explorer with infinite resources at hand [8]. In this work approach, we have theorized few ways by which the BYOD concept can be leveraged [5, 6, 16] (Fig. 1).

1.1 Audience Response Systems

A part of the BYOD application in student interaction is the use of audience response systems (ARSs). It will act as a tool in our application for a passive response system during the student–teacher interaction scenario, during classroom lectures or online sessions [1, 18]. The audience response system finds its use in a multitude of scenarios where the cumulative, passive and anonymous response from audience is taken to be viewed by presenter. They will consist of input devices for students (portable interactive devices—smartphones, wearable devices and tablets), and a central server/cloud platform as well as a software, for the communication and the calculation of the results. The teacher will ask questions (commonly multiple-choice questions) or ask for feedbacks or repeats, with the ARS and all students can answer, react and share live review—on the input device. The central server will fetch all results and process them immediately to be presented to the teacher, to modify as

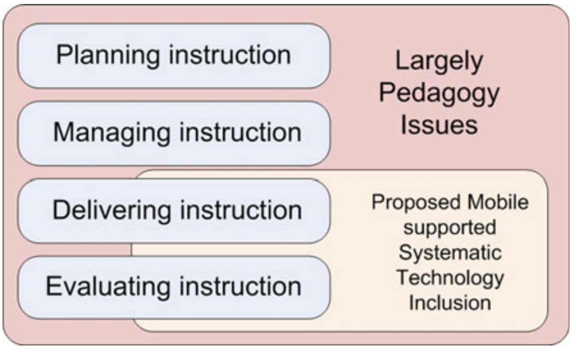


Fig. 1 Proposed technology inclusion with traditional pedagogy

well as improvise the content based on the reactions fetched. ARSs have engaged researchers since the 1960s. However, since the technology was not as highly developed as nowadays, these ARSs were very complex and expensive systems using wiring of all the response devices together for collecting data as the concept of PCs and Internet were still vague. One very famous ARS was classtalk, which was developed in 1996 [12]. This system offered new ways and possibilities of interacting with students during lectures. ARSs have changed, since the rise of wireless technologies. Currently, the input devices are no longer needed to be connected via cable but via Internet to a central server or a cluster of servers/clouds based on load. The system became portable and could be used in different lecture halls instead of one. In the last years, the usage of mobile Internet devices such as smartphones, tablets and laptops has increased considerably also the increasing bandwidth has allowed users to engage in applications using large chunks of data at a very high rate, applications like virtual reality, live video streaming and live polling. Students bring their own devices to lectures—this policy is called bring your own device—BYOD. Despite the risks of BYOD, it seems to be obvious to use the devices of the students as input devices for ARSs. ARSs provide students a way of interaction in massive courses.

1.2 Attention and Usage Monitoring

The second way BYOD can be useful in increasing the quality of classroom lectures is to utilize the inbuilt hardware of the portable interactive devices in conjunction with the devices feature of allowing interaction between different processes [3, 4, 11]. With the consent of the students, data on the time spent on applications (educational, semi-educational and non-educational) can be collected and stored for analysis to create a formative prediction on the students interest in the topics taught in the classroom or topics inline to be taught. Any case of repeated deviation into related fields can indicate the development of interest attained through lecture and teacher can be notified for assistance or improvise on classroom content to tailor to that student. Moreover, sensors like accelerometers and proximity sensors can give live feed about body movements, device usage, along with non-lecture related app activities during the lecture to the central server. In the central server, we can deploy a machine learning-based model that will be able to analyze the attention span of the students in the classroom environment [17]. Their affinity toward the topics taught and their level of interaction with the course can be identfied in formative manner that will be helpful to both teacher and student to make necessary changes in teacher taught activities [9].

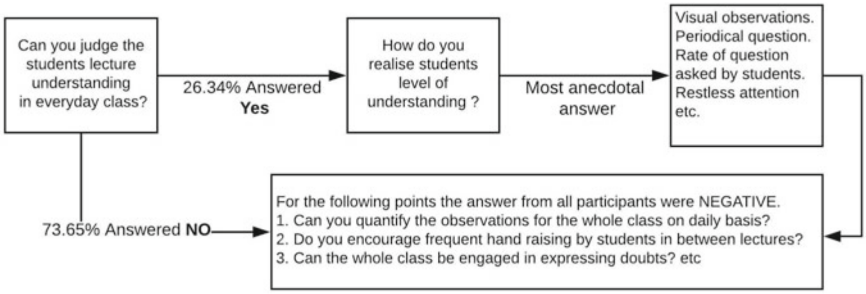


Fig. 2 Pretest questionnaire and anecdotal report to validate requirements

1.3 Learning Assistance

The third way BYOD platform can be beneficial to the students in classroom [14] is by providing live learning aid [13]. With the lecture video being directly fed to the student devices, links to technical dictionaries for the detailed description of keywords discussed in lectures can be provided. References to terms can be provided via hyperlink. For ease of understanding, subtitles in multiple languages can be provided, markers across the live stream to point doubts to keep as a reminder to be discussed later on, or if the number of students having the doubt at the same point in lecture is considerable. The teacher can be notified to discuss in detail. This publication covers the topic of theorizing a Web-based BYOD-based classroom architecture, which uses mobile Internet devices of students as input devices. A state-of-the-art analysis shows the most crucial facts, which must be kept in mind when developing such a system [7, 20]. Out of these points, different requirements will evolve. The conclusion and future work will follow (Fig. 2).

2 Implementation Plan

The BYOD student–teacher interaction platform and behavioral analysis, can be realized both as a cloud-based or centralized server-based coupled with user interfaces for teacher and students. As shown in Fig. 1 the user interface can be realized as a Web application and will run on all modern browsers or a Web-based app for Android/iOS devices can be developed. This approach fulfills the requirement to provide cross-platform capabilities. BYOD comes along with a heterogeneous device environment policy because mobile Internet devices include laptops, smartphones, tablets, smart watches, etc. To support the different screen resolutions and aspect ratios on these devices, the responsive design approach can be used. With responsive design, the appearance can be optimized for different devices based on their typical screen resolution with the use of Bootstrap framework. This approach needs less coding effort compared to writing a separate user interface for different devices without using Twitter Bootstrap. HTML along with Bootstrap user interface (UI) library, which supports responsive design by default, is the best solution for all devices that can browse over Internet, thus avoiding specialized apps for different platforms. The business/application logic on the client side generally uses HTML and JavaScript while the backend framework is realized through plain PHP, Python or JSPs. But to have a cleaner, easily deployable framework an MVC platform would be suggested. There are two popular model view controller paradigms (MVC) in market, viz.—Django, MEAN Stack (Mongo DB/MySQL/ MSSQL, ExpressJS, AngularJS, NodeJS). The model view controller paradigm (MVC) can be realized either with Django—a very popular backend application interface that allows quick deployment, asynchronous connectivity as well as ease of scalability, it also provides a huge range of open-source app library for quick development and deployment. Django is based on Python, another similar Web backend framework based on JavaScript that can be utilized is Node.js. The Bootstrap framework utilizing jQuery will make client-side application cross-browser/cross-screen size capable. ExpressJS can be used as a view templating engine combined with body. Parser module of NodeJS will help us in realizing a dynamic Single Page Application (SPA) Platform. The last remaining component of the MEAN (Mongo DB/MySQL/MSSQL, ExpressJS, AngularJS, NodeJS) Stack is AngularJS that can be utilized further in implementing client-side business/application logic. Angular JS Modules gives us the ability to implement routing within the Single Page Application (SPA) though it can be done using through NodeJS too, Angular directives give us the ability to present data served by the persistence layer dynamically within the static HTML page. Moreover, Angular gives us two-way data-binding and controller modules to do further processing on the client side. With the increasing processing capacity of the client-side devices, we can distribute the processing load between the client and the server. Unlike past legacy frameworks, Angular gives us the ability to do complex and modular processing with its dependency injection feature. Now the front end-clients of both the teacher and the students can be implemented by hosting an application server or an application server cluster. The server thus realized through NodeJS if a MEAN stack is considered will have the role of persisting the data and serving the data to the front end and further realization of the business logic for behavioral analysis would be done here. It will also give the live stream of the feedback data to the teacher, learning aids and lecture feed to the students, and moreover, Node gives us an advantage in serving such requests with its inherent load balancing and asynchronous request handling abilities. For alternate implementation plans of the server, we do have Python programming language, J2EE aided with Java Server Pages (JSPs), PHP (another popular language) coupled with Apache server framework; Django coupled with Jinja view serving module is exact alternate of the MEAN Stack, unlike aforementioned stacks but the drawback being dependency on JavaScript again for client-side processing. Not utilizing a complete framework to create the server will lead to many problems in the future maintenance, scaling, enhancement and patching. Thus, a complete framework like Django or MEAN will help us in realizing a fast, easily maintainable, secure and modular server. The persistence layer can be realized by a combination of relational and NoSQL databases, since the data fetched for behavioral analysis and student feedback and other metadata would be a combination of relational and cluttered data, though the post-processed data would contain a higher quantity of relational data, requirement of NoSQL data would still be there. Thus, using frameworks like Django and MEAN coupled with MongoDB/PostgreSQL will give us the ability to scale both the application server as well as database cluster horizontally. The client-side of out application (student and lecturer interface) can communicate through a RESTful server interface as shown in Fig. 1. The process of registration on the application will be on need-basis, the lecturer can register to see feedback from students and uploading or modifying teaching aids or feedback parameters. The students can register but can use features of feedback anonymously. The student can register to analyze their progress, downloading proprietary content. Thus, the process of registration can be optional only for use of advanced features, but the realization of more one-to-one environment between teacher and student would require the registration of both parties. An agile development process can be used for developing the system (Figs. 3 and 4).

3 Advantages and Disadvantages

The BYOD concept is as revolutionary as the term controversy is associated with it. The BYOD concept reduces the cost of the institution in providing every individual with a specialized device. The opportunity to leverage the handheld hardware already in place helps in concentrating on the software for a more focused approach toward the target. This is beneficial for both institutions and members—reducing cost and efforts on the part of the institution, getting more hours of commitment from the members due to increased interaction and flexibility of the members. Few of these issues have been discussed.

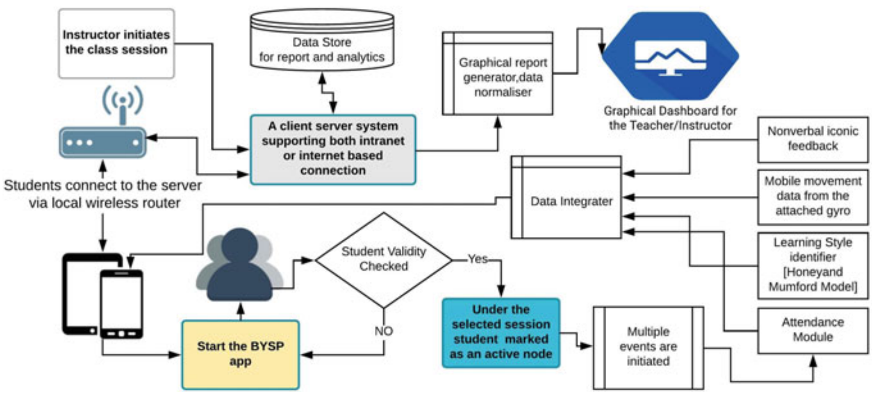


Fig. 3 System architecture to accomplish a closed-loop interaction

3.1 Familiarity with Devices

The institution can provide devices for implementing this interactive platform, but it increases the overhead of allocating funds for procuring the devices and there is an additional overhead of training the students on a common platform supporting the application. On the other hand, both the teacher and the student community in today's time are using portable computing devices that support either platform-based applications or cloud-based applications through clients. Since the community are already familiar and habituated with the personal devices, a cross-platform application coupled with BYOD will enable the students to focus more on the lessons and feeding data to the platform on behavioral analysis. The financial and time overhead for deployment and training would be removed.

3.2 Student's Affinity to the Device Will Be Utilized Devices

Nowadays the increasing diversity and quantity of entertainment applications, social networking the student community is already glued to their devices. Using this affinity to their devices to impart lesson as well as to take feedback both conscious response as well as subconscious biometric feedback, will be a great way to inject knowledge in the way they like it. This will be much more effective than the traditional many to one, selective feedback existing in the classroom environment now. The enhanced level of interactivity rather than simply watching and listening a monotonous classroom feed.

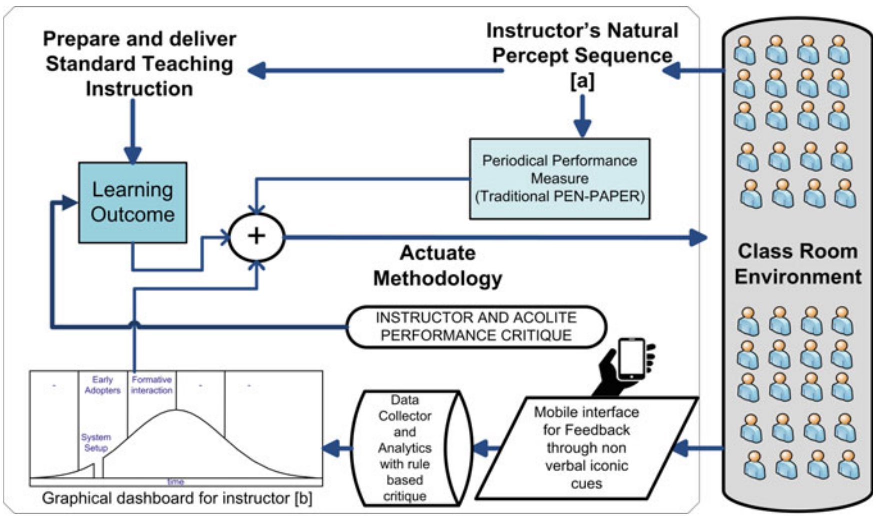


Fig. 4 BYOD supported many to one interaction with feedback and background analysis

The ability to interact with the study material provides feedback and controlling the content of study with your interest and behavior will be sure to increase the attention span in a classroom environment.

3.3 Current Technology Helps Assimilating Content Easier

The sheer advancement of technology has led to the introduction of many portable devices with higher processing capability, connectivity, interactivity and memory. These combined with the wide array of sensors, viz.—capacitive touch, proximity sensor, accelerometer, GPS, camera, etc., have increased the field of application of the devices beyond imagination. All these things with software applications implementing augmented reality, demonstration videos, tasks and 3D visualization of objects are creating space to have learning aids that help students realizing the concepts easier and retain the lessons for longer period of time. All these things when assigned to BYOD, learning can be regulated at their own pace rather than limiting lessons to classroom time bindings and pace. Although BYOD offers a great deal of advantage in formative learning experience, this platform has its disadvantages too. Lets have a look at them.

3.4 Distraction from Other Application in the Device

It can not be denied that the education platform thus introduced through BYOD will be much beneficial, but the question arises—with all other applications with push notification mechanism, it is really easy to be distracted from the actual goal of formative learning. Even though we may decide to restrict access to few Web sites through firewalls, we cant deny the fact that the students would be quite tech savvy to bypass such mechanisms using VPN, moreover we cant restrict their access to applications that do not require network connectivity. So, with proper consent, if their usage of the device is monitored during classroom hours, this disadvantage can be abolished as per our proposal.

3.5 Security Concerns

Security, another major concern in the present context of BYOD implementation. The network of the institution has to be properly secured. With so many devices and so much data movement, the major security concern that every organization is facing is data theft. Moreover, with every student implementing and owning multiple devices in this BYOD platform, it has become tougher to monitor the activity of each device, register their physical address. Moreover, anonymity and data tapping have become an increasing threat. Thus, before the implementation of BYOD, an institute must revamp its security measures quite well.

4 Conclusion

It is quite evident that students are moving with their smartphone in hand. Mostly, the Internet connected these devices are a trigger for distraction. BYOD framework in conjunction with regulated application applying various computer-mediated interactions definitely enhances the learning experience of the students. This cost effective and less infrastructure demanding solutions transform the classroom environment to a one-to-one teacher–student interaction session. This will also open up the pragmatic designs, predictive devices that will make the process of learning effortless.

References

1. American Psychological Association. How to Build a Better Educational System: Jigsaw Classrooms (2003)
2. Barry, B.E., Jonathan Bodenhamer, M.: Understanding Your Students Nonverbal Communication: A Primer for the New Engineering Educator. American Society for Engineering Education (2011)
3. Beatty, I.D., Leonard, W.J., Gerace, W.J., Dufresne, R.J.: Question driven instruction: teaching science (well) with an audience response system. In: Banks, D.A. (ed.) Audience Response Systems in Higher Education Applications and Cases, pp. 96–115. Idea Group Inc. (2006) (chapter 7)
4. Beatty, I.D., Gerace, W.J.: Technology-enhanced formative assessment: a research-based pedagogy for teaching science with classroom response technology. J. Sci. Educ. Technol. **18**(2), 146–162 (2009)
5. Carnegie Mellon University. What is the Difference Between Formative and Summative Assessment? (2017)
6. Crowin. What are the components of effective instruction. https://www.corwin.com/sites/default/files/upm-binaries/8955_Ch_1_ex_from_Algozzine4_Rev_Final_Pdf.pdf (2017)
7. Dohen, M., Schwartz, J.-L., Bailly, G.: Speech and face-to-face communication an introduction. Speech Commun. **52**(6), 477–480 (2010)
8. Drummond, T.: A Brief Summary of the Best Practices in College Teaching (2016)
9. Dufresne, R.J., Gerace, W.J., Leonard, W.J., Mestre, J.P., Wenk, L.: Classtalk: a classroom communication system for active learning. J. Comput. High. Educ. **7**(2), 347 (1996)
10. Elmes, J.: Six Significant Challenges for Technology in Higher Education (2017)
11. Haggarty, L., Postlethwaite, K.: Strategies for improving communication between teachers and school students about learning: a university/school collaborative research project. Educ. Act. Res. **10**(3), 449–478 (2002). https://doi.org/10.1080/09650790200200195
12. Hattie, J., Timperley, H.: The power of feedback. Rev. Educ. Res. **77**(1), 81–112 (2007)
13. Kelsey, S.: Premier reference source. Computer-Mediated Communication: Issues and Approaches in Education: Issues and Approaches in Education. Information Science Reference (2011). ISBN: 978-1-61350-078-1
14. McDaniels, N.: How technology in the classroom is destroying education. https://marquetteeducator.wordpress.com/2010/12/15/how-technology-in-the-classroom-is-destroying-education/ (2010)

15. Mishra, P., Koehler, M.J.: Technological pedagogical content knowledge: a framework for teacher knowledge. Teach. Coll. Rec. **108**(6), 1017–1054 (2006)
16. Radford, K.W.: Observing the class. Educ. Can. **30**, 36–39 (1990)
17. Saa, P., Moscoso-Zea, O., Lujan-Mora, S.: Bring your own device (BYOD): students perception privacy issues: a new trend in education? (2017)
18. Stanley, A., Erin Porter, M.: Engaging Large Classes: Strategies and Techniques for College Faculty. ISBN: 978-1-119-11124-5
19. Summerlee, A.J.S.: Challenge of engagement inside and outside the classroom: the future for universities. http://www.portlandpress.com/pp/books/online/wg85/085/0067/0850067.pdf (2010)
20. Walther, J.B.: Selective self-presentation in computer-mediated communication: hyperpersonal dimensions of technology, language, and cognition. Comput. Hum. Behav. **23**(5), 2538–2557 (2007)

Factors Affecting Sprint Effort Estimation

Melvina Autar Ramessur and Soulakshmee Devi Nagowah

Abstract Scrum projects consist of multiple iterations, known as sprints during which user stories are implemented. Estimating user stories of an iteration accurately is important in order to provide clarity and help management to control the project successfully. This leads to the necessity of identifying factors that impact the accuracy of a sprint effort estimation. This paper therefore aims to identify factors and reasons for inaccurate effort estimation of a sprint. For this purpose, a survey is conducted in a scrum environment with professionals with proven agile expertise. The survey is based on 15 completed small-scaled agile projects in a well-known Mauritian company. Results of the survey are summarized in the paper, and recommendations are made.

Keywords Effort estimation · Scrum · Sprint · Factors

1 Introduction

Effort estimation is a crucial activity in a software project as it helps to establish realistic and credible plans to implement the project requirements and satisfy commitments. The failure or success of the project depends hugely on the effectiveness of the estimation done [1]. By improving the accuracy of effort estimation, the project planning becomes efficient and provides various benefits for an organization. In order to achieve accurate estimations, it is important to cater for the risks to which a project is or may be exposed during its implementation. Thus, this study tries to identify factors that impact estimation accuracy at the level of a sprint. The project is assured to be completed as planned and within expectations with a good sprint planning and estimation.

M. A. Ramessur · S. D. Nagowah (✉)
Software and Information Systems Department, University of Mauritius, Reduit, Mauritius
e-mail: s.ghurbhurrun@uom.ac.mu

M. A. Ramessur
e-mail: melvina.autar1@umail.uom.ac.mu

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_43

For an agile project, when features planned to be completed in one iteration spill over onto the next, delays are triggered in subsequent iterations which consequently increase the delivery time of the project. Therefore, having a good initial sprint-level estimation avoids this problem. The paper therefore aims to identify the relevant factors that directly affect sprint effort estimation. These factors will then be considered in the next phase of the research which consist of developing a model for predicting sprint effort estimation. The remainder of the paper is structured as follows: Sect. 2 gives an overview of agile effort estimation techniques and the factors that affect effort estimation. Section 3 presents results of a survey carried out to identify factors that affect sprint effort estimation. Discussions about the results are presented in Sect. 4. Future works are highlighted in Sect. 5. Finally, Sect. 6 concludes the paper.

2 Agile Effort Estimation

Software estimation is the process of predicting the amount of effort, resources, time and money required to build a software system. Inaccurate effort estimates lead to wrong decisions in controlling and managing software projects. There are various software effort estimation models that have been proposed in the past decades [2]. The popular ones are COCOMO and Function Point Analysis (FPA) [2–6]. The traditional estimation methods, COCOMO and FPA, are not effective when dealing with agile software estimation due to varying requirements and incremental development [7]. This prompted the introduction of other estimation methods for estimating agile projects.

The application of effort estimation methods in agile projects is very challenging. This is because the scope is continuously adjusted throughout the project and new tasks are discovered even during later stages of software development [7]. Customer requirements are subjected to change based on changing technology and domain, budgets and political influence. Due to uncertainty in requirements, there may not be adequate information to estimate upfront and thus estimating the effort of the complete agile project becomes more complex and difficult [5].

Many effort estimation methods in agile software development, such as analogy, expert judgement, planning poker and story points, use subjective expert effort estimation [7]. Analogy-based method depends on historical data of previous projects. The new project is compared to similar projects for which cost, time and effort data are known [8]. Expert opinion approach requires opinion and experience of multiple experts from different areas of software development to estimate the size and time of the stories of the project. Experts' estimates are based on intuition [3]. Story point-based method is the most currently used estimation technique in agile software development. A story point is an estimate that includes the effort involved, the complexity and the inherent risk in developing a feature. It is relative in nature; that is, a three-point value is assumed to take trice the effort than a single point value story [5]. User stories are normally estimated for release and sprint planning [8, 9]. A developer needs to have some experience of estimating and access to historical data to set an effort estimate for developing a user story. Soliciting expert opinion to determine effort is also useful. It is, however, difficult to find suitable experts with a variety of domain skills [5]. Planning Poker is one of the recent expert estimation techniques that is widely used in agile methods for software development, especially scrum and extreme programming [10]. It is based on story point for estimating the size of user stories and developing release and iteration plans [10]. This technique encourages the involvement of the whole development team during the estimation session, thus making it more effective [10]. It is assumed to provide better estimates than individual experts as participation of people with different backgrounds helps in reducing the over-optimism of expert judgment-based estimates by identifying more issues affecting implementation [10]. Table 1 presents the different factors that affect agile effort estimation.

3 Sprint Effort Estimation

Many agile methodologies have evolved over a period of time. Scrum remains the most popular and practiced one. Scrum is an iterative and incremental framework for application or product development which is achieved through iterative cycles known as sprints [11]. A sprint is relatively a short period of about 2–4 weeks in which a number of user requirements in the form of user stories are completed by the development team [12]. At the start of each sprint, a cross-functional team selects items from product backlog and commits to complete the items by the end of that particular sprint. The product backlog contains the requirements of the project and is refined and updated through the lifetime of the project by the product owner. Every day, the team gathers for a short meeting to discuss the progress and blocking issues, if any. At the end of the sprint, the team reviews the work product with stakeholders and demonstrates what has been built. The feedback is then incorporated in the subsequent sprint. At the end of each sprint, scrum emphasizes that the integrated working software is fully tested and potentially made shippable. The sprints are strictly time-boxed and occur sequentially.

A survey was carried out in a company in Mauritius to determine the list of factors that have an impact on sprint effort estimation. A questionnaire is therefore designed for conducting a survey on 15 completed small-scaled agile projects in the company which is well known in providing information technology, outsourcing and consulting services. The company implements an agile environment in which a variety of projects are developed by distributed cross-functional teams. Most of its projects employ scrum methodology which is successfully introduced in teams and guided by high-skilled scrum masters. Most existing effort estimation systems consider factors that affect the estimation of a whole project. However for this study, the primary focus is to determine whether factors affecting the estimation of a whole project are same for a sprint and if not, which factors that affect the estimation of a project's sprint. The questionnaire is distributed to each project's scrum master

Table 1 Factors affecting agile effort estimation

Factors	Details
Stakeholders	One factor that has to be considered when estimating effort in an agile team is stakeholders. Often, team's work gets delayed when stakeholders are not responsive to their requests for information [13]
Team dynamics and commitment	Team familiarity is beneficial in reducing the duration and effort of a project as it helps team members to coordinate, communicate efficiently and share knowledge among them [14]. However, team performance may be affected with excessive familiarity between team members as this can cause team errors. Team errors often occur in the interactions of members in the team [15]. Additionally, the absence rate and separation rate are low for a team with high commitment levels [16]
Communication	Communication skills (writing, reading, listening, speaking, inter-debate, etc.) are important to reduce the duration in interacting with others. Team members work efficiently and effectively when there is a good communication as they take less time to understand each other's work [15]
Staff experience and technical ability	Technical skills, commitment, experience and qualifications of staff are important factors to consider for estimation. Turnover of qualified staff may affect estimation previously done [17]
Experience of previous project	The estimation of effort is more likely to increase if team members do not have past knowledge and proficiency in similar previous projects [17]
Responsibilities outside of the project	The productivity of team members who work on more than one project may reduce when shifting between projects [13]
Configuration	Configuration for the proper functioning of software in terms of hardware/software requirements and Internet requirements requires special efforts [3]
Security	Security includes network security, functional security, code security and documentation security. Based on customer requirements, security may increase the complexity as well as the effort of the project development [15]

(continued)

Table 1 (continued)

Factors	Details
Defects and changes of previous implementation	Defects can be a result of changing, missing/undiscovered or incorrect requirements, poor requirement specification or quality, misinterpretation of tasks due to poor communication, non-functional requirements when testing the system limits, integration defects and logical errors due to misunderstanding of domain. Defects or changes from customer introduce rework or re-implementation, thus affecting the effort and efficiency of sprint planning [18]
Quality of requirements	Quality is characterized as functionality, usability, effectiveness, maintainability and reliability and portability [15]. Special attention should be given to quality during implementation so that the final work meets customer's requirements and expectation [15]. It is also important that customer's requirements are well defined, clear and accurate so as to ensure that implementation is done in the right direction [18]
Complexity of requirements	It refers to the technical complexity to develop requirements such as the technologies being used, the number of technical interfaces to be implemented and so on. Complex processing in the form of scientific rules or commercial rules or remote accessing of computers may be required [15]
Volatility of requirements	Requirement volatility is inevitable in an agile project. Requirements tend to change due to several factors such as changes in technologies and customer needs or errors in original requirements. The development effort and project duration are significantly impacted with changes in requirements [19]
Data transaction	Data transaction is the transfer of data from one machine to another. The amount of data and the frequency of data transfer are important aspects to consider for data transaction [15]. Special efforts in design and implementation may be needed for high volume of data transfer [3]
Operation ease	It helps in increasing the usability of a product by making it user-friendly, and this consequently increases customer satisfaction. Better GUI provides easy understanding of functionalities to users [15]

(continued)

Table 1 (continued)

Factors	Details
Process	Changes in processes such as agile methods, builds, releases, fall under the category of variable/dynamic forces that decelerates the velocity of a team which consequently affects the effort estimation of a sprint [13]
Management	The tipping point between potential success and failure is determined by top management when developing and implementing business continuity management projects [17]
Environmental factors	These factors are friction forces that affect the velocity of a team. They cannot be eliminated but can be reduced. The infrastructure/condition of the environment such as uncomfortable seating and desks, and poor ventilation may worsen health of the team, thus reducing capacity to work [13]
Working time	Working time means "*any period during which the worker is working, at the employer's disposal and carrying out his activities or duties, in accordance with national laws or practice*" [15]. It is the normal working hours that a person spends at paid labor. Working time is also a factor that affects effort estimation

to collect factors that affected the project's sprints, together with justifications. The results of the survey are presented in Fig. 1.

It has been observed that not all factors that affect effort estimation in an agile project affect sprint effort estimation. Out of the factors described in Sect. 2, the following factors have less impact on a project's sprint: *data transaction*, *hardware and software requirements*, *operation ease*, *process*, *management*, *environmental factors* and *working time*. *Staff experience and technical ability* and *complexity of requirements* are the two most impacting factors in all projects. Factors *stakeholders*, *team dynamics and commitment*, *communication*, *staff experience and technical ability*, *experience of previous project*, *responsibilities outside projec*t, *configuration*, *security*, *defects and changes of previous implementation*, *quality of requirements*, *complexity of requirements*, *volatility of requirements* are vital factors that need to be catered in sprint effort estimation of agile projects.

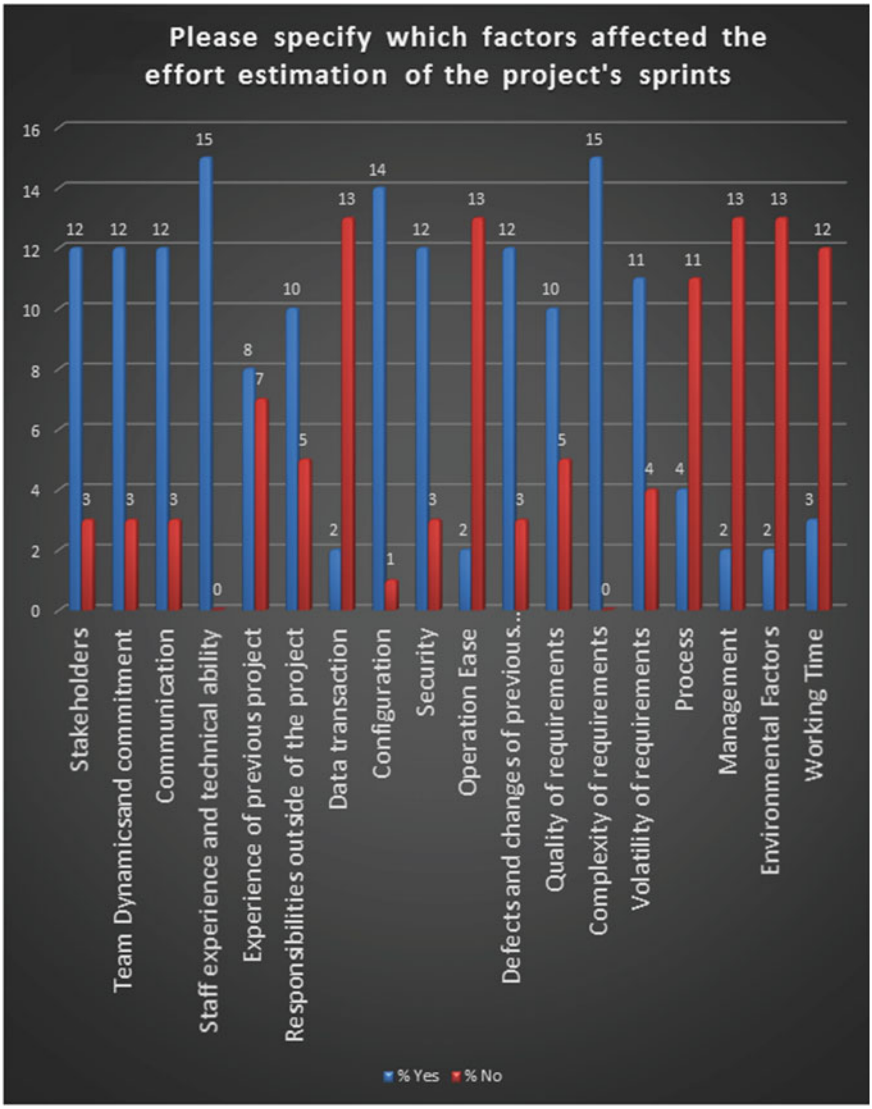


Fig. 1 Survey results

4 Discussion

This section presents justifications for the need of specific factors in sprint effort estimation.

- **Stakeholders**

Based on the justifications of most projects' scrum masters, responses from stakeholders get delayed when they are absent during a specific period of the ongoing sprint due to other commitments or holidays. Many scrum masters also share the opinion that this factor impacts a project if the product owner (PO) who is a project's key stakeholder does not understand the principles of agile or is reluctant to invest himself in the project's activities. The PO is responsible not only for representing the business on an agile project but also for well maintaining the product backlog and working in collaboration with the team so as to maximize value of the project.

- **Team dynamics and commitment**

According to scrum masters, lack of team dynamics and commitment reduces team effectiveness and team performance, thus affecting effort estimation. Lack of team dynamics and commitment often occurs in projects due to conflicts between team members, personal issues, health problems and lack of motivation.

- **Communication**

Most projects' scrum masters respond that communication is a major factor when a project is being developed with distributed teams of different cultures and at different sites. The team spends more time in communicating with each other for the synchronization of activities.

- **Staff experience and technical ability**

Effort estimation is hugely impacted with the turnover of qualified staffs in an ongoing sprint. Scrum masters find that teams with good technical expertise work more efficiently by saving time in problem solving. From the survey, many projects' teams consist of novice people who require time to rise in competence.

- **Experience of previous project**

All new projects are impacted by this factor as team members consume time in doing investigations during implementation and this reduces their working pace.

- **Responsibilities outside of the project**

Some team members may have additional tasks outside the project such as contributing to other activities of the company. It may also happen that skilled developers need to help in other projects or train novice team members.

- **Configuration**

Most answers from scrum masters are based on the configuration of a project for its deployment on multiple environments after its implementation. Configuring a project for deployment consumes time.

- **Security**

Most sprints include a level of security that needs to be catered, thus increasing the sprint effort. Functional security like user authentication, user and administrator view are common in all projects. Code security is imperative so as to reduce the security risks and vulnerabilities that occur due to coding errors. It is also determined that this factor includes database security which is essential to prevent data breaches of the project. Security of sensitive information is the primary concern for customers.

- **Defects and changes of previous implementation**

Scrum masters justify that defects in previous sprints may cause changes in implemented functionalities and this increases the current sprint's effort. The team needs to correct defects in implemented functionalities in order to prevent extensive reworks or other functional defects from arising.

- **Quality of requirements**

Effort estimation is affected when the team misunderstands customer's requirements which lack clarity. The team spends time in implementing features that are out of scope, and this finally results in rework. In most cases, it also happens that customer realizes pitfalls in requirements when seeing the working software.

- **Complexity of requirements**

A requirement can have a complexity of level low, medium or high for its development, depending on its nature. All requirements have a level of complexity of development, thus affecting all sprints.

- **Volatility of requirements**

During a sprint, the scope is kept constant and all changes are treated as new requirements. If a change means more value to the customer, the change is accepted if it is small (up to a number of hours) and has been identified at least 2–3 days prior to the end of the sprint. The change should be well aligned with the value proposition of a story in the sprint. Major changes in scope are treated as new stories in subsequent sprint.

- **Data transaction**

According to scrum masters, the team often caters for data transaction as part of technical complexity when estimating a sprint. Thus, this factor is combined in the technical complexity effort estimation of a sprint. Among the 15 selected projects for the survey, it is deduced that this factor has less impact.

- **Operation ease**

Design of interfaces is normally done before the implementation phase. The effort needed for its implementation is catered as part of technical complexity of requirements for a sprint. Further enhancement in terms of GUI asked by customer is taken as new requirements in future sprints.

- **Process**

Companies with well-established processes rarely change. This factor has less impact on its projects. According to scrum masters, any changes in processes should be considered after the completion of a sprint.

- **Management**

Top management affects mainly decisions regarding a whole project, for example funding of the project, technologies, resources and processes. The effort estimation for the implementation of requirements of a sprint is rarely impacted by this factor.

- **Environmental factors**

Scrum masters are of the opinion that these factors are more or less constant. Nowadays, most companies integrate ergonomics into all their operations for the health, safety and well-being of employees. Employees too adopt a more flexible approach to their workplaces and try to adjust to their settings.

- **Working time**

Many scrum masters state that this factor impacts projects when team members have to work beyond office hours. Productivity is decreased when team members have to work for longer hours to meet deadlines.

5 Future Works

Based on the identified factors, a predictive model will be developed for sprint effort estimation using machine learning techniques. The following methodology will be used to implement the model:

- **Literature Review**: An in-depth literature review will be carried out so as to understand different approaches for software effort estimation. Systems integrating machine learning techniques in agile effort estimation systems will be studied.
- **Analysis**: The analysis phase will consist of determining the requirements of the prototype. A critical analysis of existing agile effort estimation systems and the problems associated with these systems will be carried out. A model will be proposed to cater for the observed problems in effort estimation. The requirements of the proposed model, including the technologies and techniques adopted for its implementation, will be described. Factors collected from the survey as discussed in Sects. 3 and 4 will be used to determine effort estimation of a sprint more precisely.
- **Design**: This phase will consist of designing the proposed model. Based on identified sprint factors, a dataset will be built to train different machine learning techniques for the estimation of effort. Machine learning algorithms and methods used for the model will be described in detail.
- **Implementation**: Based on the design, the model will be implemented and the estimates obtained from different machine learning techniques will be gathered and analyzed.
- **Testing**: The model will be evaluated by using different performance measures, and the results obtained will be carefully analyzed. The results will be represented with the help of graphs and interpreted accordingly. Further improvements for the model will also be discussed.

6 Conclusion

In today's software industry, it is observed that most projects are adopting agile methodologies due to its enriched practices. Effort estimation of such projects is challenging. There are various recent studies that have been carried out for agile effort estimation based on factors. These studies focus on the correctness of project estimates. This paper identifies twelve key factors among many more factors that influence sprint effort estimation. The selected factors for a sprint are strongly justified by professional scrum masters. It is important for estimators to understand the reasons for integrating these factors in effort estimation of an agile project's sprint in order to achieve good accuracy of estimates. As future work, a predictive model will be developed for sprint effort estimation based on the identified factors using machine learning techniques.

Acknowledgements Consent was obtained from all participants who participated in the survey. They were informed that the data provided would be kept anonymous and no personal or health-related information would be recorded, prior to conducting the survey.

References

1. Popli, R., Chauhan, N.: Research challenges of agile estimation. J. Intell. Comput. Appl. (2012)
2. Kuan, S.: Factors on software effort estimation. Int. J. Softw. Eng. Appl. **8**(1), 23–32 (2017)
3. Bhalerao, S., Ingle, M.: Incorporating vital factors in agile estimation through algorithmic method. IJCSA **6**(1), 85–97 (2009)
4. Borade, J.G., Khalkar, V.R.: Software project effort and cost estimation techniques. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **3**(8) (2013)
5. Coelho, E., Basu, A.: Effort estimation in agile software development using story points. Int. J. Appl. Inf. Syst. (IJAIS) **3**(7) (2012)
6. Nagowah, S.D., Rumjaun, S.S., Gutteea, K.A., Nagowah, L.: Effortest—an enhanced software effort estimation by analogy methods. ADBU J. Eng. Technol. **5**(2) (2016)
7. Munialo, S.W., Muketha, G.M.: A review of agile software effort estimation methods. Int. J. Comput. Appl. Technol. Res. **5**(9), 612–618 (2016)
8. Sapre, A.V.: Feasibility of automated estimation of software development effort in agile environments. Doctoral Dissertation, The Ohio State University (2012)
9. Haugen, N.C.: An empirical study of using planning poker for user story estimation. In: Agile Conference, p. 9. IEEE (2006)
10. Mahnič, V., Hovelja, T.: On using planning poker for estimating user stories. J. Syst. Softw. **85**(9), 2086–2095 (2012)
11. Hurbungs, V., Nagowah, S.D.: A practical approach to teaching agile methodologies and principles at tertiary level using student-centred activities, pp. 355–389. In Agile and Lean Concepts for Teaching and Learning. Springer, Singapore (2019)
12. Choetkiertikul, M., Dam, H.K., Tran, T., Pham, T.T.M., Ghose, A., Menzies, T.: A deep learning model for estimating story points. IEEE Trans. Softw. Eng (2018)
13. Ziauddin, S.K.T., Zia, S.: An effort estimation model for agile software development. Adv. Comput. Sci. Appl. (ACSA) **2**(1), 314–324 (2012)
14. Vakkalanka, S., Engu, R.: Influence of team familiarity on team performance in distributed teams. Int. J. Mod. Eng. Res. (IJMER) **2**(4), 2549–2551 (2012)

15. Popli, R., Chauhan, N.: Agile estimation using people and project related factors. In: International Conference in Computing for Sustainable Global Development (INDIACom), pp. 564–569. IEEE (2014)
16. Jing, W., Ling, C.: Study on the relationship between the team commitment, knowledge sharing and performance. In: International Conference on Logistics, Informatics and Service Sciences (LISS), pp. 1–4. IEEE (2016)
17. Potdar, S.M., Ingle, A., Puri, M., Potdar, M.: Factors influencing on cost estimation for software development. Glob. J. Adv. Eng. Technol. **3**(2) (2014)
18. Badampudi, D.: Factors Affecting Efficiency of Agile Planning: A Case Study (2012)
19. Nurmuliani, N., Zowghi, D., Williams, S.: Requirements volatility & its impact on change effort: Evidence based research n software development projects. In Verified OK. University of South Australia (2006)

Mauritius as a Smart Tourism Destination: Technology for Enhancing Tourism Experience

Randhir Roopchund

Abstract This research paper provides an overview of how technology is shaping the pathway for Mauritius to become a Smart Tourism destination. Mauritius a small island economy recently hosted an international conference entitled 'Digitalisation and Sustainability' in the tourism sector. The conference emphasised on the need for enhancing competitiveness of Mauritius as a tourism destination through technology and innovation. Mauritius has also invested in Smart Cities to provide technology-driven facilities to businesses and customers. A qualitative research approach is adopted based on content analysis, literature review and reference will be made to the challenges for the country to emerge as a Smart tourist destination. The research analyses the megatrends in digitalisation that emerged from the conference and further analyses different indexes for ICT to assess the readiness of Mauritius as Smart Tourism destination. Different Mauritian tourism websites and metrics are used to gather data for purposes of analysis. The research shows that the hospitality industry is investing in new technological tools due to increasing customer sophistication. The Hospitality.mu and Mari Deal.mu portals are clear examples of the digitalisation trend in the tourism industry with the increasing use of the internet by consumers at all levels of the value chain. The research will also showcase some examples of real applications of technology in the tourism sector being adopted at the national and international level.

Keywords Smart Tourism · e-Tourism · Digitalisation · Competitiveness · Technology · Destination management

1 Introduction

Tourism is one of the biggest and rapidly growing and promising sectors in the world that is triggering growth, creating jobs and also helping to reduce global poverty. According to the UNWTO (2017), the tourism sector accounts for 10% of

R. Roopchund (✉)
Beau Plan, Université Des Mascareignes, Pamplemousses, Mauritius
e-mail: rroopchund@udm.ac.mu

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_44

global GDP, 10% of total employment worldwide and 7% of the world's exports equivalent to USD 1.4 trillion. However, the tourism industry is currently dealing with a number of challenges such as climate change, sustainability and digitalisation which will shape the future of the industry. Digitalisation and tourism success are inherently and intrinsically linked (Tralac 2018). The rapid progress in Information and Communication Technology (ICT) is fast developing the tourism sector creating both opportunities and challenges.

Contribution to Knowledge

Though the concept of Smart Tourism has been well established in developed countries, small island economies still have enormous progress to make. Hence, this study fills the research gap by analysing the progress made by Mauritius as a tourism destination. Wilson [1] uses the example of how Fogo Island makes use of innovative policies and sustainable tourism for fostering economic growth and development. Mauritius has the highest internet penetration rate in Africa (ITU 2017) and consequently it is important for the key economic sectors to adopt Smart technologies and solutions to pave the way towards a high-income economy. This study, therefore, makes a strong case for the adoption of new technologies and innovations in the tourism sector to improve our destination image. A study by Wang et al. [2] demonstrated that the website quality of a hotel is strongly linked to the e-trust which also mediates the relationship between website quality and online booking intentions. The study is also of significance as it increases the awareness of stakeholders about the importance of technology in the overall customer behaviour.

Technological advances are creating major disruptions in tourism by empowering tourism actors to engage in new markets, offerings, management practices and competitive strategies [3]. The Mauritian tourism sector is facing a number of challenges with globalisation and increasing competition globally triggering the need for the use of technology for improving the value chain. The research puts emphasis on the ICT applications at different levels for example for destination management and improving the overall quality of services in the tourism sector.

2 Background and Context of Study—The Mauritian Tourism Sector

The Mauritian Tourism industry is the third pillar of the economy and contributes significantly to economic growth and has been a major driver in the overall development of Mauritius [4]. The tourism sector contributes around 25.6% to the overall GDP of the Mauritian economy. The World Travel and Tourism Council has forecasted that the GDP will increase by 4% in the year 2027. During recent years about 67% of the tourist arrivals are of European origin, with France and Great Britain originating almost half of the visitors. The Minister of Tourism in the conference for digitalisation explained that 'intend to seize the potential of new technology to advance the tourism industry'. With social media and new products like crypto-currency and blockchain, technology can hold a lot of promise but it can also be disruptive. Consequently, the use of ICT and digitalisation represents significant opportunities and challenges for the Mauritian economy. Hence, the Mauritian tourism sector will surely benefit from the exploitation of innovative digital practices such as the use of destination management and the use of innovative apps. It is important to evaluate the potential of digitalisation as it may help companies to improve their overall efficiency and at the same time ambition to become more sustainable.

2.1 Mauritius as ICT Hub

The aim of the Mauritian government is to position the country as one of the leading ICT destinations, being a model in Africa, which supports the new e-Global age [5]. There are six hundred ICT companies operating in Mauritius in different business activities ranging from software development to business process outsourcing. There are some major players such as Microsoft, IBM and Accenture amongst others. The drive to emerge as an ICT-BPO Hub has resulted from a plethora of factors such as the changing global economic environment, changing government policies and the Mauritian bilingual population (Fig. 1).

Mauritius is a small island with around 1,280,000 population. However, we have an internet penetration rate of 63% (highest in Africa). Some 720,000 people use different social media such as Facebook, Instagram and Twitter. There are 1.74 million subscribers to mobile services indicating that we have more mobile phones are sim cards than the actual population size. In addition, around 48% of mobile users are active mobile social users. These statistics clearly demonstrate that Mauritius is poised to emerge as Smart economy where technology plays a key role in economic

Fig. 1 Key digital statistical indicators. *Source* Hootsuite (2018)

growth and development. The government has also adopted the e-government initiative where it plans to digitalise different Ministries for improving effectiveness and efficiency.

Research Objectives

- Study the importance of Smart Tourism for a Small Island economy as Mauritius
- Analyse the Smart Tourism initiatives at the national and organisational level for improving the destination competitiveness
- Consider the use of innovative technologies for improving tourism experience
- Develop recommendations for improving the Smart Tourism concept in the Mauritian Context.

3 Literature Review

3.1 Smart Tourism Concept

The Smart Tourism destination (SD) idea that is attracting a lot of interest from researchers and may be considered as a significant development in the tourism field [6]. The hypothetical development is still limited and the 'destination concept' is complex, evolving, socially-developed and multi-layered, as reflected in literature (e.g. Pearce [7]; Saarinen 2004; Saraniemi and Kylänen [8], and so on). Smart destinations have been largely influenced by some earlier conceptualisations, for example, 'e-Destinations'. E-Destinations stress the usage of ICTs to give data and to end up an instrumental part of all transactions along the value chain [9], in smart destinations technology is centrally embedded in all elements thanks to new developments, such as the Internet of Things (Koo et al. 2016).

Many countries are feeling increasing pressure to realise Smart Tourism to influence economic growth and development. In Asia, there are more collaborative efforts to drive the Smart Tourism agenda forward. Governments in China and South Korea are supporting initiatives in developing the technological infrastructure that supports Smart Tourism Hwang et al. [10]. In Europe, many of the Smart Tourism initiatives were born out of smart city projects and, as a consequence, Smart Tourism destinations are increasingly making an appearance in the European tourism landscape. The focus in Europe is about developing smart end-user applications that support enriched tourism experiences using already existing data combined and processed in new ways based on innovative approaches Lamsfus et al. [11]; Boes et al. [12, 13]. In the Mauritian context, the government is encouraging Smart city schemes for building intelligent buildings for improving the sustainability of the economy. There are a number of Smart City projects that have been developed across the island—Medine Smart City, Moka Smart City and that of Beau Plan amongst others. In 2015, the Finance Minister proposed the creation of eight smart cities and five 'technopoles',

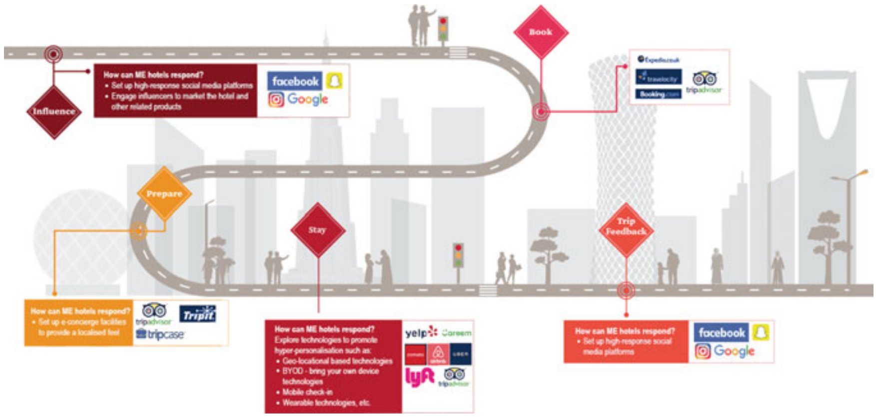


Fig. 2 Technology interactivity in hotel value chain

with an overall investment of Rs.120 billion and requiring 7000 acres of land. In Australia, the emphasis has shifted to smart governance and use of open and big data. Based on the discussion, transformative power of smart technologies is being universally accepted not only in terms of the economic potential but also the social and experiential dimensions (Fig. 2).

The above diagram shows how technology may influence the tourists during the whole value chain in hotels affecting the overall tourism experience. At the pre-purchase stage, tourists may be influenced by social media posts and use of search engine optimisation where technology helps to make hotels become more visible to customers. Nowadays, many customers book their air tickets and hotels using online platforms such as Expedia, Booking.com amongst others. These agencies also send customised communications to the customers to influence their purchase decision. During the stay, technologies may be used to provide high degree of personalisation such as mobile check-in or also provide some updates about services available during the stay. After the stay, customers may still use technologies to provide feedback on their overall level of satisfaction or any complaints regarding the quality of services being offered.

3.2 *Smart Tourism versus E-Tourism*

Smart Tourism is defined as 'tourism supported by integrated efforts at a destination to collect and aggregate/harness data derived from physical infrastructure, social connections, government/organizational sources and human bodies/minds in combination with the use of modern technologies to transform that data into on-site experiences and business value-propositions with a view to improve the overall efficiency, sustainability and experience enrichment'.

Table 1 ICT tools in tourism

Tools	Meaning	Tourism area/application
Destination management system	DMS is used information management, marketing [27], exchange with different tourism stakeholders including tourist education	Destination e-Metrics
Intelligent transport system	Provides real-time transport information to customers [28]	Visitor experience at Acadia park
Location-based services	Provide real-time information to tourists at different geographic locations Berger et al. [29]; [30]	Tourism industry at large
Global positioning system	GPS is identified for both tracking and analysis of tourist movements [31] and also identifying tourists location	Possibility of tracking tourists—places of visit

Hence, Smart Tourism is much broader in scope compared to E-tourism in terms of use of technologies for enhancing the overall tourism experience. The use of big data and smartphones for providing customised interactions and solutions is becoming very common in the tourism sector. For example , the use of geo-localisation software to analyse the tourist's attractions preferred by tourists is an example of Smart Tourism practices.

3.3 ICT Tools for Destination Management

There are many ICT tools to gather data about tourists' movements and conduct at the destination is mainly based on GPS, WiFi or video technologies. The system is about computing the number of people at different access points such as airports or at the entrance of specific sites such as theme parks and so on.

Some of the tools that may be used for destination management are (Table 1).

3.4 Technology and Tourists Buying Behaviour

Tourism technology research has primarily focused on the role of technology on customer buying behaviour (selection, purchase and evaluation of customer experiences). Information and Communication technologies directly or indirectly influence the consumer decision making processes and very important the overall tourism experiences. For example, tourists check consumer ratings and feedback before booking for hotels Sagala [3]. However, what is more, important is that consumers are empowered so that can take part actively in the decision making processes known as active co-creators (prosumers) of their hospitality experiences. Several research studies Sigala and Gretzel [3]; Stephen and Lamberton [14] illustrate how tourists 'become co-designers, co-marketers, co-advertisers, co-promoters, co-distributors of tourism experiences through user-generated-content, customer review platforms, blogs, wikis, participation in innovation contests and toolkits, crowdsourcing practices'. Crowdfunding platforms also allow tourists to sponsor and support the implementation of tourism experiences such as concerts, tours or other events.

However, technology is also influencing the motivation of tourists to choose their destinations as many people are being influenced by social media pictures and selfies being posted. Travelling may also be used as an act of self-promotion and identity development, however, social media advances have further escalated and digitised this social practice. For some travellers [3], the quality of the tourism experience itself does not determine their satisfaction. Tourists are not experiencing their environment as much of the time is spent with the camera and their emotions depend much on the content shared on social media.

Many companies are making use of social media as communication tools for Destination Management Organisations (DMOs) and National Tourism Organisations (Leung et al. 2013; Hays et al. 2013). However, many organisations are not effective in managing communication to reach their targeted audience. The companies need to work on their communication strategies so as to achieve their set objectives (Huertas and Marine-Roig 2014).

The concept of Travel 2.0 was developed to respond to the emergence of the concept of Web 2.0. The main idea is not about getting the cheapest trip but more about improving the overall tourism experience and satisfaction [15]. Conrady [16] characterises Travel 2.0 by suggesting that it consists of five important branches namely transparency, collaboration, better basics, speed and predictability. It is also the realisation by the tourism industry of the Web 2.0 referring to a cultural travel change (Miguéns et al. [17], p. 2) and includes the harnessing of different technological applications like social media in the travel context [18, 19].

3.5 The Mega Trends in Mauritian Hospitality

The Mauritian Hospitality Website (Hospitality.mu) makes reference to the use of three technologies in the tourism sector which are namely cloud systems, Social Media and mobile platform. There is increasing disintermediation and direct bookings through the technological platforms. Social media plays an important role in driving sales and improving the hotel's reputation. This has given rise to innovative hotel management structures in this twenty-first century, and these have become totally 'Guest-Centric'. In other words, it's all market-driven in real-time, and guests dictate how quickly these changes are adopted by hotels. The inability to identify these basics may drive hotels out of their markets.

Table 2 Summary of technological megatrends

Use of customer-centric solutions—changing business environment from business to business, business to customer and now peer to peer environment
Artificial intelligence for linking CRM tools to yield management tools—use of big data and precision data may help to re-invent Mauritian tourism with more customised and smart solutions to customers
Need for greater adaptation of the content such as the use of videos and online streaming
Compliance with global data security standards and adherence to general data protection regulation
Increasing use of mobile marketing and Smart Apps to influence tourism behaviour
Possibility of greater crowdsourcing possibilities
Hospitality black chains—use of inter-connected supply chains
Cloud-based training programmes for employees in the hotel sector

Source Adapted from Hospitality.mu [20]

The Hospitality.mu [20] site makes reference to the following applications of technology in the Mauritian context for achieving smart and sustainable tourism (Table 2).

Boodnah et al. [21] referred to the use of Polwi by Light and Polwi by Nature as one of the examples of Smart Tourism in Mauritius. Each year, Porlwi (Mauritian Capital) chooses a theme that encapsulates a key element of the city's regeneration. The use of lights and technology is used to make the capital look very beautiful based on a particular concept.

3.6 Tourism Destination, Competitiveness and Technological Advancement

Ritchie and Crouch (2003) explained that tourism competitiveness is a multidimensional concept that includes economic, political and technological competitiveness as well. As explained earlier, the European Commission and the Mauritian government are multiplying the use of technology for improving the overall competitiveness of the tourism sector and achieve a higher degree of sustainability. The technological megatrends in Mauritius show how the Mauritian tourism industry is using technology for reducing costs and also streamlining the different processes so as to enhance the overall customer value.

3.7 Challenges of Smart Tourism for SIDS

The implementation of Smart Tourism in SIDS requires the right ecosystem with the active participation of the major stakeholders, namely the government, hotels, promotion agencies and the customers. The challenges for a small island country is largely different from that of developed economies due to resource availability, the socio-cultural landscape and also exogenous factors affecting the destination competitiveness. Kaudeer et al. [22] referring to Mauritius as an example, make reference to priority areas that should be considered when creating smart SIDS. Mauritius as a Small island economy suffers from geographical distance, climate change and resource constraints for the Mauritian tourism sector.

3.8 Research Method

The research approach used is qualitative and analyses the trends in Smart Tourism at two levels namely the national and organisational level. The research uses different tourism websites at the national level to gather information about Smart Tourism initiatives. Some of the important websites used are Hospitality.mu, Ministry of Tourism, Mauritius Tourism Promotion Authority and the website used for hosting the digitalisation conference amongst others. Reference is also made to the World Economic Forum Global Tourism competiveness and ICT indexes to makes inferences about the readiness of ICT in the travel and tourism sector. For the section on Smart Apps in the tourism sector, the data was sourced from Mauritius Telecom which is the major player in the Telecommunications sector. In order to consolidate the Smart Tourism perspective in the Mauritian context, use is made of Azuri Mauritius as a Case study to analyse some smart initiatives including the analysis of the social media platform where it uses technology for enhancing customer engagement. Buhalis and Amaranggana [23] are of the view that case studies are suitable for researching the best practice of STDs and generating in-depth knowledge on this subject. As the study is qualitative, the main objective of the study is simply to understand where Mauritius stands as a Smart Tourism destination. The ontological perspective adopted is that Smart Tourism may be highly beneficial for the country to achieve its marketing objectives. The main hypothesis is whether Mauritius may be considered as a Smart Tourism destination. The research is of significance as the country is a small island economy that does not have natural resources which may be used to achieve competitive advantage. The main limitation is, therefore, the fact that the findings may not be used for purposes of generalisation. The main constructs used for doing the analysis are Smart Tourism initiatives, digitalisation, ICT readiness, development of smart tools which is influencing the tourist satisfaction during their visit to Mauritius amongst others. These constructs are grounded in the literature review discussed earlier. The diagram below outlines the research model adopted for the present research:

4 Analysis and Findings

4.1 Government Involvement and Smart Tourism

It is important to assess the overall ICT readiness of our country to become a Smart Tourism destination. As a starting point, it is important to indicate that Mauritius ranks first in ICT Development Index in Africa and consequently bolsters our ambition to become an ICT hub in Africa. The organisation of the first international conference on digitalisation and sustainability in the Tourism Sector has been another major step towards positioning Mauritius as a Smart Tourism destination. Zhu et al. [24] discuss the important role of government in developing Smart Tourism. They are of the view that 'Smart Tourism can serve as guider on tourism informatization in the designing stage. In macro level, government not only needs to encourage the construction of tourism informationisation in the form of policies and regulations, but also needs to concretely standardise the framework of Smart Tourism at a national scale'. Koo et al. [25] stress the importance of developing the right ecosystem for the development of Smart Tourism based on mutual support and collaboration. The executive members of the conference also signed a Memorandum of Understanding (Ministry of Tourism 2018) with a view to further enhance the process of digitalising the tourism sector for improving our brand image and identity. Some of the key points in the roadmap towards digitalisation include the need:

- 'to create a working group on Digital Platforms aimed at identifying, analysing and proposing a balanced approach, exchanging best practices and helping in developing regulatory framework and policies to create a level playing field for tourism service suppliers';
- to promote and diversify sustainable tourism practices by promoting the development of 'ecotourism, agro-tourism, medical tourism and cultural tourism';
- to develop a national regulatory framework to protect the privacy of visitors.

The ambition of Mauritius to be an ICT hub and also enhance our visibility of the Tourism sector has been further consolidated by the announcement of the Prime Minister to invest in 'new digital platforms will be created to provide information on safety, costs of inland travel, road maps, dining, shopping and exchange rates of currencies, amongst others' (National Budget 2018).

4.2 ICT Readiness and Tourism Competitiveness Index

The World Economic Forum has developed the Travel and Tourism Competitiveness Index which measures 'the set of factors and policies that enable the sustainable development of the travel and tourism sector, which in turn, contributes to the development and competitiveness of a country'. Mauritius ranks 65th amongst 136 countries on the global competitiveness index and one of the sub-index is ICT readiness where

Table 3 ICT readiness and global ITU index

Countries	ICT readiness index (2017)	Global ITU index (2017)
Mauritius	4.54	5.88
Seychelles	NA	5.03
South Africa	4.4	4.96
Madagascar	2.1	1.68
India	3.2	3.03

Source WEF (2017) and global ITU (2017)

Mauritius has scored 4.54 in 2017. Mauritius has improved its ICT readiness by 7.6% in two years (ICT Readiness Index being 4.22 in 2015). It is important to compare ICT readiness Index of Mauritius to the neighbouring countries such as Seychelles, Madagascar, South Africa and India (Table 3).

Hence, based on the two comparative indexes Mauritius seems to be well poised in terms of ICT readiness at the regional level to be a Smart Tourism destination. Another element which is of critical importance is the IT security issues. Based on a survey carried out by IBM Mauritius ranks first in Africa in terms of security levels.

4.3 Analysing Some Smart Tourism Metrics

The analysis also seeks to analyse some metrics that may be used to evaluate the overall readiness or trend of Smart Tourism in Mauritius.

From Table 4, it may be analysed that Mauritius has made significant strides to emerge as a potential Smart Tourism destination due to the high internet penetration, use of social media and based on the tourism booking profile. However, the online booking rate is not so strong in the Mauritian landscape.

4.4 Development of Smart Apps for the Tourism Industry

The Mauritius Telecom has recently devised a few Smart Apps that may of helpful in the endeavour to become a Smart Tourism destination. These apps include Mauritius Tourist Guide, Chake, Myweather and NouMoris which may be used by tourists during their visit to Mauritius. The role of these tools is to empower tourists to take decisions and make them part of the services being provided. A small brief on the different Apps and technologies are provided in Table 5.

Table 4 Smart Tourism metrics

Smart Tourism metrics	Statistics	Analysis
Digitalisation indicators	63% internet penetration 57% social media utilisation	With the highest internet penetration Mauritius ambitions to become a Smart Tourism destination (Source: Internet World Stats 2018)
Official my tourism twitter profile	Tweets, current page 2,681 following 1,009 followers 23.1 K likes 205	Though there are many followers yet the very small number of likes indicate little engagement with Mauritius tourism twitter account
Mauritius tourism facebook page	14,178 likes for FB page	This shows that there is higher engagement on FB page
Air Mauritius	Online booking-around 40%	Use of traditional travel agencies may explain such a behaviour
Tourism booking profile	Haywantee and Nunkoo [32] Statistics from survey based on 877 respondents from Europe Chi square test values Travel agents 132.54 Friends 99.91 Internet 127.97 Information leaflets 82.6	Results indicate that nationality and existing product knowledge largely influence travellers' information search behaviour. The results of the study indicate that internet was an important source of information for tourists. However, first-time travellers preferred traditional sources such as travel agencies, friends, leaflets and national tourism offices as compared to repeat visitors

Source Author (Statistics from various sources referenced)

4.5 Smart Living in Tourism: Case Study of Azuri Mauritius

Hotels in Mauritius are also adopting Smart Tourism initiatives with the increasing use of digital tools and applications for achieving greater sustainability. Azuri Mauritius has been set on a preserved beach of the north-eastern coast that frames the Azuri village, Radisson Blu Azuri Resort and Spa delivers guests a distinctive getaway experience. Some of the smart initiatives being adopted by the hotel are outlined based on case study research. The hotel has installed fibre optic network across the village and currently awaiting connection from the service provider. This will boost up speed connectivity throughout Azuri and improve the working aspect of Azuri. The hotel has also developed the 'Azuri app'. This has a huge potential in serving its citizens and visitors through information distribution and interaction as well as managing 'municipal' services. The company also makes use of blogs to create interests of customers on its different services (an example is provided below).

Table 5 Smart Apps for tourism in Mauritius

Apps	Smartphone-driven applications for tourists
Mauritius tourist guide	Provides tourists with different information related to tourists attractions and places of interests
Chake app	Provides updates on different events happening in Mauritius. The chake app provides users with a map view and detailed information on different activities (such as art galleries, theatre, sports)
NouMoris	The NouMoris app provides a complete schedule of national events and activities on the smartphone
https://taxiservicemauritius.com/mobile-app/	Book taxi and tours around the island through mobile app
https://www.myt.mu/mobile/mytweather-app	Enables tourists to check weather predictions and plan their holidays accordingly

Source Mauritius telecom website [33]

KI BON BAZ

Barefoot Beach Shack

TAPAS PIEDS DANS L'EAU

Le restaurant Ocean One, spécialiste du fine dining au cœur du Radisson Blu d'Azuri, s'est ouvert aux repas sur la plage. Convivial et rafraîchissant, le concept fait des heureux.

Source Azuri Facebook (2018)

Besides thinking of the effective use of resources and services the group is carefully addressing its current and future demands. Azuri is also considering other initiatives such as the development of education campus, designing and building a professional business hub and introducing Wifi free public zones as part of its future development plans. The company is also very active on social media for increasing customer engagement. Research shows that social networking sites are pressurising suppliers and buyers who are being largely influenced by the views and opinions of travellers. The information generated on these social media sites provide a range of information that are being shared and circulated which helps to educate customers about the products and other service issues [19]. An example of customer feedback and engagement is provided on the social media page of Azuri:

> 'The best hotel service by far, staff was very courteous, very welcoming and even upgraded our room. Oh had a great birthday cake as well and was serenaded by the band. Rooms are very clean. Cheers to the staff members for your great hospitality. I'll be sure to recommend it to friends and hopefully come back.'… (Feedback 1)
>
> Lovely place to relax.. beautiful apartment with all that you wud need for your stay… only toaster cud not be used as it kept on tripping the electricity. thanks Kris for the warm welcome n check out… deff will be back… (Feedback 2)
>
> Amazing place… excellent food and favoured place which has a lot of qualities that i really appreciate about Azuri #AK (Feedback 3)

From the above comments, it is clear that social media may be used for value creation and starting in a conversation for higher customer engagement.

4.6 Policy Implications

Te purpose of the study is to analyse the role of technology at national and organisational level to make Mauritius become a Smart Tourism destination. The literature shows that tourism is a major pillar of our economy and the government is interested in making Mauritius a Smart island and destination by supporting the digitalisation process. The analysis shows that Mauritius ranks high in ICT readiness and internet penetration important for the process of digitalisation. The government through the Mauritius Telecom has developed a number of applications that may be used by tourists and the public at large. The research findings are in line with the literature which shows that Smart Tourism may be a source of differentiation, a way to reduce costs of operations and more important a strategic marketing tool.

In addition. The case study of Azuri portrays different smart initiatives for improving tourists' engagement at all levels of the value chain. Hotels are using websites, social media, blockchain technologies to be more competitive and improve the overall visibility of the hotel industry. At the national level, the use portals such as Hospitality.mu, Marideal.mu and Mauritius Tourism are clear landmarks of the significant progress for harnessing technology at national and international levels in the Mauritian tourism industry.

4.7 Recommendations for Tourism Sector

Based on the analysis, some recommendations can be made to improve the image of Mauritius as a Smart Tourism destination. The institutional support from the Ministry of Tourism and MTPA are instrumental for improving the global visibility of Mauritius. The global conference on Digitalisation and Sustainability has been very useful to promote Mauritius as a Smart Tourism destination and also provides an overview of how technology is being used in the hospitality and tourism sector. The government should continue to support the local actors to adopt new trends such as destination management system, location-based services and also destination management system (destination e-metrics). This year the government has sponsored the World Artificial Intelligence Show in Mauritius with a view to transform the way businesses are being conducted and to portray Mauritius as leading Smart City island in Mauritius.

Mauritius can play a leading role to showcase Smart Tourism for African economies in the years to come. The country can also partner with some African economies such as Nigeria to collaborate on tourism technological platforms. There is a huge potential of using technology for marketing Mauritius as a strong tourism destination with the increasing number of flights to African economies. Mauritius can also seek help from technologically advanced countries such as India and UK to improve its destination image.

Hotels should provide information to tourists about Smart Apps that may be highly useful for improving the overall customer experience. The hotels may also develop their own apps for a smart interaction in terms of booking, provision of information about tours/activities and cultural information about the country. The Case Study of Azuri shows that hotels are taking the right initiatives to provide the right tourism experience.

In the first instance, it is important to improve the broadband internet connectivity so that tourists will not face any problems in using the Smart Tools and Apps. Hotels should also employ high profile IT Managers so that they can develop appropriate tools for enhancing the overall tourism experience. Hotels may also use online links for assessing customer satisfaction and also interact with customers through portals. The study also shows that social media may be used for enhancing customer engagement and building a good image of the company. Consequently, hotels should invest in social media marketing to start the dialogue and build a long term profitable relationship.

5 Conclusion

The research paper presents very interesting findings for the tourism sector to further developing the Smart Tourism concept. It is important to include all the actors of the tourism industry to enhance the overall competitiveness of destination economies such as Mauritius. There is need for greater efficiency in the use of resources through the deployment of right technologies and also improving the overall service quality through innovative service processes. The Mauritian government is supporting the process of digitalisation through capacity development and by providing technology diffusion schemes. Mauritius needs to engage in sustainable business practices so that it may develop both as a Smart island and Tourism destination.

References

1. Wilson, M.I.: Smart and sustainable: lessons from Fogo Island. In: Stratigea, A., Kyriakides, E., Nicolaides, C. (eds.) Smart Cities in the Mediterranean. Progress in IS, Springer, Cham (2017)
2. Wang, L., Law, R., Guillet, B., Hung, K., Fong, D.K.: Impact of hotel website quality on online booking intentions: eTrust as a mediator. Int. J. Hosp. Manage. **47**, 108–115 (2015)
3. Sigala, M.: New technologies in tourism: from multi-disciplinary to anti-disciplinary advances and trajectories. Tourism Manage. Perspect. 1–5 (2018)
4. Ministry of Tourism. Ministry of Tourism. Website, Available from http://tourism.govmu.org/English/Pages/default.aspx (2017). Accessed 16 May 2018
5. MICT. Available from http://ictbpoconference.govmu.org/English/Mauritius%20ICT%20Hub/Pages/default.aspx (2018). Accessed 24 July 2018
6. Jovicic, D.Z.: From the traditional understanding of tourism destination to the smart tourism destination. Curr. Issues Tourism (2017). https://doi.org/10.1080/13683500.2017.1313203
7. Pearce, G.D.: Toward an integrative conceptual framework of destinations. J. Travel Res. **53**(2), 141–153 (2014)
8. Saraniemi, S., Kylanen, M.: Problematizing the concept of tourism destination: an analysis of different theoretical approaches. J. Travel Res. **50**(2), 133–143 (2011)
9. Buhalis, D.: eTourism: information technology for strategic tourism management. Pearson (Financial Times/Prentice Hall) (2003)
10. Hwang, J., Park, H.Y., Hunter, W.C.: Constructivism in Smart tourism research: Seoul destination image. Asia Pac. J. Inf. Syst. **25**(1), 163–178 (2015)
11. Lamsfus, C., Martín, D., Alzua-Sorzabal, A., Torres-Manzanera, E.: Smart Tourism destinations: an extended conception of smart cities focusing on human mobility. In: Tussyadiah, I., Inversini, A. (eds.) Information and Communication Technologies in Tourism 2015, pp. 363–375. Springer, Heidelberg, Germany (2015)
12. Boes, K., Borde, L., Egger, R.: The acceptance of NFC smart posters in tourism. In: Tussyadiah, I., Inversini, A. (eds.) Information and Communication Technologies in Tourism 2015, pp. 435–448. Springer, Google Scholar, Heidelberg (2015)
13. Boes, K., Buhalis, D., Inversini, A.: Conceptualising smart tourism destination dimensions. In: Tussyadiah, I., Inversini, A. (eds.) Information and Communication Technologies in Tourism 2015, pp. 391–403. Springer, Heidelberg (2015)
14. Lamberton, C., Stephen, A.T.: A thematic exploration of digital, social media, and mobile marketing: research evolution from 2000 to 2015 and an agenda for future inquiry. J. Mark. **80**(6), 146–172 (2016)
15. Wolf, P.: Travel 2.0 confronts the establishment. Retrieved 12 Mar 2014, from http://www.phocuswright.com/research_updates/travel-20-confronts-the-establishment (2006)
16. Conrady, R.: Travel technology in the era of Web 2.0. In: Trends and Issues in Global Tourism 2007, pp. 165–184). Springer (2007)
17. Miguéns, J., Baggio, R., Costa, C.: Social Media and Tourism Destinations: TripAdvisor Case Study. Adv. Tour. Res., Aveiro (2008)
18. Ling, J.: Intention to use online social networks for travel purposes: a case study at three Bangkok-based universities. AU-GSB e-J. **3**(2), 50–59 (2010)
19. Xiang, Z., Gretzel, U.: Role of social media in online travel information search. Tourism Manage. 179–188 (2010)
20. Hospitality.mu website: Available from https://hospitality.mu/ (2018). Accessed 28 June 2018
21. Boodnah, K.D., Armoogum, V., Jaunky, V.C., Armoogum, S., Towards smart tourism: an individual appreciation of porlwi-by-light festival. In: 2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), pp. 323–328. Balaclava (2016)

22. Sahib-Kaudeer, N.G., Jhummun, D.S., Gobin-Rahimbux, B.: What is 'smart' for small Island developing states? In: 2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), pp. 13–19. Balaclava (2016)
23. Buhalis, D., Amaranggana, A.: Smart tourism destinations. Inf. Commun. Technol. Tourism **2014**, 553–564 (2013)
24. Zhu, W., Zhang, L., Li, N.: Challenges, Function Changing of Government and Enterprises in Chinese Smart Tourism, Tourism tribune (2014)
25. Koo, C., Gretzel, U., Hunter, W.C., Chung, N.: The role of IT in tourism. Asia Pac. J. Inf. Syst. **25**(1), 99–104 (2015)
26. Azuri Facebook Page. https://www.facebook.com/AzuriLuxuryResidences/ (2018). Accessed 15 July 2018
27. Horan, P., Frew, A.J.: Destination eMetrics. In: Frew, J.A., (ed.) Proceedings of the Travel Distribution Summit, Europe Research Conference 2007, pp. 25–44. Axon Imprint, London (2007)
28. Diagle, J.J., Zimmerman, A.C.: The convergence of transportation, information technology and visitor experience at Acadia National Park. J. Travel Res. **43**, 151–160 (2004)
29. Berger, S., Lehman, H., Lehner, F.: Locationbased services in the tourist industry. J. Inf. Technol. Tourism **5**, 243–256 (2002)
30. Liburd, J.J.: Sustainable tourism and innovation in mobile tourism services. Tourism Rev. Int. **9**(1), 107–118 (2005)
31. Shoval, N., Isaacson, M.: Tracking tourists in the digital age. Ann. Tourism Res. **34**, 141–159 (2006)
32. Ramkissoon, H., Nunkoo, R.: Information Search Behaviour of European Tourists Visiting Mauritius, pp. 7–21. Tourism (2008)
33. Mauritius Telecom Website. Available from https://www.myt.mu/mobile/moreapps/ (2018). Accessed 25 June 2018

A Review of Machine Learning Techniques for Software Quality Prediction

Sanjeev K. Cowlessur, Saumendra Pattnaik and Binod Kumar Pattanayak

Abstract Successful implementation of a software product entirely depends on the quality of the software developed. However, prediction of the quality of a software product prior to its implementation in real-world applications presents significant challenges to the software developer during the process of development. A limited spectrum of research in this area has been reported in the literature as of today. Most of the researchers have concentrated their research work on software quality prediction using various machine learning techniques. Another aspect pertaining to software quality prediction is that the prediction must be achieved in the earlier stages of software development life cycle in order to reduce the amount of effort required by the developer in course of the development of a software product. In this paper, we carry out a comprehensive review of machine learning techniques which have been used to predict software quality.

Keywords Software development · Software quality · Machine learning · Quality prediction · Quality assessment

S. K. Cowlessur (✉)
Department of Software Engineering, Université des Mascareignes, Beau Bassin-Rose Hill, Mauritius
e-mail: scowlessur@udm.ac.mu

S. Pattnaik · B. K. Pattanayak
Department of Computer Science and Engineering, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, India
e-mail: saumendrapattnaik@soa.ac.in

B. K. Pattanayak
e-mail: binodpattanayak@soa.ac.in

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_45

1 Introduction

The quality of a software product can be defined as the measure of performance of a system on which the software is implemented in terms of execution time, memory capacity utilized and probability of errors, etc. In addition to this, the amount of effort contributed by the software developer also represents a key factor while assessing the quality of a software product. The quality of a software product can be considered to be internal as well as external. The internal quality of a software can be assessed in course of development during software development life cycle (SDLC); whereas, the external quality can be measured during its implementation and can be assessed with respect to its level of functionality. The external quality also depends upon its internal quality. In order to assess the external quality of a software product, quality models can be devised that represent a function of the internal quality attributes. In order to achieve this, first of all the internal attributes must be identified and then the relationship existing between the internal and external quality attributes must be identified. A number of software quality prediction models have been proposed by various authors. However, the machine learning approach to devising such a model appears to be more popular and more effective as claimed by the authors. We have been motivated by this aspect to carry out a review of the machine learning approaches for software quality prediction models.

In this work, we mainly focus on the following machine learning techniques: evolutionary algorithms, artificial neural networks (ANN), Bayesian networks (BN), fuzzy logic (FL), decision tree (DT), support vector machine (SVM) and case-based reasoning (CBR). We provide a detailed description of different models used for the prediction of software quality as proposed by various authors in each of the techniques mentioned above. The rest of the paper is organized as follows. Section 2 includes the discussion of various software quality models proposed by different authors using various machine learning techniques mentioned above. Section 3 concludes the paper along with probable future extensions.

2 Machine Learning Approaches to Software Quality Prediction

In this section, we discuss different machine learning approaches used by various authors for prediction of software quality.

2.1 Evolutionary Computing Technique-Based Approaches

In this section, we bring forth a few approaches as proposed by different authors to predict software quality making use of evolutionary computing techniques such as genetic algorithm (GA), ant colony optimization (ACO) and several other techniques as well.

2.1.1 Software Faulty Module Identification Using GA

In practice, during implementation of a software product, often errors are encountered for the reason that the software may consist of faulty modules which need to be identified, and such a task is tedious for a software professional. Moreover, this problem intensifies in case of problem-solving software. GA, being more of a problem-solving algorithm, can presumably be useful for identification of such faulty modules in a software. A GA-based approach for faulty module identification of open software was proposed by the authors in [1], wherein the following sequence of steps needs to be followed:

Step 1: Data collection of data from source code;
Step 2: Evaluation of collected data in terms of the metrics such as lines of code (LOC), depth of inheritance (DOI), number of children (NOC), lack of cohesion (LOCM), coupling between objects (CBO), number of public methods (NPM), response for a class (RFC) and weighted methods per class (WMC);
Step 3: Identification of appropriate metrics for fault prediction by virtue of filtration of data refined in step 2;
Step 4: Processing of data obtained from step 3 using GA;
Step 5: Assessment of performance via confusion metrics and analysis of prediction.

Here, fault-prone classes along with metrics are identified using GA. The authors claim a high accuracy of fault prediction using GA as compared to other approaches even though it principally relies on evolutionary computation techniques.

2.1.2 Fault Table for Fault Identification Using GA

In this approach, the faults in the current version of a software product can be comfortably anticipated with reference to the history of fault occurrence pertaining to the previous version of the same software that is maintained in a fault table. In the absence of such a table, the classification of faults characteristic to that particular software can be useful. A fault classification technique is proposed in [2] where the authors use GA for fault classification using a supervised learning method. Here, authors identify the most appropriate attributes of the software in consideration and then classify the attributes as faulty and non-faulty ones using GA in the context of an example following which as claimed by the authors, the faulty attributes need revision.

2.1.3 Software Quality Prediction Using ACO

It becomes difficult to predict the quality of a software during the early stages of the SDLC due to the fact that many of the software attributes at this point in time remain immeasurable. However, this can be derived from the attributes that are measurable at that point in time. A quality prediction model can serve this purpose. An ACO-based approach is proposed by authors in [3] where they assume adaptation of an existing model for prediction of software quality that is designed for one data set to a new data set. The authors conduct their experiments on an object-oriented system thereby claiming that it can also be used for various other software systems too. Authors conclude that such an approach can provide much better results as compared to other machine learning approaches.

2.1.4 Software Behaviour Estimation Using SPS

There are several software tools available that can successfully be used for estimating the behaviour of a software product. Software project simulator (SPS) belongs to such class of tools which can necessarily estimate the possible outcomes from the decisions taken by the developer and thus the behaviour of the software as a whole. The model proposed in [4] aims at running an evolutionary algorithm on SPS that is capable of yielding in the necessary decision rules used by a developer in course of development of a software product which can presumably facilitate the estimation of software behaviour during its implementation. As per this approach, SPS is expected to generate the database for a specific project that would represent the input to an evolutionary algorithm which subsequently provides the management rules to be taken up by the software developer during development. A management rule, thus, can predict a probable outcome from every action that is executed by the developer. It can help the developer to pick one or other action as per requirement, and subsequently the software behaviour can be usefully estimated further.

2.2 Artificial Neural Network (ANN)-Based Approaches

In this section, we discuss various artificial neural network-based approaches to software quality prediction.

2.2.1 An ANN Approach to Software Reliability Assessment

The quality attributes of a software product can be useful in verifying various functionality-related aspects of it that are mostly related to the internal organization of the system. These attributes can necessarily reveal the reliability of the software system. Artificial neural network (ANN) approach can be very useful in this regard. A software reliability growth model is proposed by the authors in [5] which relies on a neuro-fuzzy inference system that can provide a significant measure of the reliability of the software.

2.2.2 ANN-Based Software Reliability Prediction Model

Conventional models used for prediction of software quality mostly function on a set of assumptions made relating to the development environment, and due to this reason, such models may not be suitable for sufficiently complex applications. To overcome such an issue, a reliability prediction model needs to be free from any assumptions, and such a model can be viable using ANN approach as suggested in [6]. This model takes as input the history of failures of the software and can successfully predict the probable future failures, and most importantly, such a model can necessarily scale with the complexity of the software.

2.2.3 ANN-Based Approach for Software Fault Prediction Models

As claimed by the authors in [7], ANN model can be useful in determining the occurrence of the anticipated number of faults in a software product during its implementation, and the authors justify the same having proposed an ANN model for software fault prediction with reference to a hypothesis that suggests that the distribution of faults in a software product mostly depends on its complexity. Using an experiment, the authors conclude that it performs better than other existing software fault prediction models.

2.3 Bayesian Network (BN) for Software Quality Prediction

In this section, we bring a review of various Bayesian network (BN)-based software quality prediction models.

2.3.1 Activity-Based BN Models

In [8], the authors proposed an activity-based quality model (ABQM) for software quality prediction as well as assessment using BN approach. Here, authors propose to split a more complex concept into simpler definitions, and for each definition, a node is created in BN. Further, BN is enriched with quality-related quantitative information that can predict the software quality with much perfection.

2.3.2 Integration Software Quality Prediction Model

Most of the prediction models for software quality are oriented towards the prediction of only a single quality feature. But in practice, any software is associated with a set of quality features such as functionality, efficiency, reliability, reusability and many others. It is necessary that a prediction model be capable of predicting the behaviour of multiple features. In [9], the authors develop a BN model that is aimed at predicting two quality features, namely reliability and usability. Here, the authors split each feature into a set of sub-features, and each sub-feature is assigned a node in the network. As claimed by the authors, this model can be scaled to more quality features, and thus, more accurate quality prediction can be achieved using such a model.

2.3.3 Quality Prediction in Extreme Programming

Extreme programming (XP) is an iterative approach to software development which comprises components called user stories where each user story is equivalent to software requirement specification (SRS) document in software development. However, quality prediction in XP processes is extremely difficult. A BN-based quality prediction model for XP processes is proposed in [10] that is capable of predicting the quality during the planning stage well before the development starts. It can also predict the time when the project will be completed.

2.3.4 Quality Prediction of Embedded Software

A Bayesian belief network (BBN) is proposed by in [11] that focuses on the quality prediction of embedded software. This model mostly relies on various relationships that exist between different processes involved in development of embedded software. Authors claim that BBN is much more efficient in the quality prediction of embedded software.

2.4 Fuzzy Logic-Based Prediction Models

In this section, fuzzy logic (FL)-based software quality prediction models proposed by various authors are detailed.

2.4.1 Triangular Fuzzy Membership Function for Quality Assessment

Prior to the assessment of quality of a software product, a thorough investigation of software metrics must be done. A FL-based quality assessment procedure is proposed by authors in [12]. Here, the authors take into consideration two metrics for quality assessment, namely error rate and rate of inspection. The authors, in their approach, have used triangular fuzzy membership function for the representation of these metrics and applied them to a set of fuzzy inference rules. They have conducted experiments over a number of modules and successfully identified the error-prone modules.

2.4.2 Transparent Model for Maintainability Prediction

Most of the software quality assessment models do not take into consideration two crucial quality aspects: imprecise linguistic knowledge regarding the software obtained from the experts and precise quantitative information obtained from historical data of implementation of the software. A FL-based transparent software quality assessment model has been developed by the authors in [13] which incorporates these two forms of knowledge in order to predict software maintainability. Here, Mamdani fuzzy inference has been implemented by the authors that demonstrates significantly better performance for maintainability prediction as compared to other machine learning techniques used for this purpose.

2.4.3 Integrated Software Quality Assessment

In order to develop a software quality prediction model, it is essential to take into account the most relevant quality attributes and their accurate quantification. However, it is associated with two major aspects: choice of the source of the relevant attributes and uncertainty associated with their accurate quantification. A fuzzy multi-criteria-based integrated software quality evaluation model is proposed in [14] with the quality attributes having been picked from ISO/IEC 9126 standard. Results justify the efficiency of this model in quality prediction.

2.4.4 Evaluation of Software Similarity

Quality evaluation between two different software products can be achieved from the level of similarity existing between them in the context of the amount of effort consumed in the development process. This aspect can be rarely found in the literature pertaining to software quality prediction. This problem of similarity evaluation has been addressed by the authors in [15], wherein the level of similarity between two software projects is attributed as the distance in quality between the two projects. Nevertheless, the similarity cannot be expressed in numerical values for which the authors have used FL to assign linguistic values to the level of similarity between two software projects which can further be defuzzified to obtain a numerical assessment.

2.4.5 Neuro-Fuzzy Model for Quality Prediction Model

A hybrid neuro-fuzzy model for software quality prediction has been proposed by authors in [16] where the authors use the software attributes and the sub-attributes which are analysed by a fuzzy logic neural network (FLNN) system developed by them. A set of fuzzy If–Then rules are used for this purpose. This model can be useful during the process of software development.

2.5 Decision Tree (DT) Approach for Quality Prediction Model

In this section, we detail the spectrum of research work carried out by different authors relating to software quality prediction making use of decision tree approach.

2.5.1 Quality Classification Using DT

By virtue of prediction of quality of a software product prior to testing can significantly reduce the quantum of work of the developer during software development. DT approach-based model for classification of software quality that relies on most relevant software metrics is capable of identifying the faulty modules in a software product. Researchers attribute such models to white box quality assessment models. DT approach-based algorithm SPRINT is used by the authors in [17] for the classification of software quality and obtains successfully the relevant classification trees that can contribute to the classification of software quality. This approach is free from any memory bottleneck that can be found in other algorithms implemented for this purpose. SPRINT algorithm is an extended version of DT algorithm CART that is also used for classification of software quality.

2.5.2 DT-Based Quality Models Using Principal Component Analysis (PCA)

Quality prediction models based on classification trees make use of metrics related to the software product, software process as well as execution of it for identification of faulty modules in a software. As observed, these models are found to be more robust as compared to other models. However, the accuracy of identification in case of such models may present a challenge. The authors in [18] have conducted a case study of four different versions of large communication systems making use of principal component analysis (PCA) DT-based models which can be further improved by virtue of transformation of quality prediction algorithms.

2.5.3 DT-Based Quality Assessment Models for Fault Prediction

Identification of high-risk modules in a software product at early stages of development can make it feasible for improvement of anticipated faulty modules that can presumably enhance the reliability of the software during its implementation. Regression tree-based models can be significantly more effective in prediction of software reliability than any other models. A regression tree approach for quality prediction models has been addressed by authors in [19], wherein a case study is conducted with large telecommunication systems with lines of code of more than 13 million making use of regression tree algorithms like CART-LS and S-PLUS, and the authors conclude that identification of faulty modules in a software can be achieved at ease using this approach. The fault prediction accuracy has been assessed in terms of the metrics such as absolute, average and relative errors.

2.6 Support Vector Machine (SVM)-Based Prediction Models

In this section, we discuss different models for software quality prediction using support vector machine (SVM) approach.

2.6.1 SVM for Prediction of Software Defects

Defects in software products can be removed during testing of the software. However, if such defects can be identified at earlier stages prior to testing during software development, then significant amount of time and effort can be saved. SVM approach can be used to identify such defects well in advance. SVM approach is used by authors in [20] for defect prediction, and the authors claim to have achieved a high accuracy in prediction. Here, authors use MATLAB as the interface and implement SVM on the ANT-1.7 data set for identification of the defects.

2.6.2 SVM Classifier for Code Fault Prediction

Machine learning classifiers and SVM classifiers, in particular, have been proved to be significantly useful in the prediction of fault in the code of a software. In order to achieve this prediction, the classifier needs to be trained in the historical data. However, if the number of features is very large, then the accuracy may be reduced. A feature selection technique has been proposed by authors in [21] which can be used for fault prediction in software code. Here, authors use Naive Bayes and F-measure metric for improvement of accuracy of prediction.

2.6.3 Prediction of Defects in Software Using SVM

Identification and removal of software defects are a time-consuming process, and if a software project is not planned properly, then the process of identification and removal of defects consumes significantly more time than the process of code development. A SVM-based technique proposed by authors in [22] for software defect prediction uses a data mining approach for the identification of software defects on the basis of existing software metrics. Authors claim significant improvement of software quality using their technique that subsequently reduces the development cost significantly.

2.6.4 Empirical Approach to Defect Detection Using SVM

An empirical approach to software defect prediction is proposed by authors in [23] that uses SVM approach. Here, authors have chosen code smells as the feature for the prediction. First of all, a smell prediction model is developed using SVM technique that is subsequently used for software defect prediction. The authors revealed in their paper that this technique was used in the development of subsequent release of the Eclipse software and claims high accuracy in prediction.

2.7 Case-Based Reasoning for Quality Prediction Model

In this section, various techniques for software quality prediction using a machine learning technique called case-based reasoning (CBR) are detailed.

2.7.1 CBR for Software Quality Estimation

Authors in [24] propose a CBR technique for software quality estimation procedure that can be conducted by human experts. This approach mostly relies on the similarity that exists between projects in the past which is achieved taking into account some key quality attributes of the software. Various similarity techniques are conducted by the authors in order to find out the optimal one that resulted in much higher accuracy as well as reliability in the estimation of software quality.

2.7.2 CBR for Prediction of Faulty Modules of Software

If the faulty modules of a software can be identified in advance, the quality of the software can be presumably improved. Authors in [25] propose a CBR-based technique for prediction of fault-prone modules within a software during the process of development. Here, authors refer to the hypothesis that a module is supposed to be fault-prone if a module developed earlier with the same functionality was found to be fault-prone as well.

2.7.3 CBR for Estimation of Development Effort

If the amount of human effort contributed in software development can be assessed, the quality of the software can be accordingly improved that consequently reduces the cost of the software. Authors in [26] operate on a CBR approach for effort estimation during development with respect to student programs. Here, the CBR model proposed by authors validates the student data and mainly concentrates on the imprecision and uncertainty related to the chosen attributes from student data. This model is mainly intended at reducing the cost of effort involved in the development of software.

2.7.4 Fault Prediction Using CBR

Most of the prediction models of software quality that use CBR approach focus on the similarity distance between developed software modules. The authors in [27] have proposed a quality prediction model based on CBR approach taking into consideration the similarity distance between software modules, wherein they combine similarity function with a solution algorithm for the exploration of the fault prediction procedure. Here, authors have conducted a wide spectrum of case studies with large telecommunication legacy systems and claim that using similarity function significantly improves fault prediction in a software product.

3 Conclusion and Future Work

In this review paper, we have conducted a thorough investigation through the literature available up to date on the machine learning techniques applied to software quality prediction by a wide spectrum of authors. Different models for software quality prediction have been proposed using machine learning techniques such as ANN, BN, FL, DT, SVM and CBR and are detailed in this paper. In most cases, the authors conclude that machine learning techniques prove to be significantly more efficient in software quality prediction in comparison with other techniques. However, still there exists ample scope for implementation of some innovative approaches pertaining to machine learning techniques to be tested for prediction of the quality of a software product. The contents of this paper can serve a useful guide to the researchers in the field of software engineering in order to carry out further research work in the development of some novel software quality prediction models.

References

1. Puri, A., Singh, H.: Genetic algorithm based approach for finding faulty modules in open source software systems. Int. J. Comput. Sci. Eng. Surv. (IJCSES) **5**(3), 29–40 (2014)
2. Adline, A., Ramachandran, M.: Predicting the software quality using the method of genetic algorithm. Int. J. Adv. Res. Electr. Electron. Instrum. Eng. **3**(2), 390–398 (2014)
3. Azar, D., Vybibhal, J.: An ant colony optimization algorithm to improve software quality prediction models: case of class stability. Inf. Softw. Technol. **53**(4), 388–393 (2011)
4. Pattnaik, S., Pattanayak, B.K.: A survey on machine learning techniques used for software quality prediction. Int. J. Reasoning-Based Intell. Syst. **12**(1/2), 3–14 (2016)
5. Kapur, P.K., Khatri, S.K., Goswami, D.N.: A generalized dynamic integration software reliability growth model based on neural network approach. In: Proceedings of International Conference on Reliability, Safety and Quality Engineering, pp. 831–838 (2008)
6. Karunanithi, N., Whitley, D., Malaiya, Y.K.: Using neural networks in reliability prediction. Softw. IEEE **9**(4), 53–59 (1992)
7. Khoshgoftaar, T.M., Pandya, A.S., More, H.B.: A neural network approach for predicting software development faults. In: Proceedings of Software Reliability Engineering in 3rd International Symposium on IEEE, pp. 83–89 (1992)
8. Wagner, S.: A bayesian network approach to assess and predict software quality using activity-based quality model. Inf. Softw. Technol. **52**(11), 1230–1241 (2010)
9. Radinski, L.: A conceptual bayesian net model for integrated software quality prediction. Ann. UMCA Informatica **11**(4), 49–60 (2011)
10. Abouela, M., Benedicenti, L.: A bayesian network based XP process modelling. Int. J. Softw. Eng. Appl. (IJSEA) **1**(3), 1–15 (2010)
11. Amasaki, S., Takagi, Y., Mizuno, O., Kikuno, T.: Constructing a bayesian belief network to predict final quality in embedded system development. IEICE Trans. Inf. Syst. **8**(6), 1134–1141 (2005)
12. Mittal, H., Bhatia, P., Goswami, P.: Software quality assessment based on Fuzzy logic technique. Int. J. Soft Comput. Appl. (IJSCA) **1**(3), 105–112 (2008)
13. Ahmed, M.A., AL-Jamini, H.A.: Machine learning approaches for predicting software maintainability: a Fuzzy based transparent model. IET Softw. **7**(6), 317–326 (2013)
14. Challa, J.S., Paul, A., Dada, Y., Nerella, V., Stivastava, P.R., Singh, A.P.: Integrated software quality evaluation: a Fuzzy multi-criteria approach. J. Inf. Process Syst. (JIPS) **7**(3), 473–518 (2011)
15. Idri, A., Abra, A.: A fuzzy logic based measures for software project similarity: validation and possible improvements. In: Proceedings of 7th International Symposium on Software Metrics, pp. 85–96. IEEE, England, UK (2001)
16. Pattnaik, S., Pattanayak, B.K., Patnail, S.: Prediction of software quality using neuro-fuzzy model. Int. J. Intell. Enterp. (IJIE) **5**(3), 292–307 (2018)
17. Khoshgoftaar, T.M., Seliya, N.: Software quality classification modelling using SPRINT decision tree algorithm. Int. J. Artif Intell. Tools (IJAIT) **12**(3), 207–225 (2002)
18. Khoshgoftaar, T.M., Shah, R.M., Allen, E.B.: Improved tree-based models for software quality with principal component analysis. In: Proceedings of Software Engineering Reliability Engineering, ISSRE 2000 of 11th International Symposium on IEEE, pp. 198–209 (2000)

19. Khoshgoftaar, T.M., Seliya, N.: Tree based software quality estimation models for fault prediction. In: Proceedings of Software Metrics of Eighth IEEE Symposium, pp. 203–214 (2002)
20. Gupta, R.: Software defects prediction using Support vector machine. Int. J. Comput. Sci. Softw. Eng. **1**(1), 39–48 (2015)
21. Shankar, K., Kannan, S., Jennifer, P.: Prediction of code fault using naive bayes and SVM classifiers. Middle East J. Sci. Res. (MEJSR) **20**(1), 108–113 (2014)
22. Selvaraj, P.A., Thangaraj, P.: Support vector machine for software defect detection. Int. J. Eng. Technol. Res. (IJETRE) **1**(2), 68–76 (2013)
23. Reshi, J.A., Sing, S.: Predicting software defects through SVM: an empirical approach. Int. J. Sci. Res. Dev. **5**, 1835–1838 (2017)
24. Rashid, E., Patnaik, S., Bhattacherjee, V.: Software quality estimation using machine learning case-based reasoning technique. Int. J. Comput. Appl. (IJCA) **58**(14), 43–48 (2012)
25. Khoshgoftaar, T.M., Ganesan, K., Allen, E.B., Ross, F.D., Munikoti, R., Goel, N., Nandi, A.: Predicting fault-prone modules with case-based reasoning. In: Proceedings of Software Reliability Engineering, The Eighth International Symposium on IEEE, 27–35 (1997)
26. Rashid, E., Bhattacherjee, V., Patnaik, S.: The application of case-based reasoning to estimation of software development effort. Int. J. Comput. Sci. Inform. (IJCSI) **1**(3), 20–34 (2012)
27. Khoshgoftaar, T.M., Seliya, N., Sundaresh, N.: An empirical study of predicting software fault with case-based reasoning. Proc. Softw. Qual. J. Springer Sci. **14**(2), 85–111 (2006)

Investigating Determinants of Profitability of Commercial Firms: Rough Set Analysis

Arpit Singh and Subhas Chandra Misra

Abstract To achieve excellence in any business venture and have an edge over the competitors, it is necessary to optimally use the available scarce resources which form the foundation of a perpetually flourishing business enterprise. This paper employs rough set theory to categorically remove any superfluous data present in the system by establishing a discernibility matrix which accommodates the elements that differentiates the objects or the equivalence classes obtained using the indiscernibility relation. The basic principle, to achieve the objective of data reduction, is to minimize the Boolean expression obtained by logically concatenating entries of the discernibility matrix. The reduced information is subjected to the standard statistical regression procedures and is found that it is statistically consistent. Finally, an artificial neural network modeling is suggested which validates the results obtained using rough set analysis for the relations between the data variables or the given information other than the linear ones.

Keywords Rough sets · Discernibility matrix · Vagueness · Reducts

1 Introduction

Dealing with a huge amount of data poses an enormous amount of problems for various technical analyses. It becomes imperative to work with the seamless amount of data in real-world scenario especially when working in an industrial setup. To analyze this huge amount of data and eventually coming up with decisions, either the betterment of the system or perhaps even formulating new policies in the interest of the organization, formally forms the backbone of almost every business venture and establishments [1].

As is evident that such kind of an analysis requires a colossal amount of data to work with, which ultimately is a daunting and time-consuming task, it therefore

A. Singh (✉) · S. C. Misra
Department of Industrial & Management Engineering, IIT, Kanpur, India
e-mail: arpits@iitk.ac.in

S. C. Misra
e-mail: subhasm@iitk.ac.in

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_46

becomes an important task for every organization to judiciously take steps which not only satiates their needs but also minimizes the time consumed and the resources required. One very direct and an obvious method in this direction seems to conspicuously choose only those data information which really has an apparent effect on the decision-making process thereby eliminating the information that actually bears very little or no significance with the output whatsoever [2–5].

To avoid these circumstances, it is always necessary to devise ways and means so as to simplify the task as much as possible. One way to accomplish the foresaid objective is to reduce the data, i.e., data reduction, which means that all the redundant or superfluous data is removed or obliterated. Redundant data means, the data which when taken out of consideration does not affect the decision making process or which have no or little use in the decision making process. By doing so, not only the data to be considered is reduced substantially but also the accuracy level of the decision evolved is better and is way more understandable [4].

In traditional statistical scenario, techniques like principal component analysis, factor analysis and clustering have been adopted to establish dimensionality reduction. These techniques, nevertheless, are extremely helpful but suffer from the inherent drawback of various suppositions at numerous levels about the data in question. One of the many assumptions required before the application of these methods is, for instance, the assumption of the distribution of the data which is a grim situation when we are unable to figure out the distribution of the data we intend to work with [2, 3].

Among the various methods available for data mining, this paper projects rough set theory approach for the same. A new technique of dealing with inconsistent and uncertain data coined in 1982 [6] is in many ways better and sophisticated than the conventional statistical methods for the same. Rough set theory not only provides data mining but also facilitates decision rules resulting from the reduced dataset. The major advantage of using rough set theory is that any prerequisite knowledge about the data supplied, i.e., regarding the statistical properties of the data, etc., which is indispensable when working with the traditional methods of data mining, is not required. Rough set theory gives detailed decision rules which are supported by the examples from the given data itself. Nothing has to be assumed beforehand when working with this approach. The decision rules so obtained are easy to understand and interpret unlike the traditional methods.

An immense amount of work has been done in the area of data mining using rough set theory. A parallel knowledge acquisition framework using MapReduce was proposed that proved instrumental in devising hierarchical decision rules based on minutest levels of granularity [7]. Classical rough set methodology does not seem to work very well with labeled datasets. To deal with limited labeled datasets, local rough sets are coined that suggest series of concept approximations and attribute reduction algorithms which surprisingly outperformed the traditional rough set methodology [8]. Sabita, Sreekumar et al. have demonstrated the attribute selection in marketing scenario using rough set theory where the decision rules were extracted from a database making use of lower and upper approximations of a given set. Durairaja and Meena have successfully displayed a hybrid prediction system using rough set approach and artificial neural networks where the original problem variables were reduced and then supplied to the artificial neural network modeling which rendered a better prediction system [9]. In their paper on site selection decision making for water reservoirs [10], Lashteh Neshaei and Pirouz suggested the use of rough set theory in deciding the feasibility of the locations for constructing the water reservoirs [11]. To deal with dynamic datasets that display high level of complexity in time/space, a progressive increasing algorithm for attribute reduction with rough sets is presented that adopts active sample selection process for attribute reduction [12]. The application of rough set methodology in finding reducts or to reduce the attributes is quite time consuming. In order to tackle this, two quick reduct algorithms have been proposed that are useful in determining the reducts relatively quickly as compared to the traditional methods. It relies mostly on neighbor inconsistent pairs to develop the new attribute reduction algorithms [13]. A modified cuckoo search algorithm is developed for high dimensionality reduction using rough sets that follows the obligate brood parasitic behavior of cuckoo birds. The results thus obtained are compared with ANOVA and other machine learning algorithms such as K nearest neighbors and support vector machines [14]. It was found that rough set methodology proves much better than the existing parametric techniques. In an attempt to overcome the limitations of classical rough set methods that operate only on categorical data, a new feature reduction algorithm based on fish swarm algorithm is devised that presents optimal feature subsets based on neighborhood rough set method [15].

The work put forth in this paper shows the inclusion of Boolean algebra in minimizing the expression obtained by indiscernibility relation on the attributes set thereby deducing reduced sets of attributes from the actual set of original attributes which can be used in place of the original set of attributes without affecting the statistical as well as analytical properties of the database [3, 4].

The concept of rough set theory along with its application to a financial dataset [16] will be dealt in this paper. The paper consists of five sections. The first section opens up with an introduction of the paper work and a brief literature review. The second section presents a prelude to the basic concepts of rough set theory [17] and its properties along with an introduction of artificial neural networks and its subsidiary multilayer perceptron. The third section outlines the algorithm employed for the data reduction using rough set theory. The fourth and fifth sections conclude the paper by citing results and conclusions of the work, and finally references are listed which were useful throughout the project.

2 Preliminaries

2.1 Rough Sets

In the following section, an attempt has been made to present the whereabouts of the technique in detail as was established [3, 5, 18]. As was suggested by Lotfi Zadeh, the vagueness in a set can be expressed in terms of the fuzzy sets which is nothing but description of the belongingness of a given element in a universe of discourse often expressed by fuzzy membership function. Since many mathematicians and philosophers have developed a keen proclivity in vagueness of the set, thus Pawlak coined a new venture in understanding vagueness of the set using rough set theory [17].

Rough set theory (RST) was given by Pawlak [5] at the institute of computer sciences, Warsaw. The basic fundamental of this technique lies in the fact that every entity been considered has some information inherently linked with it, as suggested by Pawlak and Skowron in the year 2007. And the entities with the similar information are considered similar in light with the available data. Now, this similarity gives rise to the *indiscernibility relation* which forms the mathematics behind the rough set analysis.

To employ rough set theory to data analysis, we need to arrange the given data in the form of a rectangular array of information framework where the horizontal lines constitute the entities, the vertical lines comprise the attributes or the factors on which indiscernibility relation is formed, and the entries are the *attribute values.* The objects which display the same features are indiscernible and form the elementary granules of knowledge. The collection of the sets of these objects is called crisp sets, and all the other sets are referred to as rough sets [19].

As rough sets are indicative of the vagueness in a set much like those of fuzzy sets, so rough sets can be understood more precisely by *interior* and *closure* operations collectively known as **approximations**.

Suppose a group of objects is considered represented by universe of discourse U and an indiscernibility relation given by R which is a subset of the Cartesian product of U. Clearly, the indiscernibility relation is a reflection of the uncertainty about the information of the universe U. As a simplifying assumption, the indiscernibility relation is considered as an equivalence relation [20]. Let X be a subset of U. After characterization of the set X with respect to R, we obtain following approximations:

- The *lower approximation* of the set X with respect to R is the set of all objects which can be **for certain** classified as X with respect to R [21].
- The *upper approximation* of the set X with respect to R is the set of all objects which can be **possibly** classified as X with respect to R [21].
- The *boundary region* of the set X with respect to R is the set of all objects which can be classified as not belonging to either set X or its complement [21].

2.1.1 Properties of Rough Sets Approximations

If the lower approximation of the set with respect to an indiscernibility relation π is given by $\pi_{\mathbf{l}}(\mathbf{Q})$ and the upper approximation by $\pi^{\mathbf{u}}(\mathbf{Q})$, then the following properties hold:

- $\pi_{\mathbf{l}}(\mathbf{Q})$ is a subset of a set of without doubt collection of entities, and set of doubtful collection of entities is a superset of $\pi^{\mathbf{u}}(\mathbf{Q})$

- $\pi_{\mathbf{l}}(\mathbf{Q}) = \pi^{\mathbf{u}}(\mathbf{Q}) = \mathbf{Q}$

- $\pi_{\mathbf{l}}(\mathbf{U}) = \pi^{\mathbf{u}}(\mathbf{U}) = \mathbf{U}$

- $\pi^{\mathbf{u}}(\mathbf{QUN}) = \pi^{\mathbf{u}}(\mathbf{Q})\mathbf{U}\,\pi^{\mathbf{u}}(\mathbf{N})$.

Rough sets can also be defined in terms of *rough membership function* as

$$\mu_Q^{\pi} : \to (0,\ 1). \tag{1}$$

$$\mu_Q^{\pi}(q) = |Q \cap \pi(q)|/|\pi(q)| \tag{2}$$

where |·| denotes cardinality.

From the aforesaid arguments, it can be summarized as follows:

- Set Q is rough with respect to π if $\pi_{\mathbf{l}}(\mathbf{U}) \neq \pi^{\mathbf{u}}(\mathbf{U})$
- Set Q is rough with respect to π if for some q, $\mathbf{0} < \mu_Q^{\pi}(\mathbf{q}) < \mathbf{1}$.

2.2 Artificial Neural Network

Neural networks propose methods for modeling nonlinear data which enables the user to develop more efficient and accurate predictive model for the given nonlinear data. Essentially, multilayer perceptron algorithm is used for designing the predictive model in this paper [22].

2.2.1 Multilayer Perceptron

The working of multilayer perceptron is analogous to that of human brain function. Just as happens in a human brain, the input signal is transmitted by synapses through the network of neurons, and finally yielding an output, the same protocol is followed in an MLP where each network has a set of input signals associated with the synapse set and the corresponding output set. The set $\{x_{0,}\ x_1,\ \ldots,\ x_n\}$ forms the vector of input signals. The input signals are then fed to the nucleus $\Phi(\ldots)$ by the synapses represented as the set $\{w_1,\ \ldots,\ w_n\}$which eventually leads to an output. There is an additional neuron parameter w_0 known as bias that can be represented as a synapse associated to an input $x_0 = -1$. The neuron output is then obtained through the activation function of the neuron $y = \Phi(x,\ w)$ where the activation function can be any function depending upon the user. The mechanism which was employed in the multilayer perceptron is a feed forward mechanism. In the analysis presented in this paper, there were two layers, namely a hidden layer and an output layer. Here, each neuron in the hidden layer is connected to each neuron in the output layer. The neural network's training is established by back propagation algorithm. This algorithm allows the network to adjust the synapse's values each time an iteration is performed. This process is repeated until a suitable interruption criterion is met. The synapse value correction is done so that the output obtained from the neural network is not way different from the actual output. The entire procedure of finding a proper neural network model by multilayer perceptron is accomplished using SPSS software. Applying the above-mentioned procedure, the multilayer perceptron network was obtained for the problem. The scatter plot revealed that there was a positive correlation between the observed and the model predicted values suggesting the appropriateness of the model found which suggested that even if the linearity assumption was foregone the reducts sets were appropriate in the prediction of the dependent variable appropriately [9, 23].

3 Methodology

The algorithm for data reduction using rough set theory is illustrated for the following example financial dataset of companies [5, 18, 24]. The information table consists of objects which is a set of 20 companies, the attributes are the parameters: number of employees (in 1000s), capital expenditures($ millions), intangible expenditures ($ millions), costs of goods ($ millions), labor and related expense($ millions), advertisement expenditures ($ millions) and research and development expense ($ millions) abbreviated as α, β, γ, δ, ε, ζ and η. The decision variable will be sales which is abbreviated as ρ as shown below. The dataset has been taken from the book Business Forecasting, Ninth Edition by Hanke and Wichern and has been provided by Dr. Lynn Stephens of Eastern Washington University [25] (Tables 1 and 2).

Step 1: Discretize the attribute values by using the formula

$$x(\text{discrete}) = \frac{x}{\text{maximum value}}. \tag{3}$$

Step 2: Generate the equivalence classes using the equivalence relationship established between the elements of the given dataset using the following property:

For all a, b in the set X, the equivalence class of "a" under the equivalence relation **R** [26] denoted by [E$_i$] is given by

$$[\mathrm{E}_i] = \{b \varepsilon X | a\mathbf{R}b\}$$

- E1: (1, 3, 14)
- E2: (2, 10, 15)
- E3: (4, 8, 12, 13, 16, 17, 20)

Table 1 Financials of the companies arranged in the form of information table

	α	β	γ	δ	ε	ζ	η	ρ
1	42	147.9	30.6	2285.2007	599.7998	118.3	28	3221.808
2	20.905	93	29.1	1057.2002	343.2	114.9	8.9	1690.601
3	39	66.867	55.86	1387.0679	661.3997	95.568	11.182	2197.274
4	23.3	59.556	69.608	1743.7952	25.632	51.917	8.5	2357.826
5	35	297	29	7423	1178	12.8	9.253	8129
6	23	394	20	10942	2556	11.653	14.6	11851
7	3.9	2.59	4.288	233.53	22.835	3.529	30.732	323.866
8	8.378	10.984	3.372	582.2649	25.625	44.999	64.873	660.486
9	50.912	102.708	217.092	4156.8671	12.836	66.264	8.779	4351.161
10	5.5	16.601	29.59	874.1287	19.5	112.386	18.365	985.837
11	39.6	206.102	157.352	2997.2703	518	139.729	16.413	3802.551
12	22.6	50.669	47.079	1885.9053	349.491	48.817	9.5	2576.044
13	28	1.312	42	84.659	35.555	22.937	8.733	106.01
14	46.881	103	31.1	4424.3007	785	141.3	18.5	5669.895
15	2.894	4.577	2.209	246.698	42.837	87	1.1	319.65
16	10.1	19.56	27	286.2288	48.999	1.87	23.652	511.727
17	22.801	58.094	33	467.4436	36.5	16.035	29.632	884.619
18	2.3	3.951	5.289	111.031	31	4.023	38.542	166.37
19	18	1.14	14.5	43.743	26.321	90.325	56.982	59.13
20	3.1	2.009	18.493	105.33	15.888	46.3	8.633	136.69

- E4: (7, 18)
- E5: (9)
- E6: (11)
- E7: (19)
- E8: (5)
- E9: (6)

Step 3: Create the discernibility matrix by using the semi-distance qualitative function given by

Every pair of objects (m, n) is assigned a subset of attributes given by $€(m, n)$ which satisfies the following properties:

(a) $€(m, m) = \Phi$
(b) $€(m, n) = €(n, m)$
(c) $€(m, o) = €(m, n)U€(n, o)$ for all $m, n, o \in U$

Step 4: Form the indiscernibility relationship between the equivalence classes and attributes using Boolean operations. For the discernibility matrix $\mathbf{M} = [\mathbf{a_{ij}}]_{\mathbf{i \times j}}$, **a**$\in$**{the set of the condition attributes} Boolean OR** operation is applied to each

Table 2 Normalized values

	α	β	γ	δ	ε	ζ	η	ρ
1	0.825	0.375	0.141	0.209	0.235	0.837	0.432	0.272
2	0.411	0.236	0.134	0.097	0.134	0.813	0.137	0.143
3	0.766	0.17	0.257	0.127	0.259	0.676	0.172	0.185
4	0.458	0.151	0.321	0.159	0.01	0.367	0.131	0.199
5	0.687	0.754	0.134	0.678	0.461	0.091	0.143	0.686
6	0.452	1	0.092	1	1	0.082	0.225	1
7	0.077	0.007	0.02	0.021	0.009	0.025	0.474	0.027
8	0.165	0.028	0.016	0.053	0.01	0.318	1	0.056
9	1	0.261	1	0.38	0.005	0.469	0.135	0.367
10	0.108	0.042	0.136	0.08	0.008	0.795	0.283	0.083
11	0.778	0.523	0.725	0.274	0.203	0.989	0.253	0.321
12	0.444	0.129	0.217	0.172	0.137	0.345	0.146	0.217
13	0.55	0.003	0.193	0.008	0.014	0.162	0.135	0.009
14	0.921	0.261	0.143	0.404	0.307	1	0.285	0.478
15	0.057	0.012	0.01	0.023	0.017	0.616	0.017	0.027
16	0.198	0.05	0.124	0.026	0.019	0.013	0.365	0.043
17	0.448	0.147	0.152	0.043	0.014	0.113	0.457	0.075
18	0.045	0.01	0.024	0.01	0.012	0.028	0.594	0.014
19	0.354	0.003	0.067	0.004	0.01	0.639	0.878	0.005
20	0.061	0.005	0.085	0.01	0.006	0.328	0.133	0.012

entry in a_{ij}, and a **Boolean AND** operation is applied to all a_{ij}'s where $i > j$. The indiscernibility relation is obtained by minimizing the expression so obtained according to the Boolean laws of algebra

$$
\begin{aligned}
&\alpha \cdot (\alpha + \zeta) \cdot (\alpha + \zeta + \eta) \cdot (\zeta + \eta) \cdot \eta \cdot (\gamma + \zeta) \cdot (\alpha + \gamma + \zeta) \cdot (\alpha + \gamma) \cdot (\alpha + \gamma + \eta) \cdot \gamma \cdot (\alpha + \gamma).\\
&(\alpha + \gamma + \zeta) \cdot (\alpha + \gamma + \zeta + \eta) \cdot \zeta \cdot (\alpha + \eta) \cdot \eta \cdot (\zeta + \eta) \cdot \zeta \cdot (\alpha + \gamma + \zeta + \eta) \cdot (\alpha + \gamma + \eta) \cdot (\beta + \delta + \zeta)\cdot\\
&(\alpha + \beta + \zeta + \delta) \cdot (\alpha + \beta + \delta) \cdot (\alpha + \beta + \delta + \eta) \cdot (\beta + \gamma + \delta) \cdot (\beta + \gamma + \zeta + \delta) \cdot (\alpha + \beta + \zeta + \eta + \delta)\cdot\\
&(\alpha + \beta + \zeta + \varepsilon + \delta) \cdot (\beta + \zeta + \varepsilon + \delta) \cdot (\beta + \varepsilon + \delta) \cdot (\beta + \varepsilon + \eta + \delta) \cdot (\alpha + \beta + \gamma + \varepsilon + \delta)\cdot\\
&(\alpha + \beta + \gamma + \varepsilon + \delta + \zeta) \cdot (\varepsilon + \beta + \zeta + \eta + \delta) \cdot (\alpha + \varepsilon).
\end{aligned}
\tag{4}
$$

Step 5: Minimize the Boolean expression and obtain the reduct sets $\alpha\ \gamma\zeta\ \eta\ (\varepsilon + \beta + \delta)$ implying three different reducts: $\alpha\gamma\zeta\eta\varepsilon$, $\alpha\gamma\zeta\eta\beta$ and $\alpha\gamma\zeta\eta\delta$ with core $\alpha\ \gamma\zeta\ \eta$.

4 Results

The decision rules obtained using the reducts are stated below.

4.1 *First Reduct:* α γ ζ η β

- If α, ζ = HIGH and β, γ, η = LOW then sales = LOW
- If α, β, γ, η = LOW and ζ = HIGH then sales = LOW
- If α, β, γ, ζ, η = LOW then sales = LOW
- If α, β, γ, ζ = LOW and η = HIGH then sales = LOW
- If α, β = HIGH and γ, ζ, η = LOW then sales = HIGH
- If α, γ, ζ, η = LOW and β = HIGH then sales = HIGH.

4.2 *Second Reduct:* α γ ζ η ε

- If α, ζ = HIGH and ε, γ, η = LOW then sales = LOW
- If α, ε, γ, η = LOW and ζ = HIGH then sales = LOW
- If α, ε, γ, ζ, η = LOW then sales = LOW
- If α = HIGH and γ, ζ, η, ε = LOW then sales = HIGH
- If α, γ, ζ, η = LOW and ε = HIGH then sales = HIGH.

4.3 *Third Reduct:* α γ ζ η δ

- If α, ζ = HIGH and δ, γ, η = LOW then sales = LOW
- If α, δ, γ, η = LOW and ζ = HIGH then sales = LOW
- If α, δ, γ, ζ, η = LOW then sales = LOW
- If η = HIGH and γ, α, δ, ζ = LOW then sales = LOW
- If α, δ = HIGH and γ, ζ, η = LOW then sales = HIGH
- If α, γ, ζ, η = LOW and δ = HIGH then sales = HIGH.

5 Comparison with Linear Regression

The linear regression model issued for the original set of attributes with the use of SPSS software is shown in Table 3.

Table 3 Output regression summary for the original set of attributes

Model	R	R square	Adjusted R square	Std. error of the estimate
1	0.998[a]	0.997	0.995	219.4103649

[a]Predictors: (Constant), η, ζ, ε, γ, α, β, δ

The linear regression model for the reduct sets obtained by rough set analysis is shown in Tables 4, 5 and 6.

We tried to introduce a statistical perspective to our study by modeling the given situation using a multiple linear regression model. We resorted to utilizing the linear regression model as was insinuated by the scatter plots that indicated linear relationships. Without delving into too much details of the models obtained, we would like to talk about the significantly high values of the R square so obtained. The high values so obtained clearly signify the relative importance and appropriateness of the attributes chosen for the model. The value of coefficient of determination for all the reduct sets is above 90% which suggests that when the original set of attributes is reduced using rough set theory, the reduct sets so obtained are statistically compliant. Moreover, the chosen variables sets or the reducts substantially explain the variance in the dependent variable which is "sales" in our case. In all the three reduct sets obtained, more than 90% of the variations in the dependent variable, i.e., sales, is explained by the variations in the independent variables. Thus, a decision table with the reduced set of attributes can be used for analytical purposes without meddling with the statistical nomenclature of the database.

Table 4 Output regression summary for the first reduct

Model	R	R square	Adjusted R square	Std. error of the estimate
1	0.962[a]	0.925	0.898	982.0348386

[a]Predictors: (Constant), η, ζ, ε, γ, α

Table 5 Output regression summary for the second reduct

Model	R	R square	Adjusted R square	Std. error of the estimate
1	0.954[a]	0.909	0.877	1079.8751191

[a]Predictors: (Constant), β, ζ, η, γ, α

Table 6 Output regression summary for the third reduct

Model	R	R square	Adjusted R square	Std. error of the estimate
1	0.998[a]	0.996	0.995	219.3389808

[a]Predictors: (Constant), δ, ζ, η, γ, α

Table 7 Correlation matrix for the first reduct

Correlations		Sales	$Sales_{predicted}$
sales	Pearson correlation	1	0.956**
	Sig. (two-tailed)	–	0.000
	N	20	20
$Sales_{predicted}$	Pearson correlation	0.956**	1
	Sig. (two-tailed)	0.000	
	N	20	20

**Correlation is significant at the 0.01 level (two-tailed)

Table 8 Correlation matrix for the second reduct

Correlations		Sales	$Sales_{predicted}$
Sales	Pearson correlation	1	0.960**
	Sig. (two-tailed)	–	0.000
	N	20	20
$Sales_{predicted}$	Pearson correlation	0.960**	1
	Sig. (two-tailed)	0.000	
	N	20	20

**Correlation is significant at the 0.01 level (two-tailed)

5.1 Predictive Models Obtained Using the Reducts

A suitable predictive model is designed using the reducts found from the dataset, and scatter plots are sketched to observe the correlation between the predicted values of sales obtained from the three reduct sets models and the predicted value of sales using the original dataset.

The scatter plots obtained along with the coefficient of correlation between the predicted values of sales of the reduct datasets with that of the original dataset are shown in Tables 7, 8 and 9.

Figures 1, 2 and 3 conclude that the predicted values of the dependent variable of the reduct datasets exhibit strong positive and a linear correlation with the predicted

Table 9 Correlation matrix for the third reduct

Correlations		Sales	$Sales_{predicted}$
Sales	Pearson correlation	1	0.937**
	Sig. (two-tailed)	–	0.000
	N	20	20
$Sales_{predicted}$	Pearson correlation	0.937**	1
	Sig. (two-tailed)	0.000	
	N	20	20

**Correlation is significant at the 0.01 level (two-tailed)

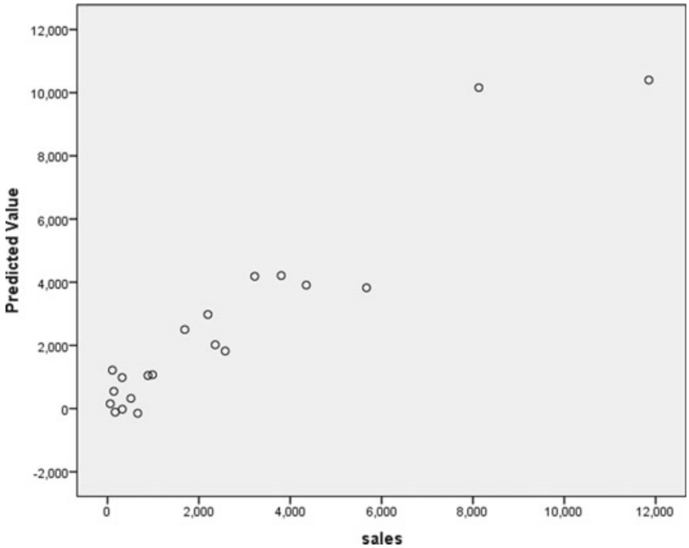


Fig. 1 Scatter plot and correlation matrix for the first reduct set

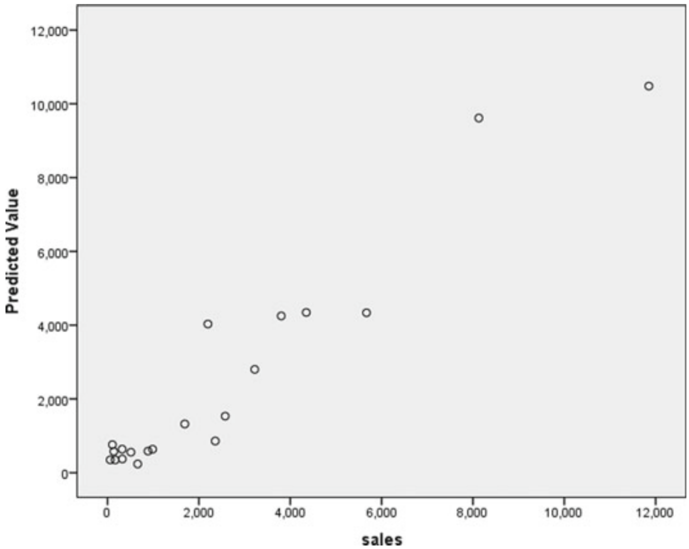


Fig. 2 Scatter plot and correlation matrix for the second reduct set

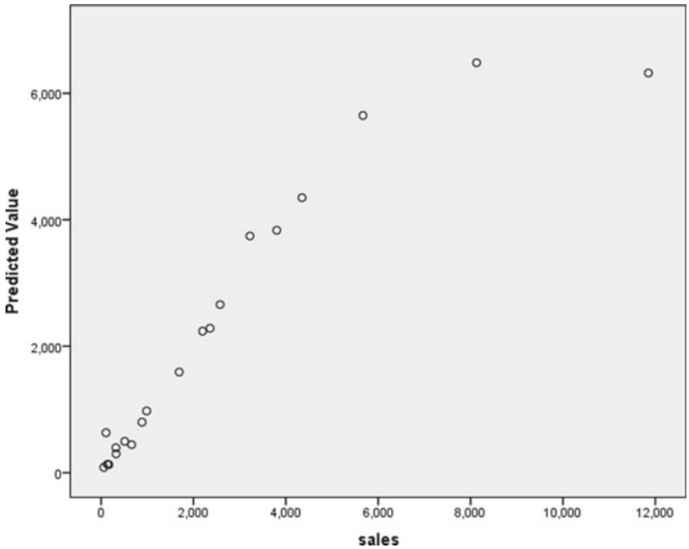


Fig. 3 Scatter plot and correlation matrix for the third reduct set

values of the dependent variable of the original dataset which suggests that the prediction efficiency of the model is not affected with the usage of the reduced sets of attributes thereby suggesting statistical validity of the rough set approach in data mining.

6 Conclusions and Discussion

This paper aimed at data reduction by applying Boolean laws of algebra in reducing the redundant information present in the discernibility matrix of the decision problem which can be considered as an extension of the rough set theory. We observe that not all the variables are needed together to determine the outcome "sales." We notice that the reduced sets of attributes also qualify to determine the level of "sales" for our dataset. The results obtained complied with the traditional statistical methods thereby proving the statistical validity of the results. The results obtained using the pairwise correlations signify strong correlations between the variables sales and the other sets of variables obtained using the rough sets analysis which further validates the results obtained using the rough set model. The multiple linear regression models taking the variables obtained using the reducts displayed reasonably high R-squared values thereby suggesting the importance the given sets of reduced attributes ascertain given the current dataset. Modeling with the artificial neural network using the reduced information obtained by this method rendered results which conformed to the statistical legitimacy for the existence of nonlinearity of the information set. This method considers primarily that the minimized form of a Boolean expression gives an expression that is self-sufficient in describing the complete logic taking place in the original Boolean expression. Thus, using this principle, a huge amount of data can be considerably reduced, and the reduced information still retains the properties which the original data contained. The methodology can be applied to any decision-making problem which can be transformed into the form of a matrix.

References

1. Sreekumar, Panda, B.: Business intelligence: an overview. J. IPM, **5**(2), 28–32 (2005)
2. Dubois, D., Prade, H., Pawlak, Z.: Foreword. Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer, Dordrecht (1991)
3. Grzymala-Busse, J.W.: Knowledge acquisition under uncertainty—A rough set approach. J. Intel. Rob. Syst. **1**(1), 3–16 (1988). Grzymala-Busse, J.W.: Managing Uncertainty in Expert Systems. Kluwer, Dordrecht (1991)
4. Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems. In: Slowinski, R. (ed.) Intelligent Decision Support. Handbook of Advances and Applications of the Rough Set Theory, pp. 331–362. Kluwer, Dordrecht (1992)
5. Pawlak, Z.: Rough Sets. Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Poland (1982)
6. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Inf. Sci. **177**(1), 3–27 (2007)
7. Qian, J., Xia, M., Yue, X.: Parallel knowledge acquisition algorithms for big data using MapReduce. Int. J. Mach. Learn. Cybernet. **9**(6), 1007–1021 (2018)
8. Qian, Y., Liang, X., Wang, Q., Liang, J., Liu, B., Skowron, A., Dang, C.: Local rough set: a solution to rough data analysis in big data. Int. J. Approximate Reasoning **97**, 38–63 (2018)
9. Durairaj, M., Meena K.: A hybrid prediction system using rough sets and artificial neural networks. Int. J. Innovative Technol. Creative Eng. **1**(7) (2011). (ISSN: 2045–8711)
10. Abed-Elmdoust, A., Kerachian, R.: Wave height prediction using rough set theory, Ocean Eng. (2012)
11. Lashteh Neshaei, M.A., Pirouz, M.: Comp. Meth. Civil Eng. **1**(1), 85–94 (2010)
12. Yang, Y., Chen, D., Wang, H.: Active sample selection based incremental algorithm for attribute reduction with rough sets. IEEE Trans. Fuzzy Syst. **25**(4), 825–838 (2017)
13. Dai, J., Hu, Q., Hu, H., Huang, D.: Neighbor inconsistent pair selection for attribute reduction by rough set approach. IEEE Trans. Fuzzy Syst. **26**(2), 937–950 (2018)
14. El Aziz, M.A., Hassanien, A.E.: Modified cuckoo search algorithm with rough sets for feature selection. Neural Comput. Appl. **29**(4), 925–934 (2018)
15. Chen, Y., Zeng, Z., Lu, J.: Neighborhood rough set reduction with fish swarm algorithm. Soft Comput. **21**(23), 6907–6918 (2017)
16. Bose, I.: Deciding the financial health of dotcoms using rough sets. Inf. Manag. **43**(7), 835–846 (2006)
17. Falc, R.: Rough set theory: a true landmark in data analysis, vol. 174, Springer Science & Business Media (2009)
18. Ziarko, W. (ed.): Rough sets, fuzzy sets and knowledge discovery. In: Proceedings of RSKD'94 Workshop (Banff). Springer, Berlin (1994)
19. Wu, S.: An algorithm for clustering data based on rough set theory. In: International Symposium on Information Science and Engineering (2008)
20. Huang, C.C., Tseng, T.L., Jiang, F., Fan, Y.N., Hsu, C.H.: Rough set theory: a novel approach for extraction of robust decision rules based on incremental attributes. Ann. Oper. Res. (2014)
21. Pratiwi, L.: An empirical study of density and distribution functions for ant swarm optimized rough reducts. In: Communications in Computer and Information Science (2011)

22. Ziarko, W., Golan, R., Edwards, D.: An application of datalogic/R knowledge discovery tool to identify strong predictive rules in stock market data. In: Proceedings of AAAI Workshop on Knowledge Discovery in Databases, pp. 89–101. Washington, D.C (1993)
23. Hou, Zhijian, Lian, Zhiwei, Yao, Ye, Yuan, Xinjian: Cooling-load prediction by the combination of rough set theory and an artificial neural-network based on data-fusion technique. Appl. Energy **83**, 1033–1046 (2006)
24. Mahapatra, S., Sreekumar., Mahapatra, S.S.: Attribute selection in marketing: a rough set approach. Sci. Dir. IIMB Manage. Rev. **22**, 16–24 (2010)
25. Hanke, J.E., Reitsch, A.G., Wichern D.W.: Business forecasting, vol.9, Upper Saddle River, NJ:Prentice Hall (2001)
26. Pawlak, Z., Slowinski, R.: Rough set approach to multi-attribute decision analysis. Eur. J. Oper. Res. (1994)

A Survey of Knowledge Capture and Knowledge Sharing Techniques in Agile Software Companies

Mohammad Anass Korimbocus, Thaveersingh Towokul and Soulakshmee D. Nagowah

Abstract Knowledge is the key resource in organisations. Nowadays, most of the industries whether public or private, have implemented this practice to maintain the effectiveness of knowledge sharing in agile software companies. Knowledge comes from the interpretation of information and data that have been processed to make appropriate decisions. It is considered as one of the fundamental parts of software development where practitioners can acquire accurate information in the right place at the right time. A survey has therefore been carried out in ten agile-based software companies in Mauritius to investigate different techniques adopted by the development teams for capturing and sharing the knowledge acquired. Results from the survey have been analyzed. Some of the main aspects have been discussed and recommendations have been made.

Keywords Knowledge · Knowledge sharing · Knowledge capture · Agile software development

1 Introduction

Knowledge is considered as a key resource in organizations [1, 2]. In the knowledge management field, knowledge is termed as the interpretation of data and information to be used to support decision making [3]. Knowledge management is about creation of values, meeting strategic requirements, transmitting the right information effectively and efficiently and managing the knowledge [4]. In these recent years, agile software companies are giving much importance to knowledge sharing strategies since information sources are constantly growing at exponential rates and this

M. A. Korimbocus · T. Towokul · S. D. Nagowah (✉)
Software and Information Systems Department, University of Mauritius, Reduit, Mauritius
e-mail: s.ghurbhurrun@uom.ac.mu

M. A. Korimbocus
e-mail: mohammad.korimbocus1@umail.uom.ac.mu

T. Towokul
e-mail: thaveersingh.towokul@umail.uom.ac.mu

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_47

has proved to increase the efficiency of operational units as well as the business process [5]. Knowledge sharing, being one of the main areas of knowledge management, helps with capturing, organizing, storing, applying and sharing knowledge to individuals across agile teams faster and more accurately [6]. Knowledge can be categorized into three forms, namely explicit knowledge, tacit knowledge and implicit knowledge [7]. Organizations tend to focus more on tacit knowledge which will bring innovation and competitive advantage to the company based on its richness of experience, competence, and commitments [8]. Therefore, the main goal of knowledge sharing is to extract knowledge from different levels of organizational knowledge resources, store the knowledge acquired and diffuse them to the different individuals and groups across the organization according to the business' requirements [9].

The main challenges that organizations face while capturing and sharing knowledge are that the process is costly and employees may hardly adapt themselves to the system [10]. Additionally, the knowledge acquired may not be universally recognized since the knowledge created and shared by one company may differ from others. Moreover, some employees prefer to keep the knowledge they have acquired for themselves so that their effort be rewarded afterwards [11]. Disgruntled employees leaking knowledge outside the company or leaving the job will eventually cause knowledge loss to the company [12].

The rest of the paper is structured as follows: knowledge capture and sharing in agile software companies are discussed in Sect. 2. Section 3 describes the aim of the research and methodology. Findings of the survey in software companies in Mauritius are discussed in Sect. 4. Recommendations and future works are discussed in Sect. 5. Section 6 finally presents the conclusion.

2 Knowledge Capture and Sharing in Agile Software Companies

Agile methodology is defined as an adaptive software development method that is widely considered among practitioners for distributed development [13]. It is based on incremental and iterative processes that involve consistent communication, planning, testing and deployment thus encouraging collaboration among stakeholders and fast decision making [14]. The benefit of using such methodology is that "*it promotes self-documenting designs and self-describing code that conforms to standards and guidelines*" [14]. Nowadays, most agile software companies work globally rather than toiling only among co-operated teams. Practitioners work in a socially interactive environment where they share explicit and tacit knowledge using different tools and techniques resulting in optimizing internal communication. Tacit knowledge is captured in the form of experience and understanding of individuals and reused afterwards for better decision making thus improving business processes.

Since knowledge is dynamic, different areas can be considered when applying knowledge sharing such as capturing and storing domain knowledge, continuous learning from different sources and team composition [15].

Knowledge residing in businesses is often ignored which eventually leads to poor business practices [12]. Moreover, communication problems like language barrier, cultural issues and time difference due to different geographical regions are often the factors which contribute to knowledge loss. Furthermore, it is difficult to extract and interpret tacit knowledge which resides in human mind if the methods for knowledge elicitation are poorly defined [16]. Another issue is that agile methodology relies mainly on iterations where team members find it difficult to demonstrate their piece of work due to tight scheduling [17]. These challenges can be surpassed by applying appropriate knowledge strategies. Coordination, communication, and collaboration play a fundamental role in sharing information and knowledge effectively where these factors complement each other in one way or the other [15].

Whenever a project has been completed, it is important to capture and store the knowledge in an appropriate format so that it can be reused afterwards. This will bring innovation, save time and improve the effectiveness of the organization. The tacit knowledge can be captured using different tools and then distributed among the team members [17]. Agile-based companies mainly share knowledge across organizations using iterations where the different phases are constantly reviewed, and feedbacks are provided for improvement and ensure that the project is being executed according to the customers' needs [18]. This will, therefore, contribute to improving the informal way of communicating between co-located and distributed teams. Electronic document system can also be used to classify and distribute relevant information where knowledge can be codified, stored and reused afterwards [19]. Other tools such as Kanban board, task boards, innovation board can even be used as visual aid to conduct virtual meetings depending on the conditions applied [20]. Some companies have their own discussion forum which allows employees to share their issues, opinions, ask questions and discuss about the best ways to tackle problems [21]. Other companies use an issue tracking software where clients address their issues simply by sending a ticket to his supplier and wait for the issue to be solved within a timeframe [22]. Knowledge can be shared across various levels namely: individual knowledge sharing where knowledge is shared from one individual to another mainly through face-to-face communication, knowledge sharing within specific groups for example, a particular functionality of the software through meetings, presentations or innovation boards and lastly organizational knowledge sharing where group knowledge combined from various subunits is processed for new knowledge creation to improve the efficiency of business processes [23].

3 Research Aims and Methodology

The main objective of this study is to determine the strategies, techniques, and tools adopted by organizations and challenges faced by them when it comes to knowledge capture and sharing practices in agile-based software companies. Previous work with respect to knowledge sharing in agile distributed teams was carried out in Mauritian companies [16]. However, the research was limited to only three companies. For this study, surveys and interviews have been conducted in ten software agile companies in Mauritius based on their existing systems, approach used and methods of knowledge sharing. On top of that team leaders, senior, and junior developers from three agile companies were interviewed. Other factors such as importance of knowledge sharing, the accuracy of information shared, amount of effort, and time required to retrieve needful information are also considered. The major challenges that team members encountered while capturing and distributing knowledge are also discussed. The research is limited to only ten companies due to lack of funding and time. Some companies could not answer specific questions because of the confidentiality of information. Nevertheless, the survey was made anonymous so that no personal information about the respondents and the organization is revealed. Results have been presented graphically to get an overview of all the responses in the next section.

4 Findings

The main findings of the survey are summarized as follows.

4.1 Agile Methodologies Adopted by the Companies

Most of the companies have shifted towards the agile methodology as this approach tends to be more practical due to continuous changes happening during the implementation of the project. From all the companies that were interviewed, and survey conducted, it can be deduced that most companies use an agile approach where 78% are adopting a scrum practice followed by extreme programming (20%), dynamic system development method (13%), 10% opting for adaptive software development, and 10% for other practices. The details are illustrated in Fig. 1.

The main reason for scrum methodology adoption is that it facilitates project management and efficient software development by providing a panoply of benefits. The benefits include daily meetings to promote efficient work track and teamwork. Team members thus get an overview of what the whole team is developing. During the meeting, they often discuss the struggles being experienced and find solutions. Moreover, in scrum methodology, tasks with respect to a project are broken down into smaller ones and are easily estimated. Furthermore, each team member is aware of task allocated to other members. Hence, they can together resolve time and budget constraints, if any.

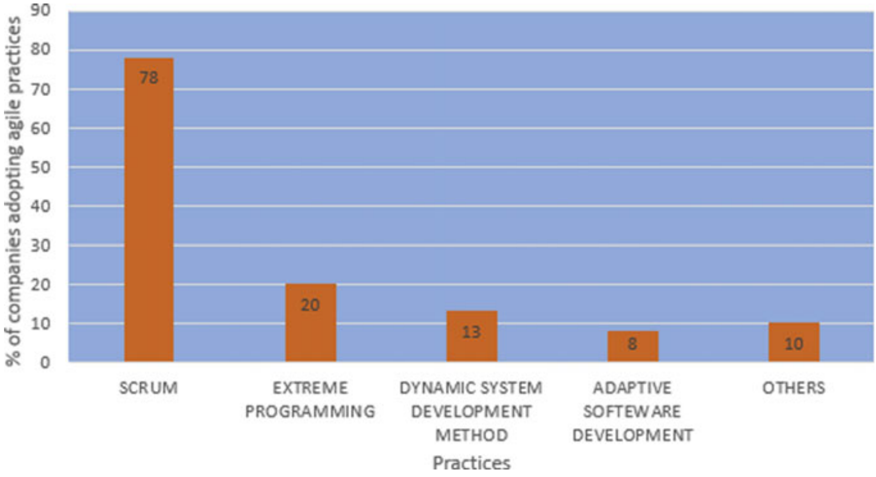


Fig. 1 Agile methodology adoption

4.2 Methods of Sharing Knowledge in Agile Teams

Numerous methods are used in different organizations for knowledge distribution among individuals. It can be deduced that most of the information is shared during meetings (80%) followed by informal conversations (50%), 40% for both formal presentations and emails and 30% through leads and other methods such as software (SharePoint, TFS, slack) as illustrated in Fig. 2.

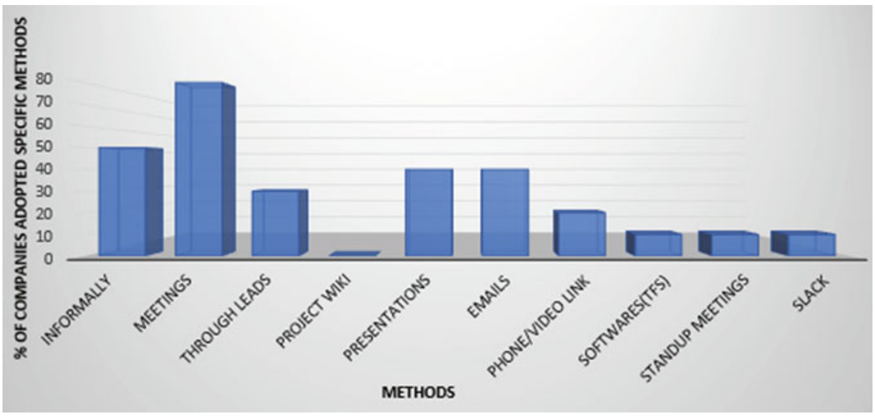


Fig. 2 Knowledge sharing techniques

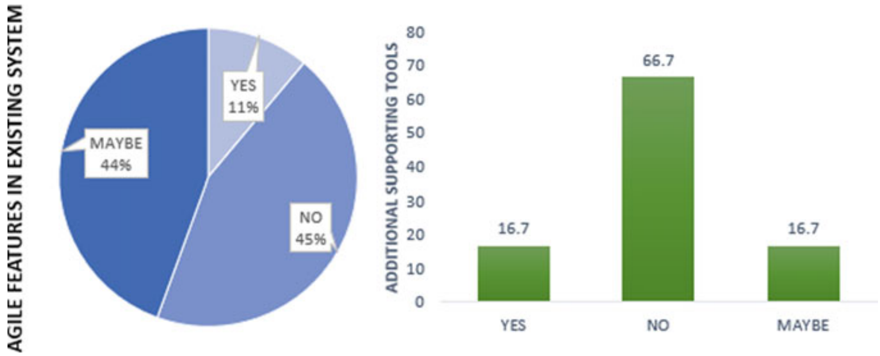


Fig. 3 Agile features in existing system and additional supporting agile tools

4.3 Agile Features Supported by Existing Systems

It is of utmost importance that companies have the appropriate software with proper features to provide better services to customers. As shown in Fig. 3, 11% of the respondents have confirmed that their systems already possess all the necessary agile features. Out of the 89% who presume that their systems lack relevant features, 16.7% stated that they use alternative tools to support the features missing in the system.

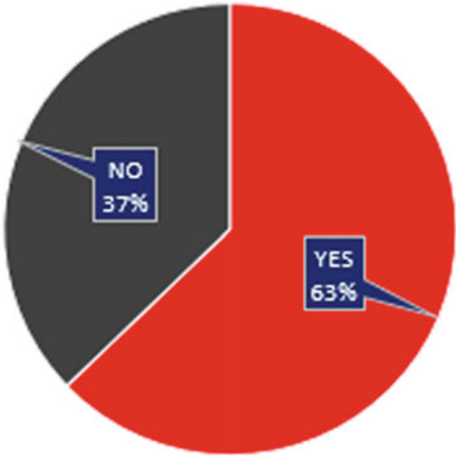
4.4 Capturing and Sharing of Information During Standup Meetings

During a standup meeting, the meeting leader must ensure that the committee has well understood the message delivered. Additionally, while convening the meeting the information must be stored in case some of the members could not attend due to some inconvenience. 63% of the respondents are in the opinion that their organizations capture the information discussed during the meetings to be reused afterwards while 37% believe the contrary as shown in Fig. 4.

4.5 Clarity of the Information Shared by the Team Members on the Existing System

Clarity of information being shared by team members is fundamental in organizations as it helps to avoid misunderstanding of information while saving time and maximizing the efficiency of business processes. 50% of respondents argue that about half of the information shared within the organization is clear to understand while 23%

Fig. 4 Information captured during standup meetings



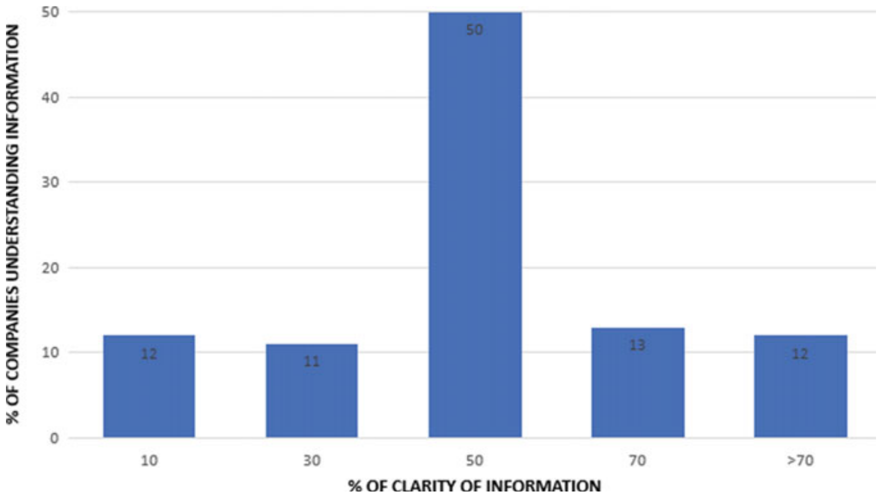


Fig. 5 Clarity of the information shared

of the companies state that less than 30% of information shared is understandable and about 25% of companies understand 70% or more the knowledge shared. The details are illustrated in Fig. 5.

4.6 *Effort and Time Required for Information Retrieval*

A lot of time and effort are spent in an organization searching for the right information. This often brings delays in projects which eventually increase the development cost. Figures 6 and 7 give an idea of the effort required and the number of hours spent on

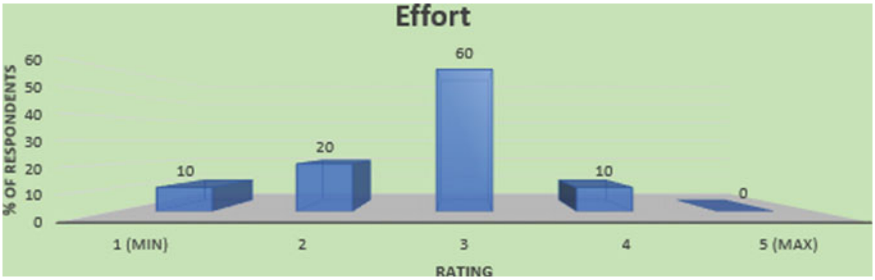


Fig. 6 Effort required for information required

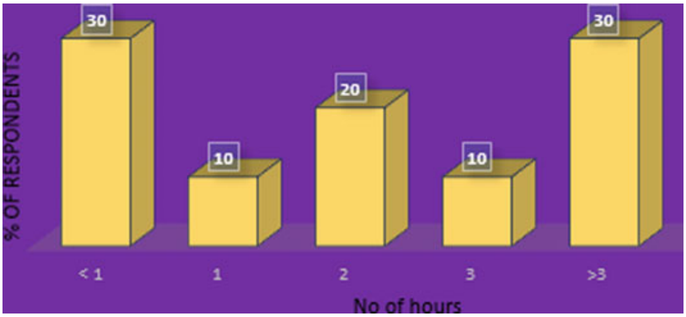


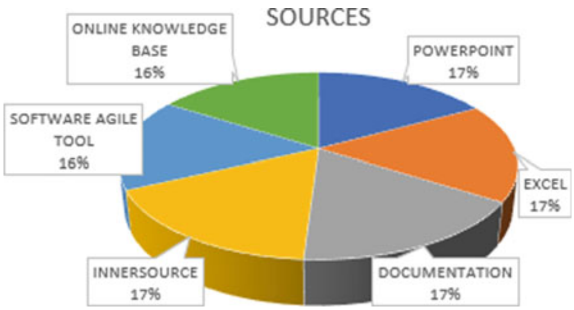
Fig. 7 Time required for information required

average for information retrieval, respectively. On a scale of 1–5, it can be deduced that 60% of respondents spend an average amount of effort searching for information while 30% easily find their information with a minimum amount of effort but still 10% of cases, people spend a great amount of effort to retrieve information. Moreover, 30% of respondents spend more than 3 h searching for information while 20% spend 2 h and 40% search their information within 1 h.

4.7 *Importance of Knowledge Sharing Across an Organization*

All the members of the companies have agreed upon the importance of knowledge distribution among team members. Some of the main reasons are resource sharing, trust, and mutual understanding among team members, better coordination and communication of the flow of resources, avoiding repetitive mistakes and getting everybody on the same level whether technical or non-technical knowledge. Others presume that knowledge sharing is fundamental to prevent loss of knowledge for example if someone leaves the organisation, the work is not halted or if one employee

Fig. 8 Data sources



absent other one can take up the responsibility of the task. Organizations can limit the impact of employees leaving and foster dynamic knowledge sharing thus improving productivity.

4.8 Storage of Lessons Learnt from Previous Projects in Organizations

Team members learn from their mistakes and improve their working conditions. Therefore, lessons learnt should be saved for future use. Some of the storage methods are shown in Fig. 8. Among the different sources, 17% of knowledge stored comes from documentation, another 17% from excel, 16% from online knowledge base, 17% from inner source, 16% from software agile tool and the rest from PowerPoint.

4.9 Challenges of Knowledge Sharing in Agile Distributed Projects

Some of the main challenges that individuals face while sharing knowledge in a project are that information is often in much detail during standup meetings. Sometimes, new recruits in organizations feel too nervous to voice out their opinions and share their ideas in meetings which will lead to loss of valuable skill set and knowledge. Language and time difference could be a major issue especially for live meetings with disperse team in different countries. Moreover, organizations might not have enough documentation to support knowledge shared by different projects and require an active risk management plan. Sometimes it is very difficult to pull everyone to the same level where people having different levels of experience or background might learn things differently. It happens that practitioners prefer to keep their knowledge to make them unique and that their potential is recognized.

5 Recommendations and Future Works

It has been observed that meeting is the main method for sharing knowledge across agile software companies where source codes and project reviews are mostly discussed followed by informal conversations and presentations. One issue that has been highlighted during the study is that there is no proper way of capturing and storing information during standup meetings and a lot of time is wasted retrieving necessary information. Moreover, companies argue that the clarity of information obtained must be enhanced and a minimum amount of effort is required to obtain the corresponding information.

Based on the survey conducted, a web-based framework has been proposed as shown in Fig. 9 to cater for the above challenges. The system shall consist of a backlog system which will contain a scrum board to display the progress of the project. A meeting support tool will be integrated within the system to facilitate knowledge capture and sharing among the practitioners where a recorder will be used to record the whole session. The recording will eventually be saved in the repository and be used as reference afterwards. All the knowledge and information captured will be then stored in a knowledge repository for easy retrieval and dissemination afterwards. The system shall provide a user-friendly web-based interface to allow users to easily interact with the system and have access to all the necessary knowledge in the minimum time delay while putting the minimum amount of effort. The proposed framework is under development. Feedback from industry will be used to improve the framework so that the latter matches the needs of the originations.

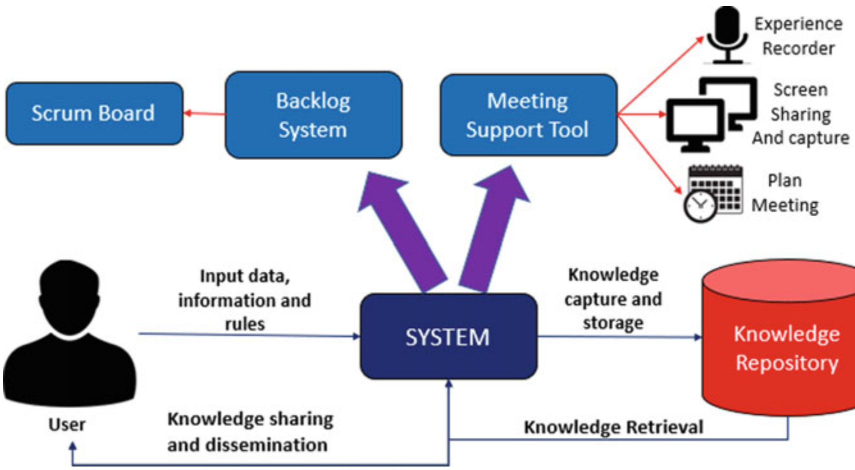


Fig. 9 Proposed framework

6 Conclusion

Knowledge capture and sharing promote the interpretation and distribution of knowledge based on experience, values, and insights. Different methods and frameworks have been used by organizations to create, capture and share knowledge. A survey has been carried out in ten agile-based software companies in Mauritius to investigate different techniques adopted by the development teams for capturing and sharing the knowledge acquired. Team leaders, senior and junior developers from the companies have also been interviewed. Based on the findings, it can be deduced that scrum is the most common practice used by these organizations and knowledge is mostly shared through meetings. A number of issues with respect to knowledge sharing and capture have been discussed in the paper. The paper proposes a framework to allow knowledge capture and sharing among distributed teams. The framework is under development.

References

1. Liyanage, C., Elhag, T., Ballal, T., Li, Q.: Knowledge communication and translation—a knowledge transfer model. J. Knowl. manage. **13**(3), 118–131 (2009)
2. Ipe, M.: Knowledge sharing in organizations: a conceptual framework. Hum. Resour. Dev. Rev. **2**(4), 337–359 (2003)
3. Leech, C.: Information-based decision making (2006)
4. Donoghue, L.P., Harris, J.G., Weitzman, B.A.: Knowledge management strategies that create value. Outlook **1**(1), 48–53 (1999)
5. Ouriques, R.A.B., Wnuk, K., Gorschek, T., Svensson, R.B.: Knowledge management strategies and processes in agile software development: a systematic literature review. Preprint at arXiv: 1807.04962 (2018)
6. Davenport, T.H., Prusak, L.: Working Knowledge: How Organizations Manage what they Know. Harvard Business Press (1998)
7. Gemma, W.: Udemy. (Online), Available https://blog.udemy.com/types-of-knowledge/
8. Tamer Cavusgil, S., Calantone, R.J., Zhao, Y.: Tacit knowledge transfer and firm innovation capability. J. Bus. Ind. Mark. **18**(1), 6–21 (2018)
9. Hajric, E.: Knowledge Management Tools. Online, Available http://www.knowledge-management-tools.net/. (2010)
10. Zammit, J.P., Gao, J., Evans, R., A framework to capture and share knowledge using storytelling and video sharing in global product development. In: IFIP International Conference on Product Lifecycle Management, pp. 259–268 (2015)
11. Nerur, S., Mahapatra, R., Mangalaraj, G.: Challenges of migrating to agile methodologies. Commun. ACM **48**(5), 72–78 (2005)
12. Desouza, K.C.: Managing Knowledge Security: Strategies for Protecting your Company's Intellectual Assets. Kogan Page Publishers (2007)
13. Stoica, M., Mircea, M., Ghilic-Micu, B.: Software development: agile vs. traditional. Inf. Econ. **17**(4) (2013)
14. Chau, T., Maurer, F., Melnik, G.: Knowledge sharing: agile methods vs. tayloristic methods. In: WET ICE 2003 Proceedings. Twelfth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, pp. 302–307. IEEE (2003)
15. Ahmed, M.A.R.: Knowledge Management in Distributed Agile Projects (2013)

16. Gervigny, M.L.I., Nagowah, S.D.: Knowledge sharing for agile distributed teams: a case study of mauritius. In: International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), pp. 413–419. IEEE (2017)
17. Kamara, J.M., Anumba, C.J., Carrillo, P.M., Bouchlaghem, N.: Conceptual framework for live capture and reuse of project knowledge. CIB Rep. **284**, 178 (2003)
18. Abrahamsson, P., Salo, O., Ronkainen, J., Warsta, J.: Agile software development methods: review and analysis. Preprint at arXiv:1709.08439 (2017)
19. King, W.R.: Knowledge management and organizational learning. In: Knowledge Management and Organizational Learning, pp. 3–13. Springer, Boston, MA (2009)
20. Fast-Berglund, Å., Harlin, U., Åkerman, M.: Digitalisation of meetings-from white-boards to smart-boards. Procedia CIRP. **41**, 1125–1130 (2016)
21. Wagner, C., Bolloju, N.: Supporting knowledge management in organizations with conversational technologies: discussion forums, weblogs, and wikis. J. Database Manage. **16**(2) (2005)
22. Bertram, D., Voida, A., Greenberg, S., Walker, R.: Communication, Collaboration, and Bugs: The Social Nature of issue Tracking in Software Engineering. University of Calgary (2009)
23. Bryant, S.E.: The role of transformational and transactional leadership in creating, sharing and exploiting organizational knowledge. J. Leadersh. Organ. Stud. **9**(4), 32–44 (2003)

hBOSOS: An Ensemble of Butterfly Optimization Algorithm and Symbiosis Organisms Search for Global Optimization

Sushmita Sharma, Apu Kumar Saha, V Ramasamy, Joy Lal Sarkar and Chhabi Rani Panigrahi

Abstract The present study proposes a hybrid metaheuristic algorithm, viz. hBOSOS with the help of two popular metaheuristics, viz. BOA and SOS. The global search ability of BOA and the local search ability of SOS are combined here to make the proposed algorithm robust and efficient. The proposed algorithm is tested with 26 classical benchmark functions and also to compare, several metaheuristics are taken from the literature. From the comparison, it can be observed that the hBOSOS is superior to the other compared algorithms.

Keywords Metaheuristics · Hybrid algorithms · BOA · SOS

1 Introduction

In science, especially in mathematics and engineering, the optimization problems are very common and vital as well. These optimization problems may be solved in two ways: exact method and heuristic method. Though the exact method can solve the problem giving an exact optimal solution, often it takes a huge amount of time which is not permitted. On the other hand, the heuristic algorithms are easy to implement over a diverse range of problem though often they do not give the exact optimal solution, rather they provide near optimal solutions. Metaheuristic is such heuristic method with high-level searching technique. Due to the simplicity of these methods, heuristic/metaheuristic methods are gaining popularity day by day.

S. Sharma · A. K. Saha (✉)
Department of Mathematics, National Institute of Technology Agartala, Agartala, India

V. Ramasamy
Department of Computer Science and Engineering, Park College of Engineering and Technology, Coimbatore, Tamil Nadu, India

J. L. Sarkar
Department of Computer Science and Engineering, Tripura University, Agartala, Tripura, India

C. R. Panigrahi
Department of Computer Science, Rama Devi Women's University, Bhubaneswar, India

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_48

2 Related Work

The present day metaheuristics are generally developed depending on some particular metaphors. Some of the metaheuristics are genetic algorithm (GA) [1], particle swarm optimization (PSO) [2], differential evolution (DE) [3], teaching–learning-based optimization (TLBO) [4] and artificial bee colony (ABC) [5].

It has already been proved that, in many a case, a combination of two algorithms (may be two metaheuristics or parts of two metaheuristics or one metaheuristic combined with another heuristic or any other method) gives better results in solving optimization problem than the component algorithms. So, for the last one decade, hybrid metaheuristic algorithms have been gaining their popularity in solving real-life optimization problems. Few of the recent hybrid metaheuristic algorithms may be cited as [6–10], etc. Very recently, imitating the behaviour of butterflies, Arora and Singh [11] have introduced butterfly optimization algorithm (BOA). To prove its efficiency, BOA has been applied to solve some real-life optimization problem. In 2014, Cheng and Prayogo [12] introduced another efficient swarm intelligence-based metaheuristic known as symbiosis organisms search (SOS). In the present study, a hybrid metaheuristic, viz. h-BOSOS, combining BOA and SOS, is developed and also this proposed method is used to solve unconstrained optimization problem. It has been seen that exploration ability of BOA is strong [13] and SOS is known for its better exploitation capabilities [14].

3 Butterfly Optimization Algorithm

In 2018, Arora and Singh [11] introduced butterfly optimization algorithm (BOA) mimicking the food foraging and mating pair searching strategy of butterflies. The described form of BOA may get from [11]. In BOA, the most important part is the procedure to control fragrance in the algorithm. It is seen that the fragrance capturing phenomenon depends on the following three special parameters, viz. (i) stimulus intensity (I), (ii) sensory modality (c) and (iii) power exponent (a). The fragrance is formulated in the following way:

$$f_i = c * I^a \tag{1}$$

where f_i is the apparent value of the fragrance of the ith butterfly.

BOA consists of two phases, viz. global and local. A butterfly moves towards the direction of the best performer among the butterflies in a particular iteration in global search phase. Mathematically, it is represented as

$$x_i^{t+1} = x_i^t + (r^2 * g^* - x_i^t) * f_i \tag{2}$$

The local phase is formulated by the following equation:

$$x_i^{t+1} = x_i^t + (r^2 * x_j^t - x_k^t) * f_i \tag{3}$$

where x_i^t is the ith butterfly in tth iteration and x_j^t and x_k^t are two randomly selected butterflies from the candidate solution and $g*$ is the best candidate of the ith in a particular stage. Also, r is a randomly taken number in the range of [0, 1] and f_i is the magnitude of fragrance for ith butterfly.

4 Symbiosis Organisms Search

In 2014, Cheng and Prayogo [12] proposed SOS mimicking the interactive relationship among the organisms in the ecosystem. The SOS is executed in four steps, viz. (i) initialization, (ii) mutualism phase, (iii) commensalism phase and (iv) parasitism phase. The details of SOS are summarized below.

4.1 Initialization

Initially, each organism or candidate solution is generated randomly within the search space by Eq. 4:

$$Org(i, m) = l_m + rand(0, 1) * (u_m - l_m) \tag{4}$$

where l_m and u_m are the lower and upper bounds of the ith organism, respectively.

4.2 Mutualism Phase

Two organisms of the ecosystem interact mutually and both get benefit from the interaction in mutualism phase. Mathematically, the updated Org_i and Org_j are determined by Eqs. 5 and 6 and updated in the ecosystem. The new organism will replace the old one if the new organism is better than its previous fitness value.

$$Org_i^{new} = Org_i + rand(0, 1) * (Org_{best} - Mutual_{Vector} * BF1) \tag{5}$$

$$Org_j^{new} = Org_j + rand(0, 1) * (Org_{best} - Mutual_{Vector} * BF2) \tag{6}$$

where

$$Mutual_{Vector} = (Org_i + Org_j)/2 \tag{7}$$

Here, Org_{best} represents the best organism or solution in the ecosystem, the value of benefit factors (BF1 and BF2) are considered as 1 or 2 randomly. $Mutual_{Vector}$ represents the characteristic of the relationship of the organism Org_i and Org_j.

4.3 Commensalism Phase

An organism, Org_i, interacts with another organism Org_j where Org_i is benefitted from the interaction but Org_j remains same. Mathematically, the updated fitness value of Org_i is determined using Eq. (8).

$$Org_i^{new} = Org_i + rand(-1, 1) * (Org_{best} - Org_j) \quad (8)$$

where Org_{best} is the best organism according to the fitness value in the ecosystem.

4.4 Parasitism Phase

On parasitism relationship of two organisms in the ecosystem, sometimes, one organism gets benefitted but the other organism gets harmed. In SOS, the parasite vector is obtained by following way:

Select one dimension (say d) from the dimensions of organism Org_i, i.e. from $1, 2, 3, \ldots, D$.

Randomly modify the selected dimension using Eq. (6) and replace the position of the dimension into Org_i.

$$Org_{i,d} = l_d + rand(0, 1) * (u_d - l_d) \quad (9)$$

If the fitness value of the $Parasite_{Vector}$ is superior to fitness value of Org_j, then $Parasite_{Vector}$ will replace Org_j in the ecosystem. If the reverse happens, Org_j is sustained in the ecosystem and the $Parasite_{Vector}$ leaves the ecosystem.

5 Proposed Methodology

In this section, a hybrid metaheuristic, viz. hBOSOS, using the two popular algorithms BOA and SOS is introduced in details. One robust algorithm is developed when the balancing of diversification (exploration or global search) and intensification (exploitation or local search) is assured. When exploitation is better than exploration, diversification becomes poor and causes premature convergence. On the other hand, when exploration is better than exploitation, then it causes poor intensification and thus algorithm's convergence rate becomes slow.

As per literature, it has been observed that selection of exploration and exploitation phases randomly based on the considered value of switching probability sometimes becomes reason to lose path and go far away from the optimal solution [13] of the BOA. It has been observed that the value of switch probability factor 'p' ($p = 0.8$) becomes the reason to make the algorithm better in exploration but not so good in exploitation.

On the other hand, it has been observed that mutualism and commensalism phases are responsible for improvement of the exploitation ability of SOS, and the parasitism phase is responsible for enhancing the exploration capability [14]. So, the algorithm becomes stronger in exploitation. Therefore, we have proposed a hybrid method that provides better solution quality as well as convergence rate by balancing both exploration and exploitation. In the hBOSOS, initially BOA is executed and then all the phases of SOS are executed.

6 Experimental Setup

6.1 Classical Benchmark Functions

To check the efficiency of the proposed algorithm, the Hybrid hBOSOS algorithm is tested on 26 different classical benchmark functions obtained from literature, viz. Sphere (1), Cigar (2), Step (3), Quartic Function with noise (4), Bohachevsky (5), Ackley (6), Griewank (7), Levy (8), Rastringin (9), Alpine (10), Schaffer (11), Rosenbrock (12), Schwefel 1.2 (13), Schwefel 2.21 (14), Schwefel 2.22 (15), Schwefel 2.26 (16), Goldstein and Price (17), Matyas (18), Powell (19), Sum Square (20), Zettl (21), Foxholes (22), Branin (23), Kowalik (24), Penalized 1 (25) and penalized 2 (26).

Here, the algorithm has been implemented in MATLAB R2015a and number of iterations is 10,000 as stopping criteria and each function runs for 30 times repeatedly. The mean and std are obtained for each function. The combination of parameters is used as mentioned in the original BOA [13] and SOS [12] paper.

7 Results and Discussion

7.1 Comparison of hBOSOS with BOA, SOS, DE, PSO, JAYA

To examine the efficiency of hBOSOS over the original BOA, SOS, DE [3], PSO [2] and JAYA [15], twenty-six classical benchmark functions (taken from literature) are used and the results obtained from simulation which contain the mean and the standard deviations of function are given in Table 3. For comparison, we have fixed the number of iterations as 10,000 and each algorithm has been executed for 30 times repeatedly. From Table 3, the performance analysis oh hBOSOS over BOA, SOS, DE, PSO and JAYA is given in Table 1 and from this table, it is seen that the hBOSOS is efficient than the other compared five algorithms.

Table 1 Analysis of performance of hBOSOS over BOA, SOS, DE, PSO and JAYA

Algorithms	Number of cases		
	Superior	Similar	Inferior
BOA	11	13	2
SOS	11	8	7
DE	14	5	7
PSO	20	3	3
JAYA	17	3	6

7.2 Comparison with Other State-of-the-Art Metaheuristic Algorithms

To examine the robustness of hBOSOS, it is again compared to five other state-of-the-art algorithms, viz. ABC [5], CS [16], FA [17], GA [1] and MBO [18] over twenty-one well-known benchmark functions. The results obtained from simulation which contain the mean and the standard deviations of function are given in Table 6. For comparison, we have fixed the number of iterations as 10,000 and all the algorithms have been executed for 30 times repeatedly. From Table 4, the performance analysis oh hBOSOS over ABC, CS, FA, GA and MBO is given in Table 2 and from Table 2, it is seen that the hBOSOS is also efficient than the other compared five algorithms.

7.3 Statistical Rank Test

Also to evaluate the simulation results shown in Tables 2 and 3 we have introduce the Friedman rank test on these two tables. The ranks are given in Tables 5 and 6.

From Tables 5 and 6, it can be observed that in both the cases, rank of proposed hybrid hBOSOS is minimum. As the benchmark functions are minimization problem, we can rank the algorithms as 1 for hBOSOS, 2 for SOS, 3 for BOA, 4 for DE, 5 for

Table 2 Analysis of performance of hBOSOS over ABC, CS, FA, GA and MBO

Algorithms	Number of cases		
	Superior	Similar	Inferior
ABC	16	2	3
CS	19	1	1
FA	13	4	4
GA	15	2	4
MBO	15	3	3

Table 3 Performance results of hBOSOS, BOA, SOS, DE, PSO, and JAYA

$f(x)$		hBOSOS	BOA	SOS	DE	PSO	JAYA
1	Mean	0	0	6.03E−173	2.6178E−165	1.76657E−51	2.6814e−39
	STD	0	0	0	0	1.1246E−50	3.4015e−39
2	Mean	0	0	0	1.4054E−172	3.10418E−53	1.9742e−26
	STD	0	0	0	0	2.10651E−52	1.6028e−26
3	Mean	0	0	0	1.3000E−01	1.31167E−31	0
	STD	0	0	0	4.8524E−01	6.653E−31	0
4	Mean	1.3308e+00	3.8917E−05	1.74E+00	2.6159E−03	3.7408	2.1049e+00
	STD	1.4810e−01	2.9019E−05	2.05E−01	1.0521E−03	1.0296	2.5492e−01
5	Mean	0	0	0	0	3.0393E−02	0
	STD	0	0	0	0	2.6192E−17	0
6	Mean	8.8818e−16	1.7183E+00	1.00E−15	1.7183E+00	1.8186E+01	4.4409e−15
	STD	0	0	6.38E−16	0	5.4812E+00	0
7	Mean	0	1.8472E−19	0	9.22E−17	4.0503E+01	4.4665e−01
	STD	0	2.6886E−20	0	2.0465E−17	4.1334E+01	7.8317e−02
8	Mean	3.7204e−04	4.4108E−01	6.73E−23	5.9824E−02	4.9297E+00	1.3498e−31
	STD	6.0855e−04	5.7467E−02	1.95E−22	2.0758E−01	5.9951E+00	6.6809e−47
9	Mean	0	0	0	1.2475E+01	1.4453E+02	2.8917e+0
	STD	0	0	0	3.3093E+00	3.7957E+01	3.3052e+00
10	Mean	0	0	0	6.90E−16	4.4402E−02	4.5172e−08
	STD	0	0	0	2.57579E−15	4.5318E−01	1.3268e−07
11	Mean	0	0	4.85E−16	0	0	5.7301e−11
	STD	0	0	2.74E−16	0	0	3.8659e−11
12	Mean	8.4837e+00	2.8837E+01	1.82E−01	2.3552E+01	2.5746E+02	5.1681e−01
	STD	5.5754e−01	3.1281E−02	1.20E−01	4.9866E+00	3.7311E+02	9.0096e−01
13	Mean	0	0	0	1.1423E−175	1.2030E+03	1.4856e−38
	STD	0	0	0	0	5.4485E+02	1.6823e−38
14	Mean	0	6.996E−153	2.27E−74	1.0201E+00	1.0586E−01	2.6664e−09
	STD	0	1.4788E−152	2.41E−74	4.5979E−01	6.5371E−02	2.4656e−09
15	Mean	0	0	1.24E−86	8.15E−91	7.28E+01	2.6828e−21
	STD	0	0	1.26E−86	1.2303E−90	2.2745E+01	2.0084e−21
16	Mean	−6.5185e+03	−2.266E+03	−4.19E+02	−4.565E+12	−7.570E+03	−5.1098e+03
	STD	8.0605e+02	4.5626E+02	0.00E+00	3.1934E+12	8.8893E+02	3.4792e+02
17	Mean	3	3	3	3	3	3
	STD	6.4311e−06	0	1.94E−15	0	0	8.4605e−04
18	Mean	0	0	2.62E−43	0	0	3.6996e−21
	STD	0	0	5.87E−43	0	0	1.4214e−20
19	Mean	0	0	7.51E−13	1.3099E+02	3.3671E+03	1.0993e−08
	STD	0	0	3.99E−12	1.8525E+02	9.5201E+02	1.0770e−08
20	Mean	0	0	1.09E−117	1.4595E−19	9.2E+02	2.4299e−81
	STD	0	0	3.94E−117	1.1767E−19	8.2704E+02	1.2564e−80
21	Mean	−3.7850e−03	−3.7912E−03	−3.79E−03	−3.7912E−03	−3.7912E−03	−3.7912e−03
	STD	1.2863e−05	0	1.76E−18	4.5714E−19	0	8.0472e−09
22	Mean	9.9840e−01	2.84	9.98E−01	9.9800e−01	1.1955	9.9856e−01
	STD	1.1343e−03	2.55	2.92E−16	0	9.1220e−01	2.8656e−03
23	Mean	3.9790e−01	4.98E−01	3.98E−01	3.9789e−01	3.9789e−01	3.9923e−01
	STD	1.5828e−05	1.47E−01	3.17E−07	0	5.9502e−08	1.6041e−03
24	Mean	3.7293e−04	6.66E−04	3.27E−04	3.0749e−04	4.0453e−04	3.7191e−04
	STD	9.9770e−05	7.52E−04	7.60E−05	1.2480e−08	2.4399e−04	1.3829e−04
25	Mean	1.8330e−01	2.27E−01	4.71E−32	4.7116e−32	3.7035e−01	6.2618e−02
	STD	1.0374e−01	8.54E−02	2.23E−47	1.6702e−47	1.0073	1.4760e−02
26	Mean	7.1572e−01	3.00E+00	1.35E−32	1.3498e−32	2.7426e−04	1.3498e−32
	STD	8.0040e−02	0.00E+00	2.78E−48	5.5674e−48	2.2034e−04	5.5674e−48

Table 4 Performance results of hBOSOS, ABC, CS, FA, GA and MBO respectively

$f(x)$		hBOSOS	MBO	FA	ABC	GA	CS
1	Mean	0	2.21E−01	7.86E−04	2.58E−01	2.20E−02	6.48E−61
	STD	0	1.41E−01	1.36E−04	5.13E−01	9.67E−03	9.76E−61
2	Mean	0	3.40E−04	1.08E−05	8.43E−03	2.19E−04	5.53E−63
	STD	0	4.79E−04	1.45E−06	2.30E−02	1.09E−04	8.92E−63
3	Mean	0	3.17E+04	0	2.23E−01	2.30E−02	1.21
	STD	0	2.57E+04	0	4.82E−01	1.05E−02	1.48
4	Mean	7.95	3.48E−03	5.95E−02	5.14E−01	2.73E−03	1.24E+01
	STD	3.25e−01	6.22E−03	2.47E−02	1.31E−01	8.50E−04	1.41
5	Mean	0	0	0	0	0	4.92E−04
	STD	0	0	0	0	0	8.66E−04
6	Mean	8.88e−16	1.31E+01	7.73E−04	3.65E+00	4.66E−02	1.70E+01
	STD	0	9.41E+00	4.80E−05	6.93E−01	1.68E−02	5.46
7	Mean	0	4.30E+02	1.07E−07	5.48E−01	5.55E−02	20.06
	STD	0	1.65E+02	1.31E−08	2.72E−01	2.94E−02	4.30E−02
8	Mean	4.47e−04	3.15E+01	1.19	3.08E−02	1.37E−04	1.62E+01
	STD	6.97e−04	6.31E+01	1.67	4.82E−02	7.65E−05	11.94
9	Mean	0	6.37E+01	3.47E+01	2.36E+01	1.64E+01	1.08E+02
	STD	0	4.29E+01	9.54	5.48	4.109	4.99E+01
10	Mean	0	1.97E−01	1.19E−02	6.05	3.14E−17	1.05E+01
	STD	0	3.77E−01	6.42E−03	6.16E−01	3.07E−16	4.23
11	Mean	0	6.6812E−04	1.4304E−02	8.6222E−16	1.68965E−17	9.8155E−01
	STD	0	8.09E−04	5.35E−03	8.54E−16	1.68E−16	7.87E−01
12	Mean	2.88e+01	7.14	1.83E+01	1.25E+02	3.78E+01	3.95E+01
	STD	1.79e−01	1.42E+01	4.24	8.77E+01	2.52E+01	3.33E+01
13	Mean	0	1.11E+05	5.24E−02	2.36E−02	2.42E−03	4.05E−02
	STD	0	1.98E+04	6.70E−02	4.18E−02	1.07E−03	6.06E−02
14	Mean	0	3.540	1.43E−04	1.00E−01	4.66E−02	3.67E−03
	STD	0	3.99	1.48E−05	6.60E−01	1.08E−02	1.58E−−03
15	Mean	0	7.01E−03	1.50E−03	1.42E−01	5.43E−02	1.11E+01
	STD	0	2.96E−03	1.13E−04	6.83E−02	1.61E−02	2.49E+01
16	Mean	−6.44e+03	−3.94E+04	−7.145E+03	−1.05E+04	−1.18E+04	−1.08E+02
	STD	7.76e+02	5.70E+04	9.09E+02	2.93E+02	2.95E+02	1.07E+01
17	Mean	3	3	3	3	3	3
	STD	4.90e−06	2.96E−03	0	6.46E−12	0	0
18	Mean	0	0	0	1.19E−04	8.38E−104	1.33E−06
	STD	0	0	0	1.30E−04	2.51E−103	2.16E−06
19	Mean	0	3.43E+03	2.39E−04	1.51E+01	3.27E−02	6.63
	STD	0	2.63E+03	4.76E−05	1.10E+01	1.11E−02	6.99
20	Mean	0	7.69E+03	1.45E−06	1.21E−02	3.42E−03	2.09E+01
	STD	0	1.07E+03	3.20E−07	1.29E−02	2.62E−03	1.15E+01
21	Mean	−3.78e−03	2.01E−01	−3.79E−03	−3.79E−03	−3.79E−03	−3.79E−03
	STD	1.36e−05	2.23E−01	4.57E−19	5.71E−18	4.57E−19	4.45E−07

Table 5 Ranks based on Friedman rank test

Algorithm	Friedman rank test	
	Mean rank	Rank
hBOSOS	2.62	1
BOA	3.25	3
SOS	2.90	2
DE	3.33	4
PSO	4.85	6
JAYA	4.06	5

Table 6 Ranks based on Friedman rank test

Algorithm	Friedman rank test	
	Mean rank	Rank
hBOSOS	2.00	1
ABC	4.17	4
CS	4.79	6
FA	2.86	2
GA	2.83	3
MBO	4.36	5

JAYA and 6 for PSO from Table 5. Evaluating Table 4 we can rank 1 for hBOSOS, 2 for FA, 3 for GA, 4 for ABC, 5 for MBO and 6 for CS as per the results given in Table 6. Hence, the proposed hBOSOS is better than the compared algorithms.

8 Conclusion

This paper presents an enhanced ensemble hBOSOS with balanced exploration and exploitation ability to improve the quality and rate of convergence of the algorithm. To examine the efficiency of newly proposed hBOSOS, a number of popular benchmark functions are collected from the literature and conducted the numerical experiments. In this study, the proposed hBOSOS is compared with BOA, SOS and also eight other popular algorithms such as DE, PSO, JAYA, ABC, CS, FA, GA and MBO. From simulation results given in Tables 5 and 6, it can be said that hBOSOS is efficient than the compared algorithms.

References

1. Holland, J.H.: Genetic algorithms. Sci. Am. **267**, 66–72 (1992)
2. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of ICNN'95—International Conference on Neural Networks, vol. 4, pp. 1942–1948 (1995)
3. Storn, R., Price, K.: Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. **11**(4), 341–359 (1997)
4. Rao, R.V., Savsani, V.J., Vakharia, D.P.: Teaching–learning-based optimization: a novel method for constrained mechanical design optimization problems. Comput. Aided Des. **43**, 303–315 (2011)
5. Yi, Y., He, R.: A novel artificial bee colony algorithm. In: 2014 Sixth International Conference on Intelligent Human–Machine Systems and Cybernetics, vol. 1, pp. 271–274 (2014)
6. Chuanwen, J., Bompard, E.: A hybrid method of chaotic particle swarm optimization and linear interior for reactive power optimisation. Math. Comput. Simul. **68**, 57–65 (2005)
7. Mafarja, M.M., Mirjalili, S.: Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection. Soft Comput. https://doi.org/10.1007/s00500-018-3282-y
8. Zhou, Y., Su, K., Shao, L.: A new chaotic hybrid cognitive optimization algorithm. Cognit. Syst. Res. **52**. https://doi.org/10.1016/j.cogsys.2018.08.001
9. Nama, S., Saha, A.K.: A new hybrid differential evolution algorithm with self-adaptation for function optimization. Appl. Intell. **48**(7), 1657–1671 (2018)
10. Aydilek, A.: A hybrid firefly and particle swarm optimization algorithm for computationally expensive numerical problems. Appl. Soft Comput. **66**. https://doi.org/10.1016/j.asoc.2018.02.025
11. Arora, S., Singh, S.: Butterfly optimization algorithm: a novel approach for global optimization. Soft Comput. https://doi.org/10.1007/s00500-018-3102-4
12. Cheng, M.-Y., Prayogo, D.: Symbiotic organisms search: a new metaheuristic optimization algorithm. Comput. Struct. **139**. https://doi.org/10.1016/j.compstruc.2014.03.007
13. Arora, S., Singh, S., Yetilmezsoy, K.: A modified butterfly optimization algorithm for mechanical design optimization problems. J. Braz. Soc. Mech. Sci. Eng. **40**(1), 21 (2018)
14. Do, D.T.T., Lee, J.: A modified symbiotic organisms search (MSOS) algorithm for optimization of pin-jointed structures. Appl. Soft Comput. **61**, 683–699 (2017)
15. Pandey, H.M.: Jaya a novel optimization algorithm: what, how and why? In: 2016 6th International Conference—Cloud System and Big Data Engineering (Confluence), pp. 728–730 (2016)
16. Yang, X., Deb, S.: Cuckoo search via lévy flights. In: 2009 World Congress on Nature Biologically Inspired Computing (NaBIC), pp. 210–214 (2009)
17. Yang, X.-S.: Firefly algorithm, lévy flights and global optimization, pp. 209–218 (2010)
18. Wang, G.-G., Deb, S., Cui, Z.: Monarch butterfly optimization. Neural Comput. Appl. https://doi.org/10.1007/s00521-015-1923-y

PSO-Based Test Case Generation: A Fitness Function Based on Value Combined Branch Distance

Rashmi Rekha Sahoo and Mitrabinda Ray

Abstract All-path coverage of software under test is a NP-complete problem. At the time of automatic test case generation, there is some path whose probability of coverage is low. To overcome this problem, we propose a fitness function, value combined branch distance function (VCBDF) on path coverage criteria and is applied with particle swarm optimization (PSO) algorithm to automate test case generation. The objective is to achieve maximum path coverage with the challenge of covering a target path. We have conducted experiment on a bench mark case study, triangle classification problem. Two more existing fitness functions, branch distance (BD) based fitness function and combined fitness function (CFF) are also applied for comparison. Experimental result shows VCBD function gives better result in terms of number of test cases generated for target path and number of iterations as compared to the said functions.

Keywords Search-based testing · Path coverage · PSO · Fitness function · Value combined branch distance

1 Introduction

Manual testing process is error prone and also exhaust 40–70% time and cost of software development process [1]. Exhaustive testing is not feasible for software under test as the number of possible test cases may range to infinite. Search-based testing (SBT) is best fitted in such type of problem scenario. Various metaheuristic techniques are applied in search-based testing [2]. It is inspired by different observations that software testing can be formulated as optimization problem [2]. The searching process of metaheuristic algorithm is guided by fitness function. In SBT, the fitness function is formulated mainly basing upon the coverage criteria. Coverage

R. R. Sahoo (✉) · M. Ray
ITER College, Siksha O Anusandhan Deemed to be University, Bhubaneswar, India
e-mail: rashmirekhasahoo@soa.ac.in

M. Ray
e-mail: mitrabindaray@soa.ac.in

B. Pati et al. (eds.), *Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1089,
https://doi.org/10.1007/978-981-15-1483-8_49

criteria are a parameter for measuring how well the software under test is exercised by the generated test cases [3]. Some of the most used criteria are statement coverage, branch coverage, decision coverage, path coverage, etc. [4]. Our main objective is to cover the target path. In our approach, the target path contains the maximum number of conditional nodes. As it is the origin of almost all paths in the control flow graph (CFG), a single fault in any node of the target path has high influence on the failure of the entire software system. The consequence may be a more number of faults or failure of the SUT. We have chosen APSO as it well balanced the exploration and exploitation. So, premature convergence can be avoided. From the existing work [5], it is also observed that genetic algorithm (GA) is not able to cover such types of path using branch distance-based functions. Branch distance and simply combined branch distance, which is the combination of branch distance and approximation level heuristics, do not well distinguish the test cases for the target path as their magnitudes are very less. In this paper, test cases are generated for path testing using PSO search-based technique applied with branch distance fitness functions. The novelty of the search technique in this paper is the proposal of branch distance-based fitness function.

Organization of the paper is as follows: In Sect. 2, we provide the background study of coverage criteria, PSO algorithm and branch distance-based fitness functions. Section 3 presents the proposed fitness function. In Sect. 4, the experimental result and its analysis are discussed. The conclusion is presented in Sect. 5.

2 Background

2.1 Path Coverage

In path coverage testing, the test cases are generated to execute each path at least once. Thus, it can find all faults lying within the piece of code. It defines the number of linearly independent paths (basis paths) which can be determined by using cyclomatic complexity metric [4]. A set of paths for a given program is called the linearly independent set of paths, if each path in the set introduces at least one new edge that is not included in any other path in the set [4, 5]. Figure 1a shows the triangle classification program and Fig. 1b shows its corresponding control flow graph to generate basis paths.

Triangle classification program (TCP): This program takes three sides as input and displays whether this is a scalene, isosceles, equilateral or not a triangle.

As shown in Fig. 1b, the following four basis paths are obtained from the triangle classification CFG.

Path 1: 1–7–8 (Not a triangle)
Path 2: 1–2–3–8 (Scalene)
Path 3: 1–2–4–5–8 (Isosceles)
Path 4: 1–2–4–6–8 (Equilateral).

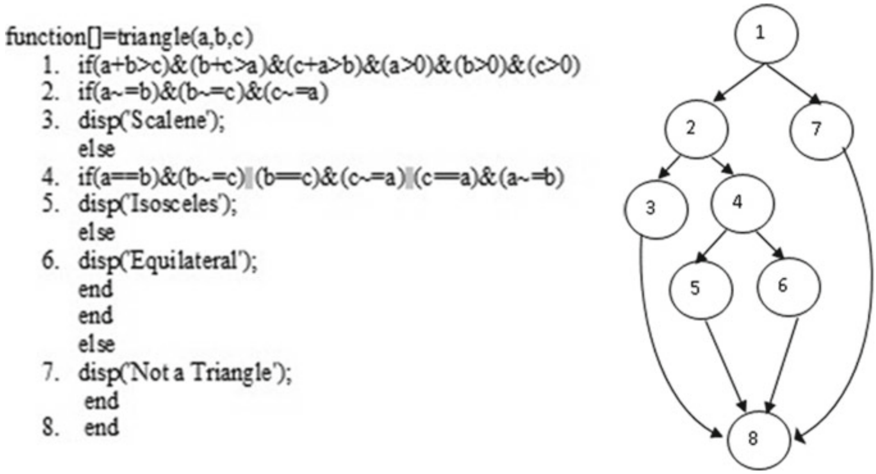


Fig. 1 **a** MATLAB code for triangle classification program. **b** Control flow graph of **a**

2.2 Fitness Function

Fitness function provides fitness value (numerical value) to each candidate solution [5]. In test case generation, test cases are candidate solution in the search space and they are evaluated by their fitness value to determine how good it is. In this paper, a target path is set to be covered. So a test case that covers a path closer to target path will have a better fitness value and considered as a good solution. To achieve such types of goal, mainly two heuristics are used for the fitness function. The first one is approximation level-based and another one is branch distance-based. For the fitness function, these two measures are linearly combined together [5, 6].

Branch Distance-Based Fitness Function (BDBFF)
Branch distance finds how close or far the generated (actual) test data from expected (required) test data to cover the target node [6]. To follow the target path, branch requires to be true or false according to the need, input data must be adjusted or optimized. Every branch contains some logical expressions. In this paper, Korel's branch distance functions [6] are used for branch distance calculation.

Approximation Level-Based Fitness Function (ALBFF)
Approximation level assesses how close an individual is to reach the target on the basis of its execution part through control structure [7, 8]. It is determined by counting the number of decision nodes not traversed by a test case while executing the target path. So the approximation level should be minimized.

Combined Fitness Function (CFF)
This is the most widely used fitness function combines both approximation level and branch distance function. In the combined fitness function, a small fixed value is

added to BDBFF [8]. Therefore, the test case closer to the target node will have a higher fitness value as compared to those that are away from the target.

2.3 *Particle Swarm Optimization (PSO)*

Premature convergence and large number of iterations can be avoided using PSO. It yields quality results in comparison to GA. In PSO, particles are the candidate solutions. Each particle keeps track of its personal best (pbest) position and global best position (gbest) [9]. PSO accelerates each particle toward its pbest and the gbest locations, with a random weighted acceleration at each iteration.

$$V_i^{t+1} = wV_i^t + c_1 \,\text{rand}\,1\,(\ldots) \times \left(\text{pbest}_i - P_i^t\right) + c_2 \,\text{rand}\,2\,(\ldots) \times \left(\text{gbest} - P_i^t\right). \tag{1}$$

where

V_i^t: velocity of agent i at iteration t,
w: weighting function,
$c_{j:}$ learning factor,
P_i^t: current position of agent i at iteration t,
pbest_i: pbest of agent i,
gbest: gbest of the group.

In our work, we take inertia weight as 0.9 and learning/acceleration factor value 2, which are considered as a standard value for the experiments [3, 10]. These values are adjustable according to the need of the experiments.

3 The Proposed Approach

Automatic testing is a challenging task. To resolve this problem many techniques have been proposed. Search-based testing (SBT) is one of them. We use PSO searching techniques to generate test cases to cover linearly independent paths of a CFG derived from the triangle classification problem. To achieve maximum path coverage with the optimum number of test cases in less time, we propose a VCBDF fitness function which is used with PSO algorithm. The steps to generate test data are shown in Fig. 2.

Value Combined Branch Distance Function (VCBDF)
Our main objective is to cover a target path. In our approach, the path with the low probability of coverage is selected as the target path. To get the more efficient test cases to cover maximum path with a target path, we propose a fitness function and value combined branch distance-based function.

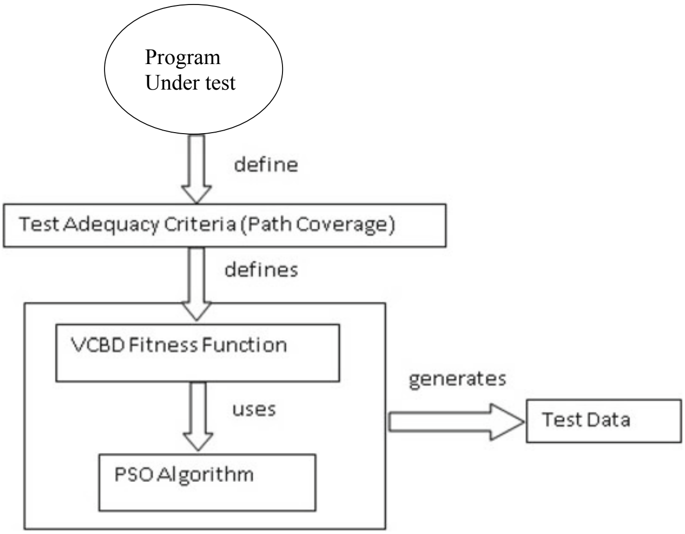


Fig. 2 Steps to generate test data

The objective is to distinguish the target path and the path nearest to the target path. Accordingly, we generate test cases to cover the target one. If the generated test cases fail to cover the target one, then also it indicates how far it is from the target path. For this reason, it is required to define distinct fitness values of the test cases. In simple branch distance-based fitness function, the branch distance value is very small which itself does not have any significance in distinguishing the target path. In approximation level-based heuristic, though we can find out how many conditional nodes are missed for traversing the target path, again its magnitude is very less which contributes to the final fitness value is ignorable. Also in approximation level, if all the nodes are traversed, then its value becomes zero, but still we cannot guarantee that the test cases are able to traverse the target path. As the last conditional node of a CFG also has two outgoing edges; one leads to target path and other one is the just diverged one.

Considering all these limitations, we improve the fitness function. In our proposed approach, the target path is assigned with a fixed large constant value to clearly distinguish among other paths. The paths are assigned with a constant value based on the distance from the target path.

Here, we calculate the distance with respect to the number of target path's conditional node covered. If all the nodes are visited and it leads to a target path, then highest value is assigned to that path. For less distance from the target path that means if maximum nodes are visited, the value assigned to the path is more. Likewise, value to the other paths is assigned in decreasing order based on the distance from the target that means less the distance, more the value and more the distance, less the value. So, there is no need to add approximation value to the final fitness. The branch distance fitness value is calculated by simply adding all the branch distances of a path and the fitness value is multiplied with the respective path values to get the final fitness. Path value and VCBD fitness computation procedures are shown below.

Steps to Evaluate Fitness Function
Input: CFG of the given code and randomly initialized test cases
Output: Fitness value of given test cases

1. Take CFG and randomly initialized test cases as inputs
2. Compute cyclomatic complexity to find the linearly independent paths
3. Find the number of conditional nodes (CN) present in each path
4. Find the path value of each path
   PV = Number of Conditional nodes × Large Constant
5. Find the branch distances of each node of a path
6. Compute the VCBD fitness of the test cases, if the test case (t) traversed the path i, then,

$$F(t) = \sum_{P_i \mathrm{CN}_j}^{n} \mathrm{BD}_{ij} \times \mathrm{PV}_i, \text{ where } j = 1, \ldots, n \quad (2)$$

Case Study (Triangle Classification Program):

Path Value (PV) Computation

1. Count the total number conditional nodes (CN) present in paths.

   In Fig. 1b, in path 4 (target path), CN = 3
   In path 1, CN = 1
   In path 2, CN = 2
   In path 3, CN = 3.
   Here, path 4 and path 3 both having three number of nodes but we choose path 4 as target node according to probability of covering a path.

2. Find the path value (PV) of each path in CFG. Path value is calculated by multiplying a large constant value (C) with the number of conditional node of target path traversed by the test case.
   For the target path, PV is obtained by multiplying twice the constant value with the number of conditional nodes it contains.

   In our work, we take $C = 1000$.
   In Fig. 1b,
   For path 1, PV = CN*C = 1*1000 = 1000
   For path 2, PV = 2*1000 = 2000
   For path 3, PV = 3*1000 = 3000
   For path 4, PV = 3*2*1000 = 6000.

VCBD Fitness Computation

1. For the true or false value of condition 1 calculate BD_1. If condition 1 is true go to step 2 else calculate PV for path 1 and go to step 5.
2. For the true or false of condition 2 calculate BD_2. If condition 2 is true, calculate PV for path 2 and go to step 5 else go to step 3.
3. For the true or false of condition 3 calculate BD_2. If condition 3 is true, calculate PV for path 3 and go to step 5 else go to step 4.
4. If condition 3 is false calculate PV for path 4 and go to step 5.
5. $F(t) = \sum_{P_i \mathrm{CN}_j}^{n} \mathrm{BD}_{ij} \times \mathrm{PV}_i, \text{ where } j = 1, \ldots, n.$

4 Experimental Result and Analysis

In this section, experimental results have been shown with analysis. Experiment has been done on triangle classification program shown in Fig. 1. Implementation has been done in MATLAB (Fig. 3). Instrumented program is taken as input to evaluate fitness function. PSO algorithm has been implemented with VCBD function and two existing fitness functions for comparison. Table 1 shows the parameter setting for PSO algorithm.

The number of iterations from Table 2 shows the minimum number of iterations required to generate test case for target path and achieve 100% path coverage. We

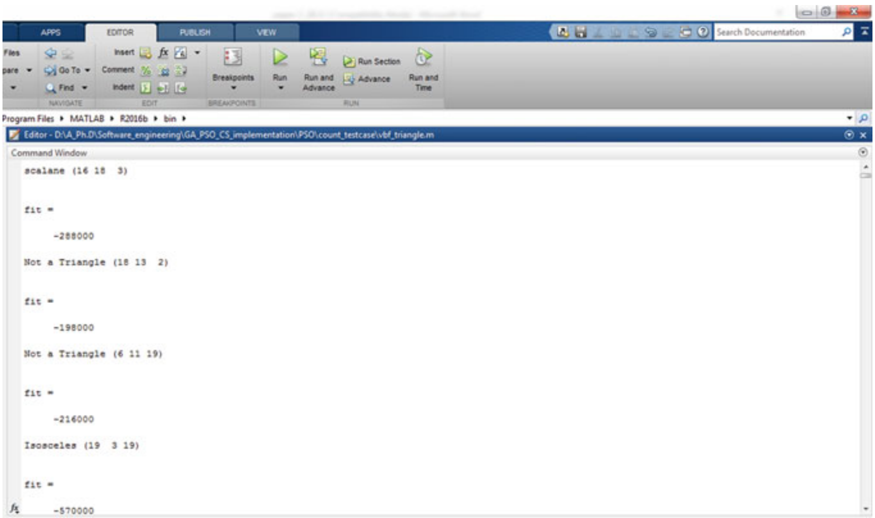


Fig. 3 Test case generation using PSO with VCBDF

Table 1 Parameter setting for PSO algorithm

Population size	40
Inertia weight (w)	0.9
Acceleration factors (c_1, c_2)	2
Lower bound	0
Upper bound	20

Table 2 Experimental result

Result						
Algorithms	No. of test cases generated for				No. of iterations	Execution time in seconds
	Not a triangle	Scalene	Isosceles	Equilateral		
PSO with branch distance-based fitness function	20	3	16	1	30	0.353586
PSO with combined fitness function	18	4	17	1	30	0.356059
PSO with VCBDBFF	15	4	18	3	10	0.225273

have plotted the convergence characteristics of PSO algorithm with the said fitness functions (Fig. 4).

From the experimental result, it is concluded that

- PSO with all the fitness functions are able to cover all the four paths. They give 100% path coverage.
- PSO with BDBFF and CFF gives nearly the same result that means approximation level value does not add any significance to the final fitness function.
- PSO is giving good results with VCBD-based fitness function than branch distance-based and simply combined fitness function.
- PSO with VCBD-based fitness function generates the optimum number of test cases to cover target path with minimum number of iterations in comparison to other two fitness functions.
- While PSO using branch distance-based and simply combined fitness function require 30 iterations to cover target path, VCBD-based fitness function requires only ten iterations with PSO.

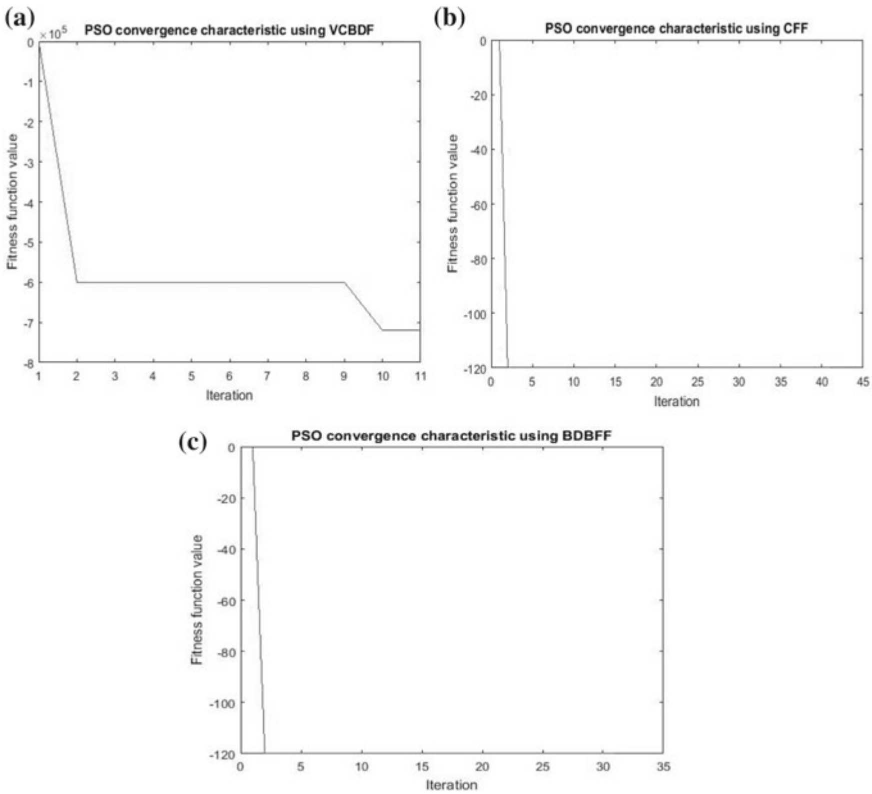


Fig. 4 Convergence characteristics using. **a** VCBDF, **b** CFF, **c** BDBFF

5 Conclusion

We propose a fitness function, which guides the PSO searching techniques to automatically generate test cases for path coverage. Using PSO with proposed VCBD function, we have achieved 100% path coverage. The significance of this fitness function is that we are reaching our target path within a few number of iterations and less time in comparison to other two existing fitness functions (BDBFF and CFF). It is observed that VCBD fitness function is efficient fitness to achieve target path in case of automatic test case generation.

References

1. Maragathavalli, P.: Search-based software test data generation using evolutionary computation. Int. J. Comput. Sci. Inf. Technol. **3**(1), 213–223 (2011)

2. Roshan, R., Porwal, R., Sharma, C.M.: Review of search based techniques in software testing. Int. J. Comput. Appl. **51**(6), 42–45 (2012)
3. Shahid, M., Ibrahim, S., Mahrin, M.N.: A study on test coverage in software testing. In: International Conference on Telecommunication Technology and Applications. Proceedings of CSIT, vol. 5, pp. 207–215 (2011)
4. Sahoo, R.R., Ray, M.: Metaheuristic techniques for test case generation: a review. J. Inf. Technol. Res. **11**(1), 158–171 (2018)
5. Baresel, A., Sthamer, H., Schmidt, M.: Fitness function design to improve evolutionary structural testing. In: Proceeding of the Genetic and Evolutionary Computation Conference, pp. 1329–1336 (2002)
6. Chen, Y., Zhong, Y., Shi, T., Liu, J.: Comparison of two fitness functions for GA-based path-oriented test data generation. In: Fifth International Conference on Natural Computation, pp. 177–181. IEEE Computer Society, Washington, DC (2009)
7. Thi, D.N., Hieu, V.D., Ha, N.V.: A technique for generating test data using genetic algorithm. In: International Conference on Advanced Computing and Applications, IEEE Conference, pp. 67–73 (2016)
8. Gursaran, A.: Program test data generation for branch coverage with genetic algorithm: comparative evaluation of a maximization and minimization approach. JSE CS & IT **4**, 443–454 (2012)
9. Chen, Y., Zhong, Y., Shi, T., Liu, J.: Comparison of two fitness functions for GA-/based path-oriented test data generation. In: Fifth International Conference on Natural Computation, pp. 177–181 (2009)
10. Harman, M., Jia, Y., Zhang, Y.: Achievements, open problems and challenges for search based software testing. In: Proceedings of IEEE 8th International Conference on Software Testing, Verification and Validation, pp. 1–12 (2015)