International Journal of Distributed and Cloud Computing
11 (1) 2023, 01-09
http://www.publishingindia.com/ijdcc

The Role of Innovation and Digitalization: Influence SMEs to Adopt Circularity in the Business Model

Sayandeep Chandra
Sales & Solution Analyst, Tata Consultancy Services; MBA, Marketing & Finance, Rajiv Gandhi Institute of Petroleum Technology, Amethi, Uttar Pradesh, India.
Email: sayandeep.chandra@gmail.com

Abstract: The circular economy can balance the environmental aspect and industrial production objectives. It defines the methodology to employ fewer resources and gain production efficiency with the help of reduce-reuse-recycle. Compared to larger organization, small, and medium enterprises (SMEs) faces technical & resource challenges to adopting circular economic practices. This lack of technical know-how and managerial myopic point-of-view towards sustainability hold back SMEs to shift to the sustainable business model from the liner structure. Here transition towards digitalization plays a pivotal role and ignites the innovative mindset. The implication of data analytics and machine-learning algorithm help to optimize the day-to-day operation. Application Artificial Intelligence empowers SMEs to understand consumer insight, risk assessment and predict the outcome. The incorporation of Internet-of-Things and Cloud Computing assist to access product usage data and user behavioural r data easily and entitle the exchange of data across multiple stakeholders.

Keywords: Circular economy, Digitalization, Innovation, R-framework, SMEs.

I. Introduction

The reorganization towards a circular economy (CE) has become a strategic focal point for global political leaders, business leaders and leading organizations. CE is the alternative model to the linear economic model that appreciates the take-make-waste principle [1]. CE discard the *"end-of-life"* approach and encourages to adoption recycle, reuse and reduction principle [2]. Transformation to the CE ensures waste generation is minimized, resource utilization is optimized and instead of scrapping, the resource is fed for reproduction. Thus the development & production processes are designed sustainably [3].

Compared to multination organizations, SMEs face challenges to transform to a circular economic model from a linear model due unavailability of data and knowledge. Here, digitalization plays a crucial role [2]. Cutting-edge technologies like Artificial Intelligence, Machine Learning, the Internet of Things, and Cloud Computing empower SMEs to harmonize with the paradigm shift [4]. Accessing the users' data, industry data and product usage data enables SMEs to gain insights into the product life-cycle pattern and it will impacts the manufacturing & decision-making process [5]. The adoption of digital technologies fuelled the importance of innovation to create a viable market that attracts non-consumers [6] and curves the avenue for revenue for SMEs.

In this article, we will try to understand the concept of innovation and CE. The first section of the literature review will explain the evaluation of the concept of innovation. Initial innovation was defined as the birth of new ideas & products, later development of new services is considered a part of innovation and in recent times market creation ability of innovation is prioritized. In the second part of the literature review, we will review the CE concept. Later we will discuss the barriers that prevent SMEs to adopt CE and how

digitalization & innovation boost the willingness to develop the CE business model.

II. Literature Review

A. Innovation and Market Creation

There are different perspectives when defining the term "Innovation". If we analyse various definitions of innovation through a period, we can witness the change in the fundamental areas of agreement. Some researchers prioritised the term "new idea", and some researchers focused on the internal & external environmental effects on the organization that cherished the innovation. On the other hand, recent definitions gave importance to value creation as a crucial factor. Victor A. Thompson defined innovation as the generation, acceptance and implementation of new ideas. From an organization's point of view, the production and utilization of new ideas should be considered innovation [7]. Becker and Whisler suggested *first and early use of an idea* that harmonizes with the organization's goal should be defined as innovation. Here they emphasized *stimulus (external or internal)* as the igniter for innovative action [8]. L. B. Mohr distinguished the concept of "*invention"* from "*Innovation"* and explained innovation as "*bringing something new into use"*. Also, L. B. Mohr highlighted once the change idea is adopted by organization stakeholders or the public, it will not be considered as innovative, no longer [9]. Innovation leads to economic growth and nourishes the job market. In the chapter *"The Institutional Analysis of Entrepreneurship: Historic Aspects of Schumpeter's Development Theory"* Alexander Ebner defined innovation as the introduction of a new good or new quality of good or new methodology of production that creates a new market which curves a different supply chain [10]. Clayton M. Christensen in his book *The Prosperity Paradox* defines innovation as a change in process that leads to the transformation of *labour, capital, material and information* into goods and services with significant value [6]. Though there is diversity in the definition of the term "innovation", the key aspects are- innovation is associated with organisational change and the objective of the innovation is to create a new market. More specifically, as stated by Clayton M. Christensen, innovation is essential to attract the "*non-consumer"* section of the market [6]. Innovation has a significant impact on the market because it leverages technology or business model to make products and services affordable and accessible to people. Abernathy and Clark [11] explained innovation can obsolete existing knowledge of a particular application with the introduction of a new product or process. Thus it develops a new market that addresses customer needs.

B. Circular Economy

Easily available resources and inexpensive and zero-impact on the environment, in that case, the linear economic model is suitable for meeting the consumers' demands. Companies extract the raw material and convert it into a product or use the core material to produce energy that directly or indirectly serves the requirement of the customers. In return, consumers pay for the product and discard the waste after usage. This is the "take, make, waste" principle of the linear economy. But the growing population and consumers demand to create pressure on the production system and most of the core resources that we use for production, are non-renewable. In simple words, a linear economy barely prioritizes the environmental aspect and zero-waste policy [12]. The business concept-three bottom line framework talks about *profit, people & planet* [13]. The said perspective links business goal with the concept of sustainable development, defined as the need of the present generation should be met without jeopardising future generations [14]. The weak version of sustainability considers *natural capital* and *manufactured capital* are substitutable and there is no difference. The primary aspect is that the aggregated *stock of capital* should be the same or increased. On the other hand, strong sustainability denies the former concept, rather it states that most natural resources are not interchangeable with manufactured capital [15]. The knowledge about sustainable development evolved the concept of CE that proposes an alternative to the "take-make-waste" model. CE model emphasizes the cyclical material flow, use of renewable energy and judicial use of waste energy. The economic aspects

of the CE are the reduction of raw mater & energy costs, waste management, emission control, risk & goodwill management and creating the market for a new product or business model innovation. From a social point of view, CE encourages towards shared economy [16]. So, CE is related to three strategies- optimized resource use for production, increase utilization period of the product and connect a post-usage product with the initial production stage [17].

III. Relevance of CE in Small, Medium Size Enterprise Ecosystems

CE approaches the win-win condition that enables an economically feasible way to reuse the product and increase the life of the goods through recycling [18]. The fundamental objective of the CE is described with various models; the *"R"* framework is the most appreciated and commonly accepted framework. The frameworks are *3R* (Reuse, Repair, Recycle), the *4R* considers Refurbish along with the formers, the *6R* includes Rethink and Remanufacture, *9R* adds Repurpose, Recover, Reuse. The said *"R"* frameworks explain the circular relationship between production and consumption [19]. The CE improves resource utilization and waste management methodology across the valuechain. In the upstream, management of resources employed for production and reduction of wastage are guided by CE, thus productivity & efficiency are boosted. It is necessary to monitor the reintroduction of the discarded product (after the end of life) back into the production process. These two features institute the closed loop for the production valuechain and ensure sustainability [20].

The transition towards the CE needs proactive engagement from various stakeholders- *policy-makers, industries, companies, and individuals* who can drive the fundamental changes associated with an industrial structure [19] [21]. Large corporations are capable to adopt the practices related to the CE, due to the stability. Proactively working on sustainable development multinational corporations strengthens the relationship with the stakeholders and embrace global ESG policies [19]. But some potential barriers hold back SMEs to adopt the circular business model. Internal stakeholders like managers, employees and owners focus on the direct market returns and benefits linked with the core business. Also, SME stakeholders have a basic level of knowledge about sustainability and CE. This is an organisational cultural barrier, creating a huddle for opportunities for CE [22] [23]. Lack of capital [24] is another aspect that obstructs the adoption of CE for the SME sector. Transform from a linear business model to a circular one requires upfront investments like supply-chain optimization, distribution, repurchase from the customers, logistics and inventory management etc. Along with the said ones the institution needs to monitor the product life-cycle and has to employ committed resources for close tracking and observation [25] [26]. On the other hand, potential investors perceive that green investment is associated with risk due to the uncertainty related to performance [27]. Technical knowledge and information related to CE are not easily accessible by the SME sectors. Whereas multinational institutions with the help of in-house research, develop technological knowledge and design ESG policies that harmonize with sustainability goals. But, SMEs do not have the desired research team, resources, and knowledge to customize the technologies that deliver environmental benefits along with economic benefits and thus they have to depend on the technologies that are available in the market [28] [29]. External support and inter-industry collaboration are essential to nourish the willingness to adopt green business practices for SMEs. But in reality, SMEs face challenges communicating with external partners regarding green business models and there are limited forums or platforms for SMEs that cultivate green innovation ideas and practices [30]. Poor organization structure leads to a knowledge gap related to green buying and supply chain and stakeholders ignore the environmental aspect for cost saving. Absence of structured training programme and technology support from the Government make SMEs myopic about the CE [31]. Customers or end-users are the primary drivers of the demand centre and their buying patterns impact the market. Cost-sensitive consumers often disagree to pay a premium for green products and thus SMEs find the adoption of green practices less potential [31] [30]. Also, the knowledge gap among consum-

ers about green products fails to develop demand pressure on the production side. Lack of demand portrays no realistic benefits of the CE to the SMEs and does not bring out the necessity of transition to CE [26] [32] [33].

Entrepreneurs act as a driver who can monetize innovation and transfer knowledge into economic growth [34]. Entrepreneurs are flexible to adopt changes and can synchronize with the market needs [35] [29]. According to The World Bank data, the SME sector accounts for 90% of business and is responsible for 50% of employment, globally. The said sector contributes ~40% of the national income in emerging countries [36]. The SMEs sector contributes a significant amount to the national income of India (~ 30% of GDP) and creates employment opportunities. Unlike a large corporation, SMEs sustains rural industrialization and socio-economic growth in India [37]. In European Union Countries (EU; 27 countries; EU-27), SMEs are the key contributor to economic growth and employment generation. In 2021, 22.8 million SMEs were active in the EU-27 and accounts for 99.8% of all enterprises in non-financial business sectors [38]. The environmental impact caused by SMEs, facts are surprising. According to the British Business Bank report, SMEs at a group level are responsible for 43-53% of business greenhouse emissions in the UK [39]. SMEs certified sustainable standard is very minor across the globe- only 5% of micro and 13% of small enterprises comply with the standard [40]. The said sector contributes to job creation and adds value to economic development Notably, SMEs are *pollution-intensive* compared to large organisations [41]. In recent years, SMEs are adopting necessary steps toward the sustainability standard and this shift is crucial for the future of mother earth. This transition is supported by eco-innovation and a change in consumers' orientation towards green products [41] [29]. The adoption of circular economic practices will support the SME sector to reduce environmental impacts.

IV. Digitalization and Innovation Boost the Adoption of the CE

Advancement in Information and Communication (ICT) technology segment, organizations adopted ICT as the key driver for business growth and a tool for market expansion. Businesses across different verticals have enabled data processing techniques and digital communication channels (like social media) to gain an advantage over the competition and bring agility into day-to-day operations (UNCTAD, 2021). Moreover, the COVID-19 pandemic has fuelled the digitalization process and connected all the stakeholders with a common thread; the thread is called data. Data is an intangible asset with a certain economic value. The most important characteristic of data is "non-rival" by nature and it allows multiple companies to utilize the same users' group data at the same time. This data is used to make a decision that impacts the economy and environment (Coyle, Diepeveen, Wdowin, Kay and Tennison, 2020). The fourth industrial revolution is nourished by the integrated information and knowledge-driven networking system. Data accumulated in every stage of production helps enhance the decision-making process and evaluate efficiency. Adoption of digital technologies helps in developing the CE and curve the path for reduction of carbon footprint. Cutting-edge digital technologies like the Internet of Things (IoT), and Big Data Analytics, are essential resources to design remanufacturing, recycling and repurposing production systems [42]. The application of these emerging technologies brings in several benefits tracking business operations, product lifecycle management, and real-time data analysis for better decision-making [43]. Moreover digitalized CE can nullify the barriers by enabling improved communication channels between stakeholders throughout the value chain and monitoring the information flow, essential for the transition to the CE [44].

The article limits its focus to a few digitalization tools- Artificial Intelligence (AI), IoT, Cloud Computing (CC) and Big Data (BD), application of these tools will be beneficial for SMEs to develop circular business models.

A. Role of Big Data (BD)

Workforce skill is an implicit resource that is invisible in the lens of the traditional measurement methods. Skills are learnt and develop with the time

[45]. Digitalization and data analytics has become the implicit skill that is the driver for modernization. This resource is complex and brings in competitive advantage by improving cost performance and operational efficiency of the organization [46]. So, BD is designed with the comprehensive dataset coming from various sources facilitating enormous amount of data and information. Then desired analytical method is applied on the dataset and data mining and machine learning algorithm is applied to discover the pattern [5]. In simple terms, Big Data Analytics (BDA) generates insight by analysing data and assists in decision-making like knowledge management, cost management, distribution etc. BDA is useful for product quality management and understanding the product lifecycle [2].

B. Role of AI and IoT

Organizations adopt AI and Machine-Learning (ML) as the latest control system that optimizes cost, manages resources and reduces response time. The autonomous decision-making ability incorporates flexibility & agility in the organization's ecosystem [46]. Implementation of AI and ML enables real-time data analytics by optimizing resource flow, predicting risk and designing a suitable system. This AI-based platform leads to improve circularity [47] [2]. AI can identify hidden patterns in the dataset and efficiently derive insights. Researchers developed an AI-based system that identifies suitable suppliers based on circularity standards, an AI-enabled decision support system that is useful in minimizing resource consumption [48] [49]. Natural Language Processing (NLP) is another efficient application of AI and it can transform methodical work into an autonomous system [50].

IoT-enabled devices or products are effective to track the lifecycle and help in maintaining, reusing, recycling and remanufacturing. Current usage of the product and future performance can be measured in IoT devices. This creates a digital profile of the product and the manufacturer can decide the circularity of the product by analysing the profile. This data-driven decision-making process is fuelled by data collection and exchange characteristics of IoT systems [51] [52].

C. Role of Cloud Computing

CC refers to the delivery of computing services over the internet. Data are stored, analysed and distributed online with the help of cloud service. The primary benefit of the cloud service it can be accessed by multiple users from various locations. CC helps the organization to save infrastructure costs by reducing the dependency on physical servers and improving operational efficiency [44]. Cloud platform helps to store IoT product data, real-time user data, organizational data etc. and make them available on a need basis. The said feature provides flexibility to run the operation smoothly and design the system with agility [53].

From the above discussion, we understood the impact of digitalization on CE. Now we need to articulate how digitalization and innovation encourage SMEs to adopt CE. CE is an example of a new way of business management that considers the environment as a key operational lever like cost-centre, profit-centre, supplier-buyer, resource management etc. To adopt the new business model SMEs need to understand Business Canvas Model (BMC) related to the industry they are serving. It is essential for SMEs to define what value is created, what value is proposed and what value will deliver to the consumers. SMEs have to structure the revenue stream and cost model for business sustainability [54]. Three types of innovation models can acknowledge SMEs to adopt CE practices. First, sustaining innovation model, second, the efficiency innovation model and lastly market-creating innovation model [6]. Sustaining innovation model focuses on improving existing products and targets consumers with purchasing power. Consumers have a positive inclination towards green products purchase and personal traits, external influencers, and knowledge work as the motivator in this case [55]. The efficiency innovation model prioritizes the process of manufacturing. The said innovation model emphasizes resource utilization and waste minimization. Reuse, Recycle, Remanufacturing, Refurbish, Repair etc. elements of CE, are linked with the efficiency innovation model and it will empower SMEs to be predominantly part of the green economy. Last innovation model-Market-creating innovation model is essential for

new SMEs or start-ups. As mentioned in the name, the said innovation model comes with new ideas and a new business model that attracts the *non-consumer* section of the market [6]. Considering some of the nine elements of the CE, entrepreneurs can develop new technology with new business models that will cater to niche sections of the market.

Innovation and digitalization are two tools for SMEs to synchronize with the circular economic model. Innovation derives new ideas, and new business models and digitalization work as a catalyst that makes the operation smooth, agile, cost-effective and data-driven.

V. Discussion and Conclusion

The fundamental concept of the CE is to produce more with minimum use of resources. CE links business operations with resource efficiency. Resource efficiency depends upon optimized production methodology-reuse, recycle, energy resource, waste reduction etc. Successful adoption of an efficient production strategy encourages cutting down production cost, minimising resource procurement cost, and energy management and finally lead to revenue generation [56] [57]. Digitalization helps in driving the CE business model and works as the key enabler. AI, ML, and BDA are essential to derive insight from the large volume of data and utilize that data to automate operations. IoT and Cloud technology are two pillars of digitalization, helping to get real-time data and manage the data easily. SMEs, especially in developing countries, need to adopt technologies to innovate business models that fit with the CE. Support from Government, policy maker and financial institutions, is desired to motivate owner-managers of SMEs to accept the transition and restructure to the eco-friendly business model.

References

[1] Z. Yuan, J. Bi, and Y. Moriguichi, "The circular economy: A new development strategy in China," *Journal of Industrial Ecology,* vol. 10, pp. 1-2, 2008.

[2] C. Chauhan, V. Parida, and A. Dhir, "Linking circular economy and digitalisation technologies: A systematic literature review of past achievements and future promises," *Technological Forecasting & Social Change,* vol. 177, 2022.

[3] F. Testa, N. M. Gusmerotti, F. Corsini, G. Pretner, and F. Iraldo, "Drivers and approaches to the circular economy in manufacturing firms," *Journal of Cleaner Production,* vol. 230, pp. 314-327, 2019.

[4] A. Chaudhuri, N. Subramanian, and M. Dora, "Circular economy and digital capabilities of SMEs for providing value to customers: Combined resource-based view and ambidexterity perspective," *Journal of Business Research,* vol. 142, pp. 32-44, 2022.

[5] S. Gupta, H. Chen, B. T. Hazen, S. Kaur, and E. D. R. Santibañez Gonzalez, "Circular economy and big data analytics: A stakeholder perspective," *Technological Forecasting & Social Change,* vol. 144, pp. 466-474, 2019.

[6] C. M. Christensen, *The Prosperity Paradox: How Innovation Can Lift Nations Out of Poverty.* New York: HarperCollins, 2019, pp. 10-12.

[7] V. A. Thompson, "Bureaucracy and innovation," *Administrative Science Quarter,* vol. 5, pp. 1-20, June 1965.

[8] S. W. Becker, and T. L. Whisler, "The innovative organization: A selective view of current theory and research," *The Journal of Business,* vol. 40, no. 4, pp. 462-468, Oct. 1967.

[9] L. B. Mohr, "Determinants of innovation in organizations," *The American Political Science Review,* vol. 63, no. 1, pp. 111-126, Mar. 1969.

[10] A. Ebner, "The institutional analysis of entrepreneurship: Historist aspects of Schumpeter's Development Theory," in J. G. Backhaus (Ed.), *Joseph Alois Schumpeter*. New York: Kluwer Academic Publishers, 2003, p. 130.

[11] W. J. Abernathy, and K. B. Clark, "Innovation: Mapping the winds of creative," *Research Policy,* pp. 3-22, 1985.

[12] P. Lacy, and R. Rutqvist, *Waste to Wealth*. London: Palgrave Macmillan London, 2015.

[13] K. Miller, "The triple bottom line: What it is & why it's important," Dec. 08, 2020. Accessed: Oct. 2, 2022. [Online]. Available: https://online.hbs.edu/blog/post/what-is-the-triple-bottom-line

[14] U. Nation, "Sustainability". Accessed: Sep. 24, 2022. [Online]. Available: https://www.un.org/en/academic-impact/sustainability

[15] P. Victor, S. Hanna, and A. Kubursi, "How strong is weak sustainability?," in S. Faucheux, M. O'Connor, J. van der Straaten (Eds.), Sustainable development: Concepts, rationalities and strategies, *Economy & Environment.* Springer, 1998, vol. 13.

[16] J. Korhonen, A. Honkasalo, and J. Seppälä, "Circular economy: The concept and its limitations," *Ecological Economics,* vol. 143, pp. 37-46, 2018.

[17] O. Rodríguez-Espíndola, A. Cuevas-Romo, S. Chowdhury, N. Díaz-Acevedo, P. Albores, S. Despoudi, C. Malesios, and P. Dey, "The role of circular economy principles and sustainable-oriented innovation to enhance social, economic and environmental performance: Evidence from Mexican SMEs," *International Journal of Production Economics,* vol. 248, 2022.

[18] N. M. P. Bocken, I. d. Pauw, C. Bakker, and B. van der Grinten, "Product design and business model strategies for a circular economy," *Journal of Industrial and Production Engineering,* vol. 33, no. 5, pp. 308-320, 2016.

[19] F. Gennari, "The transition towards a circular economy. A framework for SMEs," *Journal of Management and Governance,* Sep. 2022.

[20] A. S. Homrich, G. Galvão, L. Gamboa, and M. M. Carvalho, "The circular economy umbrella: Trends and gaps on integrating pathways," *Journal of Cleaner Production,* vol. 175, pp. 525-543, 2018.

[21] A. Chizaryfard, P. Trucco, and C. Nuur, "The transformation to a circular economy: Framing an evolutionary view," *Journal of Evolutionary Economics,* Oct. 30, 2020.

[22] M. Ormazabal, V. Prieto-Sandoval, R. Puga-Leal, and C. Jaca, "Circular economy in Spanish SMEs: Challenges and opportunities," *Journal of Cleaner Production,* vol. 185, no. 4, pp. 157-167, 2018.

[23] F. Takacs, D. Brunner, and K. Frankenberger, "Barriers to a circular economy in small- and medium-sized enterprises and their integration in a sustainable strategic management framework," *Journal of Cleaner Production,* vol. 362, Aug. 2022.

[24] A. Trianni, and E. Cagno, "Dealing with barriers to energy efficiency and SMEs: Some empirical evidences," *Energy,* vol. 37, no. 1, pp. 494-504, 2012.

[25] K. Besch, "Product-service systems for office furniture: Barriers and opportunities on the European market," *Journal of Cleaner Production,* vol. 13, no. 10-11, p. 1083-1094, 2005.

[26] V. Rizos, A. Behrens, and et al., "Implementation of circular economy business models by small and medium-sized enterprises (SMEs): Barriers and enablers," *Sustainability,* vol. 8, no. 11, 2016.

[27] C. Ghisetti, S. Mancinelli, M. Mazzanti, and M. Zoli, "Financial barriers and environmental innovations: Evidence from EU manufacturing firms," *Climate Policy,* vol. 17, pp. s131-s147, 2017.

[28] OECD, "Environmental policy toolkit for SME greening in EU eastern partnership countries," 2018.

[29] J. A. Purwandani, and G. Michaud, "What are the drivers and barriers for green business practice adoption for SMEs?," *Environment Systems and Decisions,* vol. 41, pp. 577-593, 2021.

[30] H. Gupta, and M. K. Barua, "A framework to overcome barriers to green innovation in SMEs using BWM and Fuzzy TOPSIS," *Science of the Total Environment,* vol. 633, pp. 122-139, 2018.

[31] S. Dhull, and M. S. Narwal, "Drivers and barriers in green supply chain management adaptation: A state-of-art review," *Uncertain Supply Chain Management,* pp. 61-76, Jan. 2016.

[32] O. Meqdadi, T. Johnsen, and R. Johnsen, "The role of SME suppliers in implementing sustainability," in *IPSERA 2012 Conference*, Napoli, 2012.

[33] E. G. Edbring, M. Lehner, and O. Mont,, "Exploring consumer attitudes to alternative models of consumption: Motivations and barriers," *Journal of Cleaner Production,* vol. 123, pp. 5-15, 2016.

[34] Z. J. Acs, S. Estrin, and T. Mickiewicz, "Entrepreneurship, institutional economics, and economic growth: An ecosystem perspective," *Small Bus Econ,* pp. 501-514, 2018.

[35] S. Benzidia, and N. Makaoui, "Improving SMEs performance through supply chain flexibility and market agility: IT orchestration perspective," *Supply Chain Forum: An International Journal,* vol. 21, no. 3, pp. 173-184, 2020.

[36] The World Bank, "Small and medium enterprises (SMEs) finance," 2022. [Online]. Available: https://www.worldbank.org/en/topic/smefinance

[37] IBEF, "India's MSME sector," IBEF, 2022.

[38] European Commission, "Annual report on European SMEs 2021/22," European Union, Luxembourg, 2022.

[39] British Business Bank, "Smaller businesses and the transition to net zero," British Business Bank Plc, 2021.

[40] The International Trade Centre (ITC), "Empowering the green recovery," International Trade Centre, Geneva, 2021.

[41] OECD, "No net zero without SMEs," OECD, 2021.

[42] S. Bag, G. Yadav, P. Dhamija, and K. K. Kataria, "Key resources for industry 4.0 adoption and its effect on sustainable production and circular economy: An empirical study," *Journal of Cleaner Production,* vol. 281, 2021.

[43] J. S. Mboli, D. K. Thakker, and J. L. Mishra, "An Internet of Things-enabled decision support system for circular economy business model," *Software Practice and Experience,* vol. 52, no. 3, pp. 772-787, 2020.

[44] A. Hedberg, and S. Šipka, "The circular economy: Going digital," European Policy Centre, 2020.

[45] S. L. Hart, "A natural-resource-based view of the firm," *The Academy of Management Review,* vol. 20, no. 4, pp. 986-1014, Oct. 1995.

[46] S. Bag, J. H. C. Pretorius, S. Gupta, and Y. K. Dwivedi, "Role of institutional pressures and resources in the adoption of big data analytics powered artificial intelligence, sustainable manufacturing practices and circular economy capabilities," *Technological Forecasting & Social Change,* vol. 163, 2021.

[47] B. Sundui, O. A. R. Calderón, O. Abdeldayem, and J. L. Gil, "Applications of machine learning algorithms for biological wastewater treatment: Updates and perspectives," *Clean Technologies and Environmental Policy,* vol. 23, pp. 127-143, 2021.

[48] B. Alavi, M. Tavana, and H. Mina, "A dynamic decision support system for sustainable supplier selection in circular economy," *Sustainable Production and Consumption,* vol. 27, pp. 905-920, 2021.

[49] I. Makarova, K. Shubenkova, and A. Pashkevich, "The concept of the decision support system to plan the reverse logistics in automotive industry," in *26th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2018.

[50] M. Tang, and H. Liao, "Multi-attribute large-scale group decision making with data mining and subgroup leaders: An application to the development of the circular economy," *Technological Forecasting & Social Change,* vol. 67, 2021.

[51] X. Sun, and X. Wang, "Modeling and analyzing the impact of the internet of things-based industry 4.0 on circular economy practices for sustainable development: Evidence from the food processing industry of China," *Front. Psychol.,* 2022.

[52] A. Cavalieri, J. Reis, and M. Amorim, "Circular economy and internet of things: Mapping science of case studies in manufacturing industry," *Sustainability,* vol. 13, no. 6, 2021.

[53] J. Kallio, M. Antikainen, O. Kettunen, and O. Korpipää, "Internet of things and cloud computing enabling circular economy," *International Journal on Advances in Internet Technology,* vol. 11, 2018.

[54] A. Osterwalder, and Y. Pigneur, *Business Model Generation*. Hoboken, New Jersey: John Wiley & Sons, 2010.

[55] Y. Joshi, and Z. Rahman, "Factors affecting green purchase behaviour and future research directions," *International Strategic Management Review*, vol. 3, no. 1-2, pp. 128-143, 2015.

[56] M. Linder, and M. Williander, "Circular business model innovation: Inherent uncertainties," *Business Strategy and the Environment,* vol. 26, no. 2, pp. 182-196, 2017.

[57] N. M. P. Bocken, S. W. Short, P. Rana, and S. Evans, "A literature and practice review to develop sustainable business model archetypes," *Journal of Cleaner Production,* vol. 65, pp. 42-56, 2014.

[58] K. Urabe, "Innovation and the Japanese management system," in *Innovation and Management: International Comparisons*. De Gruyter, 1988, p. 3.

[59] C. L. Wang, and P. Ahmed, "The development and validation of the organizational innovativeness construct using confirmatory factor analysis," *European Journal of Innovation Management,* vol. 7, no. 4, pp. 303-313, 2004.

[60] G. J. Tellls, and R. K. Chandy, "Organizing for radical product innovation: The overlooked role of willingness to Cannibalize," *Journal of Marketing Research,* vol. 35, pp. 474-487, 1998.

[61] UNCTAD, "Digital economy report: Cross-border data flows and development: For whom the data flow," United Nations Publications, New York, 2021.

[62] D. Coyle, S. Diepeveen, J. Wdowin, L. Kay, and J. Tennison, "The value of data: Policy implications," Bennett Institute, 2020.

[63] V. Blok, T. B. Long, A. I. Gaziulusoy, and et al., "From best practices to bridges for a more sustainable future: Advancesand challenges in the transition to global sustainable production andconsumptionIntroduction to the ERSCP stream of the special volume," *Journal of Cleaner Production,* vol. 30, pp. 1-12, May 2015.

[64] M.-L. Tseng, R. R. Tan, A. S. F. Chiu, C.-F. Chien, and T. C. Kuo, "Circular economy meets industry 4.0: Can big data drive industrial symbiosis?," *Resources, Conservation and Recycling,* vol. 131, pp. 146-147, 2018.

[65] V. Parida, D. Sjödin, and W. Reim, "Reviewing literature on digitalization, business model innovation, and sustainable industry: Past achievements and future promises," *Sustainability,* vol. 11, no. 2, 2019.

[66] N. M. P. Bocken, I. d. Pauw, C. Bakker, and B. van der Grinten, "Product design and business model strategies for a circular economy," *Journal of Industrial and Production Engineering,* vol. 33, no. 5, pp. 308-320, 2016.

International Journal of Distributed and Cloud Computing
11 (1) 2023, 10-22
http://www.publishingindia.com/ijdcc

Digital Transformation Model in the State Universities in Cameroon Based on Actor Network Theory

Tatiana Eyenga Ovono[1], Ghislain Abessolo Alo'o[2*] and Roger Atsa Etoundi[3]
[1]University of Yaounde I, Cameroon. Email: tatianaeyengao@gmail.com
[2]University of Yaounde I, Cameroon. Email: ghislain.abessolo@facsciences-uy1.cm
[3]University of Yaounde I, Cameroon. Email: atsa.etoundi@ict4d.cm
[*]Corresponding Author

Abstract: The number of students increases every year in all universities in Cameroon. The facilities in each university do not allow efficient teaching and training of students. For a given course, instead of delivering one class per week in a normal situation, the same class is delivered many times per week to allow students to try the course. For some classes that can't be scheduled multiple times a week, students attempt the class by standing outside the amphitheater when it is already full. In this chaotic situation, it becomes extremely difficult to schedule an internship that would allow students to gain experience. As a result, the failure rate is very high, students who managed to graduate did not have the necessary skills to scope with real world problems related to the environment or applied area of study. Now, the main question is how to scope with a larger number of students in the case of insufficient teaching and training facilities. More specifically, how to facilitate digital transformation in state universities in Cameroon? This is the objective of this thesis. To address this issue, the Actor Network Theory is used. The result is the definition of the digital transformation model and its implementation to facilitate the adoption of technologies in the teaching and learning process in state universities in Cameroon.

Keywords: Actor network theory, Business process modeling, Digital transformation, M-learning, Quality of service.

I. Introduction

The government of Cameroon had launched the E-National Higher Education Network project for the digital transformation of the higher education system to scope with the problem of overcrowded students on campus. This project is designed to assist state universities in scoping with the issue of overcrowded students on campus in universities. From the project, IT infrastructures have been updated to provide students with free Internet access and online access to course materials [12]. Digital development centers were built to facilitate the digitization of teaching, and a large number of portable computers were offered to students in general. Cameroonian universities are no longer last in the search for academic excellence and competitiveness. These students must use these computers to facilitate their learning practice and acquire new knowledge anytime and anywhere. With this act, the state has laid the solid foundation of the digital campus. Reason why the mobile learning project is in gestation in the state universities in Cameroon to find out how it is possible to perform their learning and teaching systems, to increase the rate of success of the students and to achieve the goals of digital governance depicted in the manual of academic norm. So, before building this kind of important project, it is necessary to assess the readiness of the universities to effectively prime mobile technology in their educational system and to bring to light the actors (human and non-human) who can positively influence this initiative. This type of initiative is called mobile learning (m-learning) [1, 16]. Many authors have defined the concept of m-learning [1, 15, 18]. M-learning is defined as learning practices that take place when learners or educators are either in a permanent, prearranged or not

in a permanent or prearranged learning environment and where mobile technologies are used as a tool to facilitate learning [2, 15, 16, 17]. In this definition, we can note that we have many different actors, learners and educators who deal with the m-learning process.

Thus, the aim of this paper is to show how to facilitate digital transformation in state universities in Cameroon to help them in scoping with the overcrowded students on campus in universities. To solve similar problem, many theories or approaches have been used and we can notice, activity theory where activity is the core concept and is defined as dialectical relationship between subject and object. This theory has been used to understand the great opportunities that mobile technologies can offer to mobile learners [1, 6]. The Unified Theory of Acceptance and Use of Technology (UTAUT) was developed by Paledi *et al.* [15]. This theory has been used by Liu [13] to emphasize and indicate that the functionality of learners' devices and the perception of support from learning material providers affect learners' satisfaction and decisions to adopt mobile learning. Many other theories have been used by information systems researchers in the specific area of mobile learning, such as Structuration Theory and Actor Network Theory, which will be the theory used in this study. In this paper, Actor Network Theory will be used. The result is the definition of the digital transformation model and its implementation to facilitate the adoption of technologies in the teaching and learning process in state universities in Cameroon.

The rest of the paper is organized as follows: in Section II, we present the existing solutions that address the highlighted problem and their limitations; Section III presents some preliminary concepts that are useful for understanding our proposal; Section IV presents our proposal; Section V presents the case study; Section VI presents the results and discussions; and Section VII concludes the paper and gives some research perspectives.

II. Related Works

Information system researchers use many theories or approaches to underpin the issues related to information and communication technology (ICT) such as theoretical aspects in information system, cultural issues, social issues, digital transformation, ICT projects and education. In the specific case of education, this section provides literature review of theories or approaches used in m-learning such as Activity Theory, Unified Theory of Acceptance and Use of Technology (UTAUT).

A. Activity Theory

The Russian psychologist Vygotsky and his disciples had introduced Activity Theory in the 1920s to provide a lens that can help researchers better understand human activity. Activity theory is based on social-cultural theory. Social-cultural theory is based on the notion that an individual cannot be viewed as separate from his or her social and cultural environment [8]. In this social-cultural environment, relationships can be formed between subject (human doer) and object (the things being done). These relations form the core of an activity.

Activity is the core concept in this theory. Activity is defined by the dialectical relationship between subject and object [14, 19]. Later, in the 1981s, Leontiev presents activity as a holistic [1, 9], high-level, usually collaborative construct, such as undertaking a work project, teaching a course, or doing a PhD, and proposes a representation of the collective activity system (Fig. 1). This representation is used by researchers as the basic model of Activity Theory [2, 4, 16].

Activity Theory is applied in research that studied the complexity of problems in the real world such as community group, places of learning, etc. The Activity Theory provides a language and a set of frameworks for making sense of what is discovered about the situation through observation, interviews, and other methods [4, 17]. To use Activity Theory in research problem, it is recommended to researchers to follow the steps below:

Step 1: Identify the significant activities of the system to be examined, along with the subject(s), object, and purpose of each activity.

Step 2: Identify the actions and mediating tools of the activity(ies), where tools may be primary, secondary, or tertiary.

Step 3: Identify the dynamics and tensions within and between the identified activities (Hasan and Kazlauskas, 2014).

These main steps have been followed by Eltyad *et al.* [8] to build the task model for mobile learning and to understand the great opportunities that mobile technologies can offer to mobile learners. In their work, learning is considered as an activity that can be performed in different contexts: independent context, formalizing context, physical context and socializing context. Fig. 1 shows a representation of a collective activity system.

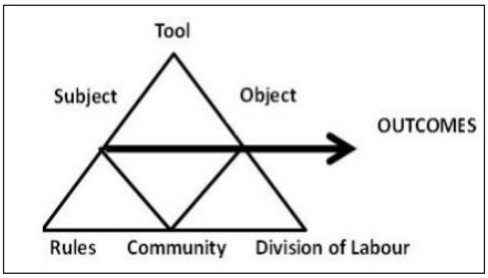


Fig. 1: The Engeström Representation of a Collective Activity System [18]

B. Unified Theory of Technology Adoption and Use

Fezil [11] developed a unified model that brings together alternative views of user and innovation acceptance. The Unified Theory of Acceptance and Use of Technology (UTAUT), which suggests that four core constructs (performance expectancy, effort expectancy, social influence, and facilitating conditions) are direct determinants of behavioral intention and ultimately behavior, and that these constructs are in turn moderated by gender, age, experience, and voluntariness of use [3, 14, 16]. The UTAUT was constructed by reviewing and integrating eight major models and theories which are: The PC Utilization Model, the Motivation Model, the Technology Acceptance Model (TAM), the Theory of Planned Behavior (TPB), a combined TBP/TAM, the Social Cognitive Theory (SCT), the Theory of Reasoned Action (TRA), and the Innovation Diffusion Theory (IDT). Fig. 2 shows the unified theory of technology acceptance and use [4, 17].

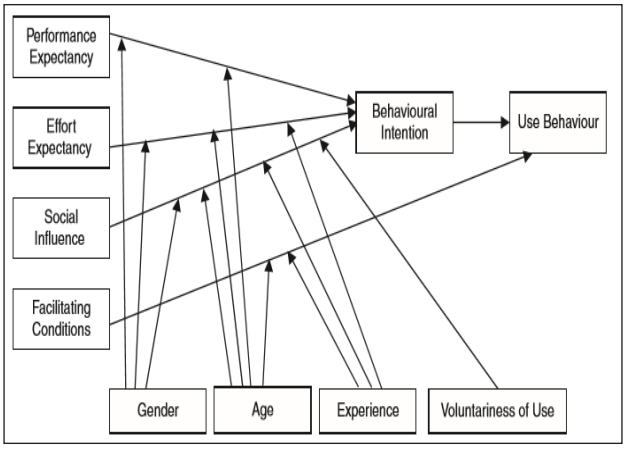


Fig. 2: Unified Theory of Technology Adoption and Use

This unified model has been used in m-learning by several authors, such as Liu [13] to develop an adoption model for mobile learning. In his research, Liu [13] indicates that the functionality of learners' devices and the perception of support from learning material providers influence learners' satisfaction and decisions to adopt mobile learning.

III. Generalities of Actor Network Theory (ANT)

Basic Concepts of ANT

A network in ANT is an abstract representation of actors (human and non-human) that interact through translation between them, have their own interests and the same goals. An actor is any entity that interacts with other entities or serves as an intermediary between them. An actor can be seen as a node in a network. A theory is a set of organized ideas, abstract concepts applied to a specific domain.

Actor Network Theory (ANT) is based on three principles: agnosticism, generalized symmetry, and free association [15]. Agnosticism means that analytical impartiality is required towards all actors involved in the project under consideration, whether human or non-human [18]. The construction of an actor network is based on "generalized symmetry," i.e., both human and non-human actors are treated equally and described with the same language [16]. The principle of free association requires the elimination and abandonment of all a priori

distinctions between the technological or natural and the social [11].

Another important concept of ANT is inscription, obligatory passage point (OPP) and translation. Inscription is an act that actors impose on other actors to shape their attitudes and characteristics [6]. OPP is a process by which all actors involved in the network satisfy the interest recognized in them by the main actor in order to make themselves indispensable to the network [17]. Translation is the process by which actors share their interests and construct lasting alliances. According to Pedri [15], this process has four stages: problematization, interessement, enrollment and mobilization. Problematization, which occurs when actors identify a problem and propose a solution [1, 14]. The second stage of translation is interessement, which can be associated with the notion of influence, where the main actor tries to convince other actors to accept the roles assigned to them during problematization [13]. The third stage, enrollment, is a political process where a high level of persuasion by the principal actor is required to convince actors to accept their roles and maintain stability [9]. The last stage is mobilization, which occurs when the solution is widely accepted [4, 6]. The use of these steps as a theoretical framework will be applied in detail in this study in the section below.

Paledi and Alexander [17], use this graphical syntax to construct the conceptual m-learning readiness assessment actor network (Fig. 3), in which actors have been identified through literature review.

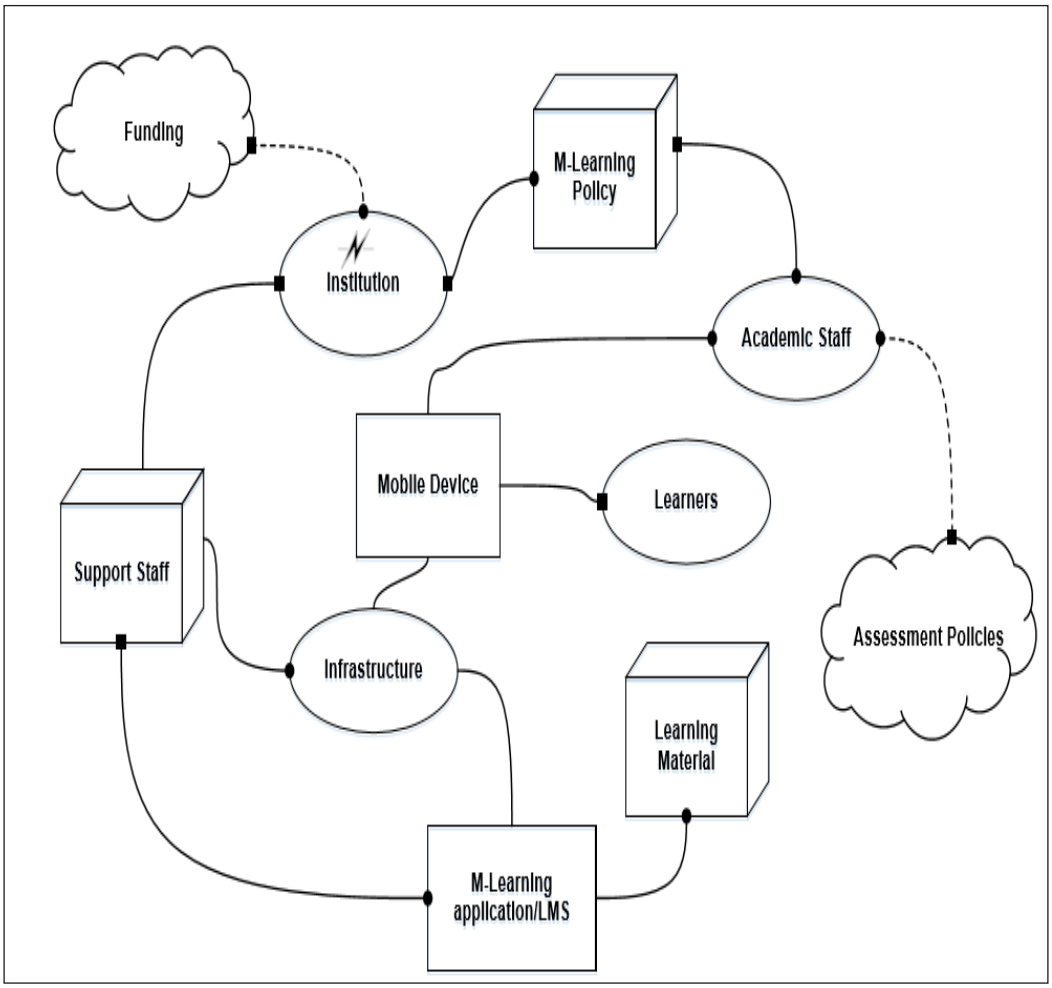


Fig. 3: Conceptual M-Learning Readiness Assessment Actor Network

IV. Digital Transformation Model

The definition of the digital transformation model is based on four stages: the translation, the obligatory passage point, the enrollment, and the mobile learning actor network. These stages are shown in Fig. 4 below:

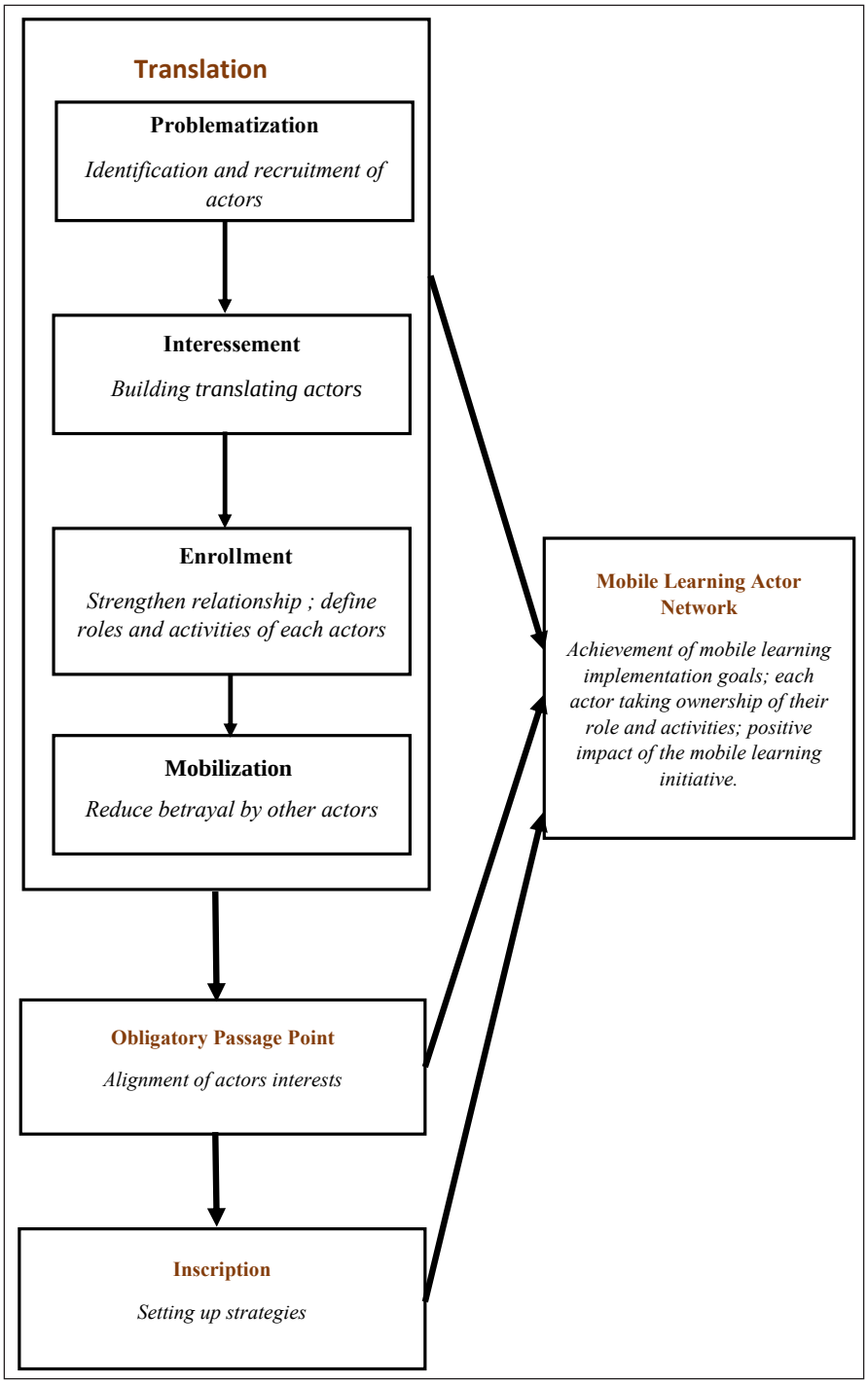


Fig. 4: Digital Transformation Model

A. Translation

Translation is a set of methods by which actors within a network try to put other actors in positions that can serve their own purposes [15]. This set of methods has four steps, as mentioned above: Problematization, Interessement, Enrollment, and

Mobilization. The following section outlines the process of problematization.

i) Problematization

The problematization step involves the identification and recruitment of actors. It is also the process of bringing the most relevant actors into the network [16]. This search provides the main actors that can be recruited in the mobile learning readiness assessment actor network to solve the problem. After identifying the problem, identifying and recruiting actors, the next stage allows to define the negotiations that will take place with other actors regarding the roles they can play within the network.

ii) Interessement

Interessement includes the strategies by which the focal actor tries to include others according to the entities and roles defined for them in the problematization [17].

To achieve these strategies, a lot of negotiation between actors has to take place. That's why interessement can also be defined as the process of building translating actors that can be placed among actors to influence their involvement and interests. These translating actors and other sets of actors that are part of the actor network are not finite. So, boundaries of the network as dynamic, actor should be added to the network to the extent that the problem is addressed [18]. At this level, it is not possible to define exactly the boundaries of any actor's network.

iii) Enrollment

Actor Network Theory is considered as a technology of description [7], where having actors, the most important thing is not what they can do, but how they can be connected to achieve some purposes. The enrollment answers the question of how to define and coordinate the roles of actors. To do this, it is necessary to analyze the activities involved to each actor. So it is important to mention the actors, their roles and activities. Note that the activities led by these actors must contribute to achieving the goals of the actor network. The enrollment process also examines how relationships within the network are strengthened or weakened. These relationships between actors should be represented.

iv) Mobilization

The key process of the mobilization moment is to reduce the betrayal by other actors to ensure that the enrolled actors abide by the commitments made by their representative or either by themselves through the enrollment moment of translation [1, 5, 19]. Pedri and Alexander [16] also argue that the alliance can strengthen or weaken over time and that the strengthening of translation and alliances with a larger group of actors is mobilization.

B. Obligatory Point of Passage (OPP)

The OPP is centered around the primary actor [16, 19]. In this model, we assume that OPP is an actor(s) that can influence the other actors to participate in digital transformation and without them, interests of some actors may be included, activities involved to the network may not take place. The consequence is that the goals of the actor network will not be achieved. Paledi and Alexander [16] argue that OPP can change over time and when the network stabilizes.

C. Inscription

The notion of inscription is influence by: a consideration of previous patterns of technological usage, the role of users, the actors' actions which are influenced by the technological functions [11, 19]. Inscription includes identifying attributes to align actors' interest in order to stabilize the actor network.

V. A Case Study at a State University in Cameroon

There are 11 public universities in Cameroon, which have more or less the same structure. Thus, the generic model of a university as seen in this study allows us to take into account the basic elements found in these universities. To this end, the following descriptions allow us to include all these state universities.

A. Problematization

In our case study, we provide the main actors that can be recruited to facilitate digital transformation in state universities in Cameroon, in order to assist them in scoping with the overcrowded students on campus in universities. For a given state university, the following actors have been identified: university, top management, student, teacher, technology, learning material, learning management system, m-learning policy, implementation plan, partner, academicals center of technology of information, and we adapt this review of actors to construct the conceptual actor network by adding state as black box actor and parent as ordinary actor. All these actors have their roles, own interests and goals that they need to achieve. These goals are linked to activities they can perform. However, they may have obstacles that prevent them from achieving their goals. That's the reason why they need to coordinate their interests with other actors, and various solution will be sought.

B. Interessement

Interest encompasses the strategies by which the focal actor attempts to include others according to the entities and roles defined for them in the problematization. Some of these strategies in our case study can be described as follows:

- Raising awareness among students and teachers about the importance and benefits of digital transformation;
- Creating a platform where students have to download their courses with their mobile devices;
- Establishing a digital transformation policy and implementation plan in state universities;
- To find the availability of sufficient funds that can support the digital transformation initiative.

Conceptually, we can see that our network concerns all the actors inside or outside the state universities that can improve or involve learning and teaching practices. We can identify the following translating actors in a given state university in Cameroon: Dashboard, communication channel, computer science department, governance strategy, administrative document and agreement.

C. Enrollment

In our defined model, we have mentioned in Table I the actors, their roles and activities that can help us to build the digital transformation actor network of state universities. Note that the activities led by these actors must contribute to achieving the goals of the digital transformation actor network.

Table I: Actor Recruitment and Activities

Actors	Activities
1. Top management (Board of Director, Senate, Rector, Vice-Rector, Head of Department)	• Define an appropriate strategy for implementing digital transformation. • Define the modality to implement an m-learning platform. • Define the session of retraining teachers in distance learning pedagogical approach. • Develop proposals for funding to support an m-learning research project. • Define m-learning policies. • Define guidelines for m-learning use and adoption.
2. Teacher	• Participate in the recycling of the distance learning approach. • Provide the content of their course in digital support as a video. • Explain m-learning to learners and increase their motivation.
3. Student	• Changing attitudes towards m-learning and understanding the principle of learner-centered approach.
4. Academic Center of Technology of Information	• Lay the groundwork for successfully enhancing learning with mobile technology. • Support key players in the implementation plan. • Ensure maintenance of student and teacher devices.

Actors	Activities
5. Technology (mobile devices and IT infrastructure)	• Availability and serviceability.
6. Learning materials (online course, pen, lab materials, etc.)	• Availability, useful and usable.
7. Learning Management System (LMS)	• Must support the exchange of information between learners and other learners, and between teachers and learners. • Availability of the platform. • Access to content.
8. M-learning policy	• Student and teacher reporting requirements.
9. Staff	• Improve the profitability of public services by leveraging mobile technology.
10. Partner	• Support research funding and sign agreement.
11. Parent	• Changing attitudes toward m-learning. • Creating an in-house digital workspace for mobile learners.
12. State	• Democratize access to the Internet. • Approval and recognition of certificates based on the pedagogical approach of mobile learning. • Install free public Wi-Fi. • Provide mobile devices to students and teachers.
13. Governance Strategy	• Provide guidance on how to operationalize and institutionalize mobile learning. • Determine the pay scale for teachers.
14. Implementation plan	• Provide guidance on how m-learning should be implemented.
15. University	• Acquisition of technological infrastructure. • Availability of funds. • Partnership with external parties.

The enrollment process also examines how relationships within the network are strengthened or weakened. These relationships between actors are shown in Table II below:

Table II: Alliances and Actor Roles

Actor 1	Actor 2	Alliances/Relationships
Top Management (Source actor)	M-learning policy (Target actor)	Define digital transformation policies; Take actions to effectively apply these digital transformation policies.
Top management (Source actor)	Dashboard (Translator actor)	Consultation
Dashboard (Translator actor)	Learning Management System (Target actor)	Integrated into LMS
Teacher (Source)	LMS (Translator actor)	Sharing information
LMS (Translator)	Student (Target actor)	Sharing information
Top management (Source actor)	Technology (Target actor)	Ensure technology availability (Internet connection).
Teacher (Source actor)	Top management (Target actor)	Apply management decisions.

Actor 1	Actor 2	Alliances/Relationships
Top management (Source actor)	Computer Science Department (Translator actor)	Supervise m-learning research.
Computer Science Department (Translator)	Student (Target actor)	Valorization of m-learning research.
Teacher (Source actor)	LMS (Target actor)	Accessibility
Parent (Source actor)	Communication duct (Translator actor)	Get information about m-learning.
Communication duct (Translator actor)	University (Target actor)	Get information about m-learning.
Academicals center of technology of information (Source actor)	Implementation plan (Translator actor)	Develop implementation plan.
Implementation plan (Translator actor)	Top management (Target actor)	Provide technological requirement.
Student (Source actor)	Learning Material (Translator actor)	Download content
Learning Material (Translator actor)	LMS (Target actor)	Accessibility
Technology (Translator actor)	Learning Material (Target actor)	Accessibility
Teacher (Source actor)	Student (Target actor)	Preparing mindsets and changing attitudes toward m-learning.
University (Source actor)	Governance strategy (Translator actor)	Approve and update governance strategies.
Governance strategy (Translator actor)	Partner (Target actor)	Provide credibility to obtain funding; Financial support; Collaborative research.
State (Source actor)	Governance strategy (Translator actor)	Take measures to evaluate the work of teachers in the new education system based on m-learning.
Governance strategy (Translator actor)	University (Target actor)	Increase institutional credibility of the university.
State (Source actor)	University (Translator actor)	Provide technology infrastructure.
University (Translator)	Student (Target actor)	Deploy Technologies (Mobile Devices).
State (Source actor)	Technology (Target actor)	Create free public wifi.
Parent (Source actor)	Student (Target actor)	Provide internet connectivity to mobile learners at home.

D. Mobilization

The research reveals two weaken relationships which are: relationship between parent, communication channel, university and top management, dashboard and LMS. The break of these alliances cannot negatively affect the achievement of mobile learning initiative goals. Top management, student, teacher, m-learning policy, state, technology, partner, university, LMS, governance strategy, implementation plan, academicals center of technology of information, and learning material are the large group of actors who have strengthen alliances among them. We can agree that in our case study mobilization occurs.

E. Obligatory Point of Passage (OPP)

Despite the fact that a state university is not autonomous and in practice human actors are those who pilot all strategies for improving the organizational process, the analysis of the existing procedures and processes of a given state university reveals that top management, technology, implementation plan and m-learning policy are the OPP of this actor network.

F. Inscription

The inscription includes the identification of attributes to align the interest of actors in order to stabilize the m-learning readiness assessment actor network. These attributes reflect the themes for stabilizing the m-learning actor network. To find these themes, we were inspired by the UNESCO policy guidelines for mobile learning and literature review. After reading and analyzing the content of UNESCO policy, considering the context of learning system of a certain state university, we have mentioned the following inscription theme, which are not discussed in this paper. We can mention: updating the m-learning policy, regularly retraining teachers in teaching techniques using mobile technologies, optimizing educational content for use on mobile devices, ensuring gender equality for mobile students, awareness strategy, reward strategy, and ease of use of the application. Thus, we obtain the following network of actors of a public university in Cameroon, on which the digital transformation model presented above is based.

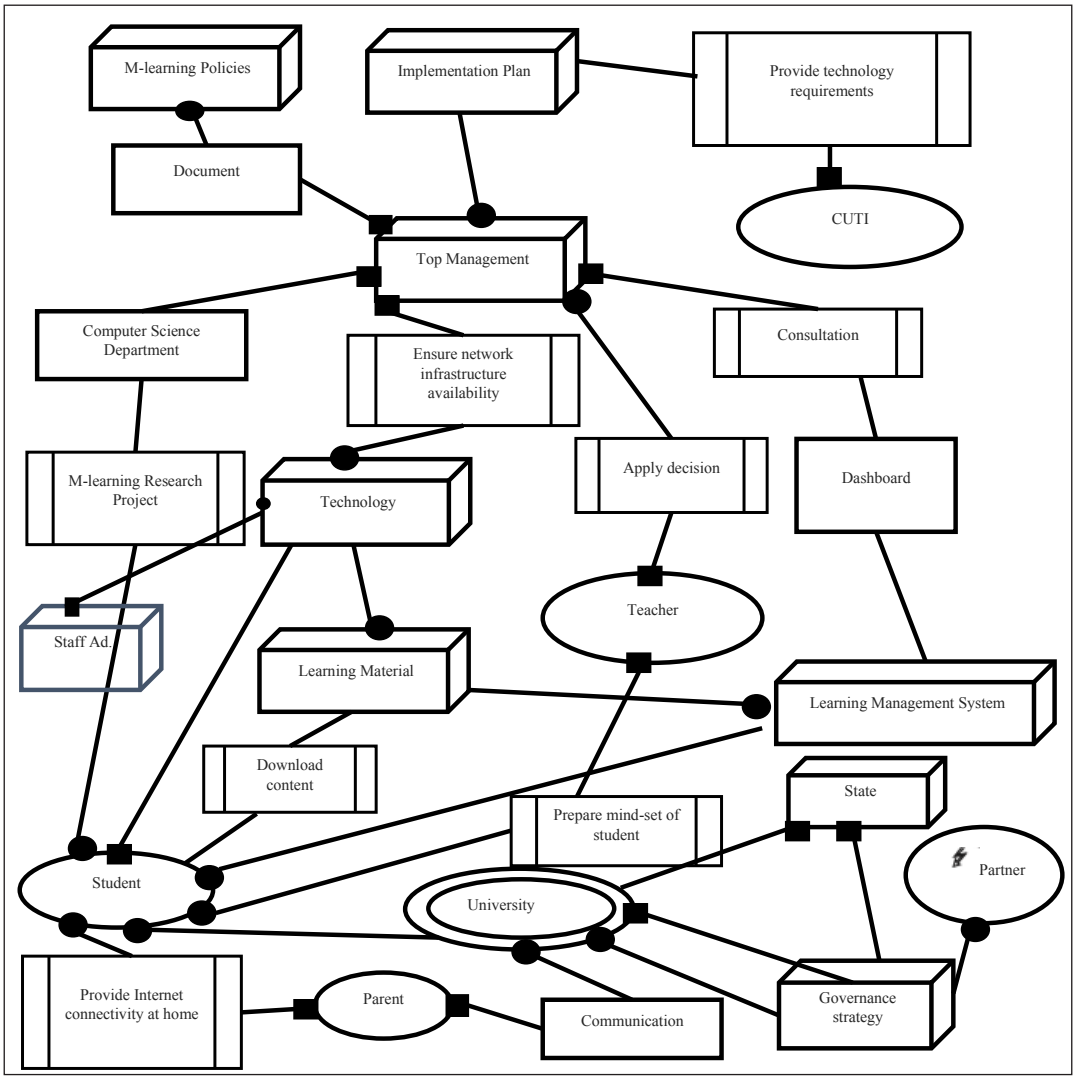


Fig. 5: Actor Network of a State University in Cameroon

VI. Results and Discussions

This part will be divided into two main ideas. First, we will explain the choice of some types of actors and second, we will analyze the types of translation between actors to better understand how this network of actors can contribute to the digital transformation in state universities in Cameroon.

Learning material is a black box actor because it includes all materials used by students or lecturers in all faculties of a state university. State is also considered as a black box actor because it takes into consideration a lot of actors that directly influence decisions that can be applied in a given university and lifelong of students and teachers. This actor includes the following actors: ministry of higher education, social media, local government, community, etc. It's also important to add parents like an actor, because he must first be prepared to accept the fact that they can see their soon should be all the time in the home or in mobile learning environment space constructed by the state; this idea, which indicates that the context in which learning occurs is clearly an important factor.

Thus, this digital transformation model of state universities gives a conceptual view of the different actors that can influence the adoption of m-learning in a given state university in Cameroon. Its main objective is to show whether a state university is ready to adopt m-learning as a pedagogical approach.

Thus, in order to implement a digital transformation model in a given state university in Cameroon, the top management of this university should make an inventory of the existing technological equipment, the level of research enhancement in this area, to see what remains to be done and to seek help from partners.

Top management also develops funding proposals to support student-led m-learning research projects through the Department of Computer Science. In addition, top management convenes recycling sessions of m-learning for teachers through mobile devices. These training sessions will be most important for teachers other than those from IT department. At the end of this retraining, data record should be collected and analyzed to see the percentage of teachers who can be m-learning ready. Another aspect in which top management can be a crucial actor in this network assessment is to revise the existing learning policies to implement m-learning policies, thence a strengthening relationship is established between top management and m-learning policies through administrative document. This revision of the existing legislative text contributes to the assessment of m-learning readiness of state universities. Further, Academicals Center of Technology of Information of each state university should provide to top management the report watchfulness of technology; with this report he can know the real level for integration of mobile technology in the university compared to other university and there, these actors should share their interest, finally to know what remains to do. Moreover, top management makes sure that IT infrastructure is available and smooth.

Academic staff is also a black box. We choose academic staff as black box actor because due to the fact that it is composed of person who render a public service, friendship relationship can be created among them. So, we use a black box to simplify the diagram. LMS aggregates all those applications and software that ensure learning process in a state university. In addition, communication channel influence on student's perception and teacher's perception towards m-learning. Communication channel is view in this actor network like translating actor because it allows parents to align their interests in the network. This actor can be represented by social media, ministerial regulation, etc.

In this actor network we can also mention the presence of some ordinary actors like teacher and student. Teachers provide the content of their courses to students through LMS and students are supposed to download these courses through LMS. These relationships are strengthened in this network and these actors have common interests: to improve the process of learning and teaching.

VII. Conclusion

This study was focused on the definition of the digital transformation model to facilitate the adoption of

technologies in the teaching and learning process in state universities in Cameroon. Thus, to carry out this work, first, we have briefly reviewed three main theories commonly applied in m-learning; but we decide to work with ANT because this theory take into account the presence of human or non-human entity to analyze the problem encountered by the organization due to technological mutation. Also, it focuses more on the role that actor should assume to perfectly achieve the goals of digital transformation imposed in organization. While activity theory doesn't really manage the impact of technology in settlement process of problem of real worlds. UTAUT is most accentuated in the behavioral aspects of people in some organization towards technology. Using ANT, we have traveled all the stages of translation by showing how these steps could be theoretically applied. Furthermore, this paper presents the basis for an effective integration of m-learning in a given state university through the digital transformation actor network. This study shows that factors that can contribute to digital transformation are: organizational m-readiness, teacher m-readiness, student m-readiness, and society m-readiness. So, in the specific context of the state universities in Cameroon, many efforts have been done concerning organizational M-readiness, but the actions remain to lead on teacher namely: the new role of teacher (e-tutor, tutor at distance, etc.), supplementary hour and grid of salary. Moreover, the remuneration of e-tutor depends on the time spent on course preparation, interaction and feedback with students. Regarding students' readiness, although majority of students have a smartphone, their m-learning perception still needs to be analyzed. M-learning regulation and large campaign of m-learning awareness for students and teachers are also necessary. The proposed digital transformation model also reveals that a state university must wake up to find the real adequacy between digital ability of students and employment; to form students to documentary research and evaluation of the quality of contents. This paper also indicates that the organigram of organization is the tool that can help us to select actors that can have important role in actor network. This study also shows that an OPP can be the focal actor. From an economic point of view, this research reduces expenses in the sense that the number of key actors that can receive funding for activities to be managed is known in advance and limited; it is not necessary to find an external mobile learning expert. However, this study doesn't study the enrollment phase of ANT in detail; this can be the perspectives of the future work.

References

[1] K. Adiyarta, D. Napitupulu, R. Rahim, D. Abdullah, and M. I. Setiawan, "Analysis of e-learning implementation readiness based on integrated ELR model," *Journal of Physics: Conference Series,* vol. 1007, pp. 1-6, 2018.

[2] M. Ally, "What is mobile learning, mobile learning policies and technologies," 2004. [Online]. Available: http://m4d.colfinder.org

[3] M. Ally, "Benefits of mobile learning." [Online]. Available: http://m4d.colfinder.org

[4] A. F. Ariani, D. Napitupulu, R. K. Jati, J. A. Kadar, and M. Syafrullah, "Testing of technology readiness index model based on exploratory factor analysis approach," *Journal of Physics*: *Conference Series,* vol. 1007, pp. 1-8, 2018.

[5] C. Baujard, "L'adoption du e-learning selon la grounded theory, facteur de coherence des pratiques de formation dans les entreprises," vol. 9, no. 4, 2004. [Online]. Available: http://dx.doi.org/10.9876%2Fsim.v9i4.164

[6] C. C. Chuva, and P. Rupino, "The social dimension of business models: An actor-network theory perspective," *Twenty-first Americas Conference on Information Systems,* Puerto Rico, 2015, pp. 1-12.

[7] A. Dussart, "Effets de la transition numérique sur le secteur de l'éducation et de la formation en termes d'activités, métiers et competences," *Le Forem, Veille, Analyse & Prospective du Marché de l'Emploi Secteur de l'Education*, pp. 1-22, 2016.

[8] H. M. Eltayeb, and M. O. Ali Hegazi, "Mobile learning aspects and readiness," *International Journal of Computer Applications*, vol. 103, no. 11, pp. 22-28, 2014.

[9] E. Silvis, and P. M. Alexander, "A study using a graphical syntax for actor-network theory," *Information Technology & People*, vol. 27, no. 2, pp. 110-128, 2014.

[10] J. Fame Ndongo et al., Normes Universitaires Applicables aux Etablissements d'Enseignement Supérieur au Cameroun, 2015.

[11] F. Odzamli, and N. Cavus, "Basic elements and characteristics of mobile learning," *Proceedia-Social and Behavioral Sciences*, vol. 28, pp. 937-942, 2011.

[12] M. Fouda Ndjodo, and A. Onana, "Rapport de recherche sur les reformes de gouvernance dans l'enseignement supérieur camerounais," UNESCO, International Institute for Educational Planning, 2012.

[13] M. Kang, B. Y. T. Liew, H. Lim, J. Jang, and S. Lee, "Investigating the determinants of mobile learning acceptance in Korea using UTAUT2," in *Proceeding of World Conference on Educational Multimedia, Hypermedia and Telecommunication*, 2013, pp. 1425-1430.

[14] H. A. Oketch, "E-learning readiness assessment model in Kenyas' higher education institutions: Acase study of University of Nairobi," *International Journal of Scientific Knowledge,* pp. 29-41, 2014.

[15] V. N. Paledi, and P. M. Alexander, "A graphical representation of an actor network: An example based on a literature review for m-learning readiness assessment," in *Proceedings of the Second International Conference on Information and Communication Technology for Africa Development*, 2017, pp. 1-16.

[16] V. N. Paledi, "Developing a mobile learning readiness assessment model within three South Africa universities based on actor network theory," PhD Thesis, University of South Africa, South Africa, 2018.

[17] V. N. Paledi, and P. M. Alexander, "Inscribed themes aligning actor's interests to influence m-learning readiness in higher education," 2018. [Online]. Available: www.researchgate.net

[18] A. Tatnall, and A. Gilding, "Actor-network theory and information systems research," in *Proc. 10th Australasian Conference on Information Systems*, 1999, pp. 955-966.

[19] Y. L. Theng, "Mobile learning for tertiary students: An exploratory study of acceptance of use," in *Proceeding of World Conference on Educational Multimedia, Hypermedia and Telecommunication,* 2009, pp. 937-946.

International Journal of Distributed and Cloud Computing
11 (1) 2023, 23-37
http://www.publishingindia.com/ijdcc

Cloud Usage Optimization through Data Analysis and Visualization: A Case Study for Company X in 2016

Samindar J. Vibhute[1*], Anil B. Desai[2] and Rohini P. Jadhav[3]
[1]Department of Computer Science & Engineering, Ashokrao Mane Group of Institutions, Vathar, Kolhapur, Maharashtra, India. Email: sjv@amgoi.edu.in
[2]Department of Computer Science & Engineering, Ashokrao Mane Group of Institutions, Vathar, Kolhapur, Maharashtra, India. Email: abd@amgoi.edu.in
[3]Department of Computer Science & Engineering, Ashokrao Mane Group of Institutions, Vathar, Kolhapur, Maharashtra, India. Email: rpj@amgoi.edu.in
[*]Corresponding Author

Abstract: Cloud computing has revolutionized the IT landscape, offering businesses scalability and cost-efficiency. To enhance cloud usage efficiency and optimize costs, companies need to analyze their cloud usage data thoroughly. This research paper presents a comprehensive analysis of a company's cloud usage data for the year 2016, focusing on CPU core, CPU capacity, Memory usage, Disk read, Disk Write, and transmitted throughput. The study utilizes a dataset containing 186 days of cloud usage, encompassing essential metrics and date-wise details. The primary objective of this research is to provide insights into cloud resource utilization patterns, identify high-usage instances, and detect instances with zero memory usage and disk throughput. The analysis leverages Google Looker Studio to create interactive charts and heat maps, facilitating a deeper understanding of the data and enabling data-driven decision-making. The paper discusses the data collection process, data cleaning, and the extraction of relevant features. The results showcase date-wise trends of memory usage and processor utilization, arranged in ascending order to highlight the most and least utilized instances. Moreover, the study demonstrates instances with zero memory usage and disk throughput, shedding light on opportunities for service optimization and resource cost reduction. The findings demonstrate the effectiveness of data analysis and visualization in optimizing cloud services for Company X. By identifying heavily utilized instances and instances with negligible usage, the company can make informed decisions to optimize resource allocation and enhance overall system efficiency

Keywords: Cloud computing, Google studio, Resource allocation.

I. Background of the Study

Cloud computing has emerged as a transformative technology, revolutionizing the way organizations manage their IT infrastructure and services. With its inherent scalability, flexibility, and cost-effectiveness, cloud adoption has skyrocketed in recent years, offering businesses the promise of enhanced operational efficiency and optimized resource utilization. However, as companies migrate their operations to the cloud, there arises a pressing need to analyze and optimize cloud usage to harness its full potential. This research paper embarks on a data-driven journey to explore and uncover valuable insights into the cloud usage patterns of Company X during the year 2016.

A. Background and Motivation

As a leading innovator in its industry, Company X has been at the forefront of technological advancements, leveraging cloud services to meet its growing IT demands. With cloud infrastructure serving as the backbone of their operations, understanding and

optimizing cloud resource allocation have become crucial for Company X's competitiveness and cost-effectiveness. The motivation behind this research stems from the need to harness the power of data analysis and visualization to unlock the secrets hidden within Company X's cloud usage data.

B. Objectives

The primary objective of this research is to delve into the vast troves of cloud usage data and extract meaningful insights that can guide Company X towards enhanced cloud usage optimization. The research aims to achieve the following key objectives:

- *Data Collection and Preparation:* Gather and preprocess the extensive dataset of cloud usage, covering 186 days of operational history. This dataset encompasses a myriad of metrics, including CPU core, CPU capacity, Memory usage, Disk read, Disk Write, and transmitted throughput, alongside date-wise records.
- *Exploratory Data Analysis:* Conduct a comprehensive exploratory analysis to uncover trends, patterns, and potential anomalies within the data. The analysis will be focused on identifying the most and least utilized instances, as well as instances with zero memory usage and disk throughput.
- *Data Visualization:* Employ cutting-edge data visualization techniques, leveraging Google Looker Studio, to transform complex data into intuitive visual representations. Interactive charts and heat maps will facilitate a better understanding of the data, enabling Company X's stakeholders to make data-driven decisions.
- *Cost Optimization Strategies:* Develop targeted cost optimization strategies based on the findings of the analysis. By identifying heavily utilized instances and instances with negligible usage, the research will offer actionable insights to enhance resource allocation and reduce unnecessary costs.

C. Data Description

The dataset acquired for this study comprises 186 days' worth of cloud usage data for Company X in 2016. Each record includes essential metrics, such as CPU core, CPU capacity, Memory usage, Disk read, Disk Write, and transmitted throughput, accompanied by a corresponding date. To ensure data accuracy and relevance, rigorous data cleaning and preparation processes have been employed, removing any redundant or irrelevant features.

D. Structure of the Paper

This research paper is organized into distinct sections to present a comprehensive analysis of Company X's cloud usage. Section I introduces the background, motivation, objectives, and data description of the research. Section II presents the literature review. Section III elaborates on the data collection process, data preprocessing, data analysis methodology. Section IV presents the findings of the exploratory data analysis, offering valuable insights into cloud usage trends. Section V showcases the data visualization results, highlighting the effectiveness of Google Looker Studio in communicating complex information visually. Section VI outlines the cost optimization strategies derived from the analysis, empowering Company X to make informed decisions. Section VII includes methods of implementation. Section VIII explains results snaps memory usage and Finally, Section IX concludes the paper with a summary of key findings and implications, as well as potential future research directions.

In the pursuit of unraveling the cloud and optimizing Company X's cloud usage, this research contributes to the broader knowledge in cloud computing, data analysis, and cost optimization. The insights gained from this research endeavor have the potential to shape industry practices and revolutionize cloud resource management for years to come.

This is an existing technique for optimizing the cloud if we get these all techniques on one platform and with no subscription module, then it will be an efficient technique for cloud optimization.

Following techniques on a single platform:

- Close/Manual monitoring
- Use discounts
- Azure advisor
- Auto start stop
- Low and High peak traffic

- Manual analysis

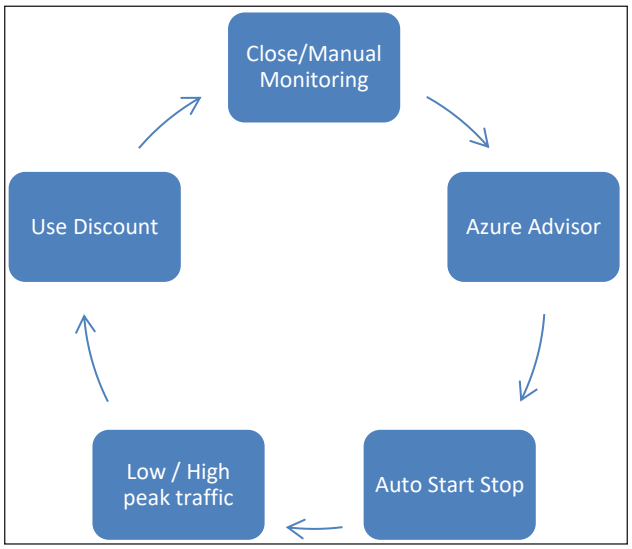


Fig. 1: Existing Techniques for Optimization

"Change Brings a Problem" refers to the phenomenon in which Cloud Computing is an innovative and impressive change for human being but there may some problems and issues also and we defeat it. Problem is about Optimizing cloud services and cost optimization. And we bring a solution for that.

II. Literature Review

Vadivel Ramasamy and Sudalai Muthu Thalavai Pillai, "An Effective HPSO-MGA Optimization Algorithm for Dynamic Resource Allocation in Cloud Environment," we review and cite relevant existing studies. If you have access to the paper you mentioned, you can extract specific information and citations from it to strengthen your review. If not, you can search online databases and journals for related papers on topics such as cloud resource allocation, optimization algorithms, and dynamic environments [1].

Xingwei Wang, Xueyi Wang and Hao Che, "An Intelligent Economic Approach for Dynamic Resource Allocation in Cloud Services," he proposed Intelligent Economic approach for Dynamic Resource Allocation (IEDA) addresses the pressing challenges in the competitive cloud market. Through a combinatorial double auction protocol, IEDA allows simultaneous trading of various resources among multiple consumers and providers. The integration of BPNN-based price prediction and a price matching algorithm enhances the auction process's efficiency. The inclusion of a reputation system fosters market integrity by identifying and excluding dishonest participants [2].

Aditya Bhardwaj and C. Rama Krishna, "Impact of Factors Affecting Pre-copy Virtual Machine Migration Technique for Cloud Computing," Virtual machine (VM) migration techniques play a vital role in optimizing resource utilization and ensuring efficient performance in cloud computing environments. The impact of various factors on the effectiveness of these techniques, especially pre-copy migration, has garnered significant research attention. This section reviews existing literature related to the factors affecting pre-copy VM migration techniques in cloud computing [3].

Isaac Odun-Ayo, "Cloud Management Optimization" the literature collectively underscores the critical role of cloud management optimization in ensuring efficient resource utilization, cost-effectiveness, and performance in cloud computing. The studies highlighted demonstrate the wide range of strategies, from resource allocation to cost management, that contribute to achieving these goals. While the specific work by Isaac Odun-Ayo on "Cloud Management Optimization" may provide deeper insights, this general literature review provides a foundation for understanding the diverse facets and challenges within this domain [4].

Ilyas Iyoob and Emrah Zarifoglu, "Cloud Computing Operations Research," A. B. Dieker topic would likely involve exploring existing research that has been conducted at the intersection of cloud computing and operations research. This could include studies on optimization algorithms for virtual machine placement, energy-efficient data center management, scheduling of cloud resources, and other related topics [5].

Eli Weintraub, "Cost Optimization of Cloud Computing Services in a Networked Environment," we focus on cost optimization from the perspective of cloud computing service customers. It highlights how cloud services are organized into hierarchies, comprising software, platform, and infrastructure

services. The current practice of bundling these services hinders cost minimization for customers and contradicts fair pricing principles. The paper envisions a competitive market where customers have the freedom to switch between service providers, leading to improved quality, open standards, and diverse service offerings. The proposed model assists potential customers in finding the most cost-effective combination of service providers. Three strategies for implementing the model are presented. The mathematical model is formulated, showcasing its advantages over existing pricing practices in cloud computing [6].

Rajnikant B. Wagh, "Priority Based Dynamic Resource Allocation in Cloud Computing," Cloud computing's popularity stems from its dynamic resource allocation, ensuring reliable services with flexible pay-as-you-use models. Efficiently meeting users' needs requires adhering to Service Level Agreements (SLAs) and utilizing parallel processing. Existing approaches have mainly focused on single SLA parameters, lacking comprehensive optimization. To address this, the paper introduces an algorithm that incorporates pre-emptible task execution and considers multiple SLA parameters like memory, network bandwidth, and CPU time. Experimental results demonstrate that in scenarios with intense resource contention, the proposed algorithm significantly enhances resource utilization, outperforming existing methods [7].

U. K. Jena, P. K. Das and M. R. Kabat, "The Hybridization of Meta-Heuristic Algorithm for Load Balancing in Cloud Computing Environment," refers to combining different optimization techniques to improve load balancing in cloud computing. Load balancing ensures that computing resources are distributed efficiently among virtual machines (VMs) to enhance performance. The hybrid approach integrates various meta-heuristic algorithms, which are problem-solving techniques inspired by natural phenomena, such as genetic algorithms, particle swarm optimization, or simulated annealing. By combining these algorithms, the hybrid approach aims to achieve better load distribution, optimize resource utilization, and enhance the overall efficiency of cloud computing systems. This paper explores how the fusion of these algorithms can provide superior load balancing solutions compared to using a single algorithm alone [8].

"A Review on Optimization Approaches in Cloud Computing Services," provides an overview of different optimization strategies used within cloud computing. The review covers various techniques aimed at improving resource allocation, performance, and cost-effectiveness in cloud services. It surveys optimization methods like genetic algorithms, ant colony optimization, and particle swarm optimization. The paper assesses the strengths and limitations of each approach, highlighting their applicability to different cloud computing scenarios. By summarizing these optimization techniques, the paper helps readers understand the landscape of optimization methods used to enhance cloud computing services [9].

N. Sujata Kumari, K. Divya Charitha Yadav, Devu Asha and M. Deepika, "A Light Weight Secure Data Sharing Scheme for Mobile Cloud Computing," with the rise of cloud computing, accessing personal data from mobile devices has become more convenient. However, data security concerns hinder its progress. Despite extensive research on enhancing cloud security, many solutions are not suitable for mobile clouds due to limited device capabilities. To address this, a lightweight data sharing scheme (LDSS) is proposed for mobile cloud applications. LDSS employs CP-ABE access control, adapting the structure for mobile clouds. It offloads computationally intensive tasks to external servers, reducing user revocation costs. The system also addresses lazy revocation using attribute description fields. Attribute-based encryption (ABE) is explored as an approach to public key cryptography. In ABE, keys and cipher-texts are based on attributes, allowing quicker encryption and decryption with fewer keys. This proposed LDSS framework aims to enhance mobile cloud security by leveraging adapted access control and attribute-based encryption [10].

Mohit Kumar and S. C. Sharma, "Dynamic Load Balancing Algorithm for Balancing the Workload among Virtual Machine in Cloud Computing," Cloud computing is a transformative IT technology

that has gained momentum over recent years due to its scalability, flexibility, and on-demand services delivered via the internet. With an increasing user base, load balancing has emerged as a challenge for cloud service providers. Various algorithms have been proposed to address this issue, each with its own strengths and limitations. Load balancing aims to optimize cloud resource utilization for enhanced throughput and performance. It achieves this by distributing tasks among nodes and monitoring virtual machines for load balance using task or virtual machine migration. Task scheduling, a crucial step, involves assigning tasks to virtual machines for execution within a minimum timeframe. However, cloud task scheduling is complex due to rapidly changing task volumes and lengths, making optimal mapping difficult. Efficient algorithms are needed to ensure balanced virtual machine loads. After task allocation, load balancing is performed to transfer tasks from overloaded to under loaded virtual machines. This three-phase process includes cloud user input, task scheduling, and cloud-level operations involving monitoring and discovering resources. The paper further discusses related work, problem formulation, a proposed load balancing algorithm, simulation tools, experimental results, and concludes with future directions [11].

Shahbaz Afzal and G. Kavitha, "Load Balancing in Cloud Computing – A Hierarchical Taxonomical Classification," a structured classification framework for load balancing techniques in cloud computing. The goal is to organize and categorize various load balancing methods systematically. This taxonomy provides a clear overview of the diverse strategies used to distribute computational workloads efficiently across cloud resources. The authors' approach aids in understanding the landscape of load balancing techniques, their characteristics, and their applicability in different cloud environments. This hierarchical classification is a valuable resource for researchers and practitioners seeking to navigate the complexities of load balancing in cloud computing [12].

Muhammad Junaid, Adnan Sohail, Rao Naveed Bin Rais, Adeel Ahmed, Osman Khalid, Imran Ali Khan, Syed Sajid Hussain and Naveed Ejaz, "Modeling an Optimized Approach for Load Balancing in Cloud," the authors propose an enhanced approach for load balancing in cloud computing. Their focus is on optimizing the distribution of computational tasks across cloud resources. The study involves developing a model that aims to achieve more efficient resource utilization and performance. The paper likely includes details about the proposed approach, such as algorithms or strategies used to achieve load balancing. This optimized approach could potentially address challenges related to uneven resource utilization, improving response times, and enhancing overall cloud system performance. The study contributes to the field of cloud computing by providing a new perspective on load balancing techniques [13].

Patryk Osypanka and Piotr Nawrocki, "Resource Usage Cost Optimization in Cloud Computing Using Machine Learning," the authors propose a method that employs machine learning techniques to optimize resource utilization costs in cloud computing. This approach likely involves utilizing machine learning algorithms to analyze and predict resource usage patterns, allowing for more efficient allocation of cloud resources. The goal is to minimize costs while maintaining performance and scalability. The paper likely presents how machine learning models are trained on historical usage data to make informed decisions about resource allocation. By leveraging machine learning, the study contributes to cost-effective resource management in cloud computing environments [14].

Xiaoyn Wang and Samee U. Khan, "A Robust Optimization Technique for Energy Cost Minimization of Cloud Data Centers," the authors introduce a method to minimize energy costs in cloud data centers. The focus is on developing a robust optimization technique that considers uncertainties and variations in factors like workload and energy prices. The proposed approach likely involves formulating mathematical models that account for these uncertainties to make efficient energy management decisions. By incorporating robust optimization, the method aims to ensure cost minimization even under changing conditions. This study contributes to improving the sustainability and

cost-effectiveness of cloud data centers by addressing energy consumption challenges [15].

III. Data Collection and Methodology

A. Data Collection Process

The data collection process for this research involved obtaining the cloud usage dataset for Company X during the year 2016 from Kaggle, a well-known platform for datasets and data science competitions. The dataset comprises 186 days of detailed cloud usage records, including crucial metrics such as CPU core, CPU capacity, Memory usage, Disk read, Disk Write, and transmitted throughput. Each entry in the dataset is associated with a specific date, enabling the temporal analysis of cloud usage patterns.

The data acquisition from Kaggle ensured that the dataset adhered to the highest standards of data integrity and authenticity. To guarantee the privacy and confidentiality of sensitive information, any identifiable information was anonymized or removed in compliance with ethical data usage practices.

B. Data Preprocessing

Before proceeding with data analysis and visualization, an essential step was data preprocessing. This phase involved handling missing or erroneous data points and transforming the data into a suitable format for analysis. Cleaning the dataset ensured the accuracy and reliability of the subsequent analysis.

During data preprocessing, any missing or null values were addressed through appropriate imputation methods to minimize data loss. Additionally, the data was converted to a consistent unit of measurement for better comparability and meaningful visualization. For instance, Disk read and Disk Write data were transformed from raw bytes to kilobytes per second (kb/sec) to provide a more intuitive representation.

Furthermore, any redundant or irrelevant features that did not contribute significantly to the research objectives were removed, streamlining the dataset for further analysis. This optimization allowed for a more focused exploration of relevant cloud usage metrics.

C. Data Analysis Methodology

This study aims to identify the cloud optimization technique and analyse data usage of cloud and giving the visual representation to user. So user can get an idea of which instances are having the more input output write applications and which are not using in that manner. Hence user can decide efficiently which instances are unused instances and which are going to stop or which are going to run on the huge manner. By using Google looker studio the user can analyze the usage data and can give a virtual representation for usage of cloud services. By these virtual views they can represent the graph, charts and heat map for the usage of cloud instances.

By using these visuals user can decide very strongly which instances are not going to use which instances are unused and which have to be stopped. By these simple and easy process of data analysis users can definitely proceed towards the optimization and giving the best results for the upcoming generation.

The data analysis methodology applied in this research aimed to uncover meaningful patterns, trends, and insights from Company X's cloud usage data. The following steps were undertaken in the data analysis process:

- *Exploratory Data Analysis (EDA):* EDA involved a systematic exploration of the dataset to gain a comprehensive understanding of its structure and distribution. Statistical measures, data visualization, and summary statistics were utilized to identify potential outliers, patterns, and relationships among different cloud usage metrics.
- *Temporal Analysis:* Given the date-wise records in the dataset, a temporal analysis was conducted to examine how cloud usage varied over time. This analysis facilitated the identification of seasonal trends, periodic spikes, and long-term usage patterns.
- *Resource Utilization Ranking:* To optimize cloud resource allocation, the instances were ranked based on their CPU core, CPU capacity, and Memory usage. This ranking allowed for the identification of the most and least utilized

instances, providing valuable insights for cost optimization strategies.

D. Data Visualization using Google Looker Studio

Data visualization played a pivotal role in this research, as it enabled the transformation of complex data into easily interpretable visual representations. Google Looker Studio, a powerful data visualization tool, was leveraged to create interactive charts, heat maps, and graphs.

The visualizations provided a comprehensive view of Company X's cloud usage, presenting key findings in a visually appealing and accessible manner. Interactive charts allowed stakeholders to explore the data further and gain actionable insights from different perspectives.

The data visualizations were carefully designed to align with the research objectives. The ascending order chart of Disk throughput and Memory usage showcased the most utilized instances on top, highlighting potential resource bottlenecks. Additionally, the chart representing zero memory usage and disk throughput helped identify instances that could be optimized or discontinued to save costs.

By utilizing Google Looker Studio's capabilities, this research aimed to empower Company X's decision-makers with valuable visual insights, driving informed actions towards cost-efficient and performance-optimized cloud usage.

In conclusion, Section III has provided an overview of the data collection process and the methodology adopted for data analysis and visualization in this research. The use of Kaggle for data acquisition and Google Looker Studio for data visualization underscored the commitment to quality, efficiency, and ethical data practices. The subsequent sections will delve deeper into the findings of the exploratory data analysis and the implications for cloud usage optimization in Company X.

IV. Findings of Exploratory Data Analysis

The exploratory data analysis (EDA) conducted in this research reveals invaluable insights into the cloud usage trends of Company X during the year 2016. By employing statistical measures, data visualization techniques, and temporal analysis, this section presents a comprehensive overview of key findings that shed light on resource utilization patterns, seasonal trends, and instances of zero memory usage and disk throughput.

A. Resource Utilization Patterns

One of the primary objectives of the exploratory data analysis was to identify resource utilization patterns within Company X's cloud environment. By ranking instances based on their CPU core, CPU capacity, and Memory usage, the analysis unveiled significant variations in resource consumption. The results demonstrate that certain instances consistently exhibit high CPU and Memory usage, while others are underutilized.

The ascending order chart of Disk throughput and Memory usage effectively illustrates the resource utilization patterns. Instances with the highest usage are positioned at the top of the chart, highlighting the most resource-intensive components. This ranking provides valuable insights into potential resource bottlenecks, which can be targeted for optimization to enhance overall system performance.

B. Temporal Analysis

Temporal analysis was conducted to explore how cloud usage varied over time during the year 2016. By analyzing usage patterns across different dates, the research sought to uncover seasonal trends, identify periods of peak and off-peak usage, and identify any recurring patterns that might impact cloud resource allocation strategies.

The temporal analysis revealed periodic spikes in cloud usage, indicating times of increased demand for resources. Such insights are instrumental in planning for capacity management and optimizing resource allocation during peak usage periods. Additionally, off-peak usage periods can provide opportunities for cost optimization by potentially downsizing resource allocation during periods of lower demand.

C. Instances with Zero Memory Usage and Disk Throughput

A critical aspect of the exploratory data analysis was identifying instances with zero memory usage and disk throughput. Instances showing no activity during specific dates may suggest underutilized or unused resources. These findings have significant implications for cost optimization, as identifying and addressing such instances can lead to potential cost savings.

The chart representing zero memory usage and disk throughput visually highlights the instances and dates with negligible activity. By flagging these instances, the research enables Company X to streamline resource allocation and potentially terminate services or scale down instances that are no longer required.

D. Significance and Implications

The findings from the exploratory data analysis carry immense significance for Company X's cloud usage optimization efforts. Understanding resource utilization patterns allows the organization to make informed decisions on resource allocation, ensuring that critical instances receive adequate resources while efficiently utilizing cloud resources during periods of low demand.

Furthermore, the temporal analysis empowers Company X to plan for capacity management and allocate resources strategically, aligning with fluctuating usage patterns. Proactively adjusting resource allocation based on temporal trends can lead to cost savings and improved performance.

Identifying instances with zero memory usage and disk throughput holds substantial cost optimization potential. By eliminating or optimizing these instances, Company X can mitigate unnecessary expenses and streamline its cloud environment more effectively.

In conclusion, Findings of Exploratory Data Analysis has provided a comprehensive overview of the exploratory data analysis findings, which offer valuable insights into cloud usage trends within Company X. These findings form the foundation for developing cost optimization strategies and making informed decisions to enhance resource efficiency and overall cloud performance. The subsequent section will delve into the data visualization results, showcasing how Google Looker Studio's capabilities were harnessed to present the analysis outcomes in an intuitive and accessible manner.

V. Showcasing Data Visualization Results using Google Looker Studio

Data visualization is a powerful tool that transforms complex data into intuitive and visually compelling representations. In this section, we present the data visualization results obtained through the use of Google Looker Studio. The interactive charts, heat maps, and graphs generated using this platform facilitate a clearer understanding of the exploratory data analysis findings, enabling stakeholders at Company X to make data-driven decisions for cloud usage optimization.

A. Ascending Order Chart: Disk Throughput and Memory Usage

The ascending order chart visualizes the resource utilization patterns in Company X's cloud environment. Leveraging Google Looker Studio's capabilities, we created an interactive chart that ranks instances based on their Disk throughput and Memory usage, in ascending order.

The chart showcases the most heavily used instances positioned at the top, while the instances with the lowest Disk throughput and Memory usage appear at the bottom. This visual representation enables Company X's decision-makers to easily identify the most resource-intensive instances that may require optimization efforts. Moreover, instances with minimal resource consumption are readily visible, providing potential opportunities for cost savings through resource downsizing or termination.

The interactive nature of the chart allows users to hover over specific data points, accessing detailed information about each instance's CPU core, CPU capacity, and other relevant metrics. This level of granularity empowers stakeholders to make well-informed decisions regarding resource allocation and capacity planning.

B. Heat Map: Temporal Cloud Usage Patterns

To analyze temporal cloud usage patterns, a heat map was designed using Google Looker Studio. The heat map provides a visual representation of cloud resource consumption over the year 2016, with each date represented along the X-axis and different resource utilization metrics along the Y-axis.

Color intensity in the heat map corresponds to resource usage levels, with darker shades indicating higher usage and lighter shades representing lower usage. This visualization technique enables the identification of recurring patterns, such as seasonal spikes or consistent periods of high resource utilization.

Company X's stakeholders can utilize the insights gleaned from the heat map to strategically plan for resource allocation during peak and off-peak periods. This proactive approach to capacity management can enhance operational efficiency and contribute to overall cost optimization.

C. Chart: Instances with Zero Memory Usage and Disk Throughput

A chart was created to highlight instances with zero memory usage and disk throughput. Using Google Looker Studio's capabilities, this chart marks the instances and dates where these resources remained idle.

By visually identifying instances with no activity, Company X can take appropriate actions to optimize cloud resources. Possible strategies include terminating unused instances, downsizing underutilized resources, or investigating potential technical issues affecting resource utilization.

D. Communicating Insights and Facilitating Decision-Making

The data visualization results presented in this section demonstrates the effectiveness of Google Looker Studio in communicating complex information visually. By leveraging interactive charts, heat maps, and graphs, the research findings are transformed into actionable insights for Company X's stakeholders.

The intuitive nature of Google Looker Studio's visualizations facilitates efficient data exploration and interpretation, empowering decision-makers to extract value from the cloud usage data. The platform's interactivity allows users to drill down into specific data points, compare instances, and comprehend usage patterns dynamically.

In conclusion, the data visualization results showcased in this section serves as a bridge between data analysis and decision-making. By harnessing the power of Google Looker Studio, Company X gains a deeper understanding of cloud usage trends and resource utilization patterns, leading to informed strategies for cost optimization and enhanced performance. The final section of this research paper will summarize the key findings and their implications for optimizing cloud usage in Company X.

VI. Cost Optimization Strategies for Informed Decision-Making

The analysis of Company X's cloud usage data, coupled with the powerful data visualization results, has provided invaluable insights into resource utilization patterns and temporal trends. In this section, we outline cost optimization strategies derived from the research findings, empowering Company X to make well-informed decisions for optimizing cloud usage.

A. Resource Allocation Optimization

Based on the ascending order chart of Disk throughput and Memory usage, Company X can prioritize resource allocation to instances that exhibit high resource consumption. By provisioning more CPU cores, CPU capacity, and Memory to these resource-intensive instances, the organization can ensure optimal performance and minimize the risk of resource bottlenecks during peak periods.

On the other hand, instances with low Disk throughput and Memory usage, identified at the bottom of the chart, can be candidates for resource downsizing or potential termination. By reallocating resources from underutilized instances to those in higher demand,

Company X can achieve cost optimization without compromising performance.

B. Capacity Planning for Seasonal Demands

The heat map showcasing temporal cloud usage patterns highlights recurring seasonal spikes in resource utilization. Leveraging this insight, Company X can develop effective capacity planning strategies to align with fluctuating demands. By provisioning additional resources during peak periods and scaling down during off-peak periods, the organization can optimize costs while maintaining optimal system performance.

C. Identifying and Addressing Instances with Zero Activity

The chart representing instances with zero memory usage and disk throughput offers a clear view of underutilized or idle resources. Company X can focus on investigating the reasons behind zero activity, which could range from instances with technical issues to services that are no longer required.

By promptly addressing instances with zero activity, Company X can release resources and reduce unnecessary costs. Furthermore, the identification of such instances can also lead to improvements in resource management practices, ensuring resources are allocated only to actively used services.

D. Periodic Review and Monitoring

Cost optimization is an ongoing process. Company X should adopt a proactive approach to regularly review and monitor cloud resource usage based on the strategies derived from the analysis. Regular assessments enable the identification of changing resource demands and potential opportunities for further optimization.

By establishing a robust monitoring system, Company X can identify deviations from expected usage patterns promptly. This approach allows the organization to make timely adjustments, thereby maintaining cost efficiency and resource optimization over time.

E. Data-Driven Decision-Making Culture

To ensure the effective implementation of the cost optimization strategies, fostering a data-driven decision-making culture within Company X is crucial. By encouraging stakeholders to rely on data analysis and visualization for decision-making, the organization can empower individuals across departments to align their strategies with cloud usage insights.

Incorporating data-driven insights into decision-making processes can lead to more efficient resource allocation, improved performance, and ultimately cost savings. Additionally, fostering a data-driven culture encourages cross-functional collaboration, enabling various teams to work together towards common cost optimization objectives.

F. Continuous Improvement and Adaptability

Cloud usage optimization is a dynamic process, influenced by evolving business needs and technological advancements. Company X should embrace a culture of continuous improvement and adaptability, regularly incorporating new findings and methodologies to enhance cost optimization strategies.

By staying abreast of the latest trends in cloud computing and data analysis, Company X can proactively respond to changing requirements, ensuring optimal resource utilization and cost-effectiveness in an ever-evolving cloud landscape.

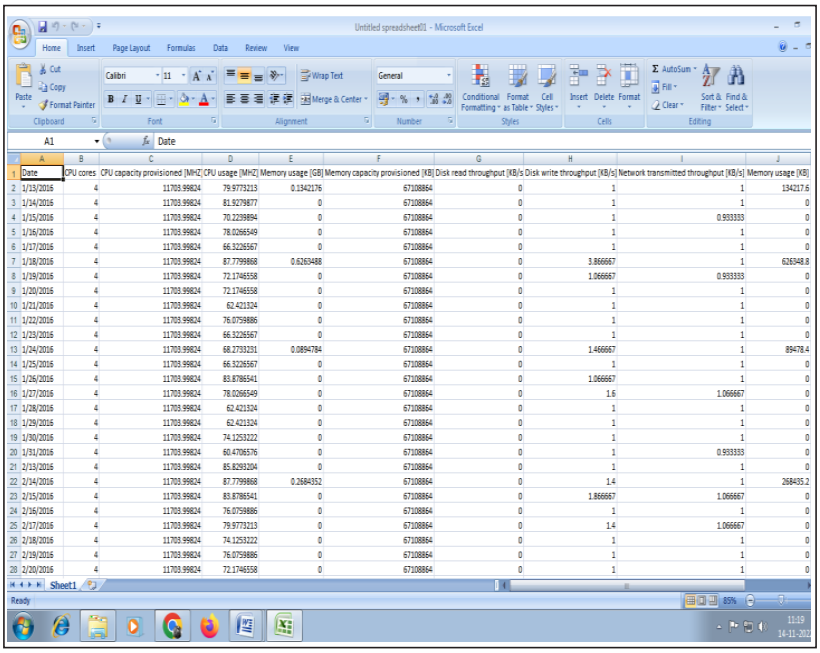
In conclusion, Section VI outlines a comprehensive set of cost optimization strategies derived from the analysis of Company X's cloud usage data. By leveraging insights from resource utilization

patterns, temporal trends, and instances with zero activity, Company X can make data-driven decisions to optimize cloud usage, enhance performance, and achieve substantial cost savings. The culmination of this research paper emphasizes the significance of data analysis and visualization in driving informed and effective cloud resource management strategies for the future success of Company X.

VII. Methods of Implementation

A. Data Fetching

We have to fetch data of cloud usage of a company for some specific range. So we can analyze it with better efficiency. We have fetch two datasets first we have to collect dataset of cloud usage of a company for specific period and second we have to collect costing for cloud as per regions for cost optimization.



Date	CPU cores	CPU capacity provisioned [MHZ]	CPU usage [MHZ]	Memory usage [GB]	Memory capacity provisioned [KB]	Disk read throughput [KB/s]	Disk write throughput [KB/s]	Network transmitted throughput [KB/s]	Memory usage [KB]
1/13/2016	4	11703.99824	79.9773213	0.1342176	67108864	0	1	1	134217.6
1/14/2016	4	11703.99824	81.9279877	0	67108864	0	1	1	0
1/15/2016	4	11703.99824	70.2239894	0	67108864	0	1	0.933333	0
1/16/2016	4	11703.99824	78.0266549	0	67108864	0	1	1	0
1/17/2016	4	11703.99824	66.3226567	0	67108864	0	1	1	0
1/18/2016	4	11703.99824	87.7799868	0.6263488	67108864	0	3.866667	1	626348.8
1/19/2016	4	11703.99824	72.1746558	0	67108864	0	1.066667	0.933333	0
1/20/2016	4	11703.99824	72.1746558	0	67108864	0	1	1	0
1/21/2016	4	11703.99824	62.421324	0	67108864	0	1	1	0
1/22/2016	4	11703.99824	76.0759886	0	67108864	0	1	1	0
1/23/2016	4	11703.99824	66.3226567	0	67108864	0	1	1	0
1/24/2016	4	11703.99824	68.2733231	0.0894784	67108864	0	1.466667	1	89478.4
1/25/2016	4	11703.99824	66.3226567	0	67108864	0	1	1	0
1/26/2016	4	11703.99824	83.8786541	0	67108864	0	1.066667	1	0
1/27/2016	4	11703.99824	78.0266549	0	67108864	0	1.6	1.066667	0
1/28/2016	4	11703.99824	62.421324	0	67108864	0	1	1	0
1/29/2016	4	11703.99824	62.421324	0	67108864	0	1	1	0
1/30/2016	4	11703.99824	74.1253222	0	67108864	0	1	1	0
1/31/2016	4	11703.99824	60.4706576	0	67108864	0	1	0.933333	0
2/13/2016	4	11703.99824	85.8293204	0	67108864	0	1	1	0
2/14/2016	4	11703.99824	87.7799868	0.2684352	67108864	0	1.4	1	268435.2
2/15/2016	4	11703.99824	83.8786541	0	67108864	0	1.866667	1.066667	0
2/16/2016	4	11703.99824	76.0759886	0	67108864	0	1	1	0
2/17/2016	4	11703.99824	79.9773213	0	67108864	0	1.4	1.066667	0
2/18/2016	4	11703.99824	74.1253222	0	67108864	0	1	1	0
2/19/2016	4	11703.99824	76.0759886	0	67108864	0	1	1	0
2/20/2016	4	11703.99824	72.1746558	0	67108864	0	1	1	0

Fig. 2: Data Fetching

We have collected data of a company usage for cloud for year 2016 which contains 186 days data which includes CPU core, CPU capacity, Memory usage, Disk read, Disk Write, memory usage, transmitted throughput with date wise.

B. Data Extractions

We have a company data of year 2016. After fetching we have to extract it and for extracting we have arrange it as per the date and memory usage. The features which we don't have any use that can be deleted and store only usable features in dataset. Date must be in a specific format which indicates dd/mm/yyyy. The disk read/write data should be in kb/sec. Memory usage should be in GB.

C. Analysing Data

We have to collect all data in Google Looker Studio. After collecting analyse as per date and usage of processor as per ascending order. We will analyse the data by extracting features and comparing with the dates.

D. Making Report

We will make report by using the Google Looker Studio and for making report we will select some features on which we need to operate and practice. We will show virtually by creating charts and heat map for better understanding.

	Date	CPU usage [MHZ]	Disk write throughput [KB/...	Memory usage [GB]	Memory usage [KB]	Disk read throughput [KB/s]
18.	5/16/2016	8,932.1	24,322.93	4.56	4,563,400	3,247.6
19.	5/18/2016	8,914.55	23,708.6	6.17	6,174,014	2,471.07
20.	5/14/2016	8,545.87	24,685.07	6.04	6,039,795	8,544.73
21.	10/28/2016	7,874.84	3.4	3.36	3,355,441	0
22.	7/28/2016	4,020.32	8.93	1.79	1,789,568	10.13
23.	10/25/2016	3,197.14	8.53	1.25	1,252,698	12.67
24.	10/18/2016	1,648.31	9.07	0.63	626,348.8	12.8
25.	10/19/2016	1,061.16	6.8	0.67	671,088	7.93
26.	7/14/2016	152.15	25.87	0.54	536,870.4	131.33
27.	8/24/2016	138.5	14.33	0.13	134,217.6	131.33
28.	1/18/2016	87.78	3.87	0.63	626,348.8	0
29.	2/14/2016	87.78	1.4	0.27	268,435.2	0
30.	3/18/2016	87.78	5.13	0.13	134,217.6	32.67
31.	3/15/2016	87.78	1.2	0	0	0
32.	9/23/2016	87.78	1.33	0	0	0
33.	2/25/2016	87.78	1.73	0	0	0
34.	2/23/2016	85.83	1	0	0	0

Fig. 3: Making Report

VIII. Results

A. Chart for Disk Throughput/Memory Usage

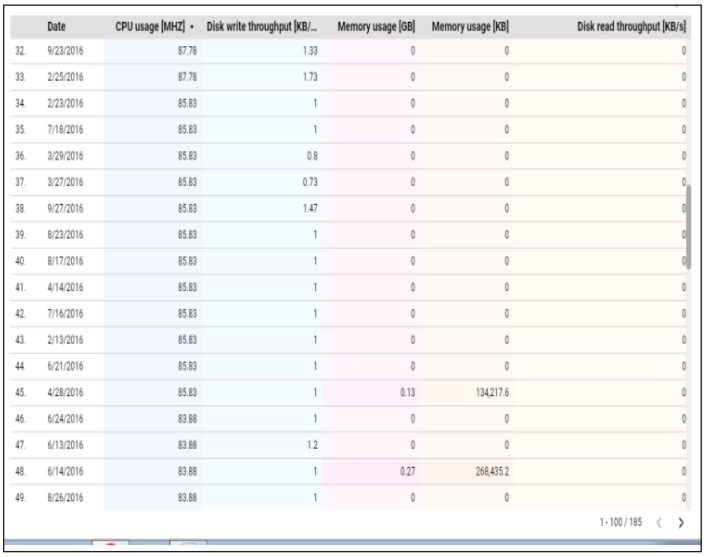
	Date	CPU usage [MHZ]	Disk write throughput [KB/...	Memory usage [GB]	Memory usage [KB]	Disk read throughput [KB/s]
1.	5/19/2016	10,796.94	15,914.87	6.49	6,487,188	115.07
2.	5/21/2016	10,713.06	8,593.87	6.93	6,934,581	0.47
3.	5/17/2016	10,711.11	15,682.2	6.49	6,487,188	392.67
4.	5/22/2016	10,691.6	9,182.73	3.8	3,802,833	0.07
5.	5/15/2016	10,662.34	14,457.13	6.31	6,308,231	259.33
6.	5/30/2016	10,590.17	7,295.2	3.71	3,713,354	0
7.	10/27/2016	10,580.41	4	3.4	3,400,180	2.53
8.	10/26/2016	10,480.93	4.13	1.92	1,923,786	2.53
9.	5/25/2016	10,371.69	9,239.2	4.83	4,831,835	965.67
10.	5/27/2016	10,319.03	7,012.13	5.32	5,323,967	2.53
11.	5/28/2016	10,287.81	7,362.93	7.07	7,068,798	3.07
12.	5/13/2016	10,215.64	21,201.13	6.04	6,039,796	1,628.33
13.	5/29/2016	10,108.35	7,309.8	6.08	6,084,535	4.33
14.	5/24/2016	9,946.45	10,051.93	4.56	4,563,401	1,710.87
15.	5/23/2016	9,755.28	8,605.67	2.86	2,863,309	5,190.8
16.	5/20/2016	9,466.58	25,983.53	7.34	7,337,234	1,114.4
17.	5/26/2016	9,277.37	7,140.4	9.08	9,082,064	251.2
18.	5/16/2016	8,932.1	24,322.93	4.56	4,563,400	3,247.6

1 - 100 / 185 < >

Fig. 4: Chart for Disk Throughput/Memory Usage

The Fig. 4 Chart for Disk Throughput/Memory Usage shows the graphical representation of 186 days data usage and shows in ascending order format which shows the more used instances first and below that lower used instance. It can be useful and efficient to use by user in better manner. We can definitely show and verify that highest used instances are on top of the chart.

B. Chart for Zero Memory Usage and Disk Throughput



	Date	CPU usage [MHZ]	Disk write throughput [KB/...	Memory usage [GB]	Memory usage [KB]	Disk read throughput [KB/s]
32.	9/23/2016	87.78	1.33	0	0	0
33.	2/25/2016	87.78	1.73	0	0	0
34.	2/23/2016	85.83	1	0	0	0
35.	7/18/2016	85.83	1	0	0	0
36.	3/29/2016	85.83	0.8	0	0	0
37.	3/27/2016	85.83	0.73	0	0	0
38.	9/27/2016	85.83	1.47	0	0	0
39.	8/23/2016	85.83	1	0	0	0
40.	8/17/2016	85.83	1	0	0	0
41.	4/14/2016	85.83	1	0	0	0
42.	7/16/2016	85.83	1	0	0	0
43.	2/13/2016	85.83	1	0	0	0
44.	6/21/2016	85.83	1	0	0	0
45.	4/28/2016	85.83	1	0.13	134,217.6	0
46.	6/24/2016	83.88	1	0	0	0
47.	6/13/2016	83.88	1.2	0	0	0
48.	6/14/2016	83.88	1	0.27	268,435.2	0
49.	8/26/2016	83.88	1	0	0	0

1 - 100 / 185

Fig. 5: Chart for Zero Memory Usage and Disk Throughput

The Fig. 5 Chart for Zero Memory Usage and Disk Throughput shows the chart for memory usage and disk throughput in which zero kb/sec data is used. Which can shows us we can delete these unused instances or we can stop these services on these dates so we can find out the unvisited instances dates efficiently.

IX. Conclusion and Future Research Directions

A. Summary of Key Findings

This research paper embarked on an analytical expedition to unravel efficiency and economies in Company X's cloud usage during the year 2016. Through rigorous data collection, exploratory data analysis, and data visualization using Google Looker Studio, we unearthed key insights into resource utilization patterns, temporal trends, and instances with zero activity.

The ascending order chart revealed resource-intensive instances that warrant optimized resource allocation. Conversely, instances with minimal resource consumption offer opportunities for cost savings through resource downsizing or termination. The heat map showcased temporal cloud usage patterns, enabling effective capacity planning for seasonal demands. Additionally, the chart highlighting instances with zero activity identified underutilized resources, indicating potential areas for optimization.

B. Implications for Cloud Usage Optimization

The findings of this research have substantial implications for optimizing Company X's cloud usage. By adopting data-driven decision-making, the organization can implement the following cost optimization strategies:

- *Optimal Resource Allocation:* Prioritize resource allocation to resource-intensive instances to maintain optimal performance.
- *Resource Downsizing:* Identify underutilized instances and reallocate resources to high-demand instances, reducing unnecessary costs.
- *Capacity Planning:* Develop proactive capacity planning strategies to align with seasonal resource demands, ensuring cost-efficient resource provisioning.
- *Instance Management:* Address instances with zero activity promptly, releasing unused resources and streamlining cloud services.
- *Continuous Monitoring:* Establish a robust monitoring system to detect changes in usage patterns, enabling timely adjustments to optimize costs.

C. Potential Future Research Directions

This research has opened avenues for potential future studies in cloud usage optimization and data analysis. The following research directions are suggested:

- *Longitudinal Analysis:* Conduct a longitudinal analysis over multiple years to examine how cloud usage trends evolve and identify long-term patterns.
- *Predictive Modeling:* Develop predictive models to forecast cloud usage trends, enabling proactive resource allocation and cost optimization.
- *Cloud Service Performance:* Explore the correlation between cloud service performance and resource utilization to optimize performance and cost-effectiveness.

- *Anomaly Detection:* Investigate anomaly detection techniques to identify abnormal resource usage and potential security threats in the cloud environment.
- *Cost-Benefit Analysis:* Perform a comprehensive cost-benefit analysis of implementing different cost optimization strategies to evaluate their effectiveness and impact on ROI.
- *Hybrid Cloud Optimization:* Extend the analysis to encompass hybrid cloud environments, addressing the unique challenges of optimizing resource allocation across multiple cloud platforms.

D. Conclusion

In conclusion, this research paper has delved into the depths of Company X's cloud usage data and illuminated essential insights through data analysis and visualization. By adopting Google Looker Studio's powerful capabilities, we have effectively communicated complex information, empowering Company X to make informed decisions for cloud usage optimization. The cost optimization strategies derived from this research offer a pathway to enhanced resource efficiency, improved performance, and substantial cost savings. By fostering a data-driven decision-making culture and embracing continuous improvement, Company X can navigate the ever-changing cloud landscape with adaptability and success. The discoveries and methodologies presented in this paper contribute valuable knowledge to the domain of cloud usage optimization, inspiring further research and advancements in cloud resource management. As the cloud computing landscape continues to evolve, the principles and insights derived from this research will guide organizations towards greater efficiency, competitiveness, and cost-effectiveness in cloud usage.

References

[1] V. Ramasamy, and S. M. Thalavai Pillai, "An effective HPSO-MGA optimization algorithm for dynamic resource allocation in cloud environment," *Cluster Comput*, vol. 23, no. 3, pp. 1711-1724, Sep. 2020, doi: 10.1007/s10586-020-03118-x.

[2] X. Wang, X. Wang, H. Che, K. Li, M. Huang, and C. Gao, "An intelligent economic approach for dynamic resource allocation in cloud services," *IEEE Transactions on Cloud Computing*, vol. 3, no. 3, pp. 275-289, Jul. 2015, doi: 10.1109/TCC.2015.2415776.

[3] A. Bhardwaj, and C. Rama Krishna, "Impact of factors affecting pre-copy virtual machine migration technique for cloud computing," 2019. [Online]. Available: www.sciencedirect.comwww.materialstoday.com/proceedings

[4] I. Odun-Ayo, and F. Agono, "Cloud management optimization postgraduate view project cloud computing view project," 2018. [Online]. Available: https://www.researchgate.net/publication/332143932

[5] I. Iyoob, E. Zarifoglu, and A. B. Dieker, "Cloud computing operations research," *Service Science*, vol. 5, no. 2, pp. 88-101, Jun. 2013, doi: 10.1287/serv.1120.0038.

[6] E. Weintraub, and Y. Cohen, "Cost optimization of cloud computing services in a networked environment," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 4, 2015, doi: 10.14569/ijacsa.2015.060420.

[7] C. S. Pawar, and R. B. Wagh, "Priority based dynamic resource allocation in cloud computing," in *Proceedings - 2012 International Symposium on Cloud and Services Computing (ISCOS'12)*, IEEE Computer Society, 2012, pp. 1-6, doi: 10.1109/ISCOS.2012.14.

[8] U. K. Jena, P. K. Das, and M. R. Kabat, "Hybridization of meta-heuristic algorithm for load balancing in cloud computing environment," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 2332-2342, Jun. 2022, doi: 10.1016/j.jksuci.2020.01.012.

[9] V. E. Jayanthi, R. Divya, and M. Jagannath, "A review on optimization approaches in cloud computing service," 2018. [Online]. Available: www.sciencepubco.com/index.php/IJET

[10] N. S. Kumari, K. Divya, C. Yadav, D. Asha, and M. Deepika, "A light weight secure data

sharing scheme for mobile cloud computing," *J Algebr Stat*, vol. 13, no. 3, pp. 3952-3956, 2022. [Online]. Available: https://publishoa.com

[11] M. Kumar, and S. C. Sharma, "Dynamic load balancing algorithm for balancing the workload among virtual machine in cloud computing," in *Procedia Computer Science*, vol. 115, pp. 322-329, 2017, doi: 10.1016/j.procs.2017.09.141.

[12] S. Afzal, and G. Kavitha, "Load balancing in cloud computing – A hierarchical taxonomical classification," *Journal of Cloud Computing*, vol. 8, no. 1, Dec. 01, 2019, doi: 10.1186/s13677-019-0146-7.

[13] Muhd. Junaid, A. Sohail, R. N. Bin Rais, A. Ahmed, I. Ali Khan, ... and N. Ejaz, "Modeling an optimized approach for load balancing in cloud," *IEEE Access*, vol. 8, pp. 173208-173226, 2020, doi: 10.1109/ACCESS.2020.3024113.

[14] P. Osypanka, and P. Nawrocki, "Resource usage cost optimization in cloud computing using machine learning," *IEEE Transactions on Cloud Computing*, vol. 10, no. 3, pp. 2079-2089, 2022, doi: 10.1109/TCC.2020.3015769.

[15] Muhd. Jawad, Muhd. B. Qureshi, Muhd. U. S. Khan, S. M. Ali, A. Mehmood, ... and S. U. Khan, "A robust optimization technique for energy cost minimization of cloud data centers," *IEEE Transactions on Cloud Computing*, vol. 9, no. 2, pp. 447-460, Apr. 2021, doi: 10.1109/TCC.2018.2879948.

International Journal of Distributed and Cloud Computing
11 (1) 2023, 38-48
http://www.publishingindia.com/ijdcc

Blockchain Technology to Combat Counterfeiting of Drugs in the Pharmaceutical Industry

Anusha Thakur
University of Petroleum & Energy Studies, Dehradun, Uttarakhand, India.
Email: anushathakur89@yahoo.com

Abstract: Counterfeiting is an ever-growing issue intensified by the increasing ease of access via e-commerce activities. The amalgamation of misinformation-driven panic, economic distress, and public health emergency contributes to the challenges related to the accessibility and supply of better and essential qualityof medicinal products, leading to price volatility, uncertainty, quality of issues, and drug shortages. The adoption of state-of-the-art technologies paves the way for the digital transformation of the pharmaceutical industry. Incorporating blockchain technology augments the safety of drugs, with easy detection of falsified drugs, thereby leading to minimized mortality rate and improved safety. Blockchain-based architectures enable the systems to offer a platform to substantially cater to the changing need for pharma-data privacy, data storage, quality assurance, and material traceability. This paper discusses the different technologies in the counterfeiting of drugs, with the prime focus being on blockchain technology. It emphasizes on the strategic decisions and models adopted by the key players in their supply chain processes to make these an efficacious one. The article also focuses on the relevant information regarding blockchain architectures and the challenges being faced in the implementation of this technology in combating counterfeiting in the pharmaceutical industry.

Keywords: Blockchain, Counterfeiting, Drugs, Pharmaceutical, Technology.

I. Introduction

Technological infusion has shown an enormous transformation in the pharmaceutical business. Major highlights of the pharmaceutical & life sciences industry include the consistent discovery of generic drugs, orphan drugs, next-generation gene & cell therapies, and biosimilars. The increasing superiority of the illicit manufacturers acts as an impetus for the latest combination of technologies and innovations.

Companies nowadays, emphasize on monitoring their suppliers and verifying the authenticity and efficacy of their supply chain process. To alleviate the risks related to health from the trade across, the organizations are focusing on technology-centric solutions.

The tryst with falsified drugs is no different, with the business of these products tarnishing the image of the companies. Counterfeiting products that are not as per the production norms for legitimate products and the regulatory standards for the same are dangerous if consumed, and pose a threat to the health of an individual. The increasing prevalence of counterfeit products can be attributed to the increasing demand for inexpensive drugs, lack of law enforcement, complex import-export mechanisms, non-stringent regulatory policies, and growing international free trade [1].

Although counterfeit and falsified medicines have increasingly affected all countries, however, the problem seems to be more visible in the countries wherein, the standard of living is low, and the process of making, supply, distribution, and sales are less scrutinized, owing to the fact, that, the forged medicines are considerably the cheapest [2] [3]. The growing investments in the pharma sector by various governments, the rise of communicable and non-communicable diseases, and the availability of

diverse medicine-tracking software/systems, among others, are set to drive the market for counterfeiting solution providers.

Growing automation is paving the way for continual upgradation of the pharmacy medicine tracking software system. These systems work in sync with the drug/medicine codes and other minute details, thereby increasing the scope of technological solutions for these drugs/medicines in modern-day situations. Changes such as the inculcation of the UDI (Unique Device Identification) system by the EU Medical Device Regulations push for sufficient traceability of medical devices throughout the supply chain [4]. The system enhances the effectiveness of the market-related activities for the devices and enables better monitoring of the products by skilled professionals, thereby minimizing errors and catering against counterfeit devices.

To combat against falsified and counterfeit medicines, several areas need to be catered to such as ensuring integrity and traceability of the medical products supply chain, securing the drugs and their packaging, easy traceability solutions for the drugs over the supply chain, increasing awareness among the individuals, and introduction of stringent regulatory measures to combat the same. Technological advancements are expected to enable the companies to adopt features such as unique packaging, and serialization, thereby creating consumer-focused awareness and content, being looked for.

The article emphasizes on the research opportunities to enable the prevailing initiatives to combat the counterfeiting of drugs such as exploring emerging technologies, possible use cases of the implementation of blockchain to minimize the market for counterfeit and falsified medicines, and the surge of the blockchain-based systems for the same.

II. Literature Review

The siloed and fragmented supply chain network of the pharmaceutical industry paves the way for numerous entry and exit points which make the products susceptible to the usage of substandard type, contamination, and adulteration processes [5].

Designing digital platforms with market-ready solutions to cater to the changing market paradigm is gaining traction. The growing emergence of organized trade, rising purchasing power, coupled with a shift towards digitalization are changing the market dynamics of customer opportunities across the different regions of the world. Problems related to a product's quality and safety are not only prevalent in developing countries but in developed countries as well. Counterfeiting products poses a worldwide issue, with the situation worsening in developing countries. At times, it becomes difficult for businesses to detect drug counterfeiting owing to their movement across the complex distributed networks, thereby creating entry opportunities for the counterfeits in the supply chain [6]. Such fraudulent activities can be valid for both generic and branded products.

Unreliable or unsafe counterfeit products are expected to negatively impact the market perception of the genuine products that are being copied, thereby reducing the potential sales of the same. Increasing consumerism is expected to bolster counterfeiting trade activities. The manufacture and sale of counterfeiting products are expected to impact retailers, distributors, and manufacturers.

In certain situations, adverse publicity, and counterfeit products might extend to the companies that manufacture legitimate products or sell the same, further damaging the industry's identity and reputation. The incorporation of different methods to curb counterfeiting across different product categories are thereby a necessity.

The rising trends towards anti-counterfeiting solutions are anticipated to add another level of complexity to the various coding necessities such as the usage of UV inks and serialization codes. Pharmaceutical companies nowadays are emphasizing on the implementation of seamlessly trackable technologies throughout the supply chains without the involvement of duplication risks. The companies are inevitably engaged in offering products that help enhance patient safety, combat counterfeiting, and cater to regulations as well as several industry concerns. Counterfeiting and traceability measures are expected to witness

an improvement due to the usage of variable data on labels, such as unique coding and marking on the packages. These measures or solutions include the application of overt and covert security coding, usage of holographic labels, and employing special color schemes to identify specific product types. The two types of anti-counterfeiting packaging strategies being incorporated include covert and overt strategies.

- The *overt* packaging technologies include uneven package designs (such as holograms and color-shift inks) which can be verified at different angles, enabling quick authentication without sophisticated devices.
- The *covert* packaging technologies require special detection devices, posing a challenge for counterfeiters to duplicate. This includes biological tracers, molecular markers, micro-text, ultraviolet inks, and taggant systems comprising of radio frequency microchips [7].

In today's scenario, the companies emphasize implementing full traceability, visibility, and serialization programs, offering them a competitive advantage over other pharmaceutical vendors. Implementation of serialization codes and barcodes is expected to help identify individual products, shipments, and packaging, thereby favorably impacting factors such as visibility and authentication. One vital way to reduce counterfeiting is to ensure that medicine packs carry unique identifiers. Serialization in the pharmaceutical industry focuses on assigning an exclusive code to every single unit or package of medicine, which is then tracked in the supply chain. Simple and effective data structure communication between logistics, sales, and production, plays a vital role in delivering a significant level of security, traceability, and accountability. The drug distribution system comprises several entities, which include wholesale distributors, manufacturers, and pharmacies before the products reach the end-users [8].

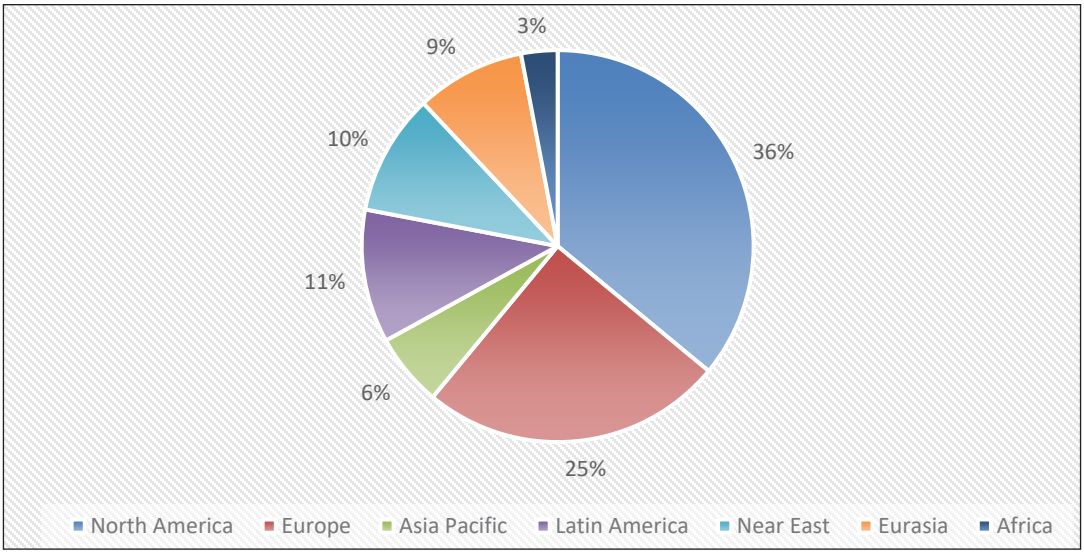
With the implementation of tracing and tracking capabilities across touchpoints in the supply chain process, various challenges and bottlenecks can be overcome. The process of traceability offers excellent benefits such as the ability to troubleshoot and investigate the issues associated with a product. A few of the issues faced by manufacturers across different verticals include product recalls and replacements. Implementation of an appropriate drug product-coding scheme can be attained by generating unique serial numbers for tracking and marking, thereby enabling a seamless tracking process across the supply chain, and blocking the progress of illegitimate serial numbers. Traceability via the entire supply chain is vital in any quality defects or manufacturing, particularly those necessitating a recall from the producer [9]. With the incorporation of efficient tracking systems, the company can easily track the finished goods, and the processes relevant to product handling in the distribution and manufacturing systems. Hence, keeping track of the entire supply chain process is one of the critical areas manufacturers should focus on.

III. Global Drug Counterfeiting Scenario

Counterfeiting of drugs has increasingly become a major health issue in both industrialized and developing countries [10] [11]. Nearly 15 percent of the drugs sold in the world are counterfeit [12].

The developing nations account for a substantial part of the forged drugs being distributed in the world. According to the World Health Organization, nearly 1 out of 10 medical products in developing countries are spurious or counterfeit [13]. Increasing incidences of drug counterfeiting can be attributed to the complexity of the drug counterfeiters due to the limited enforcement capability and regulations in developing countries. People in underdeveloped countries emphasize on reasonably priced alternatives due to the unavailability and high pricing of medicines.

Increasing incidents of counterfeiting concerning pharmaceutical companies across the world, region-wise, are illustrated in the graph mentioned below:



Source: [14]

Fig. 1: Number of Counterfeit Incidents Region-Wise, 2022

Counterfeiting of the drugs in different segments is gaining significant traction across the different regions of the world. For instance, the drugs fitting in the therapeutic categories are significantly being forged across the globe. The anti-infective drug category, such as antivirals, antibiotics, antimalarials, and antifungals account for most of the drugs being counterfeited, ranging from approximately 10 percent to nearly 50 percent [15]. Further, in the anti-infective drug segment, the antibiotics are expected to be highly falsified [16].

Although low-income and developing countries are witnessing a surge in counterfeiting incidents, developed nations such as the regions of the US, Europe, and Canada also hold a significant share in the same, thereby increasing the need to combat the same.

IV. Impact of Counterfeiting on the Global Economy

Scammers nowadays are increasingly replicating the products at a larger level, thereby turning illegal fraud into an enterprise-level industry. The unawareness of supply chains with forged goods is expected to make the companies unattractive and volatile targets for investors, thereby leading to severe damage to the corporate Environmental, Social, and Governance (ESG) aspects.

Counterfeit products contribute nearly 3.3 percent of the value of the global economy annually [17]. The rising counterfeiting and piracy issues on legitimate economic activities increase the economic costs, which further drives fiscal losses. This refers to manufacturers facing competitors that are taking away their intellectual property (IP) rights without complying with the regulatory standards or payment of taxes. Counterfeiting is also expected to impact the foreign direct investment (FDI) scenario of a country, thereby failing to attract multinational innovators to that region.

V. Technological Innovations to Combat Counterfeiting

Increasing advancements in simple field tests and forensic chemical analysis are expected to augment drug quality monitoring, access to low-cost genuine medicines, cooperation with determined political management, and stringent law enforcement, as well as support the drug regulatory authorities, to combat the counterfeiting threat.

Using a forged medicine is significantly substandard, comprising of inactive or fewer ingredients in an appropriate measures leading to treatment failure [18]. The concern to reinforce intellectual property rights and ensure brand protection for pharmaceutical companies, with the possibility for counterfeit drugs

to cause mortality is witnessing an increase. The overtly visible markers on a drug's packaging are gaining traction owing to its potential to identify fake products. However, distinguishing markers and holograms, when applied to the paper substrates, film, or blister foils can be copied to a higher extent with much accuracy. The rising sophistication of illegal manufacturers thereby, plays a pivotal role in the launch of advanced technologies. To avoid counterfeiting, the manufacturers, and suppliers link on a single blockchain platform with the usage of exclusive cryptographic identifiers "smart tags" to track and approve the provenance and position of each item. The tags can be of different forms, suchas security labels with unique RFID, and QR codes that comprise of a customized software component. A few technological innovations can be illustrated as:

Smartphone Applications: Increasing smartphone applications enables customers to instantly check the authenticity of the products before making a purchase. The applications allow the brand owners to recognize, avert, and track brand invaders from selling counterfeit products. Advancements in smartphone technologies are expected to boost the integration of anti-counterfeiting solutions in terms of product authenticity, thereby enabling the owners to prevent infringing on any type of drug. For instance, Unique Product Identifier (UPI), in retail outlets helps the consumers comprehend the authentication of the product, and accordingly inform the customers about the counterfeiting of the same.

Radio Frequency Identification (RFID): Technologies such as RFID offers strategized labeling solutions, inventory control, and automatic tracking. It generates significant impacts with a unique selling proposition, such as secure supply chains, quality monitoring, cold chain, anti-counterfeiting, and improved information for the decision-makers [19]. Technology also plays a crucial role in the packaging of pharmaceutical products, protecting them from getting tampered with, and, defending against monetary crimes [20]. Incorporation of RFID in the pharmaceutical industry thereby enables businesses to seamlessly cater to difficulties such as diversion, theft, and counterfeiting and ensures the legitimacy of the medications arriving at the pharmacies.

Companies are increasingly emphasizing on enabling safe, secure, and transparent sharing of information, thereby saving businesses from offering forged products. In addition to this, in the pursuit to gain continual confidence in the exported drugs, the regulatory bodies of organizations, nowadays are focusing on improvising on mechanisms to attain real-time visibility into all the pharmaceutical products being manufactured and exported. Various advanced technologies such as RFID, image processing, blockchain, pattern recognition, and others are thereby, significantly gaining traction to address the changing scenario.

However, businesses incorporating state-of-the-art technology such as blockchain, offer, unprecedented scalable solutions catering to the existing challenges in the dynamic marketplace. In addition to collaborating the QR codes and Distributed Ledger Technology, the blockchain solutions also poses to be significantly compatible with other technologies, such as GPS, RFID, or NFC, based on the contactless relocation of data.

VI. Blockchain Technology to Combat Drug Counterfeiting

Blockchain technology enables a revolutionary level of supply chain transparency in businesses. A blockchain-based supply chain of drugs depends on the inclusion of entities such as pharmacists, distributors (3PL), and manufacturers on a reliable system powered by blockchain. Organizational and multi-stakeholder involvement poses to be an integral part of the drug supply chain.

A supply chain comprising of partners, distributors, vendors, and suppliers is created, at the time when the drug bottles are labelled, scanned, and documented on the blockchain at every stage, from the plant to the pharmacy. The distribution points in the supply chain processes supply the medical products to the customers in bulk quantities. There are several unsecured, and unverified points of contact, which significantly pose as an entry point for the lethal forged drugs. The medical product supply chain process comprises of several parties such as wholesalers, patients, pharmacies, supermarkets,

suppliers, and manufacturers. Understanding the way how counterfeit drugs make their way in the supply chain processes, plays a crucial role in the implementation of the technology. This includes the ways:

- The probability of the key players being corrupt in the supply chain processes (mislabelling, substituting, or contaminating).
- Possibility among the manufacturers of not being aware of the source of the ingredients.
- The drug or ingredient manufacturers posing as legitimate companies due to incorrect papers [21].

With no secure or robust system and several points of entry, tracking the drugs in supply chain processes is significantly challenging. Blockchain, based on DLT empowers the storage of data in a decentralized manner, rather than in a central database. Every entry is independently encrypted and every change needs to be verified by all the distributed nodes in the system, making it virtually unbearable to address a blockchain. When stored, the data becomes difficult to be tampered with, owing to the easy identification of weak points and detection of fraudulent activities, thereby offering enhanced counterfeiting protection to the companies [22].

Keeping track of the activities associated with medical products poses to be considerably difficult with the increasing complexity of the processes, thereby limiting the data provenance and transparency checks for the same [23].

Blockchain helps in tackling the issues of counterfeiting by focusing on the identification of proof-of-origin of the product and offering a trusted and secure tracking system from one end to the other end of the supply chain. The pharmaceutical products are systematically serialized and assigned the security features, which are easily differentiated and verifiable by the consumers from counterfeits. Serialization of the drugs caters to the security threat of false medications, via digital signatures, health information, and blockchain codes [24]. Incorporation of this technology enables the security via transparent and chain code-based transactions. This can be largely attributed to the features offered by blockchain technology, such as decentralization, immutability, decentralization, distribution, and transparency, which can address the issues of trustless between parties, tracking and monitoring all processes of the drug supply chain.

VII. Blockchain-Based Architectures for Drug Traceability

Counterfeit drugs contribute to major health concern severely impacting the treatment outcomes and lives of the individuals. Growing access to the medicines via unauthorized distributed channels and online pharmacies makes it difficult to guarantee the safety of products in the supply chain process.

The traceability of drugs based on blockchain technology helps create a distributed shared platform for a transparent, accountable, trustworthy, and immutable system in the supply chain process. The two potential blockchain-based architectures include hyperledger fabric and hyperledger besu, which offer a higher degree of transparency, decentralization, trust, data integrity, scalability, modularity, security, deployment, and privacy in comparison to the other platforms based on blockchain. These act as the enablers for generating private permissioned ecosystems, wherein, the pharmaceutical end-users and stakeholders are listed, and controlled by a group of stakeholders.

Hyperledger Fabric Architecture: This architecture is a platform offering disseminated ledger solutions, derived by a modular architecture remitting higher degrees of scalability, resilience, and confidentiality. This enterprise-grade DLT is based on technology that uses smart contracts to carry out the trust between several parties and offers higher transaction throughput, up to nearly a thousand transactions per second [25].

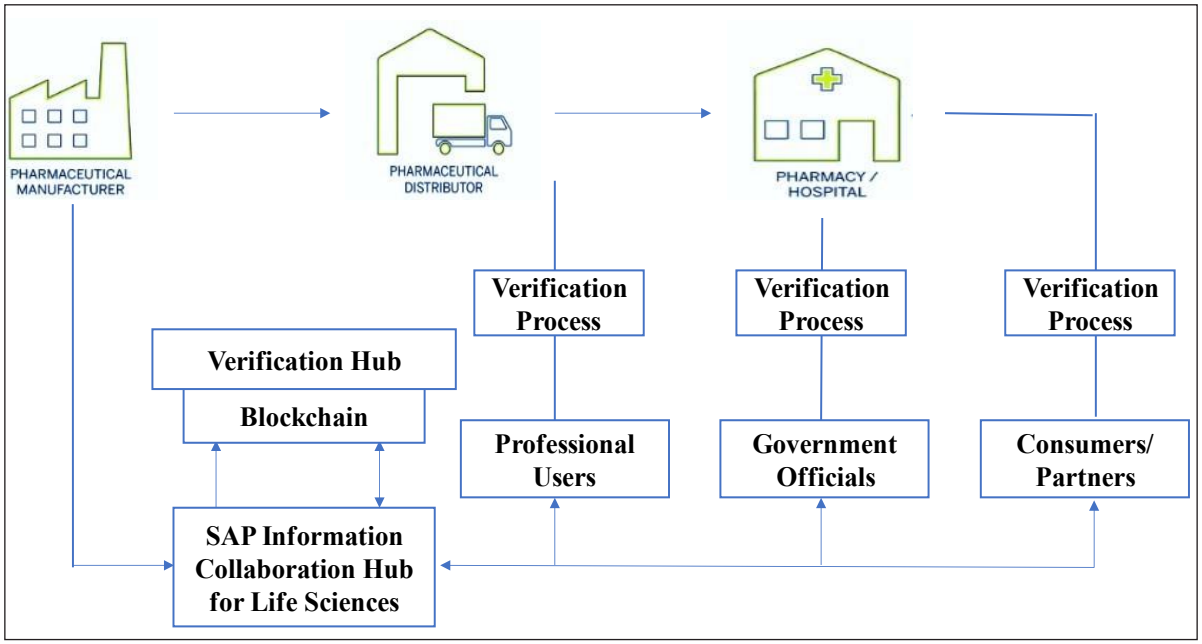
Hyperledger Besu Architecture: This architecture has gained significant traction among enterprises, owing to the fact, that it backs the building networks supporting both integration with the public blockchains (Ethereum) and private transaction processing while upholding the high transaction throughput and architectural flexibility.

VIII. Use Cases of Blockchain in Drug Counterfeiting

In 2019, in reaction to the FDA's call to form trial projects, in order to test the automated inter-operable systems, the pharmaceutical companies, announced the development of a blockchain platform such as MediLedger Network, to trace forged drugs and prescriptions and eliminate them subsequently [26].

- *HCL Technologies Limited*

HCL Technologies Ltd. announced the launch of strategies to combat the counterfeiting of drugs with blockchain and SAP. In 2017, SAP teamed up with leading pharmaceutical firms such as Boehringer Ingelheim, AmerisourceBergen, GlaxoSmithKline, and, Merck to develop distributed ledger and blockchain-based decentralized solutions to cater to legal compliance necessities. In 2019, SAP announced an "Information collaboration hub for Life Sciences blockchain solutions", supporting the eradication of counterfeit drugs through the supply chain processes in the pharmaceutical industry. The developed solutions are designed to minimize drug counterfeiting as per the government by keeping to the directives necessitating the serialization of drugs. The SAP ICH offers its partners/customers with different connectivity choices as per the technological specifications and business needs.



Source: [27]

Fig. 2: Model for SAP – Blockchain Solution by HCL Technologies Limited

The SAP information collaboration hub for the life sciences connects the supply chain partners, such as Contract Manufacturing Organization (CMO), Marketing Authorization Holder (MAH), Third-Party Logistics service provider (3PL), and wholesaler.

SAP announced the launch of the solution known as "Advanced Track and Trace for Pharmaceuticals (ATTP)" which creates exclusive identifiers for the packaging of the drugs. The solutions help store the information into four subsets: batch number, serial number, expiry date, and item number on the basis, of the GSI standard. Immediately upon the shipment of the box, the items get registered, and stored.

The items get registered on the blockchain when a manufacturer ships a box, and this is stored transparently. With the specific code recognition technology, SAP has the capability to track and monitor the packages efficiently at any point, thereby making the process seamless.

Benefits Offered: The benefits offered by the implementation of SAP blockchain solutions in businesses include transparency to enhance

accountability, end-to-end traceability of drug products, minimal losses related to counterfeiting, and efficient recall management.

- *Vault Security*

Vault Security, the US-based company is engaged in addressing the security challenges with cutting-edge applications backed by technologies such as Blockchain. The ivault solutions offered by the company are incorporated into the businesses to augment the logistics management and supply chain networks, with advanced security.

For Businesses: The company offers ivault.verified, which is a smart anti-counterfeiting technological solution incorporated in businesses, to enhance the future of blockchain in supply chain processes. The solutions offered are sustainable, scalable, affordable, and effective, that caters to the needs of the individuals effortlessly.

For Customers: The company is the provider of solutions that protect individuals from being duped in the purchase of fake or counterfeit medications, leading to death, illness, or other harmful effects [28].

The ivault™ blockchain technology is engaged in applications such as anti-counterfeiting technology, distributed computing, blockchain architecture, cloud elasticity, and data warehousing, thereby making the supply chain process an efficacious one. Incorporation of this technological solution, offered by Vault Security offers a revolutionary level of transparency in the supply chain process, thereby ensuring complete protection of the same.

- *Pfizer*

Pfizer, a pharmaceutical giant, is increasingly emphasizing towards the adoption of blockchain technology, to augment medical services. The company announced its partnership with Aimedis, a healthcare startup, to implement the latest technological innovations in the healthcare industry. The company emphasizes on researching and developing the latest therapies, and solutions to improve the lives of individuals across the world. Pfizer operates with nearly 86,000 blockchain projects, to track the health of patients and improve pharmacovigilance [29].

Pfizer's Anti-Counterfeiting Program: Pfizer has announced the introduction of a focused and aggressive campaign to deter, disrupt, and detect the key manufacturers and suppliers of counterfeit drugs in the industry. The company is engaged in working with law enforcement agencies, customs offices, pharmacies, and wholesalers to augment the surveillance of re-packagers and distributors, monitor the distribution channels, and increase inspection coverage. Pfizer also emphasizes on efficacious conduct and management of proactive investigations from the professionals, patients, along with confidential information related to the sales trends and volumes from the enforcement authorities [30]. The anti-fraud measures offered by Pfizer are patient-centric and are designed to identify and disrupt the fraud affecting the operational activities and accordingly address the breakthroughs to save lives.

- *Novartis*

Novartis announced the adoption of blockchain technology in 2016, with various proofs of Concepts (PoCs) [31]. The company emphasizes on securing the supply chain processes, with the tamper-evident and serialization features on the medical products.

Authentifield, under the brand name Novartis, offers anti-counterfeit solutions for testing and rolling out a new generation of mobile spectrometers, to efficaciously monitor the portfolio of drugs, and prevent individuals from being misled to purchase the falsified products [32]. The company is an industry leader in establishing a 29-member blockchain consortium for the European pharmaceutical industry and the European Union, known as Pharmaledger [33].

IX. Challenges in the Implementation of Blockchain Technology to Combat Drug Counterfeiting

The notable challenges in implementing blockchain based track and provenance solutions in the pharmaceutical industry are illustrated as:

Interoperability: The prevailing drug traceability solutions such as e-pedigree, bar codes, serialization, blockchain-based platforms, and solutions, as well as RFID tagslack full interoperability, considering the fact, that, there are no homogenous solutions to make the implementation, adaptability, and integration of the same a seamless one [34]. The various platforms under the hyperledger umbrella are catering tothe issues to ensure maximum adaptability, scalability, and interoperability as well as enable external and internal communication between the organizations.

Implementation and Cost: Designing of an application based on blockchain poses to be a difficult task, because majority of the prevailing solutions are under development. The interoperability, scalability, and privacy tests contribute significantly to this problem. The energy and implementation costs are one of the leading problems being faced by the supply chain processes of the drugs. In addition to this, the existing legacy software systems and platforms are centralized and poses to be inefficient at the time of executing the transactions, thereby leading to higher maintenance and implementation costs. For instance, the hyperledger fabric owes the potential to execute nearly 3500 transactions per second with less power consumption in comparison to Ethereum, owing to the different consensus protocols [35].

X. Conclusion

Counterfeiting acts as a menace to the businesses, with the need to be countered effectively and actively [14]. Lack of stringent regulations and growing international free trade contribute significantly to the rise of counterfeit products. The pharmaceutical segment nowadays, is increasingly being vulnerable to the counterfeiting issues.

Combating these challenges is thereby gaining traction. The drug companies across the different regions in the world are inevitably engaged in offering products which helps in enhancing the patient safety, combating counterfeiting, and caters to regulations, as well as several industry concerns. Cutting-edge technologies such as blockchain unfold unusual opportunities to minimize the scam in the pharmaceutical industry, owing to its unparalleled features such as transparency, distribution, decentralization, and immutability [36].

Several studies have been conducted on a decentralized supply system with immutable and secure tracking capabilities via blockchain technology. This article discusses the different technology-based approaches, that help in averting the counterfeiting of drugs. It emphasizes on the ways blockchain technology helps in combating the counterfeiting of drugs in the pharmaceutical industry, with the inclusion of a few use cases of the businesses incorporating the same.

References

[1] K. S. Ziavrou, S. Noguera, and V. A. Boumba, "Trends in counterfeit drugs and pharmaceuticals before and during COVID-19 pandemic," *Forensic Sci International,* vol. 338, p. 111382, 2022.

[2] M. Amusa, and B. Oluwade, "A historical background of some basic iCT tools used in counterfeit drug control," *African Journal of Computing,* vol. 13, no. 1, pp. 52-61, 2020.

[3] H. Isah, "Information and communication technology in combating counterfeit drugs," *Int. J. Eng. Technol.,* vol. 2, 2012.

[4] European Union, "Unique device identification (UDI) system: Under The EU medical devices regulations, 2017/745 and 2017/746," *European Union,* p. 6, 2020.

[5] P. Sikri, "Counterfeit and fake medicines: challenges for supply chain," Jul. 20, 2022. [Online]. Available: https://www.financialexpress.com/healthcare/pharma-healthcare/counterfeit-and-fake-medicines-challenges-for-supply-chain/2600555/

[6] F. Jamil, L. Hang, K. Kim, and D. Kim, "A novel medical blockchain model for drug supply chain integrity management in a smart hospital," *Electronics 2019,* vol. 8, no. 5, p. 505, 2019. [Online]. Available: https://doi.org/10.3390/electronics8050505

[7] S. V. Nair, "How technology can tackle counterfeiting in the pharmaceutical industry," Apr. 19, 2022. [Online]. Available: https://

www.healthcareradius.in/features/pharma/how-technology-can-tackle-counterfeiting-in-the-pharmaceutical-industry

[8] S. Deakins, "Serialization: Reducing counterfeit drugs and increasing sales," Aug. 3, 2021. [Online]. Available: https://www.pharmtech.com/view/serialization-reducing-counterfeit-drugs-and-increasing-sales

[9] S. Kabra, "Fighting the issue of counterfeiting in india's pharmaceutical industry," May 8, 2022. [Online]. Available: https://timesofindia.indiatimes.com/blogs/voices/fighting-the-issue-of-counterfeiting-in-indias-pharmaceutical-industry/

[10] O. Buowari, "Fake and counterfeit drug: A review," *Afrimedic Journal,* vol. 3, no. 2, pp. 1-4, 2012.

[11] A. Chika, S. O. Bello, A. O. Jimoh, and M. T. Umar, "The menace of fake drugs: Consequences, causes and possible solutions," *Research Journal of Medical Sciences,* vol. 5, no. 5, pp. 257-261, 2011.

[12] "Fighting fake drugs: The role of WHO and pharma," *The Lancet,* vol. 377, no. 9778, p. P1626, May 14, 2011.

[13] D. Bagozzi, and C. Lindmeier, "1 in 10 medical products in developing countries is substandard or falsified," Nov. 28, 2017. [Online]. Available: https://www.who.int/news/item/28-11-2017-1-in-10-medical-products-in-developing-countries-is-substandard-or-falsified

[14] R. Pathak, V. Gaur, H. Sankrityayan, and J. Gogtay, "Tackling counterfeit drugs: The challenges and possibilities," *Pharmaceutical Medicine,* vol. 37, pp. 281-290, 2023.

[15] P. Newton, M. D. Green, F. M. Fernandez, and N. P. John Day, "Counterfeit anti-infective drugs," *The Lancet Infectious Diseases,* vol. 6, no. 9, pp. 602-613, 2006.

[16] A. Delepierre, A. Gayot, and A. Carpentier, "Update on counterfeit antibiotics worldwide; public health risks," *Medecine et Maladies Infectieuses,* vol. 42, no. 6, pp. 247-255, 2012.

[17] R. Jones, "The impact of counterfeits on environmental, social and corporate governance agendas," May 10, 2022. [Online]. Available: https://www.forbes.com/sites/forbesbusinesscouncil/2022/05/10/the-impact-of-counterfeits-on-corporate-environmental-social-and-governance/?sh=88bd85f4dc21

[18] C. Hall, "Technology for combating counterfeit medicine," *Pathog Glob Health,* vol. 106, no. 2, pp. 73-76, 2012.

[19] I. Ungurean, C. Turcu, V. Gaitan, and V. Popa, "An RFID-based anti-counterfeiting track and trace solution," in *Designing and Deploying RFID Applications*, IntechOpen, 2011.

[20] S. Hsia, "How is RFID used for anti-counterfeiting?," Mar. 4, 2022. [Online]. Available: https://nexqo.com/2022/03/rfid-for-anti-counterfeiting/

[21] D. Nandan, "How blockchain can help fight counterfeit drugs in India," Jul. 26, 2018. [Online]. Available: https://www.forbesindia.com/blog/technology/how-blockchain-can-help-fight-counterfeit-drugs-in-india/

[22] P. Sinistra, "Counterfeit drugs 2021 statistics and facts or: How blockchain can fight the pandemic," Feb. 4, 2021. [Online]. Available: https://vaultsecurity.io/counterfeit-drugs-2021

[23] V. Mani, M. Prakash, and W. C. Lai, "Cloud-based blockchain technology to identify counterfeits," *Journal of Cloud Computing,* vol. 11, 2022, Art. no. 67.

[24] W. Alshahrani, and R. Alshahrani, "Assessment of blockchain technology application in the improvement of pharmaceutical industry," in *International Conference of Women in Data Science at Taif University (WiDSTaif'21)*, 2021.

[25] C. Gorenflo, S. Lee, L. Golab, and S. Keshav, "FastFabric: Scaling hyperledger fabric to 20,000 transactions per second," in *International Conference on Blockchain and Cryptocurrency (ICBC'19)*, 2019.

[26] N. A. Beaver, J. D. Lanza, G. B. Solomon, E. Wang, and K. E. Wegrzyn, "Get real: Preventing counterfeit product with blockchain," Oct. 13, 2021. [Online]. Available: https://www.foley.com/en/insights/publications/2021/10/preventing-counterfeit-product-with-blockchain#:~:text=In%20response%20

to%20the%20FDA%E2%80%99s%20call%20in%202019,globe%20in%20hopes%20of%20eliminating%20counterfeit%20medicines.%2011

[27] S. Mahajan, "Combating counterfeit drugs with SAP and blockchain," Apr. 1, 2021. [Online]. Available: https://www.hcltech.com/blogs/combating-counterfeit-drugs-sap-and-blockchain

[28] V. Security, "Vault security," 2018. [Online]. Available: https://vaultsecurity.io/ivault-blockchain-applications

[29] W. McCormick, "Pfizer is latest multinational giant to explore blockchain," Jul. 25, 2018. [Online]. Available: https://cryptobriefing.com/pfizer-multinational-explore-blockchain/

[30] Pfizer, "Combating counterfeit medicine," [Online]. Available: https://www.pfizer.com/products/medicine-safety/counterfeiting

[31] N. Morris, "Novartis explores blockchain's potential for pharmaceuticals," Aug. 22, 2018. [Online]. Available: https://www.ledgerinsights.com/novartis-pharma-blockchain/

[32] Novartis, "Novartis." [Online]. Available: https://live.novartis.com/article/fighting-fake-medicines/fighting-fake-medicines-4-digital-technology-boost#!

[33] A. Takyar, "Blockchain in pharma supply chain - Reducing counterfeit drugs." [Online]. Available: https://www.leewayhertz.com/blockchain-in-pharma-supply-chain/#blockchain_supply_chain_use_cases

[34] G. G. Dagher, J. Mohler, M. Milojkovic, and P. Babu Marella, "Ancile: Privacy-preserving framework for access control and interoperability of electronic health records using blockchain technology," *Sustainable Cities and Society,* vol. 39, pp. 283-297, 2018.

[35] M. Uddin, K. Salah, R. Jayaraman, S. Pesic, and S. Ellahham, "Blockchain for drug traceability: Architectures and open challenges," *Health Informatics Journal,* vol. 27, no. 2, 2021.

[36] A. Quzmar, Mohd. Qatawneh, and S. Al-Maaitah, "Reducing counterfeit drugs with blockchains: A survey," in *International Conference on Information Technology (ICIT'21)*, Amman, Jordan, 2021.

International Journal of Distributed and Cloud Computing
12 (1) 2024, 01-07
http://www.publishingindia.com/ijdcc

Performance Analysis of Cloud and Fog Computing at Emporium

Aadhi Mahishi J.[1], Deepthi B. V.[1], Dharini B. N.[1], Varsha B. R.[1] and Samara Mubeen[2*]
[1]Student, Department of IS&E, JNNCE, Shimoga, Karnataka, India.
[2]Assistant Professor, Department of IS&E, JNNCE, Shimoga, Karnataka, India.
Email: samaramubeen@jnnce.ac.in
*Corresponding Author

Abstract: Cloud computing refers to manipulating, configuring, and accessing the hardware and software resources remotely. It offers online data storage, infrastructure, and application. Cloud computing offers platform independency, as the software is not required to be installed locally on the PC. Hence, the Cloud computing is making business applications mobile and collaborative. Fog computing facilitates the operation of compute, storage and networking services between end devices and cloud computing data centers. Fog computing, also known as fog networking, is a decentralized computing infrastructure in which computing resources and application services are distributed in the most logical, efficient place at any point along the continuum from the data source to the cloud. In a fog environment, the processing takes place in a data hub on a smart device, or in a smart router or gateway, thus reducing the amount of data sent to the cloud. It is important to note that fog networking complements not replaces cloud computing; fogging allows for short-term analytics at the edge, and the cloud performs resource-intensive, longer-term analytics. Emporiums generate a large amount of data from thousands of sensors that it can be useful for monitoring and analyzing, transferring data from sensors to cloud and then giving a feedback to end user so, this could be a problem at emporium because of high delay. Fog computing consider to be temporally near to the sensor; thus will decrease delay. The performance analysis of both architecture is done using iFog simulator in order to know which is the best.

Keywords: Cloud, Fog, iFog simulator.

I. Introduction

The popular trend in today's technology driven world is 'Cloud Computing'. Cloud computing can be referred to as the storing and accessing of data over the internet rather than the computer's hard drive. This means that one person does not need to access the data from either computer's hard drive or over a dedicated computer network (home or office network). Cloud computing means data is stored at a remote place and is synchronized with other web information.

Fog computing is a term created by Cisco that refers to extending cloud computing i.e., fog nodes at different locations are connected to single cloud. Fog computing facilitates the operation of compute, storage and networking services between end devices and cloud computing data centres. Fog computing, also known as fog networking, is a decentralized computing infrastructure in which computing resources and application services are distributed in the most logical, efficient place at any point along the continuum from the data source to the cloud.

Emporiums generate a large amount of data from thousands of sensors that it can be useful for monitoring and analysing. An unprecedented volume of data can crash storage system. Cloud computing could provide storage on demand and processing of system, but cloud could be anywhere far from system, as well as transferring data from sensors to cloud and then giving a feedback to end user so, this could be a problem at emporium because of high delay. Fog computing consider to be temporally near to the sensor thus will decrease delay. The performance

analysis of both architecture is done in order to know which architecture provides best network usage, energy consumption and cost of execution.

Paper is organized as follows: Section I introduction, literature review in Section II, performance metrics considered for studies in Section III, system design in Section IV, implementation in Section V followed by conclusion in Section VI.

II. Literature Review

In order to support the maximum number of user and elastic service with the minimum resource, the Internet service provider invented the cloud computing within a few years, emerging cloud computing has become the hottest technology [1]. Cloud computing has evolved to emerge the most topical IT paradigm in recent times. Cloud computing is rapidly transforming the IT landscape. On a pay-as-you-use basis, cloud consumers can access resources, applications and infrastructure provided by cloud providers. Such access could be in form of applications already deployed by cloud providers for use by the cloud users. It could be in form of the capability to develop and deploy user applications using services of a cloud provider [2]. Cloud computing systems have proven to be effective in meeting the varying demands for computing distributed sharing resources and services. Cloud computing is based on the fundamentals of distributed computing networks, service orientation, virtualization and grid computing system. Further, the usage of grid technologies has also been introduced as a method for accessing cloud computing resources, enabling the monitoring and management of cloud resources [3]. VCC is a new hybrid technology that has a remarkable impact on traffic management and road safety by instantly using vehicular resources, such as computing, storage and internet for decision making. Compare the mechanism with normal cloud computing (CC) and discuss open research issues and future directions [4]. Fog computing provides the data and computation close to end users which reduces the amount of data sent to cloud. Design goals of fog computing platform are to provide low latency to end users and efficient utilization of resources and energy. Some of the challenges in fog computing platform are choice of virtualization technology, fight with latency, security and privacy, network management. Comparison of fog and cloud in terms of latency and bandwidth reveal that fog computing is advantageous for clients in terms of low latency and high bandwidth [5]. The main idea of fog computing is to utilize the processing and storage resources at the edge of network by deploying IoT services on available edge devices to reduce latency and processing cost. The framework facilitates the communication between the devices, fog device orchestration, IoT service deployment, and dynamic resource provisioning in the fog landscape. In addition to these main functionalities, the framework is able to react on various system events, e.g., device accidence, device failure, and device overload [6]. Fog computing is an emerging technology to address computing and networking bottlenecks in large scale deployment of IoT applications. It is a promising complementary computing paradigm to cloud computing where computational, networking, storage are deployed at the edge and network layers in a multi-tier, distributed and possibly cooperative manner. These elements may be virtualized computing functions placed at edge devices or network elements on demand, realizing the "computing everywhere" concept [7]. The primary motivation to employ fog computing in the proposed approach is to minimize the latency as well as network usage in the overall smart car parking system. For demonstrating the effectiveness of the proposed approach for reducing the lag and network usage, simulations have been performed in iFogSim and the results have been compared with that of the cloud-based deployment of the smart car parking system. Experimental results exhibit that the proposed fog-based implementation of the efficient parking system minimizes latency significantly. It is also observed that the proposed fog-based implementation reduces the overall network usage in contrast to the cloud-based deployment of the smart car parking. Deployment models of both cloud and fog architecture is constructed and by using iFogSim simulations have been performed to compare which is efficient [8].

III. Performance Metrices Considered for Studies

A. Energy Consumption

It is defined as the energy consumed while transmitting the data from end devices to cloud.

B. Network Usage

It is defined as the amount of network used while transmitting the data from end devices to cloud.

C. Cost of Execution

Cost spent to complete the process for both cloud and fog architecture.

IV. System Design

Fig. 1 depicts the use case diagram of cloud computing. In this case, there are outlets like Max, Zara and Peach and Lily. Raw data is captured from the cameras in outlets and stored in server. This data is completely sent to cloud for long-term storage.

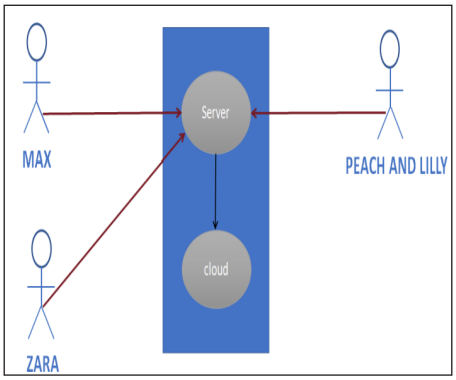


Fig. 1: Use Case Model for Cloud Computing

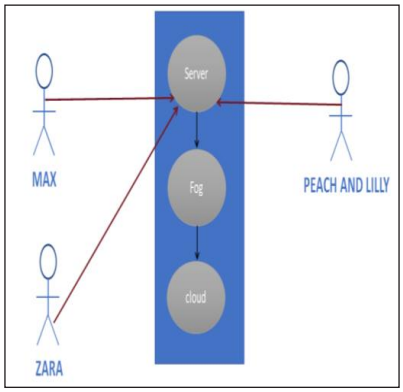


Fig. 2: Use Case Model for Fog Computing

Fig. 2 depicts the use case diagram of fog computing. In this case, there are outlets like Max, Zara and Peach and Lily. Raw data is captured by cameras in outlets and stored in servers. This data is sent to fog node which is close to source and the data is analyzed locally, filtered and then sent to the cloud for long-term storage.

Fig. 3 shows the proposed three-tier architecture of cloud computing. The first tier consist of various stores, the second tier consist of server used for moving the information which is going to be saved on the cloud which is present in the third tier.

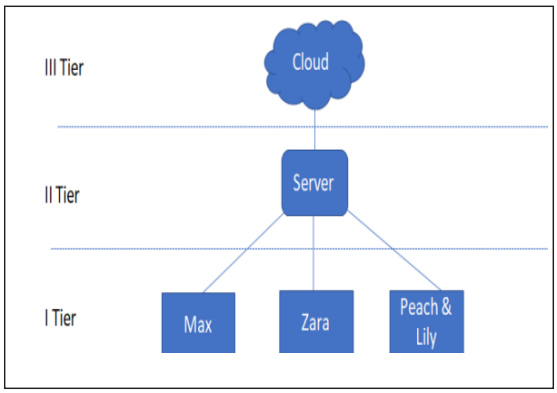


Fig. 3: Three-Tier Architecture of Cloud Computing

Fig. 4 shows the proposed fog architecture of emporium. In addition to above explained cloud architecture it consists of fog node in first tier. Information of every store is stored in single fog node.

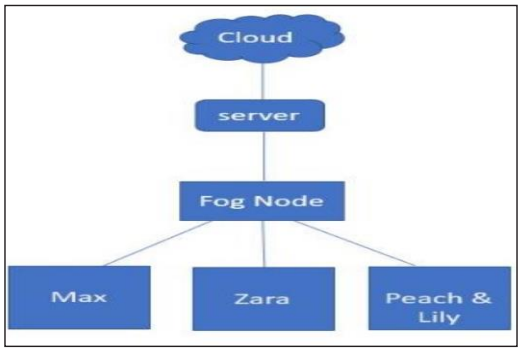


Fig. 4: Fog Architecture of Emporium

Fig. 5 shows the proposed fog architecture of various stores. In addition to above explained fog architecture of emporium, here each store is connected to separate fog nodes. Information of each store is collected and stored at their respective fog nodes.

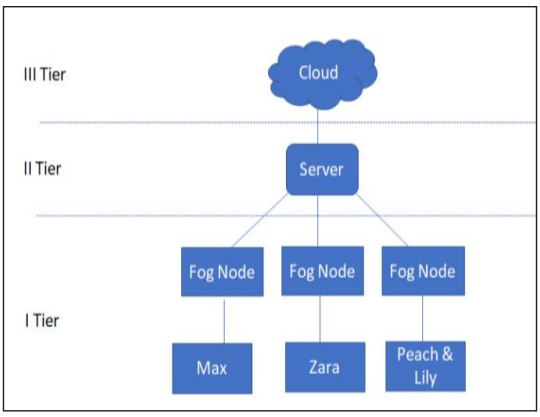


Fig. 5: Fog Architecture of Various Stores

V. Implementation and Result Analysis

Fig. 6 shows the work flow diagram of cloud and fog computing. In cloud computing data captured by camera at every store is collected and sent to the server, from server 100% of the collected data is uploaded to the cloud. Whereas in fog computing collected data of each store is uploaded to the fog node, frequently accessed data is remained in the fog node. Through the server the data which is not stored in fog node is uploaded to the cloud.

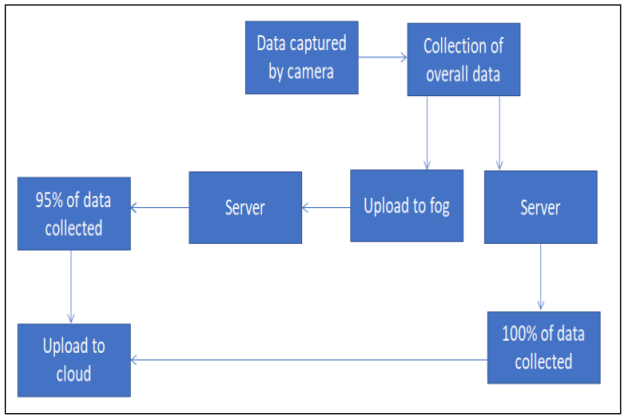


Fig. 6: Work Flow Diagram of Cloud Computing and Fog Computing

Case 1: Performance analysis of an emporium In this case performance analysis of cloud and fog in an emporium is done for the parameters such as energy consumption, network usage and cost of execution. Values of these parameters are compared in order to know which architecture gives better performance Table I shows the energy consumed by cloud and fog nodes for different number of cameras.

Table I: Performance Analysis of Overall Energy Consumed

	Iterations	No. of cameras	Cloud	Fog
Energy Consumed	1	2	2706542.91	2713616.65
	2	4	2714027.03	2727338.67
	3	6	2717184.49	2727338.67
	4	8	2728592.09	2741060.69
	5	10	2739999.69	2754782.70

Fig. 7 shows the graphical representation of cloud and fog for the parameter energy consumption. This graph portrays energy consumed by fog is greater than cloud for different number of cameras.

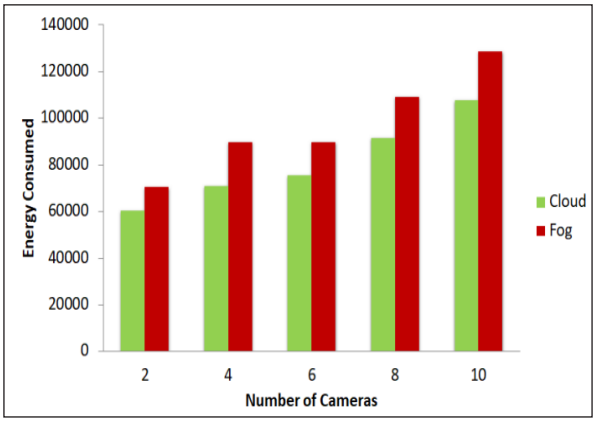


Fig. 7: Graphical Representation of Energy Consumption of Cloud and Fog

Table II shows the network usage by cloud and fog nodes for different number of cameras.

Table II: Performance Analysis of Overall Network Usage

	Iterations	No. of cameras	Cloud	Fog
Network usage	1	2	23722.3	19813
	2	4	47624.9	39626
	3	6	71799.5	59439
	4	8	96246.1	79252
	5	10	120964.7	99065

Fig. 8 shows the graphical representation of cloud and fog for the parameter network usage. This graph portrays network usage by fog is lesser than cloud for different number of cameras.

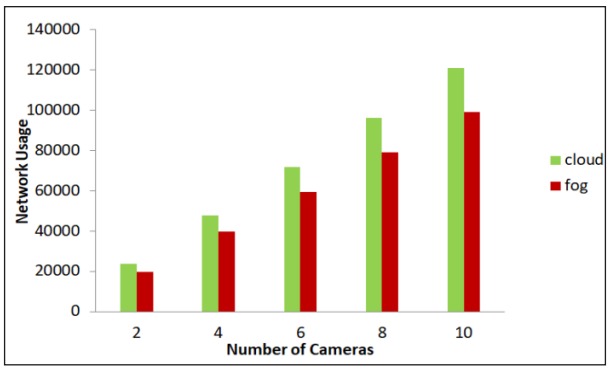


Fig. 8: Graphical Representation of Network Usage of Cloud and Fog

Table III shows the cost of execution by cloud and fog nodes for different number of cameras.

TABLE III: PERFORMANCE ANALYSIS OF OVERALL COST OF EXECUTION

	Iterations	No. of cameras	Cloud	Fog
Cost of Execution	1	2	60314.0	70342.6
	2	4	70924.4	89796.6
	3	6	75400.7	89796.6
	4	8	91573.5	109250.6
	5	10	107746.3	128704.6

Fig. 9 shows the graphical representation of cloud and fog for the parameter cost of execution. This graph portrays cost of execution by fog is greater than cloud for different number of cameras.

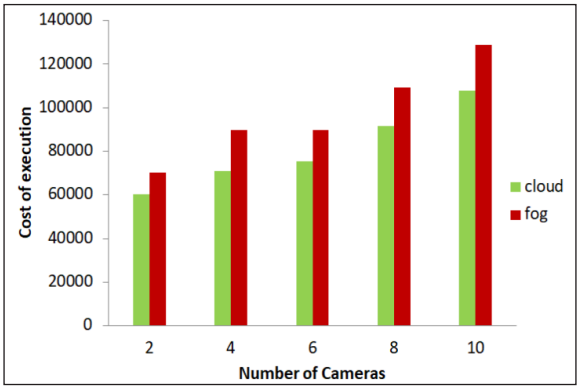


Fig. 9: Graphical Representation of Cost of Execution of Cloud and Fog

Case 2: Performance analysis of individual outlets In this case performance analysis of cloud and fog of various stores is done for the parameters such as energy consumption, network usage and cost of execution. Values of these parameters are compared in order to know which architecture gives better performance. Table IV shows the energy consumed by cloud and fog nodes for various stores like Zara, Max and Peach & Lily.

TABLE IV: PERFORMANCE ANALYSIS OF ENERGY CONSUMPTION OF VARIOUS STORES

	Iterations	No. of cameras	Zara		Max		Peach & Lily	
			Cloud	Fog	Cloud	Fog	Cloud	Fog
Energy Consumed	1	2	177974.59	185500	179410.66	185500	262863.47	292500
	2	4	278914.99	291500	277237.43	291500	360290.40	397500
	3	6	332628	344500	334034.1	344500	422391.73	450500
	4	8	585739.4	609500	492779.5	518500	477395.73	503500
	5	10	641475.74	662500	624543.17	655500	531407.62	556500

Fig. 10 shows the graphical representation of cloud and fog for the parameter energy consumption of Zara. This graph portrays energy consumed by fog is greater than cloud for different number of cameras.

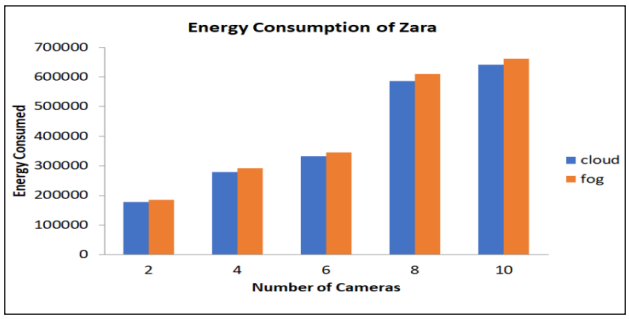


Fig. 10: Graphical Representation of Energy Consumption of Zara

Table V shows the network usage by cloud and fog nodes for various stores like Zara, Max and Peach & Lily.

TABLE V: PERFORMANCE ANALYSIS OF NETWORK USAGE OF VARIOUS STORES

	Iterations	No. of cameras	Zara		Max		Peach & Lily	
			Cloud	Fog	Cloud	Fog	Cloud	Fog
Network Usage	1	2	1631.2	1351	1631.3	1351.5	2426.4	2014.5
	2	4	2023.1	1733.5	2023.1	1733.5	3065.5	2600.5
	3	6	2414.9	2115.5	2414.9	2115.5	3463.5	2982.5
	4	8	2806.7	2497.5	3986.3	3466.5	4671.1	4027.5
	5	10	3198.5	2879.5	5913.8	5098	5913.8	5098

Fig. 11 shows the graphical representation of cloud and fog for the parameter network usage of Zara. This graph portrays network usage by fog is lesser than cloud for different number of cameras.

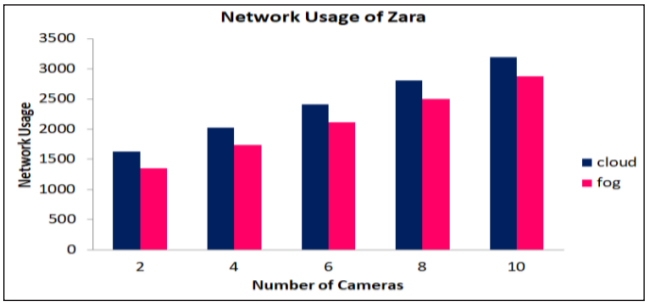


Fig. 11: Graphical Representation of Network Usage of Zara

Table VI shows the cost of execution by cloud and fog nodes for various stores like Zara, Max and Peach & Lily.

TABLE VI: PERFORMANCE ANALYSIS OF COST OF EXECUTION OF VARIOUS STORES

	Iterations	No. of cameras	Zara		Max		Peach & Lily	
Cost of execution			Cloud	Fog	Cloud	Fog	Cloud	Fog
	1	2	91029.6	380900.4	231252.7	573709.39	79959.9	194366
	2	4	111101.5	387900.4	253108.8	623709.39	80947.8	170710
	3	6	155073.5	480900.4	311324.8	573709.39	95780.6	170710
	4	8	213117.6	551709.3	411829.6	687860.1	112100	136829
	5	10	255432.8	573709.3	413337.26	707873.84	113266	157874

Fig. 12 shows the graphical representation of cloud and fog for the parameter cost of execution of Max. This graph portrays cost of execution by fog is greater than cloud for different number of cameras.

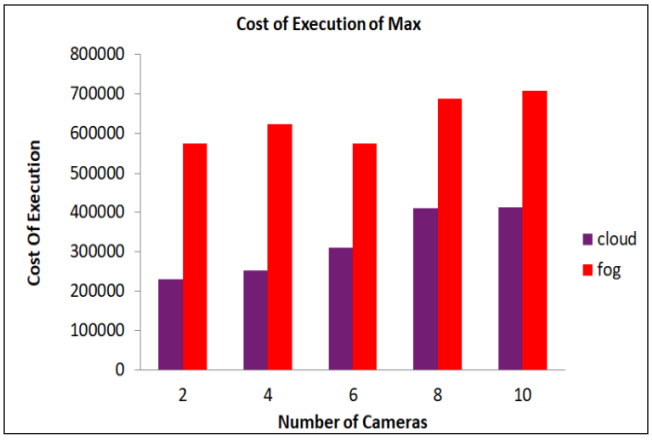


Fig. 12: Graphical Representation of Cost of Execution of Max

VI. CONCLUSION

Cloud Computing is on-demand availability of computer system resources, like computing power, without direct active management by the user. Fog Computing is an extension of cloud computing where fog nodes deployed between data source and cloud. Literature survey of papers on cloud and fog computing is carried on. A gap was identified to find the performance of both computing architecture, by considering the emporium details in iFog simulator. Result analysis concludes that cloud gives the better performance for parameters energy consumption, cost of execution and fog gives better performance in terms of network usage. In future this work can be implemented in different sections in single store and product monitoring can also be done which gives shop owner an alert message when product is going out of stock.

REFERENCES

[1] L. Qian, Z. Luo, Y. Du, and L. Guo, "Cloud computing: An overview," In M. G. Jaatun, G. Zhao, and C. Rong (Ed.), Cloud Computing: CloudCom 2009, *Lecture Notes in Computer Science (LNCS),* vol. 5931, pp. 626-631, 2019.

[2] I. Odun-Ayo, M. Ananya, F. Agono, and R. Goddy-Worlu, "Cloud computing architecture: A critical analysis," *2018 18th International Conference on Computational Science and Applications (ICCSA),* Covenant University, Ota, Ogun, Nigeria, 2018, pp. 6256-6260.

[3] N. S. Aldahwan, and Muhd. S. Ramzan, "Descriptive literature review and classification of community cloud computing research," *Scientific Programming*, pp. 459-508, 2021.

[4] Md. Whaiduzzaman, M. Sookhak, A. Gani, and R. Buyya," A survey on vehicular cloud computing," *Journal of Network and Computer Applications*, pp. 135-142, 2016.

[5] S. Yi, Z. Hao, Z. Qin, and Q. Li, "Fog computing: Platform and applications," *2015 Third IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb),* 2015, pp. 501-520.

[6] K. Bachmann, "Design and implementation of a fog computing framework," *Software Engineering & Internet Computing*, pp. 612-630, 2017.

[7] P. Habibi, Mohd. Farhoudi, S. Kazemian, S. Khorsandi, and A. Leon-Garcia, "Fog computing: A comprehensive architectural survey," *IEEE Access*, vol. 8, pp. 69105-69133, 2020.

[8] K. Sattar Awaisi, A. Abbas, M. Zareei, H. Ali Khattak, Muhd. U. Shahid Khan, M. Ali, I. Ud Din, and S. Shah, "Towards a fog enabled efficient car parking architecture," *IEEE Access*, vol. 7, pp. 159100-159111, 2019.

International Journal of Distributed and Cloud Computing
12 (1) 2024, 08-16
http://www.publishingindia.com/ijdcc

A Comparative Assessment of Load Balancing Techniques in Cloud Computing: A Review

John Abiodun Oladunjoye[1] and Egwom Onyinyechi Jessica[2*]
[1]Senior Lecturer, Department of Computer Science, Federal University, Wukari, Nigeria.
Email: oladunjoye.abbey@yahoo.com
[2]Assistant Lecturer, Department of Computer Science, Federal University, Wukari, Nigeria.
Email: egwom@fuwukari.edu.ng
[*]Correspondence Author

Abstract: Cloud computing is a resilient framework that enables users and organizations to procure services tailored to their needs. This model encompasses many services, including storage, deployment platforms, seamless access to web services, and more. However, load balancing poses a common challenge in the cloud, impacting the ability to uphold application performance aligned with the required quality of service (QoS) expected by enterprises from cloud providers. Achieving an equitable distribution of workloads among servers proves challenging for cloud providers. An effective load-balancing technique should optimize resource utilization in virtual machines, ensuring high user satisfaction. This paper presents a comprehensive review of various load balancing techniques in terms of static and dynamic and the overall performance of these algorithms using some popular performance metrics. Based on the performance measures, Cuckoo Search and Firefly was the most effective algorithm. This research makes a contribution by describing current methodologies and by helping researchers identify research issues connected to load balancing, to decrease response times and prevent server failures.

Keywords: Algorithms, Cloud computing, Load balancing, Performance metrics, Resource management, Virtual machine.

I. Introduction

Cloud computing technology is undergoing significant growth in network technology, propelled by advancements in communication technology, widespread Internet usage, and its capacity to tackle large-scale challenges [1], [2]. The term cloud computing according to S. Kaur, and V. Pandey [3] refers to the distribution of various computing services over the internet (cloud), encompassing servers, databases, networks, and software. Cloud computing facilitates the development of new applications and services, data backup and recovery, hosting of websites and blogs, streaming of audio and video, and the analysis of data patterns for predictive purposes [4]. Cloud users can leverage hardware and software applications as resources through the Internet.

The adoption of cloud computing has brought about a significant shift in how businesses perceive their available resources. These computers are dynamically allocated and presented as a unified computing resource or resources, guided by negotiated service-level agreements between the service provider and consumers [5]. Cloud computing also referred to as distributed computing, introduces innovative perspectives for both service providers and end consumers. To accommodate these evolving

aspects, novel design strategies are implemented, communicating concepts through three service delivery models: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) [6] as shown in Fig. 1.

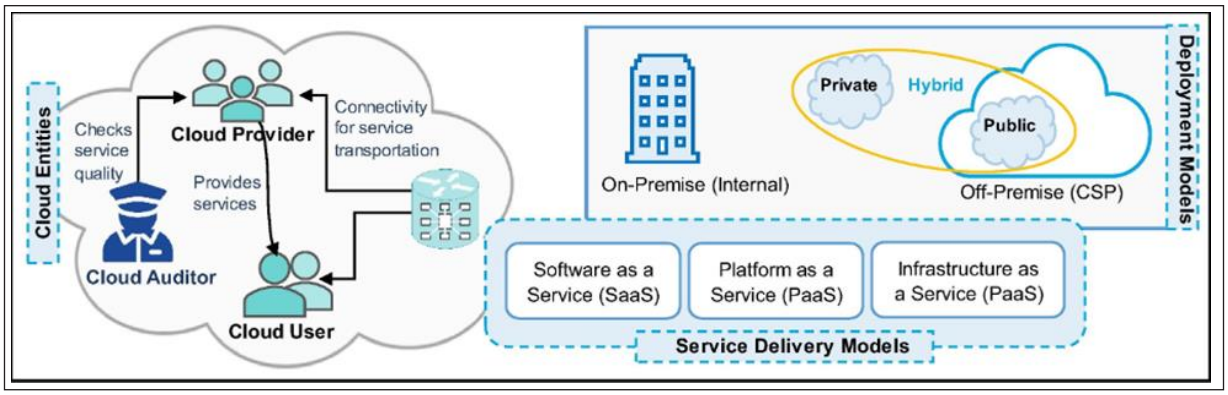


Fig. 1: A Cloud Computing Model [5]

It is important to have a reliable and consistent network connection to deliver services to clients, utilizing these cloud-based platforms effectively. Efficient and scalable characteristics of cloud computing may be attained by implementing effective cloud resource management [5], [7]. As a result, a lot of researchers have increased their focus on the load balancing of these resources in cloud computing. Load balancing refers to redistributing the overall workload among the individual nodes within a collective system which aims to optimize resource utilization and enhance job response time, while simultaneously addressing the issue of certain nodes being overloaded while others remain underutilized [5].

The problem of load balancing remains a significant challenge within cloud computing. Hence, the need for this research to provide a review of load-balancing techniques using different performance metrics. The purpose of this study is to conduct a comprehensive examination of load-balancing algorithms in the context of cloud computing. This study focuses on evaluating several load-balancing approaches utilizing multiple performance metrics providing an in-depth review of the latest advancements in load balancing and encouraging future investigation in this field. The rest of the research is organised as follows: Section II: discusses the concept of load balancing. In Section III: the various methods of load balancing, including static and dynamic load balancing algorithms, are discussed. Section IV: analyzed and showed a performance comparison of the algorithms based on certain specified performance metrics. Section V: provides the study's conclusion.

II. Concept of Load Balancing

Load balancing is a crucial strategy used to evenly spread the burden among different resources participating in computing operations inside a network, resulting in improved performance [8]. This workload includes memory utilization, CPU capacity, and network traffic. Within the domain of cloud computing, load balancing becomes a crucial issue, requiring the fair allocation of tasks among the available resources to achieve shorter response times [9]. This guarantees that the Virtual Machines (VMs) in the system are never overloaded at any one moment. Load balancing is seen as the coordination of a collection of servers engaged in the same service, executing identical tasks to enhance service quality by managing fluctuating customer loads over time [10].

According to E. J. Ghomi, A. M. Rahmani, and N. N. Qader [11] The fundamental goal of load balancing is to improve the availability, throughput, and dependability of a system, while also ensuring system stability, optimizing resource usage, and providing fault-tolerant capabilities. Careful incident handling is required since the chance of failures grows with the number of servers and high availability refers to the capability of maintaining uninterrupted service even in the presence of many faults occurring

simultaneously [11]. Load balancing aims to achieve optimal utilization of resources, which is a crucial objective for enhancing the efficiency of the cloud model. Attaining a high level of data processing capacity, which is crucial for a system that performs well, can only be accomplished when the tasks and available resources are fairly allocated throughout various nodes with the primary objectives encompassing attaining a minimal reaction time and averting bottlenecks [12], [13]. Within the domain of cloud computing, efficient load balancing enables data centers to effectively manage issues such as excessive or insufficient workload distribution among virtual machines as shown in Fig. 2. To keep up the performance of cloud applications an efficient Load Balancer must be provided an Unequal load distribution have many disadvantage one of them been Task Scheduling. Without proper task scheduling, the resources will not be efficiently utilized [9].

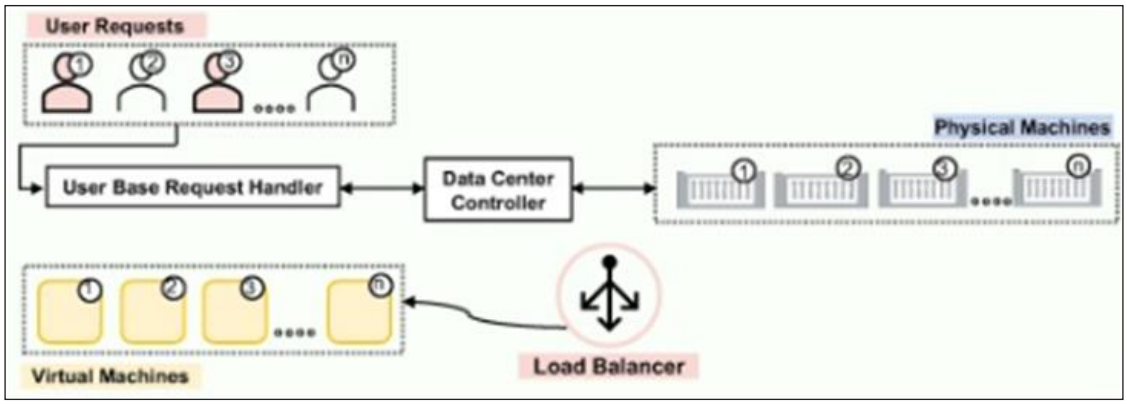


Fig. 2: Workflow in Cloud Computing [9]

The implementation of load-balancing techniques faces significant problems due to the influence of different physical and logical elements on the success of the strategy. Challenges arise from several factors, including the dispersion of graphical nodes, the migration of virtual machines, the complexity of algorithms, the presence of heterogeneous nodes, the risk of single points of failure, and the scalability of load balancers [11], [12], [13]. To address these issues, several load-balancing techniques have been proposed in the domain of cloud computing.

III. Classification of Load Balancing Techniques

The categorization of load balancing (LB) algorithms is contingent upon the operational state of the system, resulting in classifications as static, dynamic or hybrid. In essence, these classifications delineate whether the load-balancing approach employed is unchanging, adaptive to real-time fluctuations, or a combination of both static and dynamic elements [7] as shown in Fig. 3.

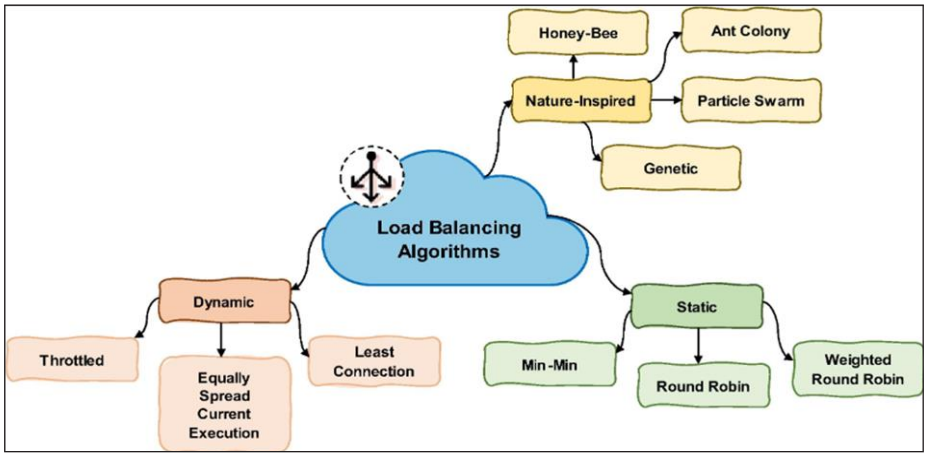


Fig. 3: Types of Load Balancing Algorithm

A. Static Algorithm

Static algorithms obtain information pertaining to the system and identify the existing resources for use before commencing thereby maintaining a consistent distribution of workloads regardless of changing conditions [4]. Static algorithms gather information about the system and identify available resources before starting. They evenly distribute workloads, regardless of changing conditions, using existing virtual machines (VMs) until a task is completed [10]. This works well when VM capabilities are similar. However, static algorithms face a challenge because they rely on fixed information and struggle to adapt to dynamic load changes in VMs effectively [14]. They are not flexible and cannot accommodate dynamic changes in attributes. These algorithms, also known as offline algorithms, need to know VM information in advance [4]. Despite their limitations, static algorithms generally achieve better overall performance than dynamic algorithms. It's important to note that when static load-balancing algorithms assign tasks to nodes, they don't consider the previous task's state and functionality of the node.

Some popular static load-balancing algorithms are presented in this section.

- *Round Robin Algorithm:* Round Robin, a static load balancing algorithm, operates without considering preceding states. Its simplicity lies in the utilization of the Round Robin Method for distributing tasks [10]. The algorithm randomly chooses the first node and then allocates jobs evenly to all other nodes using the Round Robin Method as seen in Fig. 4. An essential advantage of Round Robin is its independence from inter-process communication [15]. This algorithm provides faster response in the case of equal workload distribution among processes However, due to the lack of information about processors' running times, certain tasks may experience heavy loads.
- *Min-Min Load Balancing Algorithm:* This algorithm is user-friendly and operates quickly. It enhances performance by completing tasks in a series [16]. It calculates the time for each task and assigns it to virtual machines (VMs) based on the shortest completion time of existing tasks. This continues until all tasks are assigned to VMs. The algorithm performs well due to numerous smaller tasks, but there's a risk of neglecting larger tasks, leading to starvation as smaller tasks are prioritized over bigger ones [6].

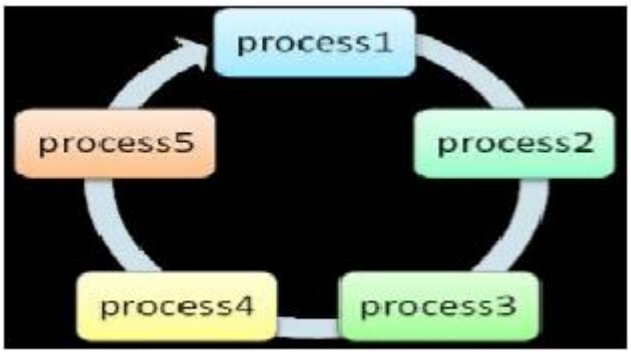


Fig. 4: Round Robin Load Balancing Algorithm

- *Max Min Load Balancing Algorithm:* This approach is similar to Min-Min load balancing, with a specific emphasis on the computation time of jobs [13]. Under this methodology, all tasks are given to the system, and the algorithm computes the duration required to complete each job. The machine is allocated the duty with the minimum time needed. The Max-Min method outperforms the Min-Min algorithm due to its ability to handle a larger job in addition to smaller tasks inside the set [9]. The Max-Min algorithm effectively manages jobs of varying durations, regardless of their length.
- *Opportunistic Load Balancing Algorithm:* The basic principles of this algorithm prevent it from comprehending the VM's activity level. Therefore, it ensures that every node in the system is actively working by randomly assigning those tasks [14]. Things go slowly, though, because it doesn't account for how long things take to complete. Consequently, it fails to provide reliable outcomes when attempting to distribute the labor evenly [14].

B. Dynamic Algorithm

Dynamic algorithms are a type of algorithm that differs from static algorithms because they exhibit

flexibility [9], [4]. These algorithms work based on rescheduling the tasks allocated to them over the available VM while carrying out the work. The main benefit is that task selection is influenced by real-time situations, leading to improved system performance [4]. Dynamic algorithms can be implemented in two ways:

- *Distributed System:* In this setup, all nodes in the system communicate with each other, and each node participates in executing a load-balancing algorithm. The responsibility of load balancing is shared among all nodes [17].
- *Non-Distributed System:* This can be centralized or semi-distributed. In a centralized system, a central node manages the load balancing for the entire system, and other nodes interact with this central node [18]. In a semi-distributed system, each cluster has a node that performs load balancing for the entire system. If the central node of a cluster fails, it affects only that cluster's functionality [18].

Some of the well-known dynamic algorithms are:

- *Throttled Load Balancer (TLB):* This algorithm generates a table containing virtual machines and their current states (available or busy). When a specific task is assigned to a virtual machine, a request is sent to the data center's control unit. This unit seeks the most suitable VM based on its capabilities for the given task [7]. The Table Allocator (TA) maintains a list, known as the index table or allocation table, containing all VMs and their corresponding states (available, busy, or idle). If an available VM with sufficient space is found, the task is allocated to that VM. If no suitable VM is available, TA returns −1, and the request is queued for prompt processing [5]. The load balancer's role is to locate a suitable virtual machine upon receiving a client request. While TA outperforms the Round Robin algorithm, it lacks consideration for more advanced load balancing requirements, such as processing time [5].
- *Ant Colony Optimization:* This ACO-based algorithm efficiently distributes workloads among cloud nodes. Drawing inspiration from ant behavior, it seeks alternative pathways when encountering obstacles, establishing new routes between nodes [18]. The algorithm begins by initializing a table, facilitating data flow, and reaching the required nodes' threshold level. As the flow passes through nodes, the algorithm evaluates each one. If a node is underloaded, it utilizes the highest Trailing pheromone (TP) to find the route for the underloaded node [18]. The table is updated until the node reaches its threshold limit. However, upon reaching the limit, the algorithm uses a Foraging Pheromone (FP) to explore new sources, updating the table until it finds an underloaded node, at which point it reallocates resources. This process repeats until the entire process is completed.
- *Honey Bee Algorithm:* This algorithm operates based on the behavior of bees searching for honey. Its primary goal is to distribute the workload on virtual machines (VMs) efficiently, avoiding both excessive resource use and underutilization [19]. The algorithm selects a VM that meets two key criteria: it receives fewer tasks compared to other VMs, and its processing time falls within the average processing time of all VMs [19].
- *Load Balancing Technique Based on Cuckoo Search and Firefly:* The cuckoo search algorithm draws inspiration from the nesting behavior of cuckoo birds. Similar to how these birds lay their eggs in the nests of other birds, the algorithm carries forward the best-quality eggs to the next generation and abandons the worst nests [20]. In the context of cloud computing, the Cuckoo Search with Firefly, a hybrid algorithm, is employed for load balancing. This algorithm efficiently schedules tasks, assigning overloaded tasks to under-loaded virtual machines (VMs). It quickly identifies the best VM, enhancing load balance efficiency and preventing task imbalances across the system [21]. The load balancing (LB) technique initiates by scheduling tasks using the round-robin method. It then calculates the capacity and load of each VM, identifying overloaded and under-loaded tasks through Cuckoo Search

with Firefly. This method ensures task migration from overloaded VMs to under-loaded VMs, preventing overall system task imbalances [21]. Notably, the approach boasts a high interchange rate and requires fewer tuning parameters compared to existing methods.

- *First Come First Serve Algorithm:* This algorithm operates by assigning new tasks to resources with the shortest waiting time, meaning those resources handle the fewest tasks. It follows a sequential execution approach, starting with the first task and progressing through the queue until completion [16]. The algorithm is chosen because it allocates tasks to virtual machines without taking into account the specifications of available virtual machines or the time required for tasks in the queue when assigning new tasks [22]. To determine the most effective load-balancing algorithm for optimal performance in cloud applications, various popular metrics are employed to measure their effectiveness.

IV. Performance Comparison of Load Balancing Algorithm

This section presents the key metrics commonly examined by authors when assessing the latest advancements in load-balancing algorithms. These metrics play a crucial role in shaping and creating effective load-balancing algorithms. They serve as indicators of an algorithm's performance in cloud applications. Researchers [16], [23], [9], [13], [5] utilize certain parameters to assess their proposed algorithms, emphasizing the need for adjustments to prevent imbalances in the cloud computing environment. From the review, the popular performance metrics that will be used in this research are:

- *Resource Utilization (RU):* The utilization of system resources such as memory and CPU is a measure of Resource Utilization (RU) within the cloud Data Center. RU gauges the extent to which resources are employed. As service demand grows, maintaining a high RU becomes crucial. A maximum RU is necessary to ensure optimal performance for the load-balancing algorithm.
- *Complexity:* An algorithm should exhibit optimal performance even when faced with unforeseen circumstances. This implies that irrespective of an upsurge in tasks and load, the algorithm should maintain scalability, ensuring consistently high performance for the load-balancing algorithm.
- *Throughput (TP):* It denotes the count of job requests completed and successfully processed per unit of time within the virtual machine. It signifies the volume of data moving from one location to another. High throughput is indicative of effective load-balancing algorithm performance.
- *Response Time (RT):* The duration the algorithm requires to address a task, encompassing waiting, transmission, and service times, is called response time. It signifies the time necessary to answer a user's inquiry, and an effective load-balancing algorithm aims for the minimum response time.
- *Speed:* The total time needed to finish all tasks and distribute resources to users within the system is known as the total completion time. This metric is critical in the scheduling process within the cloud environment, providing a measure of the time required to process a specific set of tasks. A good load-balancing algorithm aims to minimise this metric, ensuring efficient task completion.
- *Network Overhead (NO):* It is the level of additional work generated during the execution of the load balancing algorithm, which may arise from a significant number of task migrations and inter-process communication. An evenly distributed system load leads to minimal overhead caused by the load-balancing algorithm.
- *Fault Tolerance (FT):* A robust load-balancing system should function effectively even in the event of failures in some components. This implies that if one virtual machine (VM) is overloaded, another available VM should be

capable of carrying out tasks. Ensuring high fault tolerance is crucial for the optimal performance of the load-balancing algorithm.

- *Migration Time (MT):* This represents the overall time required to transfer a task from one virtual machine (VM) to another, and this migration process should occur without impacting the system's availability. The efficiency of this process is closely tied to the virtualization concept in the cloud. A low migration time is crucial for optimal performance in a load-balancing algorithm. Table I represents the performance of the Load Balancing Algorithms reviewed in this research.

Table I: Performance of the Load Balancing Algorithms

Performance Metrics	Load Balancing Algorithms								
	Round robin	MIN-MIN	MIN-MAX	Opportunistic	Cuckoo Search and Firefly	Throttled Algorithm	Ant Colony	Honey Bee	First Come First Serve
Throughput	Low	Moderate	High	Low	High	High	Low	Low	Moderate
Complexity	Low	Low	Moderate	Low	Moderate	Moderate	Moderate	High	Moderate
Response time	Low	Low	Low	Moderate	Moderate	High	High	High	High
Speed	Low	Low	High	Low	High	High	Low	Moderate	Moderate
Network overhead	Low	High	Moderate	Low	Moderate	Low	High	High	Low
Fault Tolerance	Low	Low	High	Low	High	High	High	Moderate	Low
Migration time	Low	Moderate	Moderate	Low	High	High	Low	Low	Low
Resource utilization	High	Moderate	High	Low	High	High	Moderate	High	High

V. Conclusion

Load balancing holds significance in cloud computing as it improves the distribution of workloads and optimises resource utilisation, consequently reducing the overall system response time. This research delves into the categorization of load-balancing algorithms, providing an overview of existing algorithms designed to balance loads in cloud computing and their performance. These performance metrics are crucial factors for designing a robust algorithm capable of thriving in a dynamic cloud environment. The Max-Min algorithm came in second place among the algorithms this study examined, after Cuckoo Search and Firefly which is the best algorithm. This research contributes by aiding researchers in identifying load balancing-related research challenges, particularly in further minimizing response times and preventing server failures, and by offering a summary of existing techniques. Future researchers can build upon this foundation by exploring the development of more dynamic, intelligent algorithms addressing fault tolerance issues to further elevate the efficiency of cloud computing.

References

[1] R. Boutaba, Q. Zhang, and M. F. Zhani, "Virtual machine migration in cloud computing environments," in *Advances in Systems Analysis, Software Engineering, and High-Performance Computing Book Series*, 2014, pp. 383-408, doi: https://doi.org/10.4018/978-1-4666-4522-6.ch017.

[2] U. B. Usman, F. U. Ambursa, E. J. Onyiyechi, and T. J. Jaremiah, "Virtual machine migration techniques in cloud computing environment: A review of load balancing algorithms," *International Journal of Artificial &*

Computational Intelligence, vol. 1, no. 3, Sep. 2020.

[3] S. Kaur, and V. Pandey, "A survey of virtual machine migration techniques in cloud computing," *International Institute for Science, Technology, and Education (IISTE) Computer Engineering and Intelligent Systems,* vol. 6, no. 7, pp. 28-34, 2015, ISSN: 2222-1719 (Paper), ISSN: 2222-2863 (Online).

[4] S. Afzal, and G. Kavitha, "Load balancing in cloud computing – A hierarchical taxonomical classification," *Journal of Cloud Computing*, vol. 8, no. 1, Dec. 2019, doi: https://doi.org/10.1186/s13677-019-0146-7.

[5] D. A. Shafiq, N. Z. Jhanjhi, and A. Abdullah, "Load balancing techniques in cloud computing environment: A review," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 3910-3933, Jul. 2022, doi: https://doi.org/10.1016/j.jksuci.2021.02.007.

[6] P. Kumar, and R. Kumar, "Issues and challenges of load balancing techniques in cloud computing," *ACM Computing Surveys*, vol. 51, no. 6, pp. 1-35, Feb. 2019, doi: https://doi.org/10.1145/3281010.

[7] S. Hamadah, "A survey: A comprehensive study of static, dynamic and hybrid load balancing algorithms," *International Journal of Computer Science and Information Technology & Security (IJCSITS)*, vol. 7, no. 2, pp. 27-26, Mar.-Apr. 2017, ISSN: 2249-9555.

[8] N. S. A. Chawla, "Efficient cost scheduling algorithm with load balancing in a cloud computing environment," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 3, no. 6, pp. 5138-5143, Jun. 2015, doi: https://doi.org/10.15680/ijircce.2015.0306026.

[9] S. K. Mishra, B. Sahoo, and P. P. Parida, "Load balancing in cloud computing: A big picture," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 2, pp. 149-158, Feb. 2020, doi: https://doi.org/10.1016/j.jksuci.2018.01.003.

[10] R. S. Sajjan, and B. R. Yashwantrao, "Load balancing and its algorithms in cloud computing: A survey," *International Journal of Computer Sciences and Engineering,* vol. 5, no. 1, pp. 95-100, 2017.

[11] E. J. Ghomi, A. M. Rahmani, and N. N. Qader, "Load-balancing algorithms in cloud computing: A survey," *Journal of Network and Computer Applications*, vol. 88, pp. 50-71, Jun. 2017, doi: https://doi.org/10.1016/j.jnca.2017.04.007.

[12] S. Talwani, and J. Singla, "A comprehensive review of virtual machine migration techniques in cloud computing," *Social Science Research Network*, Jan. 2020, doi: https://doi.org/10.2139/ssrn.3564971.

[13] A. A. Alkhatib, A. Alsabbagh, R. Maraqa, and S. AlZu'bi, "Load balancing techniques in cloud computing: Extensive review," *Advances in Science, Technology and Engineering Systems Journal*, vol. 6, no. 2, pp. 860-870, Apr. 2021, doi: https://doi.org/10.25046/aj060299.

[14] W. Duan, X. Tang, J. Zhou, W. Jing, and G. Zhou, "Load balancing opportunistic routing for cognitive radio ad hoc networks," *Wireless Communications and Mobile Computing*, vol. 2018, pp. 1-16, Dec. 2018, doi: https://doi.org/10.1155/2018/9412782.

[15] M. A. Shahid, N. Islam, M. M. Alam, M. M. Su'ud, and S. Musa, "A comprehensive study of load balancing approaches in the cloud computing environment and a novel fault tolerance approach," *IEEE Access*, vol. 8, pp. 130500-130526, Jan. 2020, doi: https://doi.org/10.1109/access.2020.3009184.

[16] A. Thakur, and M. S. Goraya, "A taxonomic survey on load balancing in cloud," *Journal of Network and Computer Applications*, vol. 98, pp. 43-57, Nov. 2017, doi: https://doi.org/10.1016/j.jnca.2017.08.020.

[17] R. Panwar, and B. Mallick, "Load balancing in cloud computing using dynamic load management algorithm," *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, Oct. 2015, doi: https://doi.org/10.1109/icgciot.2015.7380567.

[18] N. R. Tadapaneni, "A survey of various load balancing algorithms in cloud computing," *Int.*

J. Sci. Advance Res. Technol., vol. 6, no. 4, pp. 484-487, 2020.

[19] W. Hashem, H. Nashaat, and R. Rizk, "Honey bee-based load balancing in cloud computing," *KSII Transactions on Internet and Information Systems*, vol. 11, no. 12, pp. 5694-5711, Dec. 2017, doi: https://doi.org/10.3837/tiis.2017.12.001.

[20] M. Yakhchi, S. M. Ghafari, S. Yakhchi, M. Fazeli, and A. Patooghi, "Proposing a load balancing method based on Cuckoo Optimization Algorithm for energy management in cloud computing infrastructures," *International Conference on Modeling, Simulation, and Applied Optimization*, May 2015, doi: https://doi.org/10.1109/icmsao.2015.7152209.

[21] K. K. Kumar, T. Ragunathan, D. Vasumathi, and P. V. S. R. B. Prasad, "An efficient load balancing technique based on Cuckoo search and Firefly algorithm in cloud," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 3, pp. 422-432, Jun. 2020, doi: https://doi.org/10.22266/ijies2020.0630.38.

[22] P. Prajapati, and A. K. Sariya, "A review: Methods of load balancing on cloud computing," *Int. J. Res. Anal. Rev.*, vol. 6, 2019, doi: https://doi.org/10.6084/m9.doi.one.IJRAR19J2346.

[23] T. Tamilvizhi, and B. Parvathavarthini, "A novel method for adaptive fault tolerance during load balancing in cloud computing," *Cluster Comput.,* vol. 22, no. S5, pp. 10425-10438, Sep. 2019.

International Journal of Distributed and Cloud Computing
12 (1) 2024, 17-29
http://www.publishingindia.com/ijdcc

An ISM-MICMAC Based Approach to Analyze Cloud Computing Adoption in Organizations

Abhinav Pandey[1*] and Sanjeev Swami[2]
[1]Assistant Professor, Department of Management, Faculty of Social Sciences, Dayalbagh Educational Institute, Agra, Uttar Pradesh, India. Email: abhinavpandey@dei.ac.in
[2]Professor, Department of Management, Faculty of Social Sciences, Dayalbagh Educational Institute, Agra, Uttar Pradesh, India.
[*]Corresponding Author

Abstract: Cloud computing adoption in any business organization, represents a complex system which can be characterized as having a large number of subsystems like legacy information technology infrastructure, the new information technology architecture being created owing to cloud adoption, the techno-managerial complexity of managing infrastructure of disparate information architectures, integration of legacy architecture with new cloud architecture, managing the emergent distributed and software system, database management and manpower planning in terms of skill orientation in view of cloud adoption. In addition to the complexities of cloud adoption, the trade-offs like security vs. scalability, control vs. agility, standardizations vs. lock-ins, puts cloud computing adoption in "pluralist" category in relationship continuum, where stakeholder have contradictory view point about a problem situation. Therefore, cloud adoption in any organization represents an unclear and complex situation with poorly articulated interactions amongst associated elements, which needs to be well-defined to understand the process of cloud adoption and is impact on adopting organization.

Keywords: Cloud computing, Cloud adoption, ISM, MICMAC analysis.

I. Introduction

Cloud Computing conceptualizes delivering traditional information technology facilities as commoditized services, similar to deliverance model of utilities such as water, electricity, gas etc.; where the users of cloud computing utilizes the information technology facilities as end user and are being billed for the usage of that service. This concept of cloud computing as commoditized service, is an extension of utility computing vision, which represents new scenario of how information technology (IT) services are developed, deployed, scaled, updated, maintained and procured (Buyya *et al.*, 2008) [1].

Cloud computing adoption enables ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction (Mell and Grance, 2011) [2]. This feature of cloud computing represents a convergence of two major trends; firstly the ability of contemporary computers being employed efficiently through highly scalable hardware and software resources, and, secondly business agility where information technology act as a competitive tool through fast preparation, parallel execution, use of compute-intensive business analytics that respond in real time to user necessities (Marston *et al.*, 2011) [3].

A firm's migration towards cloud computing begins with the consolidation of server, storage, and network resources, that works to control redundancy, decrease wasted space, and increase equipment usage [4]. Hardware such as servers, storage

devices, and other components are treated as a pool of resources rather than a discrete system, thereby allowing the allocation of resources on demand [5]. Virtualization forms a solid foundation for all cloud architectures. It enables the abstraction and aggregation of all data resources, thereby creating a unified resource that can be shared by all the information technology application. Automation optimizes a firm's virtualized resources, through a transformative procedure and critical information technology processes become more dynamic and flexible management is achieved by trustworthy policies. In addition to automation, utility which include both self-service and metering, and provide feedback about the cost of the resources allocated, is necessary requirements in creating a cloud service.

A cloud infrastructure consists of hardware and software that are necessary to support the cloud services being provided [4], [5]. The hardware resources are typically server, storage and network components and also referred as physical layer. The abstraction layer consists of the software deployed across the physical layer to manifests the essential cloud characteristics and sits above the physical layer (Meel and Grance, 2011). Through cloud internetworking, disparate cloud systems can be linked in such a way, so as to accommodate both the unique nature of cloud computing and the running of legacy information technology workloads.

Cloud computing adoption in any organization, represents a complex system which can be characterized as having a large number of subsystems like legacy information technology infrastructure, the new information technology architecture being created owing to cloud adoption, the techno-managerial complexity of managing infrastructure of disparate information architectures, integration of legacy architecture with new cloud architecture, managing the emergent distributed and software system, database management and manpower planning in terms of skill orientation in view of cloud adoption. As organization looks to adopt cloud computing, the adoption process will stimulate many loosely structured interactions amongst these subsystems, the outcome of which is not predetermined.

Along with the above discussed complexities of cloud adoption, the trade-offs like security vs. scalability, control vs. agility, and standardizations vs. lock-ins [6], [7]; gives rise to a situation where basic interests of cloud service provider and adopting organization are compatible i.e. to adopt cloud offerings; but they do not share the same values and beliefs. Therefore, cloud adoption in any organization represents an unclear and complex situation with poorly articulated interactions amongst associated elements, which needs to be well-defined to understand the process of cloud adoption and is impact on adopting organization.

Interpretive structural modelling, which is defined as the process that transforms unclear and poorly articulated mental models into visible well defined models useful for many purposes (Satsangi, 1985); and enables intervention in ill-structured problem situations by providing a hierarchal model of the associated elements with reduced complexities under the given contextual relationship, to allow holistic understanding of the problem situation , becomes an ideal technique to analyse cloud adoption in business organization.

II. Literature Review

The review discourse on cloud computing can primarily be divided into two broad category technological and business issue [8]. Technological review discourse contemplates cloud computing in performance subcategory, which includes studies such as performance comparisons across different clouds (Iosup *et al.,* 2011), to improve dynamic resource allocation (Streitberger and Eymann, 2009; Warneke and Kao, 2011), to estimate performance of cloud network with nodes failure (Lin and Chang, 2011), and to improve interoperability across different clouds [9]. Data Management which includes data consistency (Vogels, 2009), data redundancy (Pamies–Juarez *et al.,* 2011), data mining algorithms and methods (Grossman *et al.,* 2010), integration of distributed data (Chen *et al.*, 2011), data encryption (Anthes, 2010), data colouring. Service Management focusing on service lifecycle in the cloud (Breiter and Behrendt, 2009) and publishing, discovering, and selecting

cloud-based services (Goscinski *et al.*, 2011) [10]. Security, which includes articles that look at general security mechanisms such as restrictions and audits (Spring, 2011a; Wang *et al.*, 2011), multi-tenancy authorization (Calero *et al.*, 2010), third-party assurance (Zissis and Lekkas, 2010), and cloud-based security services (Li *et al.*, 2011) [11]. Business Issues which includes cost comparison between the cost of leasing cloud services and that of purchasing and using a local server cluster (Walker, 2009), techniques to estimate and monitor costs for cloud services (Truong and Dustdar, 2010), algorithms for finding minimum cost storage strategy (Yuan *et al.*, 2011) [12]; Pricing, articles in this subcategory mainly focus on the pricing strategies of cloud providers. A common approach for studying this topic is to compare different pricing strategies and analyze the pros and cons in terms of acceptance of customers. Legal Issues, which examines legal issues associated with cloud computing including policy, governance, and law (Kaufman, 2009), general legal risks of adopting cloud computing (Joint *et al.*, 2009), as well as addressing specific topics such as digital forensic investigation in cloud computing systems (Taylor *et al.*, 2010) [13]. Adoption, this subcategory explores topics related to cloud-computing adoption in businesses including general businesses by providing ROI (Return on Investment) models for firms to decide on the suitability of adopting cloud computing (Misra and Mondal, 2011).

The review of published journal articles indicates that most studies empirically evaluate the general concept of cloud computing adoption like performance, cost, security, cloud providers, and, are the most frequently used to explain cloud computing adoption factors and processes. Although empirically tested theories/frameworks/models are dominant in existing literature on cloud adoption, but majority of articles have adopted some form of theoretical perspectives to gain insights about cloud computing adoption factors and processes, and, then adopted empirical methods to establish some facts and figure about cloud adoption. The emphasis on theoretical aspects to understand cloud adoption implies that cloud computing adoption phenomenon is multifaceted and requires further exploration to understand the relationship and influence amongst factor which influence cloud adoption.

III. Research Objective

Objective 1: To identify integrated elements for implementation of Cloud Computing.

Objective 2: To develop a hierarchical structural model under relevant contextual relations of the identified elements.

IV. Research Methodology

In the present research, interactive management (IM) technique of system sciences have been deployed. IM process has three phases, the Planning Phase, the Workshop Phase, and the Follow-up Phase. Planning Phase consists of two steps; namely the information gathering phase and expert identification phase. Information gathering phase was executed by researcher in the through extensive review of Literature to find established factors impacting cloud computing adoption in business organization. In the expert identifications phase, experts from academia and industry were identified and contacted to get their insights about the problem statement. The workshop phase was performed through a series of workshop, brainstorming sessions and focused group discussions.

A. Research Design

Phase 1 - Identifying Integrated Elements for Cloud Computing Implementation

In this phase, the researcher has identified the relevant factors affecting cloud computing adoption in business organization through:

- Review of literature.
- NGT workshop to enlist the factors and using RII to generate factors having highest impact on cloud computing adoption in business organizations.

Phase 2 - Development of Hierarchical Structural Model

In this Phase 2, the researcher has developed the interaction matrix amongst identified factors and developed hierarchical structural model under

relevant contextual relations of cloud computing adoption in business organization though:

- Creating SSIM matrix and deploying VAXO framework for identifying the relationship amongst factors affecting cloud computing adoption in business organizations through workshop using brainstorming.
- ISM technique being deployed to develop hierarchical model of identified factor affecting cloud computing adoption in business organizations.

B. Sampling

The number of experts required for any interactive management session should be between 10 and 15; for proper cross fertilization of ideas which encourages contributions from everyone and facilitates quick agreement on the relative importance of issues, problems, or solutions. Both participants and the facilitator are required to be subject matter experts of the topic under discussion.

The experts were identified under three categories namely policy maker, academic experts and industry practitioner. Following is the profile mix of experts who were contacted for data collection:

- Policy Maker from Government-01
- Academia-16
- Industry-12

In order to justify the minimum requirement of experts, the number of respondents in all the workshops was between 10 and 15. The sampling technique adopted was non-probability Judgment Sampling.

V. Finding and Analysis

To identify integrated elements for implementation of Cloud Computing, researcher has adopted the concept of open coding which was propounded by Wolfswinkel *et al.,* in 2011. Open Coding is the mechanism of generating categorical labels named codes, to capture the basic themes that appear in a review article [1-19]. Researcher has adopted this methodology, because it is in complete synchronization with the fundamental principle of interpretive structural modelling, which looks to identify variables and establish hierarchical model based on the contextual relationship.

The concept of opening coding resulted in 40 codes, whose frequency of occurrence in selected reviewed article is as follows:

Table I: List of Codes Generated Through Open Coding Mechanism with Occurrences in Selected Review Article

Keywords	Matching Occurrences in Selected Article for Review
Strategic Importance of IT	20
Top Management Support	25
Geographical Diversification	10
Impact of Structure	4
Innovativeness	10
Number of Employees	4
Conformity with Work Culture	10
Resistance to Change	5
Personality Traits	10
Net Worth	10
Integration with In-House and Existing System	20
Institutional Influence	10
Adequate User and Technical Support by Vendor	15
Vendor Choice	15
Market Scope	10
Nature of Industry	10
Trialability of Cloud Services	30
Compatibility with Existing System	30
Improved Collaboration with Channel Partners	10
Reduction in Time and Resource on IT Management	15
Access to Technical Expertise	15
Scalability	30
Faster Deployment	30
Total Cost of Ownership	20
Capex to Opex	20

Keywords	Matching Occurrences in Selected Article for Review
Ecological Benefit	15
Business Process Improvement	15
Agility	20
Reduced Time to Market	15
Service Level Agreement	20
Expertise of Employee	20
Internet Penetration	10
Technology Diffusion	30
Type of IS Function	10
Standardization	10
Interoperability	10
Lock In	10
Switching Cost	10
Data Security	25
Data Privacy	25

Researcher then conducted the Nominal Group Technique workshop, where the initial responses of experts were captured; with the triggering question "mention one factor affecting cloud adoption in business organization". The forum was then opened for data filtration based on the captured responses and gross list of identified factors of cloud adoption obtained through open coding mechanism as listed in Table I. A more generic statement to capture the intent of factors in adoption of cloud computing were framed, which are presented in Table II.

Table II: List of Generic Statements to Capture Factors Affecting Cloud Adoption

Sr. No.	Factors Affecting Cloud Adoption
1	Impact of level of technical knowledge of adopting organization seeking cloud services.
2	Impact of service level agreement on adoption for adopting organization.
3	Impact of quality of technical support being offered by vendor of cloud services.
4	Impact on agility (responsiveness) on adoption for adopting organization.
5	Impact on business process improvement on adoption for adopting organization.
6	Impact of level of compatibility with existing information system of adopting organization considering adoption.
7	Impact of degree of complexity of integration with in-house and existing information system on adoption for adopting organization.
8	Impact of ease of information system management (reduction in time and resource) on adoption for adopting organization.
9	Impact of ecological benefit (less carbon footprint) on adoption for adopting organization.
10	Impact of level of innovativeness in adopting organization.
11	Impact of rate of technology diffusion on adoption.
12	Impact of conformity with work culture on adoption for adopting organization.
13	Impact of level of expertise of employee in organization seeking cloud adoption.
14	Impact of geographical spread of organization seeking cloud adoption.
15	Impact of structure (team based, project based, departmental etc) of organization seeking cloud adoption.
16	Impact of government regulations on adoption for adopting organization.
17	Impact on collaboration with channel partners on adoption for adopting organization.
18	Impact of improvement in market scope on adoption for adopting organization.
19	Impact of data privacy concern on adoption for adopting organization.
20	Impact of data security concern on adopting for adopting organization.
21	Impact of fast deployment of cloud services as compared to traditional method on adoption.
22	Impact of concern regarding interoperability of cloud services.
23	Impact of concern regarding lock in (only one service provider) of cloud services.
24	Impact of scalability of cloud services as compared to traditional method on adoption.

Sr. No.	Factors Affecting Cloud Adoption
25	Impact of standardization (standard adopting architecture) of cloud services on adoption.
26	Impact of concern regarding switching cost of cloud services on adoption.
27	Impact of type of industry (Manufacturing, I.T., Banking etc) on adoption.
28	Impact of network ecosystem (internet speed, performance etc) on adoption for adopting organization.
29	Impact of employee's strength (number of employees) on adoption for adopting organization.
30	Impact of personality traits (decision maker) on adoption for adopting organization.
31	Impact of reach to market (time to market) on adoption for adopting organization.
32	Impact of resistance to change (employee) on adoption for adopting organization.
33	Impact of strategic importance of information technology services in adopting organization.
34	Impact of top management support on adoption for adopting organization.
35	Impact of total cost of ownership (capital expenditure versus operational expenditure) on adoption.
36	Impact of trialability of cloud services.
37	Impact of type of information system function (strategic function or support function) being adopted on cloud services.
38	Impact of ubiquitous access (organization accessibility to resources) on adoption.
39	Impact of vendor (selection) offering cloud services on adoption.
40	Impact of status of company (start-up, MSME, multinational) on adoption of cloud services.

Researcher floated the table obtained above in form of a questionnaire to a select set of experts from industry, to rank them on a scale of 1 to 5, where 1 represented least impact on adoption of cloud computing in business organization and 5 indicated maximum impact on adoption. Relative Importance Index (RII) were the indices of the variables are ranked according to their relevance was then deployed to rank the factors. In Relative Importance Index (RII), the indices ranging from 0.71 - 1.00 are classified as high, 0.41 - 0.60 as medium, and, Less than 0.40 as being low. The top ranked items rated 0.71 or above, indicating a high importance, were deemed to be of importance and therefore included in the research.

Table III: List of Shortlisted Factors Affecting Cloud Adoption Having Composite Index Score of More Than 0.71 As Per RII Calculation

Sr. No.	Factors Affecting Cloud Adoption	C.I.
1	Impact of level of compatibility with existing information system of adopting organization considering adoption.	0.96
2	Impact of data privacy concern on adoption for adopting organization.	0.96
3	Impact of data security concern on adopting for adopting organization.	0.96
4	Impact of top management support on adoption for adopting organization.	0.96
5	Impact of ubiquitous access (organization accessibility to resources) on adoption.	0.96
6	Impact of fast deployment of cloud services as compared to traditional method on adoption.	0.93
7	Impact of total cost of ownership (capital expenditure versus operational expenditure) on adoption.	0.93
8	Impact of concern regarding interoperability of cloud services.	0.92
9	Impact of scalability of cloud services as compared to traditional method on adoption.	0.92
10	Impact of trialability of cloud services.	0.92
11	Impact of standardization (standard adopting architecture) of cloud services on adoption.	0.90
12	Impact of type of information system function (strategic function or support function) being adopted on cloud services.	0.90
13	Impact of service level agreement on adoption for adopting organization.	0.89
14	Impact of degree of complexity of integration with in-house and existing information system on adoption for adopting organization.	0.89

Sr. No.	Factors Affecting Cloud Adoption	C.I.
15	Impact of concern regarding switching cost of cloud services on adoption.	0.89
16	Impact of concern regarding lock in (only one service provider) of cloud services.	0.88
17	Impact on agility (responsiveness) on adoption for adopting organization.	0.85
18	Impact of quality of technical support being offered by vendor of cloud services.	0.84
19	Impact of vendor (selection) offering cloud services on adoption.	0.84
20	Impact of ease of information system management (reduction in time and resource) on adoption for adopting organization.	0.82
21	Impact of network ecosystem (internet speed, performance etc) on adoption for adopting organization.	0.82
22	Impact of reach to market (time to market) on adoption for adopting organization.	0.80
23	Impact of rate of technology diffusion on adoption.	0.78
24	Impact of government regulations on adoption for adopting organization.	0.77
25	Impact of geographical spread of organization seeking cloud adoption.	0.76
26	Impact of strategic importance of information technology services in adopting organization.	0.76
27	Impact of type of industry (Manufacturing, I.T., Banking, etc.) on adoption.	0.73
28	Impact of ecological benefit (less carbon footprint) on adoption for adopting organization.	0.72
29	Impact of level of innovativeness in adopting organization.	0.72
30	Impact of structure (team based, project based, departmental etc) of organization seeking cloud adoption.	0.72
31	Impact of status of company (start-up, Msme, multinational) on adoption of cloud services.	0.72

Researcher than conducted a Follow up workshop, to discuss the findings of Data Collection. A consensus emerged in the workshop , that to have a comprehensive understanding of cloud adoption the numbers of elements needs to further reduce by summing up relevant terms in more generic terms. Inspirit, the essence of the terms which are merged must be adopted in the combined term, and, the combined term should satisfy the intended concept the present, which is to identify list of integrated elements impacting cloud computing adoption in business organization. The following terms were acronym and there conceptual definitions were formulated as presented in Table IV.

Table IV: List of Acronym and Conceptual Definition of Conceived Acronyms Affecting Cloud Adoption

Sr. No.	Acronym	Definition	Elements Combined From (Table C)
1.	AGI	Agility which relates to organization responsiveness to business opportunities, due to faster and on demand deployment of resources.	6,17
2.	CAB	Cloud Associated Benefit relates to the concept of ubiquitous access to resources post cloud adoption and the ecological impact of cloud adoption models.	5,28
3.	COM	Compatibility which relates to integration of cloud services with current I.T. products, interoperability and heterogeneity of cloud offerings.	1,8,14
4.	DM	Data Management which relates to data privacy, data security once data resides on cloud service provider.	2,3
5.	INDI	Industry Index relates to the kind of industry the cloud adopting organization belongs and type of cloud adopting organization.	27,31

Sr. No.	Acronym	Definition	Elements Combined From (Table C)
6.	INNOI	Innovation Index relates to rate and level of innovation and disruptive technology percolation.	23,29
7.	OS	Organization Status which relates to geographical spread of organization, structure and I.T. strategy.	25,26,30
8.	SCA	Scalability relates to elasticity of I.T. resource management and ease of I.T. management for adopting organization.	9,20
9.	SR	Statutory Regulation which relates to government regulations regarding operational side of cloud offerings which include security, standardization and lock-in concerns of cloud offerings.	11,16,24
10.	STM	Superior Time to Market relates to rate at which the offerings of cloud adopting organization will reach target market as compared to traditional organization.	22
11.	TCO	Total Cost of Ownership relates to the cost aspect of cloud offerings which converts capital expenditure to operational expenditure due to metered services and also includes the switching cost of changing cloud service provider.	7,15
12.	TE	Technological Ecosystem relates to internet network speed, performance and penetration and also the stage of cloud adopting organization on cloud computing maturity model.	12,21
13.	TMS	Top Management Support.	4
14.	TRI	Trialability of cloud offering for cloud adopting organization.	10
15.	VQ	Vendor Quality which relates to service quality which includes service level agreement, customization capability of product offerings as per need and the lead time to accomplish the same.	13,18,19

The hierarchical structural model was established using Interpretive Structural Modelling. Following are the steps that are typically followed to create an interpretive structural model [19].

- Problem Identification
- Identification of elements
- Determine Contextual Relationship
- Develop Structural Self-Interaction Matrix (SSIM)
- Develop Initial Reachability Matrix
- Develop Final Reachability Matrix
- Level Partitioning
- Conical Matrix
- Digraphs
- Interpretive structural Model

Researcher has completed the problem identification step in introduction section and identification of elements in objective 1. Contextual relationship of "intent", in which the relationship between two variables A & B, can be stated as "A should help achieve B"; is clearly suitable for this topic, as the objective of this research is to ascertain the relationship amongst identified elements which helps to achieve cloud computing adoption in business organizations.

Structural Self-Interaction Matrix (SSIM) of the identified elements is shown in Table V.

TABLE V: VAXO FRAMEWORK USED TO CREATE SSIM

Variables	TMS	VQ	TRI	AGI	TE	CAB	OS	STM	COM	SR	DM	TCO	INOI	SCA	INDI
TMS	X	V	V	V	V	V	V	V	V	V	V	V	V	V	V
VQ		X	V	V	V	V	V	V	V	X	V	X	V	V	V
TRI			X	V	V	V	V	V	X	A	X	A	V	X	V
AGI				X	V	V	V	X	A	A	A	A	V	A	V
TE					X	V	X	A	A	A	A	A	X	A	X
CAB						X	A	A	A	A	A	A	A	A	A
OS							X	A	A	A	A	A	X	A	X
STM								X	A	A	A	A	V	A	V
COM									X	A	X	A	V	X	V
SR										X	V	X	V	V	V
DM											X	A	V	X	V
TCO												X	V	V	V
INOI													X	A	X
SCA														X	V
INDI															X

Researcher then uses "SMART ISM" software to create the interpretive structural model, which is shown in Fig. 1.

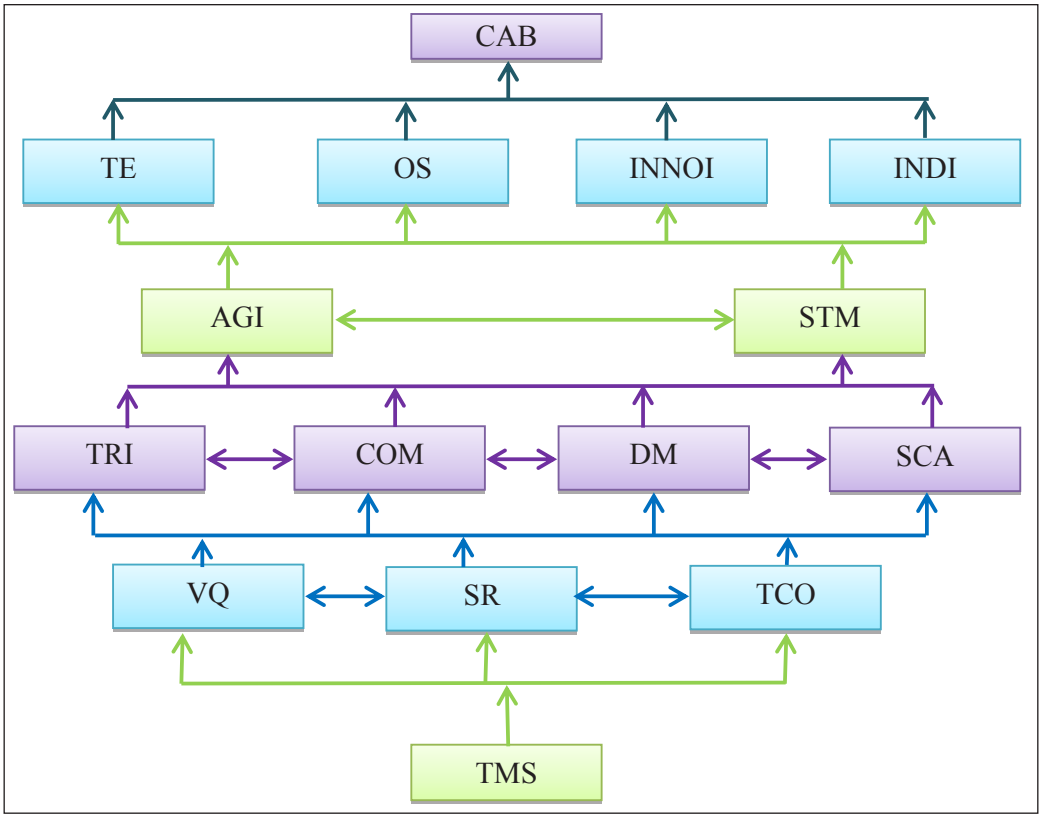


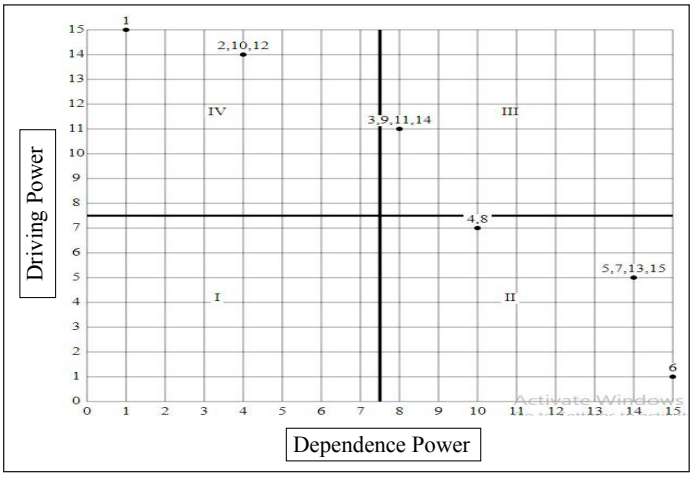
Fig. 1: ISM of Factors Affecting Cloud Adoption

MICMAC (Cross-Impact Matrix Multiplication Applied to Classification) analysis, helps in asserting, the degree to which each variable influences and gets influenced by the other variables in the problem under study. Driving Power which defines that how many elements are being achieved by the element under consideration, and, dependence powerwhich defines that how many elements helps achieve the element under consideration is calculated by summing up the 0s and 1s in the columns and rows corresponding to each variable in the Conical Matrix. The combination of the driving power and dependence power give rise to four categories of variables as following:

Table VI: MICMAC Analysis Categories

Driving Power	Dependence Power	Variable Category	Characteristic
Low	Low	Autonomous	These variables due to their low driving and dependence power do not affect the system to a great extent. They more or less remain extraneous variables.
Low	High	Dependent	Strongly influenced by the system but does not influence the system. These variables are sensitive to the action taken on the influent variables.
High	Low	Independent or Influent	Strongly influences the system but is not influenced by the system. Any action on these variables will affect other variables which are dependent on these variables.
High	High	Linkage	These variables cause instability of the systemas their power of affecting and getting affectedby the other variables is high. Any action on these variables will cause volatility in the system.

The above combinations of MICMAC analysis is understood more clearly using the graphical visualization and are depicted below to better analyse the categories of variables and their implication on cloud computing adoption.



I – Autonomous Variables
II – Dependent Variables
III – Linkage Variables
IV – Influent Variables

Fig. 2: MICMAC Graph

VI. Conclusion/Discussion

A closer analysis of cloud adoption model which has emerged using ISM and thereon the MICMAC analysis reveals that, the role of Influencer, which have high driving power , is being enacted by TMS, VQ, SR, and TCO in the developed structural model. TMS is having the highest power of influence, which impacts cloud adoption in business organization. The highest power of influence understandably emerges form the decision making authority of top management and will be crucial for implementation of any business initiative to percolate in the echelon of business organization.

If any organization is seeking to deploy cloud offerings, the critical factors which will influence its decision will be cost consideration, types of cloud product offerings, lead time of cloud adoption, service quality of cloud offering and the regulatory framework requirement to be followed. Vendor Quality (VQ) incorporates lead time, service quality and portfolio of cloud offerings; Statutory Regulation (SR) incorporates regulatory framework requirement,

and, Total Cost of Ownership (TCO) incorporates the cost consideration of capital expenditure to operational expenditure and cost of switching cloud service provider if needs arises. Therefore, from the hierarchical model and MICMAC graph we can see that the next level of influencer are VQ, TCO, SR; if the organization wants to adoption cloud computing based on the decision of top management.

The interaction of VQ, SR, TCO comes out to be counter reinforcing, where each variable is effecting and getting effected by each other. This situation can be easily understood by the example of Digital India, given the impetus of government regulation, the types of cloud offering by vendors, there service quality and there time of deliverance will improve in order to penetrate the given market condition. As organization will come across offering which reduces their capital expenditure based on the inherit property of cloud offerings of consumption based billing of IT resources, they will look to offload there IT resource requirement to cloud based solutions, which will only increase the quality of vendor offering and at prices which may more or less be comparable. The impact of these will also be seen in the context of government regulations which will further look to address issues such as data security, interoperability and lock-ins of cloud offerings, and will bring a more comprehensive statutory framework. Again the statutory requirements will affect the vendor offerings which in turn will impact the cost of cloud adoption and the cycle will keep on dynamically repeating.

The interaction amongst variables of level five of the structural model is posing adoption issues which are present in the next level of developed structural model and are called linkages factor; owing to their power of affecting and getting affected by the other variables. The impact of influencer variables in cloud adoption will depend upon the next layer of variables which are TRI, COM, DM, and SCA; as the adoption of cloud offerings will be impacted by the performance of this layer as they have high driving power which influences the next layer of dependent variables.

As evident that the interaction between level five variables results in scenario which requires further deliberations in terms of standardization of cloud offering and data security. The variable of level four perform are precisely performing the same function. Trialability (TRI), talks about cloud offering being devised in such a manner which permits easy early adoption, will require the features of compatibility (COM) like integration with existing systems, interoperability and heterogeneity of cloud offerings, to achieve its intended purpose of early adoption. Scalability (SCA) which talks of elasticity or the ease of opting in or out of cloud services will further require cloud offering to be developed in such a manner, which is easy to implement. As more and more cloud offerings will be adopted by business organization data management (DM), which talks about data security and privacy will be a major concern.

On the other hand in today's age of digitalization, the ocean of data being generated and the IT resource being required will require cloud offering to be adopted in organization. The moment organization seeks cloud offering trialability, compatibility and scalability of cloud offering, will become the deciding factor for adopting cloud computing. These interactions amongst TRI, COM, SCA, DM clearly indicates a reinforcing relationship amongst these variables and cloud adoption benefits will only manifest if they are successfully integrated in cloud offerings. Influencer as discussed represents level six and level five of interpretive structural model, and, can broadly be called as representing cloud computing ecology; whereas, linkages are the level four of structural model, researcher and linkages which can broadly be called as strategic success factors of cloud offering must work in cohesion for successful cloud adoption by business organization.

Level three of structural model models represent the benefits of cloud adoption which the organizations can realize in terms of responsiveness to market situations and time to market of product and services. Agility (AGI) relates to organization responsiveness to business opportunities, due to faster and on demand deployment of information and technology resources. If any organization requires an excess of storage capacity due to its open day sale, will benefit the most from this aspect of cloud offering

of excess storage capacity as per the demand of the organization, which requires no prior set up and can be released post consumption. This represents a typical example of agility where the organization can respond to sudden customer demand and satisfy them. As the responsiveness of organization to prevalent market condition improves, the time taken by organization to offer its products or services to the market will also be reduced. Superior Time to Market (STM) relates to enhanced rate at which the offerings of cloud adopting organization will reach target market as compared to traditional organization. But the dependence on cloud computing ecology and strategic success factors of cloud offering maps these factors as dependent factors as per MICMAC analysis.

To realize the full potential of cloud adoption like ubiquitous access to resources on heterogeneous devices; along with the ecological impact of having less carbon footprint and on demand information technology resource availability, which in this research has been coined as cloud associated benefits (CAB), leading to sustained competitive advantage. The organizations must work on their technological ecosystem (TE) which includes their stage in cloud computing maturity model as discussed in chapter of introduction; the network ecosystem which include speed and penetration of intranet and also the excess to internet. As organization will work towards achieving cloud adoption, the speed of adoption will depend upon the openness of organization towards innovation and rate at which the upgrade themselves with respect to any disruptive technology, which in this research has been coined as innovation index (INNOI). The state of technological ecosystem and rate of innovation index will be driven by the kind of industry the cloud adopting organization belongs and type of cloud adopting organization for example MSME, start-up etc. coined as industry index (INDI). The three terms TE, INNOI, INDI forms reinforcing loop where one will help achieve the other two and vice versa; for example, more competitive the industry indexes, faster will the rate of innovation index and better will the state of technological ecosystem. On the other hand, more advanced technological ecosystem will fasten up innovation index which in turn will increase the scenario of industry index. Therefore, the terms appear in same level of interpretive structural model and along with organization status (OS) will determine the extent to which cloud adoption benefits are realized in any organization.

Organization Status (OS) which relates to geographical spread of organization, structure and information technology strategy will also have a reinforcing relationship with technological ecology, innovation index and industry index. A multi-location presence, along with project or matrix based structure of organization and the strategic role of information technology will drive the status of technological ecosystem and innovation index of any organization. As more and more organization will adopt cloud computing within a given industry, the landscape of industry will change, which in turn will drive the aspects of organizational status. In the developed hierarchical model, this interaction amongst TE, OS, INNOI and INNDI has been captured at level two and as per MICMAC analysis fall under the category of dependent variable.

References

[1] N. Abokhodair, H. Taylor, J. Hasegawa, and S. J. Mowery, "Heading for the clouds? Implications for cloud computing adopters," in *Proceedings of the 18th Americas Conference on Information Systems, AMCIS'2012,* 2012.

[2] F. A. Alali, and C.-L. Yeh, "Cloud computing: Overview and risk analysis," *Journal of Information Systems*, vol. 26, no. 2, pp. 13-33, 2012.

[3] Y. Alshamaila, F. Li, and S. Papagiannidis, "Cloud computing adoption by SMEs in the North East of England: A multi-perspective framework," *J. Enterp. Inf. Manag.,* vol. 26, no. 3, pp. 250-275, 2013.

[4] F. Amato, A. Mazzeo, V. Moscato, and A. Picariello, "A framework for semantic interoperability over the cloud," in *2013 Workshops of 27th International Conference on Advanced Information Networking and Applications (WAINA)*, 25-28 Mar. 2013, 2013, pp. 1259-1264.

[5] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," Technical Report No. UCB/ EECS-2009-28, 2009.

[6] M. G. Avram, "Advantages and challenges of adopting cloud computing from an enterprise perspective," *Procedia Technol.,* vol. 12, pp. 529-534, 2014.

[7] A. Azarnik, J. Shayan, M. Alizadeh, and S. Karamizadeh, "Associated risks of cloud computing for SMEs," *Open International Journal of Informatics*, vol. 1, pp. 37-45, 2012.

[8] A. Bahga, and V. K. Madisetti, "A cloud-based approach for interoperable electronic health records (EHRs)," *IEEE J. Biomed. Health Inform.,* vol. 17, no. 5, pp. 894-906, Sep. 2013.

[9] A. Bajenaru, "Software-as-a-Service and cloud computing, a solution for small and medium-sized companies," *Bulletin of the Transilvania University of Brasov. Series V: Economic Sciences,* vol. 3, 2010.

[10] M. Bellamy, "Adoption of cloud computing services by public sector organisations," in *Proceedings of the IEEE 9th World Congress on Services*, IEEE, 2013, pp. 201-208.

[11] S. S. Bharadwaj, and P. Lal, "Exploring the impact of cloud computing adoption on organizational flexibility: A client perspective," in *Proceedings of the International Conference on Cloud Computing Technologies, Applications and Management (ICCCTAM 2012)*, IEEE, 2012, pp. 121-131.

[12] J. M. Bhat, "Adoption of cloud computing by SMEs in India: A study of the institutional factors," in *Proceedings of the 19th Americas Conference on Information Systems,* 2013, pp. 1-8.

[13] K. Bogataj, and A. Pucihar, "Business model factors influencing cloud computing adoption differences in opinion," in *Proceedings of the BLED 2013*, 2013, pp. 1-14.

[14] H. P. Borgman, B. Bahli, H. Heier, and F. Schewski, "Cloudrise: Exploring cloud computing adoption and governance with the TOE framework," in *Proceedings of the 46th Hawaii International Conference on System Sciences*, IEEE, 2013, pp. 4425-4435.

[15] N. Brender, and I. Markov, "Risk perception and risk management in cloud computing: Results from a case study of Swiss companies," *International Journal of Information Management*, vol. 33, no. 5, pp. 726-733, 2013.

[16] M. Carcary, E. Doherty, and G. Conway, "The adoption of cloud computing by Irish SMEs - An exploratory study," *Electronic Journal of Information Systems Evaluation*, vol. 16, no. 4, 2013.

[17] C. G. Cegielski, L. A. Jones-Farmer, Y. Wu, and B. T. Hazen, "Adoption of cloud computing technologies in supply chains: An organizational information processing theory approach," *Int. J. Logist. Manag.*, vol. 23, pp. 184-211, 2012.

[18] B.-Y. Chang, P. H. Hai, D.-W. Seo, J.-H. Lee, and S. H. Yoon, "The determinant of adoption in cloud computing in Vietnam," in *Proceedings of the International Conference on Computing, Management and Telecommunications (ComManTel 2013)*, IEEE, 2013, pp. 407-409.

[19] R. Attri, N. Dev, and V. Sharma, "Interpretive structural modelling (ISM): An overview", *Research Journal of Management Sciences*, vol. 2, no. 2, pp. 3-8, 2013.

International Journal of Distributed and Cloud Computing
12 (1) 2024, 30-40
http://www.publishingindia.com/ijdcc

Unveiling the Futuristic Trends of 5G Connected Cars in the Automotive Industry

Anusha Thakur
University of Petroleum & Energy Studies, Dehradun, Uttarakhand, India.
Email: anushathakur89@yahoo.com

Abstract: The concept of "vehicle to everything" has completely revolutionized the maker's and user's experience. The whirlwind of disruptions is buffeting the automotive world, creating a huge value-chain disconnect. The amalgamation of advanced in-vehicle capabilities and cellular network data services leverages innovations in the automotive sector. The introduction of connected vehicles is expected to change user-vehicle interaction and communication with the world around them. Several countries have initiated the 5G trials, with the results being under evaluation. The rapid increase in enhanced user experience and affordability for various services are increasingly impacting the adoption of various innovative solutions. The advent of this state-of-the-art technology is not only a generational step, but the possibility of opening several opportunities for the industries. The article focuses on the evolution of 5G services and their impact on the global economy. The development of 5G is expected to bolster digitalization trends, significantly impacting the automotive industry, particularly connected cars. The paper discusses the impact of the revolutionary 5G technology on future mobility and reviews the use cases of 5G in the automotive sector. Further, other major areas of study emphasize on the market determinant analysis (trends, drivers, and challenges) of 5G services in the automotive sector.

Keywords: 5G, Automotive, Communication, Connected cars, Networks, Services.

I. Introduction

With a shift from trial to early commercialization, the 5G technology acts as the game-changer for life-changing innovations [1]. 5G technology is a myriad of innovations, wherein, the amount of information and number of subscribers they consume, is no longer a restraining factor, however, it enables an infinite number of applications and devices to process, synthesize, and exchange enormous amounts of information for investments to advance the underlying infrastructure [2]. Its network architecture for mobile technology progresses immensely upon the past architectures [3]. Enormous leaps in performances are empowered by large cell-dense networks. The architecture of 5G wireless services provides enhanced security in comparison to the 4G LTE networks. Moreover, 5G wireless technology systems, offer faster speed, in comparison to the earlier generations, such as those delivered through fiber-optic cables. This increased speed is expected to necessitate more spectrally competent technologies with additional spectrums. With telecom companies emphasizing on enhanced infrastructural construction, 5G-enabled applications are expected to have an energizing impact on the economy over the projected period.

Increasing market demand for 5G services is significantly associated with the mobile networks accomplished of provisioning the services comprising of tearing down and launching of networks as per the customer requirements. These requirements can be accomplished by network function softwarization which is independent of hardware and is better to manage the networks [4]. New intermediaries and business models in the value chain processes are also garnering positions nowadays to cater to the rising demand in offering differentiated connectivity services to the end-users in the market. Facts concerning 5G services are

now becoming a reality, piloting in the era of new opportunities in different business verticals [5].

The transition from digital-first to digital throughout unleashes newer revenue streams and applications via innovative business models, impacting diversified industries. 5G, or the fifth generation of cellular technology, leads to reliable and fast connectivity, enabling seamless connectivity for billions. The lightning-fast connection offered by 5G, contributes significantly to the growing advancements in intelligent automation; in terms of blockchain, driverless cars, the Internet of Things (IoT), digital reality, Artificial Intelligence (AI), and breakthroughs untapped yet [6]. Low latency, faster speeds, and the capability to back up a plethora of device termination points make the technology a versatile one.

With the advent of four major trends such as connectivity, diverse mobility, autonomous driving, and electrification, the gap between automotive and high-end industries is narrowing down. As the convenience and comfort quotient among the users witnesses an increase, these trends are expected to revolutionize the integrated automotive value chains [7]. The evolution in wireless technologies bolsters conventional mobility to the current connected car environment with seamless vehicle operations to be driven into an autonomous climate [8].

Purpose of the Article: This article contemplates the current state of the global 5G wireless services market with a major emphasis on different determinants (market trends, drivers, opportunities, and challenges in implementing the technology) impacting the automotive market growth. It emphasizes on the economic contribution of 5G wireless services along with the key focus on future research plans and strategic implications for 5G automotive businesses.

The proposed article further intends to offer an outline for 5G wireless services along with different market-related information that enables the vendors to gain a competitive edge and profitability in the industry. A holistic competitive analysis thereby, helps the businesses deploying these wireless services across, to discover their potential to compete in the market.

Findings in the Article: Adoption of 5G technology in the automotive industry paves the way for futuristic developments such as vehicle-to-vehicle communication and real-time traffic optimization in connected vehicles and intelligent transportation systems. This article explores how 5G technology is transforming the automotive industry and driving the sector towards a connected and safer future.

II. Literature Review

The introduction of 5G technology enables manufacturing companies to get an alternative for communication processes for several Industry 4.0 applications which includes wireless control such as robots, machines, and mobile tools. 5G aids the implementation of specialized functions and tasks, such as connected cars, and remote precision medicine, along with augmented and virtual reality experiences. Advancements linked with 5G offer seamless coverage, lowered latency, with enhanced ability to back up many devices. With exponentially high download speeds of nearly 20 gigabits per second, users with 5G wireless services are expected to download the contents within seconds. On the basis, of, trials to date, the 5G networks are expected to offer higher performance in various application segments such as indoor hotspots along with dense urban areas as well. A few innovative usages of 5G technology include remote monitoring of patients in hospitals, increased usage of smart ambulances that can communicate with doctors on a real-time basis, digital wallets to connect wearables, phones, and cars, with smart devices to offer unified financial transactions, 5G-aided factories with connections being sustained among sensors and few more.

Rising emphasis on building partnerships by an amalgamation of 5G wireless service providers with different telecom operators is expected to bolster the implementation of these services in different industry verticals. Connection of the cellular antennas to data centers with fibers ensuring real-time speeds are anticipated from 5G. Although the fiber optic infrastructure prevalent in today's scenario is being used by the 4G systems, more infrastructural development will be required to support the widespread usage of 5G. Factors such as fiber-optic infrastructural development are expected to bolster the execution of 5G to certainty.

Analyzing a wide range of micro-market indicators and statistics contributing to the prospects and dynamics of the market strategically plays a vital role. The technological trends and critical domains of 5G wireless services need to be catered upon, to illustrate its evolution and projected quantum leap for profitability to the end-users. The advent of 5G thereby, focuses on advancing societies, enhancing experiences, transforming industries, and opening avenues for businesses. Identifying the research needs and priorities in terms of strategy-making to support the easy and cost-effective deployment of wireless technologies and solutions are gaining significant traction across the world.

III. 5G and the Automotive Sector

In the automotive sector, 5G is envisioned to augment the aspects of automated cars and in-car entertainment. The evolution of the connected-car ecosystem impacts different value chains, comprising those for semiconductors, automotive, software, and telecommunications. Inclusion of the specifications, on the cellular Vehicle-to-Everything (C-V2X) in Rel-16, helps boost the realization of connecting technology, designed to connect road users, roadside infrastructure, cloud services, and vehicles-to-vehicle services to enhance the overall transportation experience [9].

Embedded connectivity is anticipated to unlock the latest capabilities and features, changing the driving experience holistically, and making it satisfying and safe [10]. To remain relevant and competitive, it is significantly necessary for industry leaders to embrace the digital revolution and be dynamic investors in coordinating, constructing, and cementing the right ecosystem for the successful adoption of 5G service. The paper focuses on the important changes transforming the automotive sector, with the implementation of 5G technology. It also examines the value captured and incorporated by the market players, in their future product segments as well as operational, and organizational capabilities. Evaluating and forecasting future opportunities and trends serve as the first step for all the business activities of any small-sized or large-sized firm.

The influence of 5G on different telecom business shareholders along with its relevance and potentiality across varied application segments is witnessing immense growth. The market players are increasingly, focusing on striving towards implementing different research and development initiatives to deploy and develop efficient technologies and cater to the global product and service demand, revenue strategies, and pricing trends.

IV. Impact of 5G Technology on the Automotive Sector

In today's scenario, 5G is exponentially changing the way businesses are working. The technology opens the door for groundbreaking developments in intelligent transportation systems along with driverless and connected vehicles.

The automotive companies are embracing the advantages of technology and rekindling the industry's creativity. Millions of vehicles are being connected to cellular networks for emergency services, connected infotainment, and real-time navigation services [11]. Leveraging 5G is expected to enable automakers to launch new business models, and augment the overall manufacturing and operational efficiencies.

Improved Infotainment and User Experience: Developers nowadays are continually striving to cater to consumer needs, owing to the increasing number of latest technology adopters and smart device users [12]. The high-speed data transfer by 5G will offer individuals with immersive car infotainment experiences. Integration of 5G with personal devices will enable passengers to use AR/VR, cloud-based gaming, and high-quality streaming services efficaciously.

Smart Roads: 5G technology acts as a major component for the implementation of smart roads. Execution of smarter road infrastructure paves the way for increased profitability with revolutionized transport networks. These smart roads incorporate the usage of sensors, and smart devices, wherein, the data is transmitted both with volume and speed, via various data points to enable real-time

communication between the infrastructure and the vehicles deploying it. The technology also augments the controlling of congestion, traffic flow, and traffic lights through automated artificial intelligence and machine learning programmes. Nearly, twenty-one percent of the crashes occur due to weather conditions [13]. The incorporation of smart road technology paves the way to easily detect nearby weather conditions with sensors and subsequently send updates to the satellite navigation systems of the adjoining vehicles, to let the drivers get an idea about the roads ahead. Further, improvements can also be witnessed in road safety and transit-route planning with the transmission of real-time information.

Universal Connectivity: 5G Connected cars enable the analytics to efficiently understand customer needs and offer subsequent predictive maintenance. The technology enables the car manufacturers to manage the cost and connectivity with the One Standard Level Agreement (SLA) contract. Worldwide connectivity, thereby acting as a solution for the efficacious connectivity and integration of the connected vehicles with the related ecosystems [14].

Optimizing the smart data downloads and uploads at off-peak locations and times enables the cost control of the OEMs. The optimization of the data transfer in terms of cost of the connection, quality, and time, helps to discern the scheduling optimal conditions and the type of traffic, thereby saving the costs for OEMs.

Enhanced Mobile Broadband (EMBB): 5G technology offers more uniform and faster experiences and speed, in comparison to the 4G technology. This is thereby anticipated to augment the higher-bandwidth use cases such as machine and real-time human interface rendering, vehicle teleoperation, and in-car infotainment [15].

Re-Imagining Vehicle Maintenance: With the vehicles becoming more connected and intelligent, the maintenance of sucha fleet of vehicles, with real-time initiated repair, problem identification, and performance evaluation becomes an essential factor. The combination of edge computing and cellular connectivity helps the technicians to seamlessly predict and subsequently fix the issues [16].

Enormous Internet of Things (IoT): With a million connections per square kilometer, the 5G networks are anticipated to efficaciously back up several concurrent connections from connected infrastructure, cars on the road, and end-user devices. This is further expected to minimize the probability of cars and other devices getting disconnected from the mobile network accidentally owing to many connections.

V. Use Cases of 5G Technology in the Automotive Industry

The automotive sector is anticipated to witness profound developments concerning the past differentiators, such as focusing on the layers of hardware and software comprising of vehicle's operating system, and powertrains. Intelligent connections between the internet, and vehicles, poses to be one of the key areas shaping the future of the industry. The industry is emphasizing on shifting to networked ecosystems with the interaction between several players and the end customers from the linear value chains with OEMs at the end of the assembly line.

Fundamental improvements in environmental impact, convenience, cost-effectiveness, safety, and experience when combined, disintegrate the diversified business models at an unprecedented scale. For instance, in the regions of Germany and the U.S., connected car startups are playing a significant role, with the major expertise sections being on mobility management systems, maintenance and diagnostics solutions, and connectivity. Advances in connectivity and computing are anticipated to augment new use cases in the automotive industry. A few of the use cases can be illustrated as:

A. 5G for Vehicle Production

5G connectivity offers services such as Automated Factory Parking (AFP), wherein, the remotely controlled vehicles are mechanically moving across the logistics hubsand production areas. The private networks offer the prospect of tailoring the correlation capabilities for highly on-premise and personalized demand. For instance:

- *BMW Brilliance Automotive Ltd.*

About the Company: BMW Brilliance Automotive, founded in 2003, is a joint venture company between the Brilliance Group and the BMW Group. The business operation segments offered by the company include procurement, research and development, sales, production, and after-sales of automobiles in China. The company emphasizes on high-quality development and is one of the leading sustainable providers of first-class mobility across the automotive industry in China [17].

Solutions Offered: BMW Brilliance Automotive (BBA), announced the introduction of 5G wireless coverage across all its manufacturing plants, thereby increasing the efficacy of the production processes. The venture enables the transfer of data at a rate of one gigabyte a second, via high-performance and small computer centers directly on-site, thereby helping seamless exchange of data on areal-time basis. Incorporating this ground-breaking technology is expected to open avenues including augmented reality and cloud-based control of human-machine interaction and autonomous transport systems with the help of smart devices [18].

B. 5G for Telematics

The automobile industry poses to be a vital technological intersection, wherein, information and communication are becoming a widespread aspect nowadays [19]. Increasing connectivity is expected to bolster the role of telematics with varied valuable insights. The different 5G telematic units offer benefits such as the application software development kit (SDK), enabling the co-development models with OEMs, along the integration of the connectivity hub with the latest wireless technologies [20]. The Obtainability of automobile telematics data from the fleet poses to be one of the significant resources for the automotive OEMs, and enables seamless monitoring of the predictive maintenance and fleet performance, thereby benefitting both the OEMs and the customers. For instance:

- *Harman International*

About the Company: Harman International, a US-based company, is developing solutions designed around passengers and drivers. The in-depth expertise offered by the company is largely redefining the overall automotive landscape, with the most immersive and intuitive communication with in-vehicle technology. Harman is a subsidiary unit of Samsung Electronics Co. Ltd.

Solutions Offered: Market players such as Harman International, announced the launch of the first-to-market 5G Telematics Control Unit (TCU), to enable users to upgrade to enhanced experiences, thereby laying the groundwork for mechanized driving. The company offers connected and integrated audio car systems and services powering billions of mobile systems and devices, enabling automakers to create an ecosystem with rich in-vehicle experiences.

C. 5G for ADAS

The Advanced Driver-Assistance Systems are the cutting-edge features assisting the drivers and subsequently augmenting road safety. Connectivity plays a pivotal role in ADAS, supplementing onboard sensors with data that comes from sources outside the vehicle, such as ADAS service providers or infrastructure sensors, or is processed in the cloud. The capabilities are extended beyond the limitations of onboard processing and sensors. For example:

- *Valeo*

About the Company: Valeo, is a French company, founded in 1923. Valeo is a motor vehicle part manufacturing company, offering innovative solutions to make mobility smarter, safer, and cleaner. The company is one of the leaders in the reinvention of the interior experience, driving assistance systems, and electrification across. Valeo operates with nearly 183 manufacturing plants, 44 development centers, in 21 research centers in approximately 29 countries across the world.

Solutions Offered: Valeo is engaged in offering transformative Advanced Driver Assistance System (ADAS) solutions integrated with state-of-the-art technologies to augment several functionalities. These solutions comprise of a range of real-time data processing, intelligent algorithms, and sensors, to offer inclusive assistance, ensuring convenient and safer rides for the car owner. These solutions include

a wide range of driving features that enhance comfort and safety on the road. With the help of features such as LiDAR sensors, radar, and cameras, enables constant and seamless monitoring of the surrounding environment, thereby assisting the drivers with real-time alerts. Incorporation of concepts such as:

- *Adaptive Cruise Control (ACC)* helps maintain a safe distance from the vehicles by minimizing the risks of rear-end collisions with automatically adjusted speed.
- The *Lane Keeping Assist System (LKAS)* exploits the cameras and lane detection algorithms to detect lane markings and subsequently corrects the deviating trajectory of the vehicles from their intended path.
- *Blind Spot Detection (BSD)* signals the drivers regarding the existence of vehicles in their blind spots, thereby averting probable accidents at the time of lane change.
- *Rear Cross-Traffic Alert (RCTA)* helps in detecting the vehicles coming from the sides when maneuvering out of tight spaces, with enhanced safety [21].

Hence, the solutions represent intelligent and comprehensive systems extending their potential beyond parking, and driving, and subsequently simplifying the daunting tasks if any.

VI. Trends of 5G Technology Shaping the Automotive Industry

The technology-driven ecosystems are gaining significant traction in the automotive industry. The industry is witnessing a continual trend of integrating the electronic control units (ECUs), and car functions into the central domain controllers, thereby pushing the growth towards more connected and integrated infotainment systems.

It is becoming the need of the hour for companies to understand the latest automotive technology stack and the issues grappling within the industry.

The trends revolutionizing the way the automotive industry develops cars can be illustrated as:

A. Autonomous Driving

Autonomous vehicles are anticipated to dramatically enhance security over the forecasted period, wherein, self-driving cars become ubiquitous. Reaching that level, however, requires widespread investments to develop machine learning models for illustrating infrastructure and traffic for low-latency wireless networks [22]. Wireless communication generates a network of advanced technologies such as high-speed information systems, networks, and sensors that enable things such as vehicle-to-network communication or vehicle-to-vehicle communication [23].

The 5G technology acts as a base for autonomous car networks, by seamlessly connecting pedestrians, drivers, infrastructure, and vehicles. With the emergence of 5G networks, autonomous driving is expected to become a reality. The real-time connectivity required for the automated vehicles to engage with the surrounding infrastructure and with one another will be backed by high reliability and low latency of 5G, thereby, increasing the responsiveness and accuracy of these systems. The different ways by which 5G is anticipated to improve autonomous vehicles by making them more safe, reliable, and connected [13].

- *HD Road Mapping:* Enhanced mapping poses to be significant for an autonomous vehicle to drive safely. For instance, the incorporation of HD road mapping is expected to enable the cars to maintain situational awareness and augment the safety of the automated vehicles by giving these, instant 3D access to roads.
- *Vehicle-to-Everything Communication:* The reliable internet connectivity along with the low latency from edge computing, is anticipated to unfold the car door to vehicle-to-everything (V2X) communication, thereby enabling a connected car to seamlessly transmit the data from other road sensors, infrastructure, and nearby vehicles. The forthcoming scenario of road transportation is projected to have cars communicating with other infrastructure, roads, and cars.

- *Edge Computing:* The concept of edge computing brings the data storage and computation nearer to the source, rather than refining the data at a central location. This enables the technology to possess lower latency and subsequently communicate with other devices and sensors on a real-time basis.

B. Shared Mobility

With the shifting customer preferences, the ownership model of the cars is, significantly evolving from renting and sharing. Nearly, one-third of the expected increase in vehicle sales from urbanization and macroeconomic growth likely will not happen because of shared mobility. Different shared mobility solutions include:

- Mobility service offerings.
- Car-sharing profile broker.
- Secondary authentication through mobile phone.
- Seamless mobility application.
- Location data to be sold to the mobility provider.

C. Connectivity

5G networks being deployed are offering low latency, wider coverage, and increased speed, required for the projected advances in the transportation sector. The automotive industry, with the broader mobility ecosystem, is engaged in the transformation, prompted by the shift bolstering in tandem with the build-out of distributed communication infrastructure providing connectivity to the car [24].

Intelligent communication outside and within the car acts as an enabler for autonomous technologies. These include the different connected services and technologies such as data monetization, fleet solutions, smart traffic, dashcam solutions, predictive maintenance and remote diagnostics, payment solutions, intrusion detection, and a few more.

D. Electrification

With the advent of technology, automakers are increasingly mastering the art of designing, building, and validating vehicles with internal combustion engines (ICEs), and catering to the changing scenario [25]. With the introduction of cutting-edge technologies, connected cars are increasingly featuring electric powertrains, thereby connecting the cars digitally and subsequently performing remote operations.

VII. Market Driver Boosting 5G Technology in the Automotive Industry

A. Growing Connected Car Ecosystem

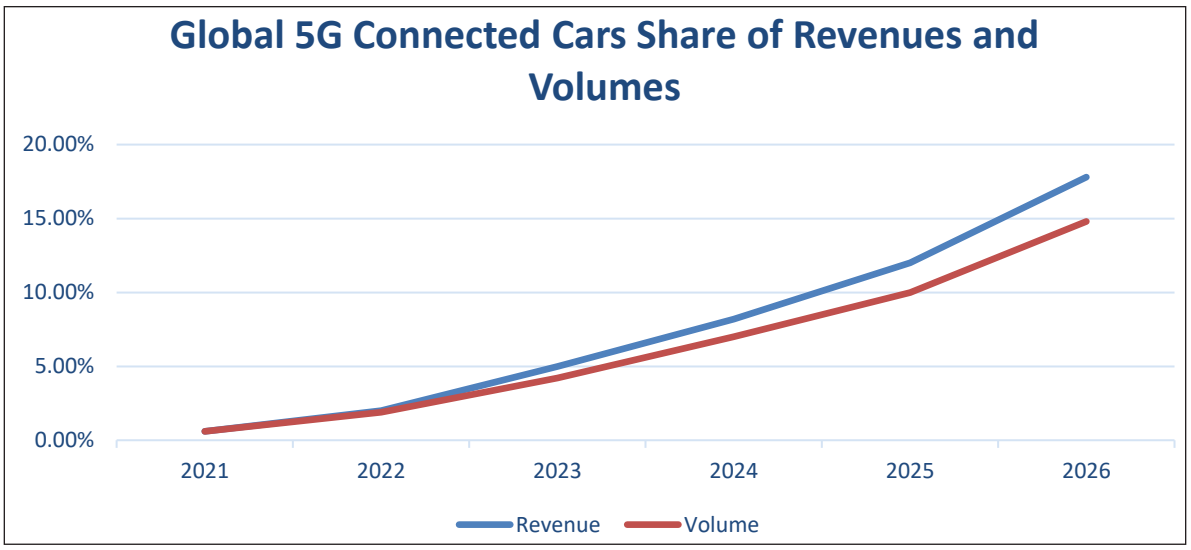
Ensuring connectivity in the ecosystem necessitates the integration of several services across the same. Shifting from a feature-driven to a holistic approach, wherein, functionalities are often upgraded on demand and play a crucial role in the automotive connected key space.

The concept of connected car technology is gaining significant traction, to cater to the increasing and changing needs of the clients, in the ever-expanding horizon. The figure mentioned below illustrates the market share for the connected cars in terms of revenues and volumes from 2021-2026.

This groundbreaking innovativeness has created an ecosystem in the automotive industry, with technological partnerships with the original equipment manufacturers (OEMs), and players, functioning outside the scope of the conventional motoring process [27].

The connected cars ecosystem has revolutionized traditional mobility by hardwiring content platforms, real-time analytics, operating systems, and telecommunications, among other factors in the driving experience. The technology-enabled cars are to be equipped with connectivity, software service, information, and data.

With the advent of pioneering technologies, connected cars are anticipated to become generic commodities over the forecast period, however, also emphasizing on the fact, that the services are largely the evolution of what the market players can offer, by posing as a differentiating criteria for the vehicle manufacturers [28]. The connected ecosystems thereby, pose an essential factor in bolstering the incorporation of 5G technology in the automotive industry.



Source: [26]

Exhibit 1: Global 5G Connected Cars Share of Revenues and Volumes

VIII. Challenges of Implementing 5G in the Automotive Industry

Compatibility and Standardization: Adhering to the standardized requirements and protocols is expected to play a significant role in the seamless integration and communication process. Ensuring compatibility in between the automobile gadgets and 5G technologies poses to be a challenge.

Security Issues: With the increasing connectivity, unauthorized access to vehicles and infrastructures along with the possibility of cyberattacks is witnessing a rise. The inclusion of 5G technology in the automotive industry poses several restraints in establishing strong security measures and thereby protecting sensitive information to prevent privacy violations and hacking attempts [29].

IX. Future of Mobility and 5G Connected Cars

The emerging innovations and collaborations are enabling the operations to become efficient. The adoption and availability of technological solutions is expected to bolster most of the key trends in the auto industry. The entire industry, from automakers to suppliers, is challenged to keep the prevailing operations beneficial, while concurrently building out the capacity to cater to these innovations.In today's scenario, the market players are emphasizing on finding the balance between the endurance of a profitable business, and at the same time paving the way to disintegrating the business models.

A few instances of innovativeness in mobility in the future can be illustrated as [30]:

- Vehicles embedded with several sensors, and internet-connected engine control units (ECUs), offering insights and data.
- Cars offering a connected platform for the latest business models.
- The capability to bundle the sales with the latest subscription-based offerings for car share, rideshare, electric vehicle (EV) charging, and parking services.
- Continual rise and expansion of the fleet services to include charge point operations.
- The algorithm-based insurance based on the data collected from the connected cars.

X. Conclusion

5G, the future of mobile connectivity, is anticipated to create great value for society as well as industries. Accelerated rollout and adoption of 5G technology, draws insight on quantifying its potential impact in

terms of economic values. As per the International Telecommunication Union (ITU), the number of connected devices on the Internet is expected to be nearly 50 million by 2028 [31]. 5G, the fifth generation or next generation of mobile technologies, is anticipated to connect various applications, people, transport systems, data, and cities in smarter networked communication environments. Implementation of 5G networks focuses on representing higher performance in different facets. With the deployment of these advanced technologies, enormous amounts of data can easily be transported, and various processes and devices with massive volumes of data can be connected, with full reliability.

In the automotive sector, the cutting-edge 5G technology benefits are expected to contribute to a wider range of edge applications. The workloads that are not defense critical, for instance - smart traffic and infotainment management, are expected to move to the edge from in the cloud or onboard. 5G connectivity thereby minimizes the latency to the point that only a few defense-critical functions begin to be amplified by the edge infrastructure, despite relying completely on the onboard systems.

Key market players along the value chain, such as hyperscalers, communication system suppliers, tier 1 suppliers, semiconductor players, and OEMs, are anticipated to increase the value capture in the growing automotive landscape [15].

Hence, the constant relocation of the spectrum towards 5G, along with the integrated features such as edge cloud, dynamic policies, network slicing, and QoS is expected to cater to a wide range of services. The amalgamation of all pioneering features, thereby, enables car manufacturers to evolve and roll out efficacious connected cars in the future [32].

References

[1] CII, "5G: The catalyst to digital revolution in India," Deloitte, Gurgaon, 2018.

[2] D. W. P. W. C. H. B. a. F. J. Littmann, "5G: The chance to lead for a decade," Deloitte, 2018.

[3] H. Remmert, "What is 5G network architecture?," Mar. 19, 2021. [Online]. Available: https://www.digi.com/blog/post/5g-network-architecture

[4] NEC, "Making 5G a reality," NEC Corporation, Tokyo, 2018.

[5] W. T. T. a. M. R. Chow, "The global economic impact of 5G: Powering your tomorrow," PwC, 2021.

[6] M. Attaran, "The impact of 5G on the evolution of intelligent automation and industry digitization," *Journal of Ambient Intelligence and Humanized Computing,* vol. 14, pp. 5977-5993, 2021.

[7] L. Chembrakalam, "What the automotive industry can learn from High-Tech," 18 Mar. 18, 2023. [Online]. Available: https://cio.economictimes.indiatimes.com/news/next-gen-technologies/what-the-automotive-industry-can-learn-from-high-tech/98746988.

[8] Future Bridge, "Next-generation wireless technologies – Changing the automotive ecosystem," Sep. 30, 2020. [Online]. Available: https://www.futurebridge.com/industry/perspectives-mobility/next-generation-wireless-technologies-changing-the-automotive-ecosystem/

[9] EY, "Role of cross-industry collaboration in India's 5G success," Ernst & Young, Kolkata, 2022.

[10] S. Sorrentino, "Ericsson," Oct. 19, 2021. [Online]. Available: https://www.ericsson.com/en/blog/2021/10/powering-connected-cars-with-5g

[11] D. Sahu, "How 5G is revolutionizing the automotive industry," Jan. 2023. [Online]. Available: https://www.wipro.com/automotive/how-5g-will-revolutionize-the-auto-industry/

[12] R. Patrick, "How 5G is transforming the world of mobile app development," Mar. 4, 2021. [Online]. Available: https://www.businessofapps.com/insights/how-5g-is-transforming-the-world-of-mobile-app-development/

[13] S. Meyer, "[INFOGRAPHIC] 11 benefits of 5G in the automotive industry," Apr. 5, 2023.

[Online]. Available: https://www.thezebra.com/resources/driving/benefits-of-5g/

[14] C. Herlitz, "Charting the future of connected cars and mobility with 5G," Jun. 8, 2021. [Online]. Available: https://www.ericsson.com/en/blog/2021/6/charting-the-future-of-connected-cars-and-mobility-with-5g

[15] P. B. O. D. J. G. G. K. A. a. W. M. Arejola, "The future of automotive computing: cloud and edge," Oct. 6, 2022. [Online]. Available: https://www.mckinsey.com/industries/semiconductors/our-insights/the-future-of-automotive-computing-cloud-and-edge

[16] P. a. S. K. Reich, "Car talk: How 5G, mobile, and cloud are transforming the automobile," Nov. 1, 2021. [Online]. Available: https://www.forbes.com/sites/googlecloud/2021/11/01/car-talk-how-5g-mobile-and-cloud-are-transforming-the-automobile/?sh=2fd625418995

[17] B. Brilliance. [Online]. Available: http://www.bmw-brilliance.cn/cn/en/pr/bba.html

[18] M. Williams, "BMW brilliance applies 5G network across China," Jul. 22, 2019. [Online]. Available: https://www.automotivelogistics.media/bmw-brilliance-applies-5g-network-across-china/38744.article

[19] H. International, "Telematics control unit (TCU) platform," Jul. 2, 2020. [Online]. Available: https://car.harman.com/solutions/connectivity/tcu-platform

[20] Valeo, "5G Premium Telematic Unit for vehicles," Aug. 11, 2023. [Online]. Available: https://www.valeo.com/en/catalogue/cda/5g-premium-telematic-unit-for-vehicles/.

[21] Jul. 31, 2023. [Online]. Available: https://www.valeo.com/en/catalogue/cda/complete-adas-solutions/

[22] J. Sanders, "Why 5G is a crucial technology for autonomous vehicles," Nov. 4, 2019. [Online]. Available: https://www.zdnet.com/article/why-5g-is-a-crucial-technology-for-autonomous-vehicles/

[23] Robustel, "How 5G autonomous vehicles will benefit from 5G," [Online]. Available: https://www.robustel.com/iot-technology-solutions-blog/how-5g-autonomous-vehicles-will-benefit-from-5g-1/#:~:text=5G%20can%20act%20as%20a,become%20smarter%20and%20more%20efficient

[24] T. Neumann, "Five automotive connectivity trends fueling the future," [Online]. Available: https://www.jabil.com/blog/automotive-connectivity-trends-fueling-the-future.html

[25] K. Robuck, "How vehicle electrification is transforming the automotive industry," Jan. 14, 2022. [Online]. Available: https://blogs.sw.siemens.com/automotive-transportation/2022/01/14/how-vehicle-electrification-is-transforming-the-automotive-industry/

[26] K. Alfalahi, "Why the future for cars is connected," Jul. 8, 2021. [Online]. Available: https://www.weforum.org/agenda/2021/07/why-the-future-for-cars-is-connected/

[27] P. Karlsson, "Connected car: A new ecosystem," May 9, 2016. [Online]. Available: https://www.ipsos.com/en/connected-car-new-ecosystem#:~:text=Connected%20cars%20have%20created%20a,the%20scope%20of%20traditional%20motoring

[28] A. Altman, "Connected ecosystems - The main driver for future success in automotive," [Online]. Available: https://www.publicissapient.com/insights/connected-ecosystems-the-main-driver-for-future-success-in-automotive

[29] S. Srivastava, "How 5G is making the automotive sector more efficient and secure," Sep. 8, 2023. [Online]. Available: https://appinventiv.com/blog/5g-in-automotive-industry/

[30] SAP, "The future of automotive and mobility," May 5, 2021. [Online]. Available: https://www.forbes.com/sites/sap/2021/05/05/the-future-of-automotive-and-mobility/?sh=7e3fb48359d5

[31] ITU, "5G - Fifth generation of mobile technologies," 2021. [Online]. Available: https://www.itu.int/en/mediacentre/backgrounders/Pages/5G-fifth-generation-of-mobile-technologies.aspx

[32] C. Herlitz, "5 trends to watch in 5G connected vehicles," Jul. 16, 2021. [Online]. Available: https://www.automotiveworld.com/articles/5-trends-to-watch-in-5g-connected-vehicles/

[33] "The future of automotive computing: Cloud and edge," Oct. 6, 2022. [Online]. Available: https://www.mckinsey.com/industries/semiconductors/our-insights/the-future-of-automotive-computing-cloud-and-edge

International Journal of Distributed and Cloud Computing
8 (1 & 2) 2020, 01-08
http://www.publishingindia.com/ijdcc

Significance of Admission Control Mechanism in Cloud Computing Environment

Nikky Ahuja[1*], Priyesh Kanungo[2] and Sumant Katiyal[3]
[1]Research Scholar, School of Computer Science and IT, Devi Ahilya Vishwavidyalaya, Indore, Madhya Pradesh, India. Email: ahuja_nikky@yahoo.co.in
[2]Professor, School of Computer Science and IT, Devi Ahilya Vishwavidyalaya, Indore, Madhya Pradesh, India. Email: priyeshkanungo@hotmail.com
[3]Professor, School of Electronics, Devi Ahilya Vishwavidyalaya, Indore, Madhya Pradesh, India. Email: sumant578@yahoo.com
*Corresponding Author

Abstract: While allocating the cloud's resources the most crucial task for any service provider is the fulfilment of SLA obligations, since, non-meeting of these obligations results in penalty cost that would ultimately increase the providers' cost and will also affect their customer's satisfaction level. This might further have an impact on the software services and may increase the administrative and resource management cost. The admission control mechanism can play a pivotal role in resolving this problem by acting as a guide for resource provisioning. Thus, one must make efforts for the establishment of a better admission control mechanism between SaaS and IaaS provider. The model and algorithm proposed in this paper is an effort to show the significance of admission control mechanism for a service provider during the process of resource allocation, so as to minimize SLA violation and it's after-effects. Findings of the simulation study show that the proposed model helps in the successful admission of tasks in Cloud, meeting QoS constraints defined in SLA.

Keywords: Admission control, Quality of Service (QoS), Service Level Agreement (SLA), Software as a Service (SaaS).

I. Introduction

Cloud Computing is a budding notion that apprehends the use of the internet for providing and administering services on pay-as-you-go model [1] [2]. Cloud is an organization of virtualized servers that facilitates the fulfilment of diverse resource demands of the user. Each and every service in an IaaS is considered as a set of Virtual machines (VMs) that are activated considering the pattern of their workflow each time a new user request arrives. Virtualization techniques are used in a cloud to facilitate the faultless positioning of the system's components which further simplifies the process of horizontal elasticity [3]. It helps in keeping a check over the system's performance for varying workloads. Also, horizontal elasticity helps in the addition or removal of duplicate VMs. The aim of admission controlling technique here is to ease the cloud performance and strengthen it. During the process of admission control, the IaaS provider considers both the basic computational and networking needs [4]. It also keeps an account of extra requirements that may arise in the run-time of the request. This makes the cloud system dynamic. In the majority of cases, these dynamic requirements are relatively large as compared to the initial requisition made, especially when multiple users are accessing the cloud resources with a high variation of a workflow.

This increases the requirement for the numbers of initiated VMs, which are added during the performance of the request. This makes it mandatory for the service provider to consider the dynamic nature of the cloud system when allocating the resources. Admission control mechanism helps the service provider in balancing the basic and dynamic resource requirement, and also in evaluating any possibility for the optimization of resources when accepting any service [5].

II. Admission Control with QoS Factors Consideration

In a cloud computing environment, the mechanism that acts as a guide for a SaaS provider as on how to prioritize tasks, applications, data flow or users to assure or meet the pre-decided level of performance is known as Quality of Service (QoS). It is one of the most important parameters that must be focused on cloud computing [6]. The standards set in QoS are frequent and are highly dependent on the final output, performance limitations and data loss rate of the application. But the major challenge faced during the fulfilment of QoS requirement is the provisioning of the resources. For the successful accomplishment of a cloud contract, it is really essential to get a thorough understanding of customer's Quality of Service (QoS) requirement and then allocating of the resources as per the users' demand. It is quite a challenging task for a service provider as they have to manage the entire request as well as

have to focus on the satisfaction of customers meeting their QoS requirements. For maintaining the QoS, a service provider is needed to give an immediate response to the customers so as to minimize request delay time [7]. This requires an effective and efficient admission control strategy that could help in the fulfilment of the basic purpose of SaaS provider. Admission control mechanism should be designed in such a way that the obligations of both, the user and the service provider are met [8]. This helps in satisfying their needs and also benefits them both.

For assuring that both the user and the service provider perform their responsibility efficiently, an agreement termed as SLA is made between them. SLA [9] is an agreement that creates a legal obligation between a service provider and user on the agreed terms of service including an implicit description of all the parameters defining a level of service, assurance of the quality. It also states the remedies for all the cases of agreement violations (if any). SLA is formed only after getting the consent of both the parties involved in the contract and is continued until the completion of all the jobs assigned by the user. The QoS constraints [10] [11] that must be fulfilled for the successful accomplishment of any service contract between the SaaS provider and the user are defined as:

Deadline: In this complex and dynamic business environment meeting of deadlines plays a pivotal role in the successful accomplishment of any cloud contract. The term deadline means the maximum time limit that is given by the client to the service provider for getting the expected service output.

Budget: The maximum limit of the amount that a client is ready to pay for the successful accomplishment of the task in the stipulated time limit is known as a budget. It also defines the amount which would be paid by the user for the renting of administrative services.

Penalty Rate Ratio: The cost incurred upon the violation of any of the term defined in SLA is termed as penalty rate ratio. It acts as a remedy for the client when the SaaS provider misses the given deadline.

Input File Size: The length of the file that contains information regarding the client and its service requirement.

Length of the Request: For the fulfilment of the user's resource-specific requirement, millions of instructions are needed to be executed. This represents the length of the cloud's service contract.

III. Literature Survey

Khoshkholghi *et al.* [12] have studied the issues of providing high-quality services to cloud service users and lowering the consumption of power in the data centre. Findings of the study show that proposed algorithms generates improved SLA guarantee up to 87% and a reduction of 28% in energy consumption as compared to conventional algorithms.

In the research work of He *et al.* [13], authors have proposed an innovative admission control mechanism, which they have named as EB-based admission control mechanism, for controlling the traffic of large volumes of concurrent resource request in a cloud computing environment. Performance evaluation of the proposed mechanism shows that a larger volume of flows could be accepted for higher delay constraint.

Choi and Lim [14] have evaluated resource provisioning strategies for SLA based cloud computing architecture. They have proposed a combinatorial auction system for the allocation of resources keeping in view the SLA constraints of the contract. Authors have suggested that the profit can be increased by reducing SLA violations and penalty cost associated with it. For cost reduction, while deciding winner for auction job urgency based deadline constraint of the service contract is examined. Authors have compared their proposed system with conventional methods and their test results show that the proposed system is more profitable and shows higher rates of successfully accomplished jobs.

For an increasing profit of cloud service providers and improving resource utilization, there is a need for effective and better admission control. This objective can be achieved by using vivid research methodologies which focuses on resource provisioning keeping in concern enhance profit and decrease makespan. Authors [11] have worked upon two different learning approaches i.e. SVM and ANN for achieving the objective of enhancing customer satisfaction level and proper resource provisioning. Experimental results show that proposed methodologies could achieve better admission control meeting customer satisfaction constraints.

Zhou and Hu [15] have presented a scheduling algorithm for tasks in the cloud that measures the minimum completion time for all assigned tasks to cloud's resources and then select minimum and maximum values to make pairs of minimum completion time. This process is continued until all tasks are scheduled. Authors have then compared their algorithm with min-min, max-min algorithms and have interpreted that for efficient task scheduling, apart from completion time we must also consider other factors like resources cost, energy consumption rate, resource utilization rate and average response time.

IV. System Model

This section describes a model of SaaS provider that comprises of the actors and components of admission control, scheduling and knowledge processing. Actors of this model consist of the service provider and the user, whereas the functions of the application layer and platform layer are the parts of the cloud system. Here user submits a request for software from the SaaS provider which consists of the description of his QoS requirements. In the next step, the platform layer utilizes the admission control mechanism and accepts a request or rejects it after doing the proper analysis. A request is filtered on the basis of the availability, price, storage capacity and load of the VMs. The decisions for admission control are based on its knowledge

collected from its knowledge processing component. On the basis of this decision, the allocation of resources is done by the scheduler. Fig. 1, shows a service model for SaaS providers in a Cloud Computing Environment consisting of three layers namely, infrastructure layer, platform layer and application layer to meet customer request [16] [17]. All services offered to customers are managed by the application layer of the model. Scheduling and mapping of customer's QoS request to infrastructure level parameters and hence, allocation of resource VM are done by platform layer. Finally, the actual initiation and release of VMs are controlled by the infrastructure layer.

A. Actors in the Model

The actors engaged in our model along with their aims are stated below:

- *Customer:* The one who actually requests for resource access to the cloud service provider and consumes them. As soon as a customer agrees to the pre-defined conditions of SLA like response time etc., an application request is sent to the application layer of a service provider with QoS requirements like type of request, product type along with contract length, deadline of the contract etc. If a customer requests the upgrade of service(s) in mid-way the contract, a service provider has to manage this request intelligently matching the SLA requirements. SLA acts as a legal contract between user and service provider and if any party violates the terms of SLA, it has to pay a penalty as per the norms defined in SLA. SLA includes QoS parameters specified by customers and pre-defined parameters of service providers.
- *Cloud Service Provider (CSP):* CSP provides services to cloud customer like storage, network etc. They actually managed their own resources. CSPs purpose is to boost profit and expand their market share by undertaking as many requests as possible.
- *Cloud Broker:* A cloud broker act as an intermediary between user and provider. The role of a broker is simply to save user time and efforts by searching required services from different vendors and also it provides information about how to use cloud computing in business to support business goals. The actual budget of serving broker is contingent with the budget of the customer and anticipated discount percentage. Broker's profit is dependent on the provider price and the budget of the customer. To maximize the profit from the margin of negotiated price between customer and service provider is the main objective of the broker.

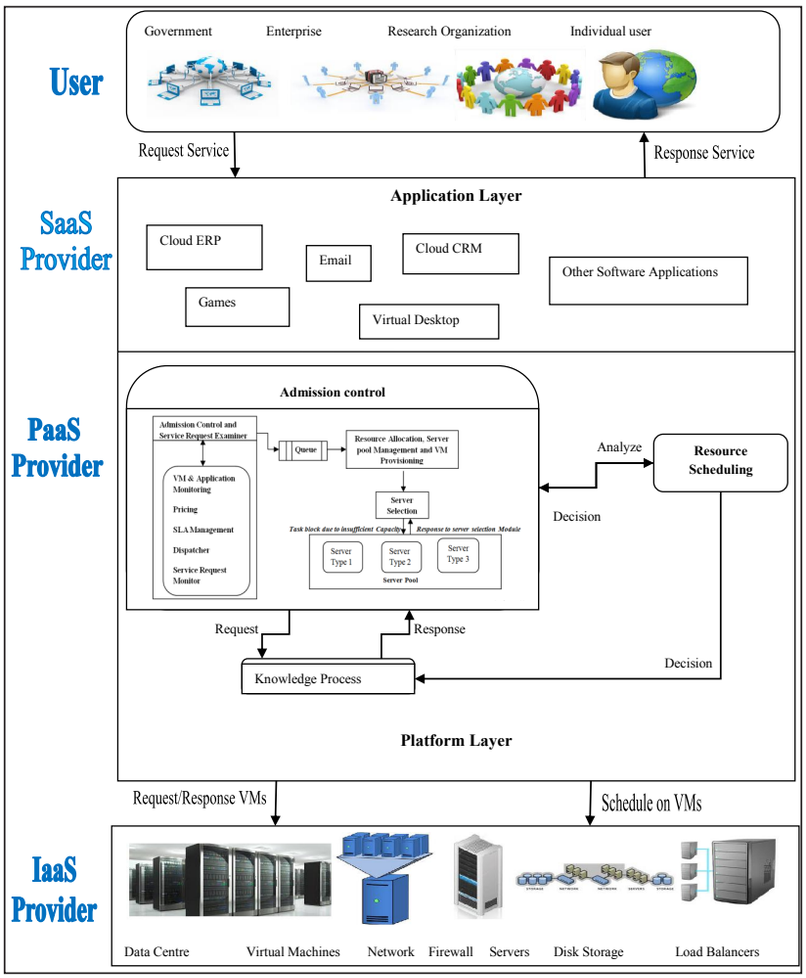


Fig. 1: Knowledge-based System Model for Admission Control and Task Scheduling in Cloud

B. Users in the Model

There are four type users in the proposed model namely: individual user, government organization, private enterprise and research organization. The request for resource allocation is initiated by the user via the application layer of the model. This request is then checked for SLA parameters and if the customer request meets provider defined SLA parameters then next checking for resource availability is done. If a resource is available, request is accepted and scheduled by a platform layer of the model. Finally, initiation of VM is done by the infrastructure layer on the basis of scheduling done and at the end of the contract infrastructure layer releases the VM thus, allocated. If at any point of contract, the customer feels a need to upgrade the contract; the whole process of request initiation is repeated. However, if constraints are met and resource is available, re-scheduling is done and a new contract is formed. Otherwise, a request is cancelled and the old contract continues. Platform layer plays a vital role in this model as major decision-making is done in it.

C. Software Services Available with the Provider

The software services available with the provider include applications that can be accessed by them from different locations and user devices. This service is provided via a thin user interface, which can either be a program interface or a web-based interface. The end-user here is not required to manage the infrastructural components like servers, operating systems, network, etc. of a cloud. Only configuration settings are customized as per user requirements. This generates the need for application software managed and initiated by the browser to meet the user demand. Software as a service plays a significant role in any enterprise model. It is helpful in managing organizations human resources, business processes and their workflow, enterprise resource program (ERP), consumer relationship management (CRM) and other integrated applications. For clients or users, it is beneficial since they are not required to make any investment in servers or buying licensed software.

D. Components of the Admission Control Mechanism

- *Service Request Examiner and Admission Control:* Before accepting or rejecting any new user request, server request examiner and admission control mechanism performs its evaluation so as to get the clarity of the QoS requirement. This helps in the minimization of SLA by keeping a check on the resource overloading as the number of available resources is limited. For making the resource allocation decisions effective and efficient, continuous monitoring of the cloud system is required. This will provide regular updates about the availability of virtualized resources using VM monitoring mechanism and average workload processing status using service request monitor mechanism. The key that ensures the successful serving of a large number of task requests from the end-user meeting the SLA norms is known as the automatic resource management technique. It also helps in the dynamic allocation of resources, so that all accepted resources can be served efficiently.
- *VM and Application Monitor:* On the basis of accepted services, the resource management system in a CCE keeps the track of the status and the performance of the available resources at different levels. Here VM monitor mechanism is used for the evaluation and monitoring of the capacity of VM that available for a new request allocation, whereas application monitoring mechanism is used for the keeping the track whether all the terms of SLA are being followed or not. If any breach in SLA is identified, then a notification is sent to SLA resource allocator which further takes the most suited action for preventing this violation.
- *Pricing Model:* Services in the cloud whether hardware or software are charged on the basis of the pay-per-use model. The greater the resources consumed the higher is the cost. This generates a necessity for the deployment of a pricing strategy that can help the provider to attain the objective of profit maximization and meeting QoS constraints as per the user demand.
- *SLA Management:* The component that keeps the record of SLA formed between users and service provider with the detailed history of their fulfilment is known as SLA management. This record helps accounting system to compute the final cost of the allocated resources by identifying their level of actual usage by the requested service. This cost is finally charged from the end-user. The information stored in SLA management records also proves to be useful for the service request examiner and admission controller for improving the quality of resource allocation decisions.
- *Dispatcher:* For the deployment of an application on the suitable virtual resource a dispatcher is used. It also performs the responsibility of creating an image of the VM and their initiation on a selective host.
- *Service Request Monitor:* The monitoring mechanism that evaluates the progress of the executed service request is known as a service request monitor.
- *Scheduler:* Scheduling of task on the available VMs is done by the scheduler. It queues the tasks after determining their priority and decides as to which VM is to be provisioned for it. For performing this task of scheduling, the scheduler requires the information about the current status of the datacenter e.g. the number of CPUs and amount of memory available for allocation, network bandwidth, etc. The load balancer provides the required information about datacenters to the scheduler. As soon as the task is accomplished, the pricing module is initiated for computing the final cost of resources used by the customer and the bill is sent to the user via SLA management.

E. Resources Available with the Provider

- *Virtual Machines (VMs):* For every accepted request dynamic allocation of multiple VMs could be started which provides flexibility for the configuration of different parts of resources on the physical host to meet the customized user requests. In a cloud system, every VM is fully isolated from another one in the same physical host.
- *Physical Machines:* The datacenters in a CCE consists of various servers for data computation which makes the required resources available for meeting the customer's service demand.
- *Datacenters:* In the present dynamic business world datacenters acts as the backbone for the IT infrastructure. There span consists of the wide range of applications varying from the distribution and production of energy, manufacturing industry, transportation, weather prediction and modelling, etc.
- *Load Balancer:* Load balancer is used to determine the actual amount of available resource (VM or PM) capacity. It identifies the over or underused resource both physical and virtual, and balances the workload among them, thus, aiding the establishment of control over resource utilization. It also helps in reducing power consumption by employing idle resources. Load balancer works together with scheduler and admission controller for balancing the workloads.
- *Load Balancing:* In a dynamic and heterogeneous cloud computing environment for controlling the unexpected changes in workload, providers are supposed to exploit the dynamic load balancing techniques. The selected technique must also put a check on the improper utilization of the cloud's resources, since the underutilization of resources may result in resource and energy wastage, whereas, overutilization increases the application's response time. Migration of VM is used as a solution to this problem.

V. Algorithm

This section illustrates an algorithm for the admission control mechanism in a Cloud environment. Here, a check is made to find whether a request is to be accepted or not. For this, the parameters to be considered include- an average of request arrival rate, the average workload of servers available for resource allocation, capacity utilization limit of servers and filtration coefficient of the request.

Algorithm for Admission Control Mechanism

1. Start
2. If (already a user)
3. login();
4. else (create an account)
5. then login();
6. loading SLA_packageinfo( );
7. put(meta info of task)
8. get request on the basis of package selected
9. for new request arrival
10. check server and VM availability
11. calculate present average request arrival rate and present average workload
12. check for the maximum limit of capacity utilization
13. if(for the present request average workload is greater than maximum capacity)
14. reject the request
15. else
16. calculate request filtration coefficient using formula 1- (Average workload-Maximum limit of capacity utilization)
17. then reject the request with the probability of 1- calculated request filtration coefficient
18. else accept the request
19. End

Whenever a new request arrives, admission control mechanism checks for the QoS requirement of the request. If it is found that resources could be allocated to the present request without violating any of the SLA norms then the request is accepted else rejected. For assuring the service quality, first of all the number of available servers and VMs is checked, then their average workload (which includes the workload of present request) is calculated along with the average rate of request arrival. If for the selected server and its VM, the average workload calculated is lower than its maximum capacity utilization limit then only request is accepted. This helps in the scheduling of tasks as per the resources available and hence, aids in reducing the number of SLA violations.

VI. Experimentation Implementation

In this work, CloudSim [18] version 3.0.3 has been used as a simulator for a cloud computing environment. An algorithm developed was tested in this simulated environment from both users' and providers' perspective. Windows provide the platform and other software used are:

- Net Beans 8.0.2
- JDK 1.8.0_40
- MySQL Server 5.1
- SQLyog Server ultimate v9.0.2.0
- JRE 1.8.0_40

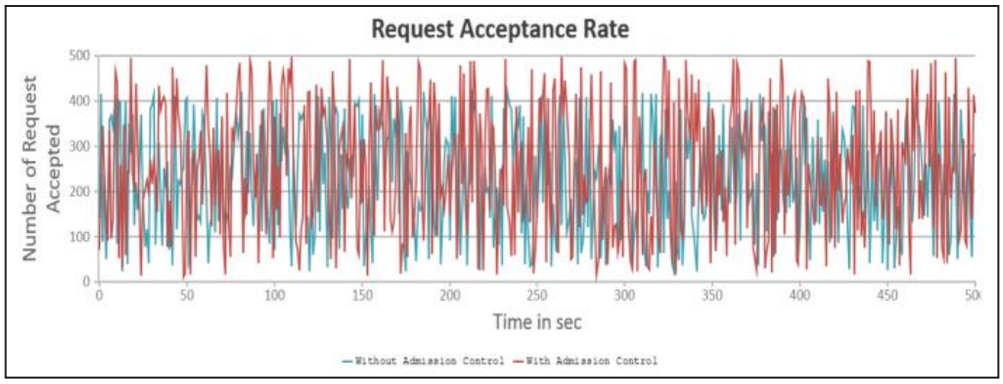
Resources used for Experimental Setup:

Our system offers four types of packages to the customers: basic, silver, gold and diamond. From these four packages, three are constant and only one package titled diamond can be customized. The configuration of the resources is as: basic pack: - 2 GB RAM, CPU core processor and Windows as the operating system, silver pack: - 4 GB RAM, CPU core processor and Windows as OS, gold: - 8 GB RAM, CPU intel core processor

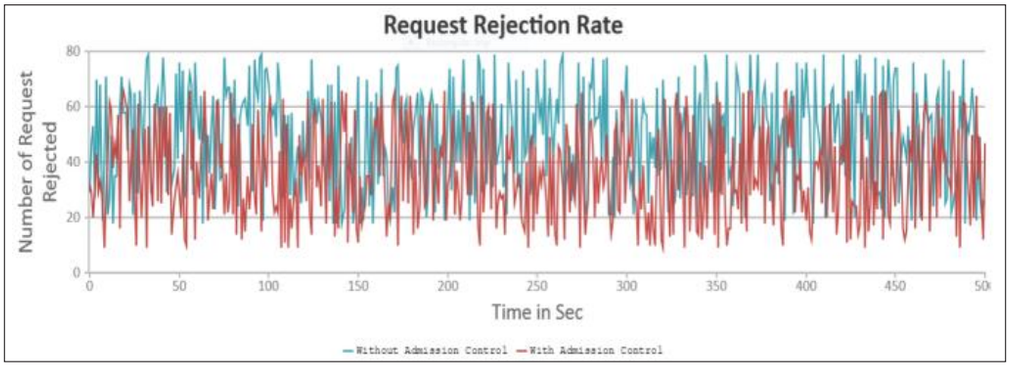
and windows as OS and for diamond configuration is tailored on the basis of a request of the user (in diamond pack local storage space is variable but the processor and RAM are constant).

VII. Analysis of Result

We have used CloudSim as a simulator of a cloud environment and have implemented the proposed model in this environment. Perspectives of both users' and SaaS providers were observed. Performance parameters of users' side are: number of requests accepted per unit of time and the speed of request processing i.e. an average of service initiation time and that of providers side are: number of VM's initiated an average profit earned. The proposed algorithm was tested on a simulator with the help of a synthetic workload. Data were synthesized randomly with the help of a suitable statistical distribution. The following graph shows the results of the model.

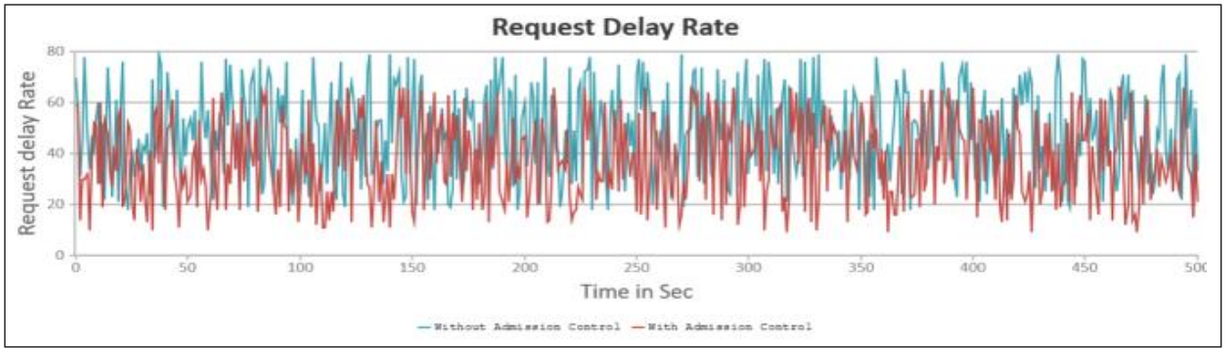


Graph 1: Variation in Number of Request Accepted



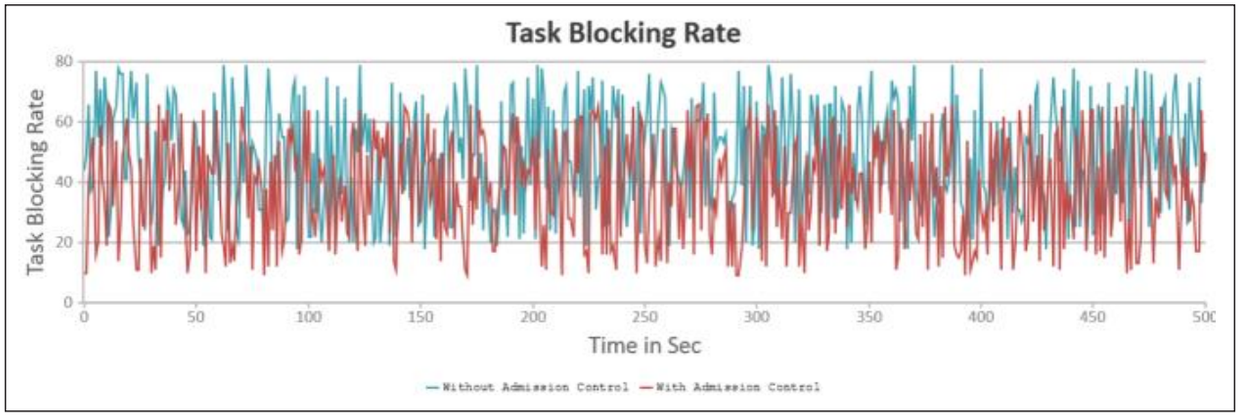
Graph 2: Variation in Number of Requests Rejected

Graph 1 and 2, shows the impact of the admission control mechanism on request acceptance and rejection rate. With the introduction of the admission control mechanism in resource allocation, the number of request accepted gets increased that ultimately results in a lower request rejection rate.



Graph 3: Average Delay Time

Graph 3, shows the impact of the admission control mechanism on request delay rate. Since every request is checked by the admission control mechanism for QoS requirements before being accepted for resource allocation, the chances of request accomplishment getting delayed get decreased.



Graph 4: Average Task Blocking Rate

Graph 4, shows the impact of the admission control mechanism on task blocking rate. Filtration coefficient is calculated to check probability with which the task is to be rejected. A task with higher probability gets rejected first. With the introduction of the admission control mechanism the blocking rate of tasks/ requests also gets reduced.

VIII. Contribution of the Research

Every new research enriches the contents of the related subject. With the help of literature review, it was found that while scheduling task in cloud one should not only consider the total completion time of all task, but must also look for some other important aspects like cost of the resources, rate of energy consumption, workload and network bandwidth, system capacity, task response time, the threshold for capacity utilization, etc. This paper presents an algorithm which considered both capacity utilization and systems average workload along with SLA constraints and resource availability to determine whether a task should be accepted or rejected. This will help in pre-filtering of task and would minimize the violation of SLA, since, if not checked properly a newly admitted task can violate the functioning of already executing a task which would ultimately increase the number of SLA violations and incur penalty cost. The model is helpful in reducing SLA violations and enhancing providers' profit.

IX. Conclusion, Limitation and Future Work

The present paper is an attempt to devise a model and algorithm for admission control mechanism so as to manage the acceptance of requests/tasks assigned by cloud users in such a way that the QoS requirement of users are met and no SLA term gets violated. Thus, the work helps in reducing the number of SLA violation and their consequences. Proposed algorithm focuses towards the admission of a task in the cloud system and in future, we would try and extend our work to devise some strategies for resource scheduling after the request has been accepted by our admission control mechanism. These strategies will focus on reducing the delay in task accomplishment and hence, the reduction in penalty cost. Work will also focus on enhancing the customer satisfaction level, controlling the cost and thereby increasing the overall revenue of the system. Different users in the cloud have different demands of quality and resource which acts as a limitation of the study since to satisfy this heterogeneous user demands with limited available resources is not an easy task.

References

[1] A. Kushwaha, P. Pathak, and S. Gupta, "Review of optimize load balancing algorithms in cloud," *International Journal of Distributed and Cloud Computing*, vol. 4, no. 2, pp. 1-9, 2016.

[2] A. Pandey, "Cloud computing implementation: An analysis using systems approach," *International Journal of Distributed and Cloud Computing*, vol. 4, no. 2, pp. 10-22, 2016.

[3] N. Ahuja, P. Kanungo, and S. Katiyal, "SLA based rescheduling of task for optimum allocation of resources in cloud computing environment," *SRIMCA*, ISSN: 0974-3308, vol. 12, no. 1, pp. 65-78, June 2019.

[4] H. N. Hoang, S. L. Van, H. N. Maue, and C. P. N. Bien, "Admission control and scheduling algorithms based on ACO and PSO heuristic for optimizing cost in cloud computing," *Research Developments in Intelligent Information and Database Systems,* pp. 15-28, Springer, 2016.

[5] M. R. Ghori, and A. A. Ahmed, "Review of access control mechanisms in cloud computing," Int. PostGraduate Conf. Appl. Sci. & Phys. 2017, Dec. 7, 2017, Johor, Malaysia. *J. Phys.: Conf. Ser.*, vol. 1049, 01209, 2018.

[6] N. Ahuja, P. Kanungo, S. Katiyal, and P. Parandkar, "Optimum management of resources in cloud computing environment: A challenge," 2015 *Int. Conf. "Comput.*

for Sustain. Global Develop." INDIACom, Organised by BVICAM, New Delhi, Mar. 2015, pp. 4.30-4.33.

[7] J. Varghese, and S. Jagannatha, "Task scheduling and VM allocation in cloud computing: A survey," *International Journal of Distributed and Cloud Computing*, vol. 7, no. 2, pp. 7-18, 2019.

[8] R. Buyya, S. K. Garg, and R. N. Calheiros, "SLA-oriented resource provisioning for cloud computing: Challenges, architecture, and solutions," *Int. Conf. Cloud and Service Comput.,* IEEE Computer Society Washington, DC, USA, ISBN: 978-1-4577-1635-5, 2011, pp. 1-10.

[9] H. Goudarzi, and M. Pedram, "Maximizing profit in cloud computing system via resource allocation," In *ICDCSW'11: 31st Int. Conf. Distrib. Comput. Syst. Workshops,* Minneapolis, IEEE, 2011, pp. 1-6.

[10] F. Motavaselalhagh, F. S. Esfahani, and H. R. Arabnia, "Knowledge-based adaptable scheduler for SaaS providers in cloud computing," *Human-Centric Computing and Information Science*, vol. 5, no. 16, pp 1-19, Springer, 2015.

[11] M. S. Lakshmi, S. P. Kumar, M. Janardhan, and K. Gayathri, "Machine learning methods for refining SLA based admission control and resource allocation in cloud computing," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 9, pp. 834-840, 2017.

[12] M. A. Khoshkholghi, M. N. Derahman, A. Abdullah, S. Subramaniam, and M. Othman, "Energy-efficient algorithms for dynamic virtual machine consolidation in cloud data centers," *IEEE Access*, vol. 5, pp. 10709-10722, 2017.

[13] Y. He, J. Huang, Q. Duan, Z. Xiong, J. Lv, and Y. Liu, "A novel admission control model in cloud computing," *Networking and Internet Architecture*, pp. 1-14, 2014. [Online]. Available: arXiv:1401.4716 [cs.NI]

[14] Y. Choi, and Y. Lim, "Optimization approach for resource allocation on cloud computing for IoT," *International Journal of Distributed Sensor Networks*, vol. 2016, Article ID 3479247, pp. 1-6, Hindawi Publishing Corporation, 2016. [Online]. Available: http://dx.doi.org/10.1155/2016/3479247

[15] Z. Zhou, and H. Zhigang, "Task scheduling algorithm based on greedy strategy in cloud computing," *The Open Cybernetics and Systemics Journal*, vol. 8, pp. 111-114, 2014.

[16] L. Wu, S. K. Garg, and R. Buyya, "SLA-based admission control for a software-as-a-service provider in cloud computing environments," *Journal of Computer and System Sciences*, pp. 1280-1299, Elsevier, 2012.

[17] L. Wu, S. K. Garg, and R. Buyya, "SLA-based resource allocation for software as a service provider (SaaS) in cloud computing environments," *11th IEEE/ACM Int. Symp. Cluster, Cloud and Grid Comput. (CCGrid),* Newport Beach, CA, ISBN: 978-1-4577-0129-0, 2011, pp. 195-204.

[18] R. N. Calheiros, et al., "CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Journal on Software-Practice and Experience*, vol. 41, no. 1, pp. 23-50, John Wiley & Sons, Inc. New York, NY, USA, 2011.

International Journal of Distributed and Cloud Computing
8 (1 & 2) 2020, 09-16
http://www.publishingindia.com/ijdcc

Hybrid Searchable Encryption Scheme for Data Security in Cloud

Prachi Gupta
M.Tech Student, Department of Computer Science & Application, Kurukshetra University, Kurukshetra, Haryana, India. Email: prachigpt.gupta7@gmail.com

Abstract: Cloud Computing is an emerging technology which provides on-demand services based on the pay-as-you-go model which reduces IT costs significantly. Cloud comes with many benefits such as Business continuity, Collaboration efficiency, scalability, resource sharing etc. But still, it constitutes several data security issues such as data confidentiality, integrity, authentication, privacy etc. Thus, to protect sensitive data, the document is encrypted at the data owner side before transmitting it to a cloud environment. Cryptography is necessary to provide data security. Although this overpowers the advantages of cloud computing, whenever data user wants to access a part of data it needs to decrypt the data before using. In this paper, a Hybrid Searchable Encryption Scheme is proposed which provide data security while supporting multi-keyword search over the encrypted data by using a Vector Space Model (VSM). Where sense embedding technique is used to create VSM for the document. To provide data confidentiality combination of AES and Blowfish encryption is used. Sha-3 and RSA are used for data integrity. Steganography is applied for secure transfer of keys and message digest. The experimental result shows that the proposed scheme achieves better data security by providing data confidentiality, integrity and authentication along with multi-keyword search at the client end.

Keywords: Cloud computing, Hybrid encryption, Searchable encryption, Sense embedding based search, Vector space model.

I. Introduction

Cloud Computing is a fast-growing technology that provides hardware as well as software services through a shared pool of resources. Even though the cloud is considered as a very good environment but still it constitutes many security issues such as data confidentiality, integrity and authentication [1]. Confidentiality is the process in which data is hidden from unauthorized access and should be interpreted only by the intended user. Integrity refers to the process in which data is secure from any tampering while transmission. Authentication is defined as the process in which only the intended user has access to data. There are three types of cloud deployment model (Fig. 1): Public Cloud, Private Cloud and Hybrid Cloud:

- *Public Cloud:* Public Cloud deals to provide cloud services to the general public. It is considered unsafe because it is accessible to the general public. A cloud service provider is an owner with all ownership of the public cloud with its own policies. Advantages of public cloud include low cost, no maintenance, scalability and high reliability.
- *Private Cloud:* Private Clouds are those in which sharing of data is done within a particular organisation only and no data shared with other organisation. Private cloud is managed by the organisation itself or by any third party. It is considered to be the safest as the number of users are limited under this category.

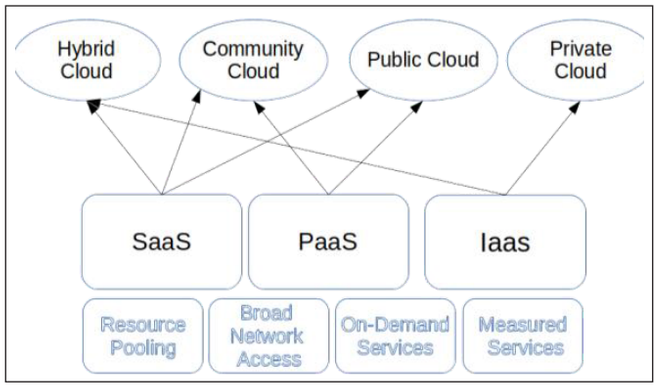


Fig. 1: Cloud Deployment Models

- *Hybrid Cloud:* Hybrid Cloud is a combination of both public and private cloud. The hybrid cloud contains the features and advantages of each cloud model.

Service models in the cloud can also be categorised into 3 categories:

- *Software as a Service (SaaS):* Using SaaS, cloud consumers can release their application on an environment hosted by cloud. The environment can be accessed through networks via different clients. Eg. SalesForce.com, Google Mail, Google Docs etc.

- *Platform as a Service (PaaS):* Paas provides a development platform which supports full software lifecycle allowing cloud users to develop cloud service (SaaS) and application. Google AppEngine is an example of PaaS.
- *Infrastructure as a Service (IaaS):* IT infrastructures such as processing, storage, networks etc are provided to cloud consumers via IaaS. Extensive use of virtualisation is done here to integrate physical resources in an ad-hoc manner. Amazon EC2 is an example of IaaS.

Data security is one of the major issues that come with the usage of the cloud environment thus, to provide security to data using cryptography is necessary. Cryptography is a technique in which data in readable form is converted to unreadable data. There are many cryptographic techniques used for providing security to data. Cryptography is divided into two main categories:

- *Symmetric Technique:* For both the process of encryption and decryption of data a single is used. It is faster because only a single key is used. AES, DES, Blowfish, Twofish are examples of symmetric encryption.
- *Asymmetric Technique:* In this, two separate keys are used one for encrypting the data and other for decrypting the data. It is slower as compared to the previous technique. RSA is one of the examples for this technique.

For maximum security of data, advantages of both symmetric and asymmetric technique need to take into account which leads to the formation of a system known as hybrid encryption. The main goal of data security is to achieve data confidentiality by using symmetric technique, integrity through message digest and authentication via asymmetric technique along with secure transfer of keys from user to client-side. For key transfer mechanism. a technique known as steganography is used. Steganography is a technique which hides the existence of data by concealing the data into a source object. An object may be an image file, audio file, video file or text file.

Whereas encrypting of data before storing to the cloud comes hinders advantages provided by the cloud environment. Each time encrypted data needs to be decrypted at the data user end to perform manipulation. Efforts have been made to conquer this issue by providing a way in which multi-keyword search could be run over encrypted data.

Searchable encryption is a method in which encrypted data is allowed to be searched in a private manner without decrypting the data. For the same, Embedding is defined which is a technique in natural language processing where documents can be represented in form of vectors.

Vector Space Model is a way in which documents are represented in form of vectors. VSM assumes that each word in a document is independent. For each word *w* present in document D. VSM of document cane represented as below:

$$D_{ij} = (w_{11}, w_{12}, w_{13}, ..., w_{nn})$$

For representing the document's features appropriately in vector space many attempts have been made. TF-IDF model is one of the most popular approaches for the same.

Word embedding is another way to represents a document in form of a real-valued vector by capturing a relationship between words present in the document. However, word embedding is unable to consider polysemous word (word having multiple meaning). Sense embedding on the other hand can be used to represent a polysemous word as well.

In this paper, a scheme proposed which provide data with three main security factor i.e. data confidentiality, integrity and authentication along with a feature to search over encrypted data which is achieved through sense embedding technique.

II. Related Work

In this section, we will describe some previously developed hybrid encryption scheme and searchable encryption scheme. Hybrid Encryption combines two or more cryptographic technique to achieve security goals. Searchable encryption is described as a procedure in which encrypted data is stored over cloud and data user is allowed to search that encrypted data.

A. Hybrid Encryption in Cloud

For increasing security of document stored in a cloud environment, a combination of cryptographic encryption scheme are used. Nagasai and Supriya [2] implemented three hybrid models on the text and their time complexity is compared. Combination of AES and RSA algorithms are used along with OTP technique to develop three different models. Olumide *et al.* [3] proposed a system in which a combination of AES and fully homomorphic encryption scheme is used to create a hybrid encryption scheme. Divya *et al.* [4] in their paper proposed a scheme which provides better security on the cloud while transmitting data. Blowfish, RSA and SHA-2 algorithms are used in combination to create hybrid encryption. Citra *et al.* [5] developed an efficient algorithm which receives data security issues data confidentiality. integrity and authentication. A combination of hybrid encryption and steganography is used to achieve security.

B. Searchable Encryption

Further in this section, the previously developed encryption scheme is introduced. In view of the single keyword search over encrypted data, the first idea was developed by Song *et al.* [6]. Symmetric searchable encryption was developed where the search time of scheme was proportional to the size of the collection of the document because searching requires a scan of the whole document. In 2006, Curtmola *et al.* [7] introduced searchable symmetric encryption (SSE). Two SSE were introduced by the author first one was Non-adaptive SSE (SSE-1) and the second one was Adaptive SSE (SSE-2). Addressing issues in previous work in which adversary is not

allowed to search over data, the author achieved a search time which is linear to the document that was matching to search query.

Public key encryption with keyword search (PEKS) implies identity-based encryption in which a user's identity is used to encrypt as well as decrypt the data. Boneh *et al.* [8] in 2004 developed PEKS, in which a user's public key was used to encrypt the data containing certain keywords for performing a search. A keyword is also encrypted which unable server to get keyword location in the document. Z. Deng *et al.* [9] developed a scheme in which multiple users were allowed to search the keyword (MSESKA). 6 polynomial-time algorithms were used to develop multi-user attribute.

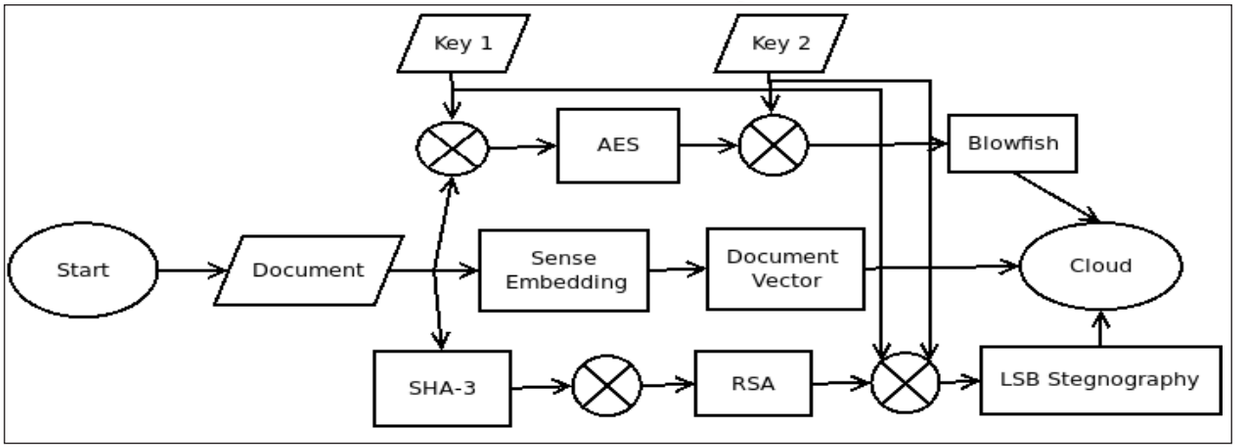


Fig. 2: Encryption Phase

In Predicate encryption, the user does not need any private key related to the public key to search on encrypted data. X. A. Wang *et al.* [10] in their paper claimed predicate encryption to be superior to traditional public-key encryption as the former scheme achieves more flexible and sophisticate functionality. Hidden vector encryption is also a type of predicate encryption. J. Karz *et al.* [11] introduced hidden vector encryption that supports a conjunctive combination of equality along with comparison and subset queries.

Schemes described above support only a single keyword search. Further studies have been conducted to extend search to multi-keyword search over data with a ranking of documents. N. Cao *et al.* [12] achieved a multi-keyword search result by using inner product similarity. Sun *et al.* [13] developed a searchable encryption scheme with document ranking using the vector space model. Aritomo and Watanbe [14] achieved a searchable ranked based scheme using a word embedding to create a vector for documents.

III. Word Embedding V/S Sense Embedding

Word embedding can be considered as a set of language modelling vocabulary words and phrases are mapped to real-valued vectors where vectors are represented in low-dimensional space. Sigmoid function of vector cosine similarity is to define the prediction probability. Two major limitations we can consider while using the word embedding 1) They don't consider the different meaning of a word and conflate the different meaning for a single word into a single vector, and 2) They ignore the wealth of information that is given by the semantic resources thus make their representations only on the basis of distributional statistics that is obtained from corpora. Several types of research have been made to eliminate these two limitations but none was successful in eliminating these two limitations at a time.

Word similarity measurement is used evaluate methods in lexical semantics and semantic similarity. For a given set of words, similarity judgement is created by the system for each pair. These judgements should be closest to those that are provided by humans.

Table I: Spearman Correlation Performance

Approaches	WS-Rel	YP-130	MEN	RG-65	WS-Sim	Average
GloVe	0.559	0.577	0.763	0.769	0.666	0.737
Word2Vec	0.476	0.343	0.665	0.732	0.707	0.644
ESA	-	-	-	0.749	-	-
PMI-SVD	0.523	0.337	0.726	0.738	0.659	0.695
Pilehvar *et al.* [21]	0.457	0.710	0.690	0.868	0.677	0.677
Our Approach	0.703	0.639	0.805	0.871	0.812	0.794

Different word similarity datsets are used to evaluate the performance of different algorithms. YP-130, RG-65, WS-353, WS-Rel and MEN. Our approach of creating sense vectors are compared with several other approaches namely GloVe, Word2Vec, ESA, PMI-SVD and Pilehvar *et al.* [21]. Table I shows the comparison between different approaches on 5 different datasets. Spearman correlation performance is done for all the considered approaches. Our approach for creating sense vector shows high reliability on both similarity and relatedness measurement task. Right most column in the table shows the average performance of all the approaches. This comparison shows that our approach is beneficial. Moving from word to sense embedding can significantly improve the effectiveness and accuracy of the representation.

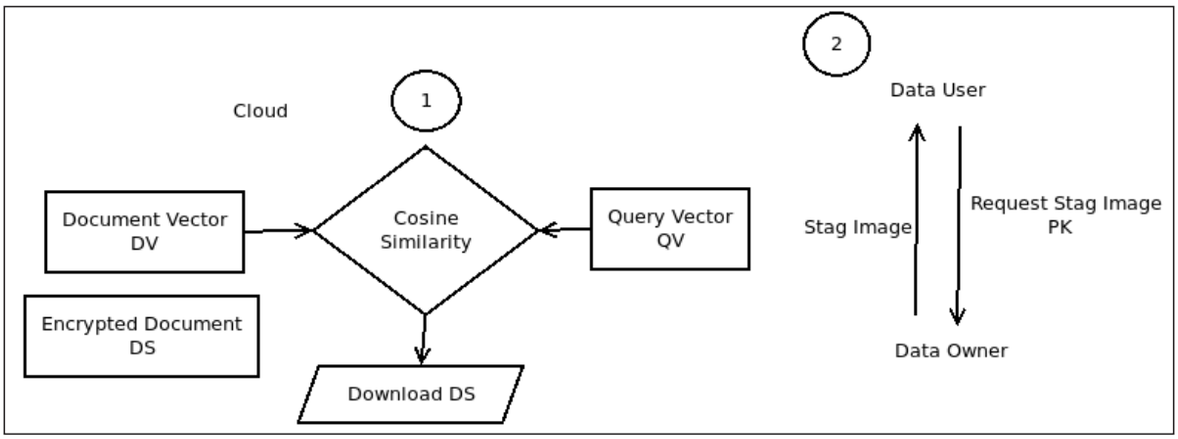


Fig. 3: Search Phase

IV. Problem Definition

A. Scheme Model

The scheme proposed contains three conscious things:

First the data owner which has the original document DC and it generates the corresponding sense embedding vector DV generated through sense embedding technique, second the data user and third one an untrusted cloud server. Document vector represents the features of the document. Sensitive data is protected from an intruder by first encrypting the document termed as secure document DS by data user before uploading it to cloud.

As shown in Fig. 2, firstly AES encryption is used and after then Blowfish encryption is applied to the output of AES encryption. To provide a document with data integrity and authentication message-digest MD of the document is created which is further encrypted using asymmetric encryption termed as secure message digest MDS. The keys (Key 1 and Key 2) and the message digest are transported using steganography technique. While the document vector DV and encrypted document DS is stored on the cloud.

In the searching session, initially, a trapdoor is created by the data used to get authorized by the data owner. The trapdoor is made which contains the query vector QV and is sent to the cloud server. The similarity between the query QV and the document vector DV is calculated by the cloud server and the calculated score is returned to the data user. The data owner according to the calculated score request for the encrypted document from the cloud server.

B. Threat Model

In the proposed scheme, the cloud server is considered, to be honest, but curious which is responsible for the introduction of searchable encryption in our scheme. This means cloud server is responsible to handle the request and pass the data to the data owner in an honest manner, but is curious at the same time. Hence, the cloud tries to recover maximum information from the document. In our scheme, we wish to let cloud infer minimum information while analyzing the data it receives.

C. Model Preludes

- *Word Embedding:* It is a technique where documents are converted into their corresponding vector. Word embedding creates a relation between the words present in the document and tries to represent those words in low dimensional real-valued vector. Doc2Vec and Word2Vec [15] are the two most popular algorithms for representing word embedding of the document. While the most common method to measure the similarity is the cosine measure in which the inner product of the vector is calculated.
- *Sense Embedding:* Word Embedding has a problem that it doesn't consider the polysemous word (word having multiple meaning). Sense Embedding in contrast to word embedding creates embedding for a document where polysemous words are also considered. There are many techniques [16], [17], [18] to do so but in our model, we have considered SensEmbed [19] to implement sense embedding to the document to create VSM.
- *Symmetric Encryption:* Symmetric Encryption is a technique which converts the plain text to cipher text

using only a single key that is a private key. It is also known as private key encryption. There are many symmteric encryptions available but AES and Blowfish are considered to be two of the best encryption algorithm. For first encryption, AES is considered and later Blowfish is applied to the output of AES encryption.

- *Asymmetric Encryption:* In contrast to symmetric encryption which only needs a single key for encryption as well as decryption, asymmetric encryption uses two separate keys each for encryption and decryption process. Asymmetric encryption is considered to be more secure as is allows user to share only the public key and not the private key. But asymmetric encryption is slower than symmetric encryptions when talking about time efficiency. RSA is one of the most popular algorithms used in this category.

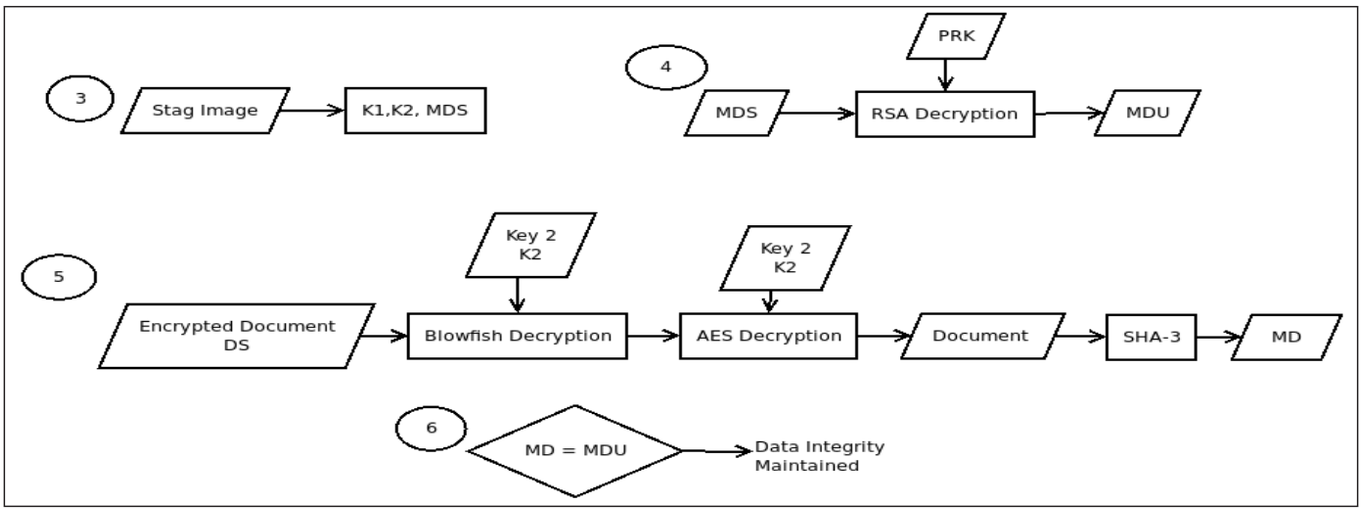


Fig. 4: Decryption Phase

- *Message Digest:* Message digest function creates a fixed-length one-way message of the document which is used to check the integrity of the document. If the document is modified in any way while the transfer from the data owner to data user then its message digest will definitely vary from the message digest of the original data. Many algorithms are there to create message-digest but one of the which SHA-3 is considered for our scheme.
- *Steganography:* Steganography is a method in which data is hidden under some other source. The source may be an image, audio or video. One of the most commonly used steganography technique is the LSB image steganography in which the transmitted data is stored in the Least Significant Bit of the source.

V. Scheme Methodology

In this section, description of the scheme methodology of plaintext is given which describes the data owner side scenario. The overview of the proposed scheme is simple and shown in Fig. 2.

A. Encryption Phase

The data user first creates a query vector QV by using same sense embedding technique as that of the data owner and send that query vector QV to the cloud server. As a measure to calculate the similarity between the query document QV and the document vector DV by the cloud server, cosine similarity between the two vectors that are query vector QV and document vector DV is generated.

Setup (K1, K2, PK): Firstly the system is initialized, two secret key K1 & secret key K2 is created by the data owner. While public key PK is obtained from the data user.

These keys are used for AES encryption, Blowfish Encryption and RSA encryption respectively. These AES and Blowfish encryption are used to encrypt the document and RSA encryption technique is used for encrypting message digest.

BuildScheme (DV, DS, MDS): The data owner runs the sense embedding technique to create sense vector DV for the document. Further, the message digest of the document is created using SHA-3 and then encrypted using RSA encryption using the key PK. The document is considered for encryption now using AES and Blowfish with the key K1 and K2 respectively.

Transport (K1, K2, MDS, IMG): The keys used for encryption that are K1 and K2 and the encrypted message digest termed as message digest secure MDS is transport using image IMG in which these three things are hidden using LSB steganography.

Table II: Notations

Q	Query
QV	Query Vector
DV	Document Vector
DS	Encrypted Document (Secure)
MD	Message Digest of Document
MDU	Message Digest Created by Data User
MDS	Encrypted Message Digest (Secure)
PK	Public Key of Data User
PRK	Private Key of Data User
IMG	Original Image
K1	Key 1 for AES
K2	Key 2 for Blowfish

B. Search Phase

Cloud server contains the encrypted document DS and the document vector DV for the document. Step 1 gets initiated when the data owner starts searching the cloud for the document of use.

Step 1: Search (QV, DV): The query vector QV which represents the vector of an interested word in query Q is received from data user and is used to calculate similarity with the document vector DV which contains the vector of polysemous words present in the document. This procedure is done without decrypting the encrypted document DS.

C. Decryption Phase

This section describes the data user's end. Where the decryption process for the required document takes place. Data user generates a query vector using the interested words as shown below:

GenQuery (Q, QV): The data user generates a query, for the keywords of interest W a query vector is generated QV which contains the interested words in query Q. QV is further used to search the plaintext vector DV. QV is sent to a cloud server for further working.

As described in part B of section IV Search (QV, DV) is generated using the DV present with the cloud server and QV provided by the Data user. If the search is successful the secure document for which Search (QV, DV) was generated is downloaded.

Step 2: The data user asks for the stag image from Data Owner as shown in Fig. 3 where the image consists of K1, K2 and secure message digest. Afterwards, Transport (K1, K2, MDS, IMG) is initiated and is transferred from Data owner to data user.

Step 3: Using the stag image, secure message digest (MDS), K1 and K2 are recovered as shown in Fig. 4.

Step 4: Secure message-digest MDS is decrypted to recover the message digest MDU using the data users own private key PRK.

Step 5: Using the keys K1 and K2 the document is recovered from secure document DS. First blowfish decryption is initiated using K2 and AES decryption process is applied to the output of blowfish decryption. Using the decrypted document, a message digest is generated using the same message digest function as that used at data owners side.

Step 6: MDU and MD are compared and considered for checking data integrity. If both MD and MDU are equal then data integrity is maintained.

VI. Performance Analysis

In this section, performance of the proposed scheme is analyzed by implementing the scheme and measuring multiple metrics. The proposed scheme was implemented in python 3.6.

- *Sense Embedding Generation:* Sense embedding for the document is created using babelfy library in python while considering the babelnet dataset. Further for creating the vector, word2vec in used from gensim python library. The time is not considered during creating the embedding for document. This scheme is open to any sense embedding technique so considering the time for one technique is not of much meaning. Desired secure document is downloaded only when similarity between the query vector and document vector is matched. If the similarity is matched then the desired secure document is download and decrypted using the AES and Blowfish Encryption.

(a) Original Image

(b) Stag Image

Fig. 5: Steganography Result

- *Data Confidentiality:* To make sure that the document which is to be stored on cloud server can be read only by data user the document is encrypted before storing to cloud. For purpose of encryption of document to make it secure, AES and Blowfish algorithms are used. Both AES and blowfish are part of crypto library in python.
- *Data Integrity:* To ensure integrity of data which means that the data is not tampered during the whole procedure while transfer from the data owner to data user. Message digest was created using SHA-3 which is also a part of Crypto python library.
- *Data Authentication:* For authenticating data user, asymmetric encryption is used which make use of data using a public key to encrypt the message digest and data users private key is used to decrypt the same. RSA is also the part of Crypto python library.
- *Key Transfer:* For transfer of keys and secure message digest between the data owner and data user, LSB steganography is used. Original image Fig. 5a and Stag image Fig. 5b is shown in Fig. 5.

TABLE III: COMPARISON BETWEEN PROPOSED SCHEME AND PREVIOUS SCHEMES

Scheme → / Parameters ↓	Proposed Scheme	Daisuke Aritomo and Chiemi Watanabe [14]	Divya Prathana and Ajit Kumar Santra [20]
Confidentiality	YES	NO	YES
Integrity	YES	NO	YES
Authentication	YES	NO	YES
Key Transfer	YES	NO	NO
Search Attribute	YES	YES	NO

The proposed scheme is compared with two of the existing scheme [14]. [20] Based on the parameters described above that are Data confidentiality, Data integrity, Data Authentication, Key transfer and Searchable encryption, the comparison between schemes is shown in Table III.

The proposed scheme protects the document from malicious users at the time of transmission and also in cloud server storage. The proposed scheme increases the difficulty level for intruders or hacker to decrypt the document through double encryption and RSA. In the scheme, the encrypted document can be searched with multiple keywords which make it convenient and time-efficient.

VII. CONCLUSION

In this paper, Hybrid Searchable Encryption Scheme in Cloud was proposed which is based on using a combination of symmetric and asymmetric encryption while providing a way to generate multi-keyword based search over encrypted data. The proposed scheme makes the following contribution: Data confidentiality of a document is maintained using a combination of two symmetric encryption algorithm that is AES and Blowfish. Data integrity is maintained by creating a message digest of the document which is encrypted using RSA encryption to ensure data authentication. Concept of steganography is used for secure transfer of keys from data owner end to data user end. Existing sense embedding are easily adopted to create a document vector. A scheme which meets requirements in such a way that the cloud server is honest but curious is proposed. A secure search scheme is proposed by creating a vector of the document using sense embedding technique which makes data user search over encrypted data.

REFERENCES

[1] A. Hendre, and K. P. Joshi, "A semantic approach to cloud security and compliance," 2015 *IEEE 8th Int. Conf. Cloud Comput.,* New York, NY, 2015, pp. 1081-1084, doi: 10.1109/CLOUD.2015.157.

[2] N. L. Kodumru, and M. Supriya, "Secure data storage in cloud using cryptographic algorithms," *2018 4th Int. Conf. Comput. Commun. Control and Automat.,* Pune, India, 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697550.

[3] A. Olumide, A. Alsadoon, P. W. C. Prasad, and L. Pham, "A hybrid encryption model for secure cloud computing," *2015 13th Int. Conf. ICT and Knowl. Eng. (ICT and Knowl. Eng., 2015),* Bangkok, 2015, pp. 24-32, doi: 10.1109/ICTKE.2015.7368466.

[4] D. P. Timothy, and A. K. Santra, "A hybrid cryptography algorithm for cloud computing security," *2017 Int. Conf. Microelectronic Devices, Circuits and Syst. (ICMDCS),* Vellore, 2017, pp. 1-5, doi: 10.1109/ICMDCS.2017.8211728.

[5] C. Biswas, U. D. Gupta, and M. M. Haque, "An efficient algorithm for confidentiality, integrity and authentication using hybrid cryptography and steganography," *2019 Int. Conf. Elect., Comput. and Commun. Eng. (ECCE),* Cox'sBazar, Bangladesh, 2019, pp. 1-5, doi: 10.1109/ECACE.2019.8679136.

[6] D. X. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," *Proc. 2000 IEEE Symp. Secur. and Privacy (S&P'2000),* Berkeley, CA, USA, 2000, pp. 44-55, doi: 10.1109/SECPRI.2000.848445.

[7] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: Improved definitions and efficient constructions," In *Proc. 13th ACM Conf. Comput. and Commun. Secur. (CCS'06),* Association for Computing Machinery, New York, NY, USA, 7988, 2006, doi: https://doi.org/10.1145/1180405.1180417.

[8] D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," In C. Cachin, and J. L. Camenisch, Eds., Advances in

Cryptology EUROCRYPT 2004. EUROCRYPT 2004. *Lecture Notes in Computer Science,* vol. 3027, 2004. Springer, Berlin, Heidelberg, 2004. [Online]. Available: https://doi.org/10.1007/978-3-540-24676-3 30

[9] Z. Deng, K. Li, K. Li, and J. Zhou, "A multi-user searchable encryption scheme with keyword authorization in a cloud storage," *Future Generation Computer Systems*, vol. 72, pp. 208-218, Jul. 2017, doi: https://doi.org/10.1016/j.future.2016.05.017.

[10] X. A. Wang, F. Xhafa, W. Cai, J. Ma, and F. Wei, "Efficient privacy preserving predicate encryption with fine-grained searchable capability for cloud storage," *Computers and Electrical Engineering*, vol. 56, pp. 871-883, Nov. 2016.

[11] J. Katz, A. Sahai, and B. Waters, "Predicate encryption supporting disjunctions, polynomial equations, and inner products," In *Annu. Int. Conf. Theory and Appl. Cryptographic Techn. (EUROCRYPT 2008),* 2008, pp. 146-162.

[12] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," In *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 1, pp. 222-233, Jan. 2014, doi: 10.1109/TPDS.2013.45.

[13] W. Sun, B. Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li, "Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking," In *Proc. 8th ACM SIGSAC Symp. Inf., Comput. and Commun. Secur.,* Association for Computing Machinery, New York, NY, USA, 2013, pp. 71-82, doi: https://doi.org/10.1145/2484313.2484322.

[14] D. Aritomo, and C. Watanabe, "Achieving efficient similar document search over encrypted data on the cloud," *2019 IEEE Int. Conf. Smart Comput. (SMARTCOMP),* Washington, DC, USA, 2019, pp. 1-6, doi: 10.1109/SMARTCOMP.2019.00020.

[15] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. arXiv:1301.3781.

[16] S. Rothe, and H. Schütze, "AutoExtend: Extending word embeddings to embeddings for synsets and lexemes," In *Proc. 53rd Annu. Meeting Association Comput. Linguistics and 7th Int. Joint Conf. Natural Lang. Process.,* Beijing, China, Jul. 26-31, 2015, pp. 1793-1803.

[17] A. Neelakantan, J. Shankar, A. Passos, and A. Mccallum, "Efficient non-parametric estimation of multiple embeddings per word in vector space," 2015. 10.3115/v1/D14-1113.

[18] A. Trask, P. Michalak, and J. Liu, "sense2vec - A fast and accurate method for word sense disambiguation in neural word embeddings," 2015. arXiv abs/1511.06388.

[19] I. Iacobacci, Mohd. T. Pilehvar, and R. Navigli, "SensEmbed: Learning sense embeddings for word and relational similarity," In *Proc. 53rd Annu. Meeting Association Comput. Linguistics and 7th Int. Joint Conf. Natural Lang. Process.,* Beijing, China, Jul. 26-31, 2015, pp. 95-105.

[20] D. P. Timothy, and A. K. Santra, "A hybrid cryptography algorithm for cloud computing security," *2017 Int. Conf. Microelectronic Devices, Circuits and Syst. (ICMDCS),* Vellore, 2017, pp. 1-5, doi: 10.1109/ICMDCS.2017.8211728.

[21] Mohd. T. Pilehvar, D. Jurgens, and R. Navigli, "Align, disambiguate and walk: A unified approach for measuring semantic similarity," *ACL 2013 - Proc. 51st Annu. Meeting Association Comput. Linguistics,* vol. 1, 2013, pp. 1341-1351.

International Journal of Distributed and Cloud Computing
9 (1) 2021, 01-12
http://www.publishingindia.com/ijdcc

Study of IoT-Based Digital Technology to Combat COVID-19 Pandemic

Anamika Rana[1*], Monisha Awasthi[2] and Sushma Malik[3]
[1]Assistant Professor, Maharaja Surajmal Institute, Janakpuri, New Delhi, India. Email: anamica.rana@gmail.com
[2]Associate Professor, Uttaranchal University, Dehradun, Uttarakhand, India. Email: monishaawasthi2011@gmail.com
[3]Assistant Professor, Institute of Innovation in Technology and Management, Janakpuri, New Delhi, India.
Email: sushmamalikiitm@gmail.com
*Corresponding Author

Abstract: The epidemic of coronavirus disease which is called COVID-19 is dispersing all over the world in a few months. The traditional health care systems face new challenges associated with the constant increase of infected persons with this disease. In the current time, the integration of smart medical sensors and devices of the Internet of Things (IoT) proposed the new health care domain which is called the Internet of Medical Things (IoMT). With this technology, accuracy, reliability, efficiency, and effectiveness are improved in the health domain. With the combination of cloud computing with IoT architecture, the smart healthcare system is designed which is ableto support real-time applications when Artificial Intelligence (AI) processing on a vast data which is collected by wearable sensors network. The emergence of IoMT has initiated the vital change in assisting the management of diseases, improving the diagnosis of disease, treatment methods and also reducing the cost and errors in healthcare. The implementations of this technology in the health domain not only improve the healthcare quality for patients but also help the healthcare workers. This paper studies the use of IoT solutions throughout this pandemic period in a health care system, starting from early health monitoring solutions from wearable sensors up to a final recovery from COVID-19. IoMT can play an imperative role during this pandemic time and facilitate to stop the spread of the virus, recognize the patient who has been infected with this virus and also help to control the coronavirus infection in real-time.

Keywords: COVID-19, Internet of Medical Things (IoMT), Internet of Things (IoT), Ontology.

I. Introduction

The pandemic of coronavirus disease which is called COVID-19 is spread globally in a few months. This novel virus and disease unknown before the outbreak initiate in Wuhan, China, in December 2019. But now COVID-19 is become a pandemic and affects a number of countries globally. This pandemic has resulted in over 20 439 814 confirmed cases and over 744 385 deaths in 216 countries in the world [1]. COVID-19 infected people will show mild to moderate respiratory problems and these patients do not require any special treatment. Serious illnesses will show in older people and also those who have any respiratory disease and these patients are required special care. The COVID-19 is a communicable disease and it spread with droplets of saliva. Till now, no specific vaccine or treatment is found for this disease. The best method to avoid and minimize the spread of COVID-19 is only to be well knowledgeable of the COVID-19 virus, its causes, and how it spread. Save yourself and loved ones from its infections only by washing hands with soap and by the use of alcohol-based sanitizer. Social distancing is also the main precaution to save ourselves and others from this disease [2].

With the development of communication technology and sensor networks technology, portable and wearable devices have become more popular globally. Wearable devices mean the device that can easily wearable by the person in the form of necklaces, belts, watches, and many more. These devices have sensors with independent computing power and monitor the physiological information of the person like heartbeat, pressure, and breathing, etc., and also monitor the environment characteristics like temperature, humidity, and air quality. Sensors implemented devices are easily connected to smartphones, tablets, and computers with technologies like Bluetooth, Wi-Fi, and the internet. Nowadays IoT devices are widely used in various domains like the health sector, sports, and fitness to monitor the health of a person [3]. The integration of health devices with the other technology which is called the Internet of Things (IoT) develops the new emerging technology which is called the Internet of Medical Things (IoMT). Medical sensors are used to design wearable devices. With these devices, constantly monitor the health of patients in real-time instead of keeping the patients under surveillance in the hospital [4]. By using IoMT, numbers of activities are easily handled by patients, health workers like remotely medical assistance, data capture, and storage for analysis, drug management, in operation tasks, patients tracking, and many more [5]. The adaptation of IoMT is increased and gain popularity in the market. It is the emerging technology of IT in the healthcare area which is used to sense, auto collected data, intelligence, and internet-oriented [6].

IoMT technology is an IoT based solution which helps to design the healthcare systems based on IoT technology that helps to monitor the vital sign of health like ECG, Heart rate and blood pressure. By the implementation of IoMT technology, health domain become e-health which given the remarkable wellness health services globally to change the healthier lifestyle. Due to this technology, healthcare services become user centric and personalized services by provide private healthcare round the clock. The main motive of IoMT enabled healthcare system is to improve the health services and quality of life by avoiding the unpleasant hospitalization. Smartphone play a vital role and act as the gateway in the IoMT edge network. With the help of smartphone, data can be received and forward on the network over the internet to the healthcare service provider for data processing and storage [7]. The numbers of medical devices are connected to produce, store, analyze and also distribute user's health data. Wearable's, remote patients monitors, sensor enabled beds, infusion pumps and tracking devices are the some IoMT devices which are used to collect the data. The main motive of IoMT technology is to provide the satisfaction of patient and also increase the quality of care which given healthcare providers [8]. IoMT technology is the amalgamation of data from medical devices that connect with healthcare IT systems through internet. The intention of IoMT technology is to combine people, data and processes with connected medical devices and mobile applications to improve patient outcomes. This technology allows patients to share health information with doctors for more accurate diagnoses with fewer errors and lower costs of care. In the current situation of COVID-19, this wireless technology plays an important role because of this no need for the patient's medical visit to hospitals which prevents the spread of coronavirus [9].

The infected patients are increasing daily globally and at that time its need to utilize the IoT in a robust, organized manner and in an efficient way to enhance the battle against the COVID-19 pandemic. This technology enables us to capture the real-time data of infected patients and locations remotely. The process involved in the implementation of IoT technology to fight COVID-19 is shown in Fig. 1 [10].

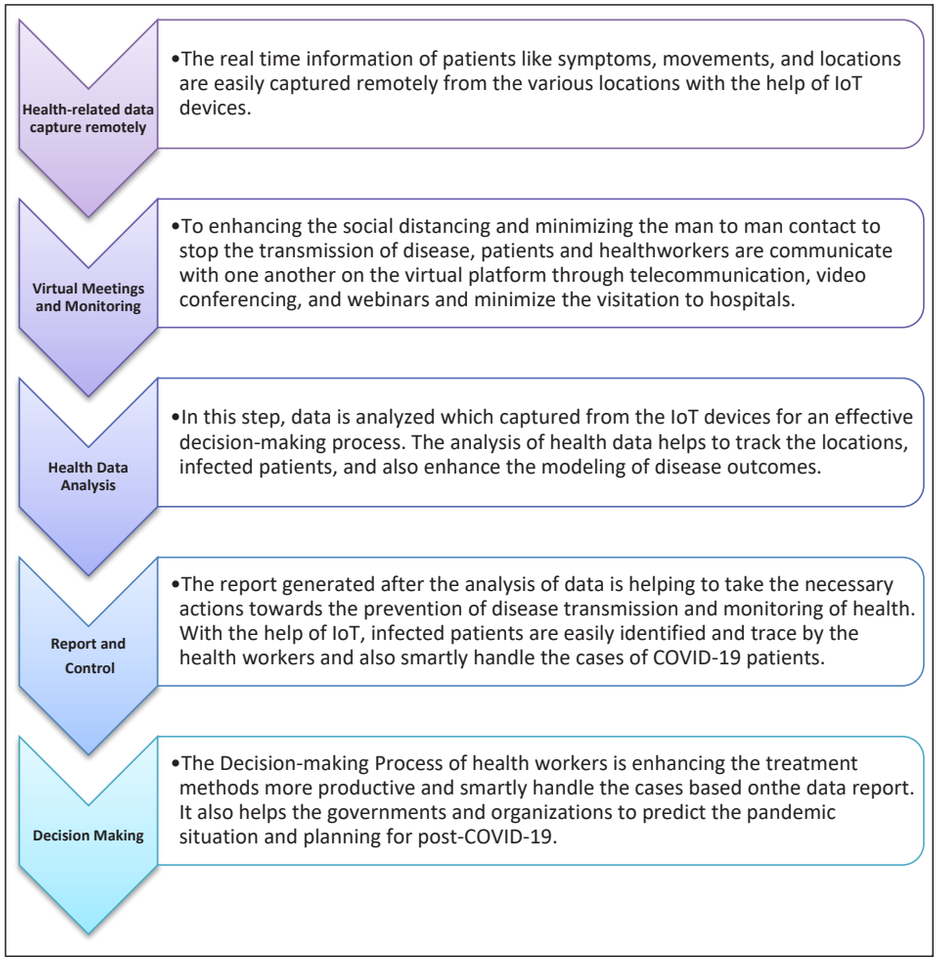


Fig. 1: Process in the Implementation of IoT to Fight with COVID-19

Major Key-Merits of IoT to Battle with COVID-19 Pandemic [11]

The major key merits of IoT technology utilization in the COVID-19 pandemic are the highlight in Fig. 2.

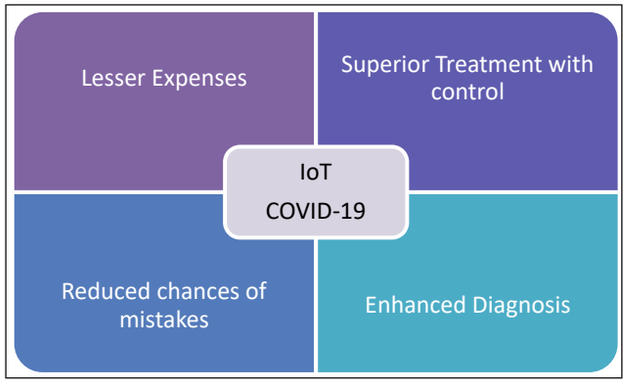


Fig. 2: The Major Key Merits of IoT to Battle with COVID-19 Pandemic

- *Lesser Expenses:* The cost of IoT implements devices is less means it is a one-time investment. The purchased devices are work for a number of years without any problem.
- *Superior Treatment with Control:* With the utilization of real data, IoT implanted devices take the right decision at the right time and increase the treatment process.
- *Reduced Chances of Mistakes:* The chances of mistakes are reduced due to the automation of IoT devices.
- *Enhanced Diagnosis:* By the utilization of IoT devices, the diagnosis process of the disease becomes easier and is able to be diagnosed in the early stage and at that time right treatment can be given to the patient.

II. Literature Review

Sr. No.	Reference	Study
1	[12]	The correct prediction of protein complex plays an essential role to understand the principal mechanism of various cellular functions and elucidates the functionality of different un-annotated proteins. The basic topological properties of a protein are studied by several complex detection algorithms without using advance topological and biological properties and it can be implemented with the association rule learners such as PART and NNGE. These association rules provide better accuracy and low computational cost.
2	[12] [13]	The main aim of the IoMT platform is to provide a better lifetothe patient by providing personalized healthcare services into 24-hour services. Edge network of IoMT has a number of security threats like confidentiality, integrity, authentication, and authorization and they may cause a significant risk to patient's privacy and safety.
3	[14]	The advancement in IT and IoMT has encouraged extensive utilization of the smart system. 24*7 healthcare services are required for effective and reliable monitoring of a patient on the daily basis. To fulfill this requirement, edge computing and cloud platforms are needed for smart healthcare systems.
4	[15]	Proposed a deep learning solution for healthcare platform which automatically identify the type of food given to the patient on the basis of disease and other attributes such as age, gender, weight, calories, protein, fat, sodium, fiber, cholesterol by implementing deep learning algorithms like RNN, MLP, GRU, and LSTM.

Sr. No.	Reference	Study
5.	[8]	Monitoring and prediction systems in the health domain are essential for saving lives mainly when patients are remotely located. By using MSSO-ANFIS, an IoMT based healthcare monitoring system for heart disease prediction is proposed which presents higher values for precision, recall, F1-score, and accuracy with the lowest values for classification error when compared with the existing techniques.
6.	[16] [7]	The extended version of IoT is called IoMT in the healthcare domain. In the current COVID-19 situation, IoMT is used to create a medical platform to provide the proper healthcare at home and remotely without any movement. The novel security method is required to preserve the security of IoMT edge networks.
7.	[17]	Developed a CMMA for cluster head selection to give effective communication to sensors enabled devices of IoMT in the health domain. This technique not only minimizes the energy utilization of edge computing-based IoMT systems but also increases the network lifetime by uniformly distributes cluster heads in the network.
8.	[18] [19]	In today's situation, healthcare systems are digitized with interconnecting of medical devices. Data integration, Privacy, Security, Real-time processing Hardware implementation, and design optimization issues are faced by the developer while implementing IoT technology in the healthcare domain. IoMT systems play a vital role during this pandemic time to track and trace the infected patients.
9.	[20]	In the interconnected e-Health devices of the IoMT network, trust management plays an essential role to provide a smooth and reliable communication process. The trustworthy communications among the devices is a tedious task to handle in a large spread network but its control is based on fuzzy logic named FTM-IoMT.
10.	[5]	A web application named IoMT-SAF is based on novel ontological to ensure security in the IoMT platform. IoMT-SAF contains the assessment attributes list that covers the essential measures of security for the IoMT users to identify and implement security on the basis of user requirements.
11.	[21]	IoMT implementation helps in the identification of brain tumors at an early stage. An association rule learner like Partial Tree with advanced feature sets is adopted to detect brain tumors.
12.	[22]	Proposed a privacy-preserving-based data collection and analysis framework named P2DCA for IoMT applications. It collects the data with mobile nodes from WMSNs which are partitioned into multiple small clusters. Both data and location coordinates of nodes are aggregated on cluster head to hide data sources before share data with mobile nodes. The cloud server performs the analysis with the help of CP-ANN to differentiate the received video into the foreground and background area to identify the moving objects after collecting the data from the mobile nodes.

Sr. No.	Reference	Study
13.	[23]	A smart healthcare system is proposed by using IoMT, Big Data analytics, and by immerging ML algorithms. IoMT technology is used to connect the wearable sensors, patients, healthcare service providers through software and ICT.
14.	[24]	To predict the vertebral tumor in the early stage and for diagnosis treatment, HMRF is proposed. This technique is used the IoMT platform with MRI images to identify the tumor.
15.	[25]	A reliable, dynamic, flexible, faster, and secure network is needed for the IoMT platform. Paper presented the new Hybrid DL-driven SDN-enabled IoMT uncovering framework which is highly usable, capable, and cost-effective.
16.	[26]	For the early detection and classification of malignant cells in breast cancer, integrate the ML and CI-based techniques in E-Health care services on the IoMT platform. The identification of malignant cells is accomplished by extracting shapes and textured-based features. Well-known classification techniques like SVM, NB, and RF are used to classify tumor cells into malignant cells.
17.	[27]	New services in health sectors like remote sensing, assisting of elder people from home, and E-Visit improve the patient's health and convenience by reducing the per-patient cost by the utilization of IoMT. The study represents the main defense mechanism for proving end-to-end privacy and security by developing approaches that protect the patients from the variety of cyber attacks and threats and also provide safeguards to the healthcare domain.
18.	[28] [29]	Presented an IoMT enabled wearable device called iGLU 2.0 for continuous glucose monitoring for diabetic patients this is based on the principle of optical detection. The device is validated on pre-diabetic, diabetic, and healthy patients from persons aged 17-80 years by using real healthcare data. iGLU 2.0 device has a great capability to provide self-care monitoring solutions for smart healthcare.
19.	[30]	Introduced the new blockchain-enabled authentication protocol for IoMT platform called BAKMP-IoMT in the medical field. This protocol provides secure data transfer management between medical devices, personal servers, and cloud servers. Users can also easily access the healthcare data from the cloud server in a secure way because whole healthcare data is maintained by the cloud servers in the blockchain.
20.	[31] [32]	Paper presented that the IoMT enabled devices to help to monitor the elder's people, by in the monitoring of elderly creating digital medical records and generate the alert message for caregivers and family members when IoMT devices detect an unusual situation. The system also produces simple and secure access toreal-time data.

Sr. No.	Reference	Study
21.	[33]	The integration of IoMT, cloud storage, and Blockchain technologies in the medical area is playing a vital role. It reduces the cost, increase speed to access data, automation, remove mediators, centralized data and also overcome the security of data issue by implementing the latest encryptions method. Introduced the new BC IoMT U6 HCS (BlockChain based Internet of Medical Things for Uninterrupted, Ubiquitous, User-friendly, Unflappable, Unblemished, Unlimited Health Care Services) layered architecture.
22.	[34]	IoMT and smart gadgets generated a large amount of data. Drift is the major challenge in the process of the huge volume of data streams. To solve this situation, the ERDDM method is introduced for the detection of drifts in data streams generated by sensors and smart devices in the medical domain. This method enhances the RDDM by minimizing the error rate.
23.	[35]	Introduced the MELINDIA architecture in the Intelligent Surveillance System, by deployed in a smart building or in a smart city with a number of cameras that generate vast valuable information. This architecture is used to produce useful data from the raw data stream by reducing the dimensions of data. It also achieves low latency by analyzing the data near the sensors and also prevents the redundant of data by reusing the data.
24.	[36]	Presented an E-ADS with the utilization of fog-cloud architecture to locate and mitigate malicious activities in the IoMT environment. It using two engines, like a traffic processing engine and an intrusion detection engine.
25.	[37]	Introduced the new real-time epileptic seizure detection model using the ordinary kriging method in an edge computing paradigm.
26.	[38]	IoMT technology is employed in the fast and safe detection of leukemia with the utilization of cloud computing. Clinical gadgets are connected with a network and the system provides real-time data for testing, diagnosis, and treatment of leukemia. It saves the time and efforts of patients and healthcare professionals.
27.	[39]	The proposed iLog model provides the monitoring systems by analyzing the eating behavior of the user and differentiate the behavior into stress eating or normal eating. This model using the edge platform, a mobile and SBC platform. Foods items are automatically detected and classified from the images and then compared with a stored database of nutrition in Firebase.

III. IoMT Architecture

IoT is the new emerging technology that effectively works in various domains. In healthcare sectors, IoT is adopted and use for efficient diagnosis and treatment of health. Medical devices are integrated with IoT technology for effective treatment and diagnosis processes in the healthcare environment. Services of IoT-based healthcare systems named IoMT are to assist the monitoring of patient data, taking decisions in emergency, reduce process cost, and increasing the facility in the medical field [40]. Fig. 3 explains the ontology-based architecture of IoMT.

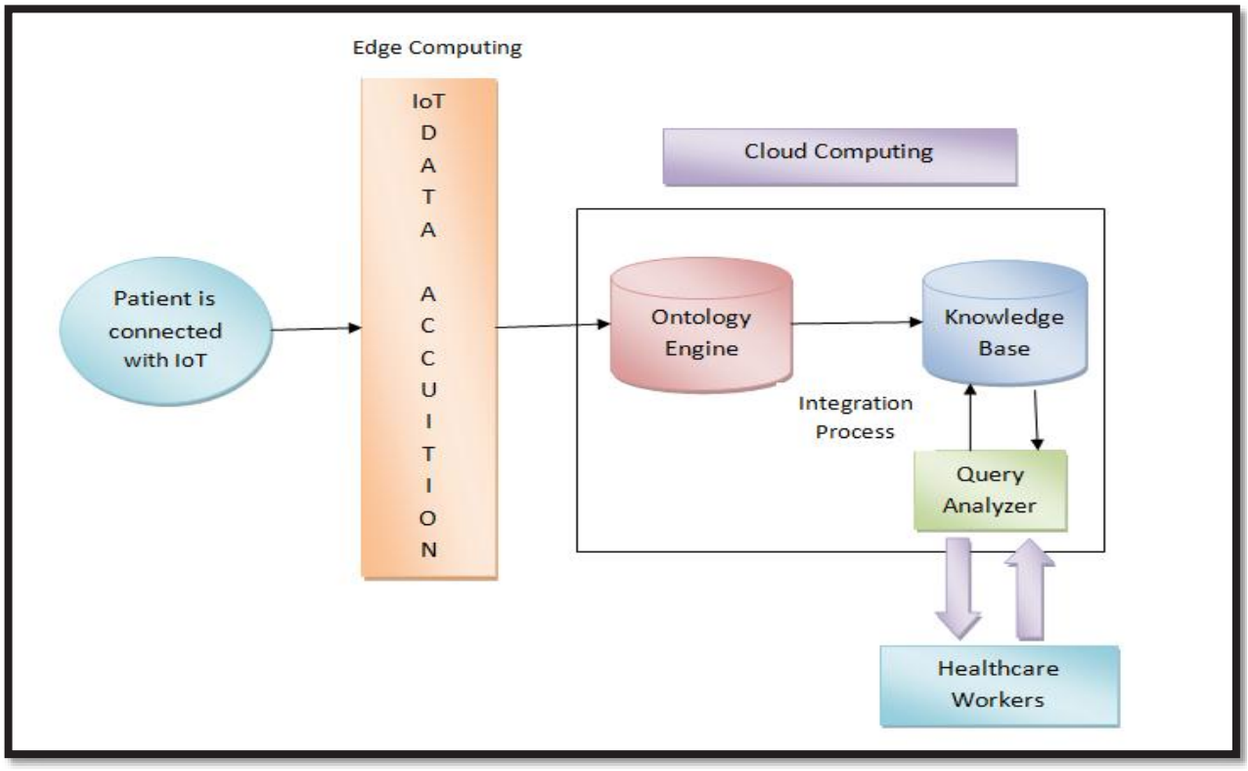


Fig. 3: IoMT Architecture with Ontology

- *Sensor Devices:* This layer consists of physical sensor devices such as medical equipment sensors based used during all activities involved in treatment and further assistance. All these equipment are capable of communicating through RFID, EPC, ZigBee, 4G-5G, Bluetooth, etc. for data transferring.
- *Edge Layer:* Edge computing has an important role in this architecture. After collecting information from the sensors, it will be stored on the cloud via the edge layer because the edge will reduce the amount of data in motion. Edge computing reducing the dependency on a remote centralized server. Due to these healthcare workers would be able to have a more agile and responsive health care system from the IT network. Now health workers don't need to wait for their data to be sent to a central repository for processing and then returned to their hospital for use and patient treatment, which is a necessity for this pandemic period.
- *Storage Layer (Cloud):* All reduced and preprocess data will be stored in a related database on servers located elsewhere. These servers are located on the

cloud and to realize the vision of the usage of IoT in medical services high-performance computing and Big Data analytics will play an important role.

- *Ontology:* During the pandemic time, fast and precise knowledge has an effective role in the management. Any kind of delay and imprecise knowledge may lead to a poor decision and detrimental action, resulting in increased mortality ratio as well as intensification. Ontology-based systems offer the advantage of flexibility and semantic interoperability for knowledge sharing in such conditions. Ontology is the framework for representing shareable and reusable knowledge across a domain. And due to this ontology have a great opportunity for performing productivity result in this pandemic time. At the time creating ontology following phase is needed in the required area.

IV. Utilization of IoT during COVID-19 Pandemic

During this pandemic time, countries of the world are integrating their all resources and try to use the new technologies to minimize the effect of COVID-19. In this situation, AI and IoT technologies are used to stop the spread of the virus and also control the infection in real-time. In the present situation, healthcare technologies have pushed themselves to innovate new technologies which become more useful to handle that situation. The IoT technology plays anessential role to stop the spread of coronavirus and also assists healthcare workers in the treatment of patients. The major application of IoT during the COVID-19 pandemic is as follows [11] [10]:

Sr. No.	Application	Description
1	Virtual consultation	Patients get a consultation from health care workers remotely with the help of connected devices.
2	Internet-connected hospital	The whole hospital premises with the implementation of IoT devices become the integrated network to support the COVID-19 pandemic.
3	Inform the concerned medical staff during emergency	With the accomplishment of an integrated network, the patients and health workers easily communicate in less time and an effective manner.
4	Rapid screening of COVID-19	The proper diagnosis is implemented with smart integrated treatment devices. With this technology, the overall screening process becomes very fast.
5	Treatment of COVID-19 become transparent	Patients get the benefit without any partiality.
6	Treatment process becomes automated	With the help of this technology, treatment process selection become more productive and assists to easily handle the cases due to automation.
7	The smartphone app and smart tracing of an infected person through wireless healthcare network	Numbers of authentic applications are available on the digital platform. An application is easily installed on Smartphone and makes the identification process of an infected person smoother and fruitful.
8	Real-time update	With the interconnectivity of devices through the internet, real-time data will be shared and help to decide on that basis and cases can handle accurately.
9	Accurate forecasting of virus	Based on available data reports, predict the situation of things in the coming time with the help of some statistical method. It is also used in the proper decision-making process and also guides the governments, health workers, etc to plan for a better working environment.
10	Track quarantine	The critical step to control the spread of COVID-19 is effectively quarantined by the infected person. But it is easier to say than implementation. IoT and GPS enabled applications are used to track the movement of an infected person and also help to take the necessary actions to restrict the movement of such a person [41] [42].
11	Cleaning and disinfecting	Cleaning, sanitizing, and disinfecting of medical facilities and all around are essential and also destroy the infectious nature of the COVID-19 virus. To accomplished this task, self-driving robots are used which develop by several companies like TMiRob, UVD, and Xenex Disinfection Services. These robots are disinfecting the surfaces with high-intensity UV light and kill the virus by destroying their DNA. Robots are Wi-Fi based and easily controlled by apps [42].
12	Reducing in-home infections	With the increase of awareness among the people, they avoid touching the surfaces like doorknobs, switches, etc, instead of it, people use the IoT enabled devices like speakers, security system to open the doors and switch on the light [42].

V. Various Issues and Future Scope of the Study

IoT technology has an essential role in the medical field but it faces a number of challenges that are listed below in Fig. 4. But with these challenges, IoT technology is mostly used in this domain. In the future, IoT implementation and usage become more useful. Data storage in acceptable format and management of data are should be targeted in future work with the development of cost-effective and affordable applications.

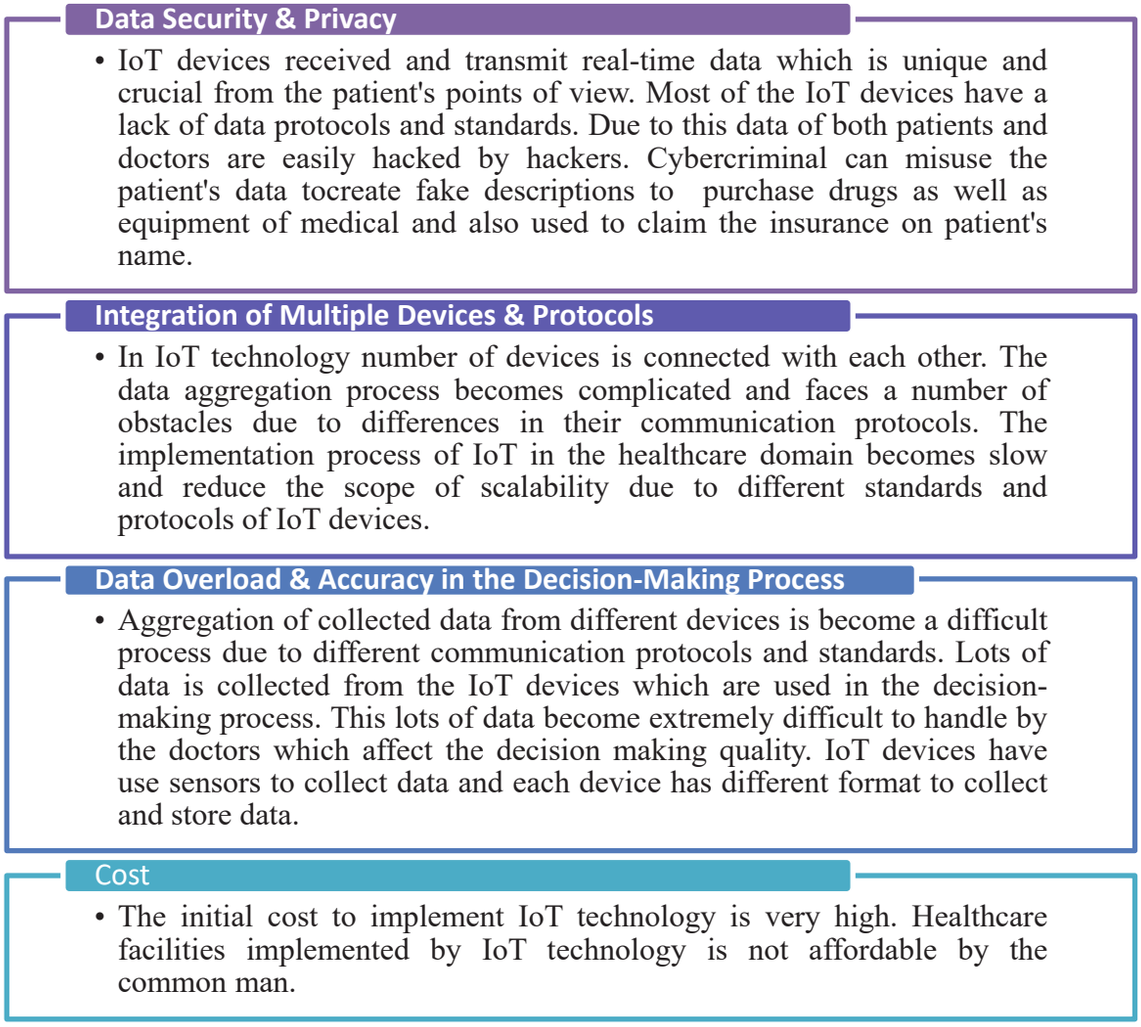


Fig. 4: Various Issues Faced by IoT during the Implementation in Health Domain [43] [44] [45]

VI. Conclusion

The internet has changed our day-to-day activities and also modified the way of living. Now worldwide, the health domain has recognized the significance of ICTs in the functioning of healthcare. IoT technology makes our living easier and pleasant with virtual interactions. The innovative use of IoT in the healthcare sector not only brings benefits to doctors, nurses, and administrative staff of the hospital as well as patients during this COVID-19 pandemic. During this pandemic time, only social distancing and precautions. IoT devices help to maintain social distancing and help in the battle with COVID-19. In this paper; we studied the influence of IoT in the health sector and act as the Internet of Medical Things (IoMT) because all the medical devices are connected with each other with the internet. Things are more and more integrated within the human body and they will communicate with healthcare and emergency services. We also represent the ontology-based architecture of IoMT. This technology plays a vital role in the health domain but it also faces challenges like accessing heterogeneous IoT data in the real environment and also creates the storage of a huge amount of IoT data accessing. We believe that IoT will greatly influence the health domain in the coming years by providing powerful applications.

References

[1] WHO, "WHO Coronavirus Disease (COVID-19) Dashboard | WHO Coronavirus Disease (COVID-19) Dashboard," 2020. Accessed: Aug. 14, 2020.

[2] WHO, "Coronavirus," 2020. Accessed: Aug. 14, 2020. [Online]. Available: https://www.who.int/health-topics/coronavirus#tab=tab_1

[3] K. Wei, L. Zhang, Y. Guo, and X. Jiang, "Health monitoring based on internet of medical things: Architecture, enabling technologies, and applications," *IEEE Access*, vol. 8, pp. 27468-27478, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.2971654

[4] J. P. A. Yaacoub, M. Noura, H. N. Noura, O. Salman, E. Yaacoub, R. Couturier, and A. Chehab, "Securing internet of medical things systems: Limitations, issues and recommendations," *Future Generation Computer Systems*, vol. 105, pp. 581-606, 2020. [Online]. Available: https://doi.org/10.1016/j.future.2019.12.028

[5] F. Alsubaei, A. Abuhussein, V. Shandilya, and S. Shiva, "IoMT-SAF: Internet of medical things security assessment framework," *Internet of Things*, vol. 8, 2019, Art. no. 100123. [Online]. Available: https://doi.org/10.1016/j.iot.2019.100123

[6] M. Irfan, and N. Ahmad, "Internet of medical things: Architectural model, motivational factors and impediments," *2018 15th Learning and Technology Conference (L and T 2018)*, 2018, pp. 6-13. [Online]. Available: https://doi.org/10.1109/LT.2018.8368495

[7] M. Papaioannou, M. Karageorgou, G. Mantas, V. Sucasas, I. Essop, J. Rodriguez, and D. Lymberopoulos, "A survey on security threats and countermeasures in internet of medical things (IoMT)," *Transactions on Emerging Telecommunications Technologies*, pp. 1-15, 2020. [Online]. Available: https://doi.org/10.1002/ett.4049

[8] M. A. Khan, and F. Algarni, "A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS," *IEEE Access*, vol. 8, pp. 122259-122269, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.3006424

[9] E. Cook, "The importance of IoMT in healthcare 2020," Digital Healthcare | Healthcare Global, Jul. 2020. [Online]. Available: https://www.healthcareglobal.com/digital-healthcare/importance-iomt-healthcare-2020

[10] J. Oyeniyi, I. Ogundoyin, O. Oyeniran, and L. Omotosho, "Application of internet of things (IoT) to enhance the fight against COVID-19 pandemic," Jun. 2020.

[11] R. P. Singh, M. Javaid, A. Haleem, and R. Suman, "Internet of things (IoT) applications to fight against COVID-19 pandemic," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 14, no. 4, pp. 521-524, 2020. [Online]. Available: https://doi.org/10.1016/j.dsx.2020.04.041

[12] M. Sikarndar, W. Anwar, A. Almogren, I. U. Din, and N. Guizani, "IoMT-based association rule mining for the prediction of human protein complexes," *IEEE Access*, vol. 8, pp. 6226-6237, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2963797

[13] D. Koutras, G. Stergiopoulos, T. Dasaklis, P. Kotzanikolaou, D. Glynos, and C. Douligeris, "Security in IoMT communications: A survey," *Sensors (Switzerland)*, vol. 20, no. 17, pp. 1-49, 2020. [Online]. Available: https://doi.org/10.3390/s20174828

[14] I. V. Pustokhina, D. A. Pustokhin, D. Gupta, A. Khanna, and K. Shankar, "An effective training scheme for deep neural network in edge computing enabled internet of medical things (IoMT) systems," *IEEE Access*, vol. 8, pp. 107112-107123, 2020.

[15] C. Iwendi, S. Khan, J. H. Anajemba, A. L. I. K. Bashir, and F. Noor, "Realizing an efficient IoMT-assisted patient diet recommendation system through machine learning model," *IEEE Access*, vol. 8, pp. 1-1, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.2968537

[16] T. Yang, M. Gentile, C. F. Shen, and C. M. Cheng, "Combining point-of-care diagnostics and internet of medical things (IOMT) to combat the Covid-19 pandemic," *Diagnostics*, vol. 10, no. 4, pp. 4-6, 2020. [Online]. Available: https://doi.org/10.3390/diagnostics10040224

[17] T. Han, L. Zhang, S. Pirbhulal, W. Wu, V. Hugo, and C. de Albuquerque, "A novel cluster head selection technique for edge-computing based IoMT systems," *Computer Networks*, vol. 158, pp. 114-122, 2019. [Online]. Available: https://doi.org/10.1016/j.comnet.2019.04.021

[18] G. J. Joyia, R. M. Liaqat, A. Farooq, and S. Rehman, "Internet of medical things (IOMT): Applications, benefits and future challenges in healthcare domain," *Journal of Communications*, vol. 12, no. 4, pp. 240-247, 2017. [Online]. Available: https://doi.org/10.12720/jcm.12.4.240-247

[19] A. Hafizah, M. Aman, W. Haslina, and S. Sameen, "IoMT amid COVID-19 pandemic: Application, architecture, technology, and security," *Journal of Network and Computer Applications*, vol. 174, 2021, Art. no. 102886. [Online]. Available: https://doi.org/10.1016/j.jnca.2020.102886

[20] A. Almogren, I. Mohiuddin, I. U. Din, H. Almajed, and N. Guizani, "FTM-IoMT: Fuzzy-based trust management for preventing sybil attacks in internet of medical things," *IEEE Internet of Things Journal*, vol. 8, no. 6,

pp. 4485-4497, 2021. [Online]. Available: https://doi.org/10.1109/JIOT.2020.3027440

[21] S. Rukh, M. Sikandar, A. Almogren, and I. Ud, "IoMT-based computational approach for detecting brain tumor," *Future Generation Computer Systems*, vol. 109, pp. 360-367, 2020. [Online]. Available: https://doi.org/10.1016/j.future.2020.03.054

[22] M. Usman, M. A. Jan, X. He, and J. Chen, "P2DCA: A privacy-preserving-based data collection and analysis framework for IoMT applications," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1222-1230, 2019. [Online]. Available: https://doi.org/10.1109/JSAC.2019.2904349

[23] L. Syed, S. Jabeen, S. Manimala, and A. Alsaeedi, "Smart healthcare framework for ambient assisted living using IoMT and big data analytics techniques," *Future Generation Computer Systems*, vol. 101, pp. 136-151, 2019. [Online]. Available: https://doi.org/10.1016/j.future.2019.06.004

[24] A. Alsiddiky, W. Awwad, K. Bakarman, H. Fouad, and N. M. Mahmoud, "Magnetic resonance imaging evaluation of vertebral tumor prediction using hierarchical hidden Markov random field model on internet of medical things (IOMT) platform," *Measurement: Journal of the International Measurement Confederation*, vol. 159, 2020, Art. no. 107772. [Online]. Available: https://doi.org/10.1016/j.measurement.2020.107772

[25] S. Liaqat, A. Akhunzada, F. S. Shaikh, A. Giannetsos, and M. A. Jan, "SDN orchestration to combat evolving cyber threats in internet of medical things (IoMT)," *Computer Communications*, vol. 160, pp. 697-705, 2020. [Online]. Available: https://doi.org/10.1016/j.comcom.2020.07.006

[26] S. U. Khan, N. Islam, Z. Jan, I. U. Din, A. Khan, and Y. Faheem, "An e-health care services framework for the detection and classification of breast cancer in breast cytology images as an IoMT application," *Future Generation Computer Systems*, vol. 98, pp. 286-296, 2019. [Online]. Available: https://doi.org/10.1016/j.future.2019.01.033

[27] G. Hatzivasilis, O. Soultatos, S. Ioannidis, C. Verikoukis, G. Demetriou, and C. Tsatsoulis, "Review of security and privacy for the internet of medical things (IoMT): Resolving the protection concerns for the novel circular economy bioinformatics," *Proceedings - 15th Annual International Conference on Distributed Computing in Sensor Systems (DCOSS 2019)*, 2019, pp. 457-464. [Online]. Available: https://doi.org/10.1109/DCOSS.2019.00091

[28] A. M. Joshi, P. Jain, S. P. Mohanty, and N. Agrawal, "iGLU 2.0: A new wearable for accurate non-invasive continuous serum glucose measurement in IoMT framework," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 4, pp. 327-335, 2020. [Online]. Available: https://doi.org/10.1109/TCE.2020.3011966

[29] A. M. Joshi, P. Jain, and S. P. Mohanty, "Secure-iGLU: A secure device for noninvasive glucose measurement and automatic insulin delivery in IoMT framework," *Proceedings of IEEE Computer Society Annual Symposium on VLSI, ISVLSI*, Jul. 2020, pp. 440-445. [Online]. Available: https://doi.org/10.1109/ISVLSI49217.2020.00-17

[30] N. Garg, M. Wazid, A. K. Das, D. P. Singh, J. J. P. C. Rodrigues, and Y. Park, "BAKMP-IoMT: Design of blockchain enabled authenticated key management protocol for internet of medical things deployment," *IEEE Access*, vol. 8, pp. 95956-95977, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.2995917

[31] P. C. Santana-Mancilla, L. E. Anido-Rifón, J. Contreras-Castillo, and R. Buenrostro-Mariscal, "Heuristic evaluation of an IoMT system for remote health monitoring in senior care," *International Journal of Environmental Research and Public Health*, vol. 17, no. 5, 2020. [Online]. Available: https://doi.org/10.3390/

[32] C. Kotronis, I. Routis, E. Politi, M. Nikolaidou, G. Dimitrakopoulos, D. Anagnostopoulos, A. Amira, F. Bensaali, and H. Djelouat, "Evaluating internet of medical things (IoMT)-based systems from a human-centric perspective," *Internet of Things*, vol. 8, 2019, Art. no. 100125. [Online]. Available: https://doi.org/10.1016/j.iot.2019.100125

[33] J. Indumathi, A. Shankar, M. R. Ghalib, J. Gitanjali, Q. Hua, Z. Wen, and X. Qi, "Block chain based internet of medical things for uninterrupted, ubiquitous, user-friendly, unflappable, unblemished, unlimited health care services (BC IoMT U6 HCS)," *IEEE Access*, vol. 8, pp. 1-1, 2020. [Online]. Available: https://doi.org/10.1109/access.2020.3040240

[34] A. A. Toor, M. Usman, F. Younas, A. C. M. Fong, S. A. Khan, and S. Fong, "Mining massive e-health data streams for IoMT enabled healthcare systems," *Sensors (Switzerland)*, vol. 20, no. 7, pp. 1-24, 2020. [Online]. Available: https://doi.org/10.3390/s20072131

[35] A. R. Neto, T. P. Silva, T. V. Batista, F. C. Delicato, P. F. Pires, and F. Lopes, "An architecture for distributed video stream processing in IoMT systems," *Open Journal of Internet of Things (OJIOT)*, vol. 6, no. 1, pp. 89-104, 2020.

[36] P. Kumar, G. P. Gupta, and R. Tripathi, "An ensemble learning and fog-cloud architecture-driven cyber-attack detection framework for IoMT networks," *Computer Communications*, vol. 166, pp. 110-124, 2021. [Online]. Available: https://doi.org/10.1016/j.comcom.2020.12.003

[37] I. L. Olokodana, S. P. Mohanty, E. Kougianos, and O. O. Olokodana, "Real-time automatic seizure detection using ordinary kriging method in an edge-IoMT computing paradigm," *SN Computer Science*, vol. 1, no. 5, 2020. [Online]. Available: https://doi.org/10.1007/s42979-020-00272-2

[38] N. Bibi, M. Sikandar, I. U. Din, A. Almogren, and S. Ali, "IOMT-based automated detection and classification of leukemia using deep learning," *Journal of Healthcare Engineering*, vol. 2020, 2020, Art. id. 6648574. [Online]. Available: https://doi.org/10.1155/2020/6648574

[39] L. Rachakonda, S. P. Mohanty, and E. Kougianos, "ILog: An intelligent device for automatic food intake monitoring and stress detection in the IoMT," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 2, pp. 115-124, 2020. [Online]. Available: https://doi.org/10.1109/TCE.2020.2976006

[40] A. Alamri, "Ontology middleware for integration of IoT healthcare information systems in EHR systems," *Computers*, vol. 7, no. 4, pp. 1-15, 2018. [Online]. Available: https://doi.org/10.3390/computers7040051

[41] R. Shaw, Y. Kim, and J. Hua, "Governance, technology and citizen behavior in pandemic: Lessons from COVID-19 in East Asia," *Progress in Disaster Science*, vol. 6, 2020, Art. no. 100090. [Online]. Available: https://doi.org/10.1016/j.pdisas.2020.100090

[42] S. Mishra, "The growing role of IoT in COVID-19 response," May 2020. [Online]. Available: https://www.iotforall.com/the-growing-role-of-iot-in-covid-19-response/

[43] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, "Artificial Intelligence (AI) applications for COVID-19 pandemic," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 4, pp. 337-339, 2020, doi: 10.1016/j.dsx.2020.04.012.

[44] R. Basatneh, B. Najafi, and D. G. Armstrong, "Health sensors, smart home devices, and the internet of medical things: An opportunity for dramatic improvement in care for the lower extremity complications of diabetes," *J. Diabetes Sci. Technol.*, vol. 12, no. 3, pp. 577-586, 2018, doi: 10.1177/1932296818768618.

[45] L. Greco, G. Percannella, P. Ritrovato, F. Tortorella, and M. Vento, "Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information," Jan. 2020.

International Journal of Distributed and Cloud Computing
9 (1) 2021, 13-17
http://www.publishingindia.com/ijdcc

Comparative Study of Cloud Computing and Edge Computing: Three Level Architecture Models and Security Challenges

Abhay Verma[1*] and Vinay Verma[2]
[1]Student, Department of Computer Science & Engineering, Vivekananda Institute of Technology, Jaipur, Rajasthan, India. Email: verma.abhay2420@gmail.com
[2]Independent Researcher, Jaipur, Rajasthan, India. Email: vnverma03@gmail.com
*Corresponding Author

Abstract: The procreation of IoT (Internet of Things), mobile internet and the attainment of cloud computing services have innovated a new computing archetype, known as edge computing. According to the International Data Corporation (IDC) forecast, worldwide data will reach about 180 zettabytes (ZB) and out of which 70% of data generated by IoT devices will be processed on the edge of the network by 2025. IDC also forecasts the IoT devices connected to reach 150 billion by 2025. Edge computing has the ability to deal with the huge volume of data and to handle the requirement of response time, data handling, bandwidth cost saving, along with privacy and data security. Edge computing enables low latency, mobility, and real-time data processing at a very faster pace. This paper is meant to introduce with the notion of edge computing, the three-level architecture of the edge model with the detailed difference between the edge computing model and the traditional cloud model. The various edge models that enable processing at the edge includes mobile edge computing, fog computing and cloudlet computing. Possessing the various advantages and models, edge computing still suffers from various security challenges. The paper concludes stating the security challenges that edge computing still need to focus upon.

Keywords: Cloud architecture, Cloud computing, Edge computing, IoT.

I. Introduction

Edge Computing is the networking philosophy which mainly focuses on bringing the computing power as closest as possible to the source of data which will enable fast processing with low latency and bandwidth utilization. It is the "Distributed Computing" epitome that conveys the computation power and data storage nearer to the user location [1]. It has been gathered and required, thus preventing the time delay taken in to communicate with the server or cloud, and hence it improves response time and saves bandwidth. *Edge* can be defined as any network or computational resources in-between the source of data and the cloud data center. The edge is generally located one hop (network segment) away from the associated end devices, which are the source of data generated [2]. Due to the nearby availability of the computing power, it is idyllic for low-latency offload infrastructure which supports evolving applications in augmented reality, public safety, connected vehicles, autonomous driving, smart manufacturing, and healthcare.

Edge Computing is combination of numerous technologies, including Cloud Computing, IoT and Grid Computing. Edge Computing reduces the latency time and enhances the quality of service for the real-time processing tasks, including data processing, video processing, artificial intelligence, and time-sensitive. The edge undergoes two-way computing stream- one is upstream "from devices to the cloud" and other is downstream "from the cloud to devices" [2]. Edge has ability to execute the computation offloading, data storing, caching, processing, and also distributes the requests and delivers the services from cloud to the user, which enables edge to act in the real-time data processing. The devices do not restrict request services and content only, but also can perform computing task from the cloud which is possible in edge computing.

A vital part of edge computing is the network edge. Network edge is where the devices, or the local network containing the device, communicate with the internet. The important takeaway is that the network's edge is geographically close to the device, unlike cloud servers, which might be far away from the source.

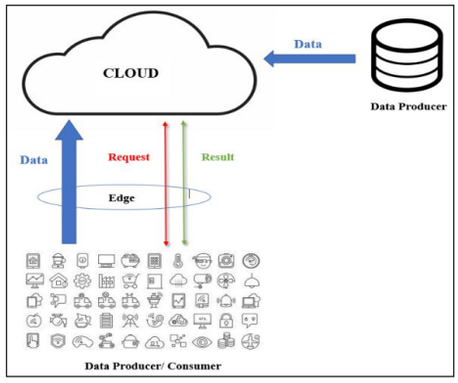


Fig. 1: Edge Computing

Edge computing has been emerging out as a new approach towards the network architecture with aim to reduce the drawbacks of the traditional cloud computing architecture. The growth of the edge, the volume of the usage has been gradually increased since 2010 and has a major impact in the technology since 2016.

II. Literature Review

This section of paper includes various comparisons of the different technologies, like IoT, edge, cloud computing. In the research paper [3], the trends of edge computing, cloud computing, fog computing and Internet of Things is analyzed on basis of scientific study. It concluded that it might be difficult to fully understand the main differences and similarities among the technology paradigms.

Another research paper [4], compared the edge and cloud computing keeping in mind the performance, data filtering, etc. and concluded that the both paradigms are likely to follow the hybrid approach and complement to each other, and suggested the real-time future study needed in the industries like manufacturing, mining, transportation to detect anomalies and send alerts.

Also, the research paper [5] made a comparison among the fog computing, cloudlet, and mobile-edge computing on theoretical model consists of node devices, context awareness, access mechanisms, etc., concluding that these have a different set of characteristics which set them apart from each other, but lacks standardization.

III. Three-Level Edge Computing Model

The three level edge computing model entails the cloud, edge and the IoT. The first level is IoT (Internet of Things) [6], which includes the end devices, like sensors, drones, mobile phones, smart watches connected in health area, smart homes, industrial internet, etc. The second level is the edge [7], which is connected to the first level using various protocols. For example, drones can be connected with 4G/LTE, and sensors can communicate with home Wi-Fi, whereas the smart watch and health devices can communicate via mobile internet. The edge here processes the data and then accordingly save it in the cloud or sends the processed data back to the user. It is capable of storing and computing the data at its end without sending and receiving from the cloud. The edge actually acts as an interface to connect the devices with the cloud with power to process data [8]. For the complex task, it utilizes large computing and data storage capabilities of the cloud. The third level is the cloud which is accountable for offering services to store, process, and analyze data.

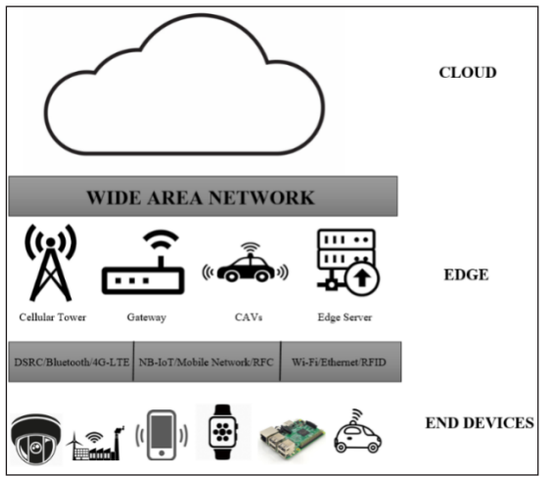


Fig. 2: Three-Level Architecture Model of Edge Computing

IV. Edge vs. Cloud

When the two terms are compared, the main difference arises over the data processing techniques. In cloud computing architecture, all the data captured and generated by the existing IoTs are processed in the cloud by the series centralized servers. While edge computing differs in way as it follows the phenomena of moving processes away from the cloud and bring them near to the end devices, which are closer to the source of origin. By now in 2020 about 45% of the worldwide data which is stored and managed in the network's edge [9]. The major drawback that the cloud faces over the edge is the amount of data processed per second is not passably supported by the cloud. Other than the issue of latency, the cloud faces the issues of wastage of resources as well. Also, when data is piled up and transferred to the cloud, it overloads the network and results in latency and high bandwidth usage. It rises challenges for the cloud-based data processing system. In the case of edge computing, the device nodes process out the data and utilize the underlying hardware of the device only, which does make use of all available resources without need of extra ones [10]. Edge Computing not only helps in minimizing the data dependency over the app or any service, but also helps in speeding up of the processing. The various different features amidedge computing and traditional cloud computing are given in Table I.

Since, edge computing is still in the growing period, whereas the cloud environment is a developed one. The theoretical comparative study provides the blurry differences among both the environments, due to which it becomes a bit difficult to get a concrete picture of both paradigms. Both the technologies are complementary to each other, instead of the substitution. One is supporting other and similarly getting the benefits out of it.

Table I: Comparison between Edge Computing and Cloud Computing

Features	Edge Computing	Cloud Computing
Latency	Highly decrease latency as the processing power is near to the source of data.	Highly increased latency as the cloud is situated at far from the source of data.
Computational capacity	Medium to low	High
Overloading	Prevents overloading of network and data center.	The resources might overload due to transfer of vast amount of data at a time.
Size and operation mode	It is smaller in size and placed over many locations.	Servers are very large in size and centralized.
Real-time data processing	Capable in processing real-time data more efficiently than cloud.	It can a longer time to process data in real-time.
Security	It demands a robust security plan that includes advanced authentication methods, and proactive tackling attacks.	It demands a less robust security plan.
Programming	Different multiple platforms can be used for programming, having different runtime.	It is made of one target platform and uses single programming language, which is best suited in cloud.
Suitable application	Suitable for operations with intense latency concerns.	Suitable for projects which deals with massive data storage.
Data handling	It has limited capacity, and cannot handle huge set of data.	It is efficient enough to store and process enormous amount of data.
Fronthaul/backhaul communication overhead	Low, since devices can get cached content directly from edge gateway.	High, since devices need to be connected to internet throughout entire duration.
Deployment	Leeway of ad-hoc deployment with no or minimal planning.	Require complicated deployment planning.

V. Types of Edge Computing Environment

Edge offers a lithe, decentralized architecture which ensures that everything is processed in the end device itself. There are mainly two types of edge computing model:

- *Cloud Edge:* In case of cloud edge, the public cloud is protracted to a series of point-of-presence (PoP) location.
- *Device Edge:* In device edge, the applications run on the local existing hardware.

The major difference that exists between them is that the cloud edge is stretched form of the traditional cloud, where the cloud provider is responsible for operation and maintenance of the whole model, while device edge exists primarily within the hardware, which enables real-time processing of data at the end devices which is very fast and accurate.

Edge computing employs three technologies:

- *Fog Computing:* Cisco has introduced the term of fog computing in 2012 and described it as a highly simulated computing environment for the migration of cloud computing data center's task to network edge devices [11]. The fog computing and edge computing techniques has many similarities, while it mainly focuses more on augmenting communication at the infrastructure level. The action involves machine-to-machine (M2M) communication or human-computer interaction (HCI). Edge computing pays attention to the computing requirements and network demand for both the end devices and infrastructure [12], which includes the association-between the end devices, edge and the cloud.
- *Mobile Edge Computing (MEG):* It is network architecture that supports the potentials of cloud computing and an IT provision environment at the users' cellular network. It mainly focuses on running services and executing processing tasks on the customers' cellular device, which reduces congestion and application performs better [13]. MEC underlines the foundation between the edge servers, cloud servers and among various edge devices to perform computing tasks, which are similar to the architecture, and regarded as an important role of edge computing.

- *Cloudlet Computing:* Cloudlet computing majorly focus on supporting resource-intensive and interactive mobile-device application by offering powerful computing resources to mobile devices [14]. It is a structural element that extends current cloud computing infrastructure by enabling mobility-enhanced minimal cloud datacenter for offloading of data that is located at the edge of network.

The three models are contributing in the development of edge computing on the various domains of the society [14], with a single motive to bring the computational power at the every smallest IoT devices which are in daily and regular use by the people.

VI. Edge Security Challenges

Due to diverse benefits and features of edge computing, like heterogeneous distributed architecture, support of massive data processing, aided parallel computing, associated low latency, location-awareness, mobile support, but the traditional data security methods [15], and privacy-safeguarding mechanism are no longer suitable for protecting massive data in edge computing. *76% of risk specialists think IoT leaves them at risk of cyber-attacks* [16]. IoT, being most vulnerable possess risk to the edge as well. The data security and privacy-safeguarding mechanisms in edge computing technique mainly challenged by the following new challenges:

- *Lightweight and Fine-Grained:* The new necessities for lightweight data encryption and decryption methods and the fine-grained data sharing based on multiple authorized parties in edge computing.
- *Distributed Identity and Access Control:* Multi-sourced assorted data distribution control and secure the data management issues in distributed computing.
- *Resource-Constrained Devices:* Security challenges that exists between the large-scale edge-based services and resource-constrained edge devices.
- *Efficient Privacy-Preserving:* There are continually growing demand of effective and efficient privacy-preserving mechanisms for number of edge services and edge computing model associated with IoT.

The various security challenges associated with the edge are the only key factors that are restricting the companies and the users from using IoT resources. Once these challenges are addressed properly, then it may be possible that IoTs are mostly used and enhancing the processing power of the traditional cloud and lead to the edge's development with a faster pace.

VII. Future Work

In this new era of computing, every electrical equipment is integrated with IoTs and sensors with more advanced features of artificial intelligence [17]. Edge computing clearly proposes a number of benefits that cannot be easily achieved in traditional cloud computing, but also outsides a lot of technical and non-technical challenges as mentioned previously. The current study lacks the standardization in terms of the actual implementation of edge computing. The algorithm and software contrivance of the edge and IoT requires collective control of cloudlets. The features mentioned in comparison could be taken towards the analytical observation by implementing both in required scenario. On the technical side, the quantitative data can be collected and compared by analyzing the various factors, like performance, latency, data accuracy & loss, amount of data storage & processing capability, level of data security offered, deployment complications, expense occurrence of both the edge and traditional cloud computing environment. The analytical comparative study could help identify the technology being more useful and effective, providing maximum benefit out of it.

For the collection of such data, edge computing and cloud computing services can be implemented in different scenarios. For example, the application of both the technologies can be implemented on the healthcare department, where data and readings are very critical. Also, the use of edge devices is growing immensely in the sector. In such case, it will become easier to calculate the efficiency and better productivity out of both technologies. The one with less latency, less data loss, more accuracy, better productivity and other factors can determine the best implementation among the both and would be more clear on the feature and implementation of edge nodes.

Another imperative area of study will be the improvement of contrivances which can compensate for the weaker perimeter security of the IoT and cloudlets, as compared to the traditional cloud computing environment.

VIII. Conclusion

Edge computing with its advanced features associated with IoT, smart & mobile connected devices, etc. are bringing processing power as close as possible to the end devices, which is enabling low latency, real-time processing, enhancing mobility, quality of service, and acting as a thin line between the cloud and the devices tier [18]. The various models enabling edge computing on the three level architecture simplifies the edge implementation and processing. Edge differs from the traditional cloud in terms of mobility, and distributed architecture. Although edge provides the advantages, but the flaws that confines the cloud computing model currently remain as the lead of the generated data that continues to increase. The technology of edge computing still need development and further studies which could help in making a clear line to differentiate among

the effectiveness of the cloud and edge computing paradigm, as the current study lacks the calibration of one technology over the other on the specific implementation field. The orchestration of edge services become more challenging as edge node can be mobile and rapid changes can cause connectivity failure and bandwidth fluctuation.

References

[1] W. Shi, G. Pallis, and Z. Xu, "Edge computing" *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1474-1481, Aug. 2019.

[2] V. Gezer, J. Um, and M. Ruskowski, "An introduction to edge computing and a real-time capable server architecture," *International Journal on Advances in Intelligent Systems*, vol. 11, no. 1&2, pp. 105-113, 2018.

[3] S. B. Calo, M. Touna, D. C. Verma, and A. Cullen, "Edge computing architecture for applying AI to IoT," *2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, 2017, pp. 3012-3016, doi: 10.1109/BigData.2017.8258272.

[4] M. De Donno, K. Tange, and N. Dragoni, "Foundations and evolution of modern computing paradigms: Cloud, IoT, edge, and fog," *IEEE Access*, vol. 7, pp. 150936-150948, 2019, doi: 10.1109/ACCESS.2019.2947652.

[5] S. Singh, "Optimize cloud computations using edge computing," *2017 International Conference on Big Data, IoT and Data Science (BID)*, Pune, India, 2017, pp. 49-53, doi: 10.1109/BID.2017.8336572.

[6] K. Dolui, and S. K. Datta, "Comparison of edge computing implementations: Fog computing, cloudlet and mobile edge computing," *2017 Global Internet of Things Summit (GIoTS)*, Geneva, Switzerland, 2017, pp. 1-6, doi: 10.1109/GIOTS.2017.8016213.

[7] S. Fayaz, and U. I. Wani, "Introduction to the edge computing architecture" *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 5, no. 11, pp. 310-313, Nov. 2018.

[8] C. M. Fernández, M. D. Rodríguez, and B. R. Muñoz, "An edge computing architecture in the internet of things," *2018 IEEE 21st International Symposium on Real-Time Distributed Computing (ISORC)*, Singapore, 2018, pp. 99-102, doi: 10.1109/ISORC.2018.00021.

[9] K. Dolui, and S. K. Datta, "Comparison of edge computing implementations: Fog computing, cloudlet and mobile edge computing," *2017 Global Internet of Things Summit (GIoTS)*, Geneva, 2017, pp. 1-6, doi: 10.1109/GIOTS.2017.8016213.

[10] B. P. Rimal, D. P. Van, and M. Maier, "Mobile-edge computing vs. centralized cloud computing in fiber-wireless access networks," *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, San Francisco, CA, 2016, pp. 991-996, doi: 10.1109/INFCOMW.2016.7562226.

[11] S. Khanagha, S.(S.) Ansari, S. Paroutis, and L. Oviedo, "Mutualism and the dynamics of new platform creation: A study of Cisco and fog computing," *Strategic Management Journal*, pp. 1-31, 2020. [Online]. Available: https://doi.org/10.1002/smj.3147

[12] S. Yi, Z. Hao, Z. Qin, and Q. Li, "Fog computing: Platform and applications," *2015 Third IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*, Washington, DC, 2015, pp. 73-78, doi: 10.1109/HotWeb.2015.22.

[13] X. Sun, and N. Ansari, "EdgeIoT: Mobile edge computing for the internet of things," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 22-29, Dec. 2016, doi: 10.1109/MCOM.2016.1600492CM.

[14] K. Dolui, and S. K. Datta, "Comparison of edge computing implementations: Fog computing, cloudlet and mobile edge computing," *2017 Global Internet of Things Summit (GIoTS)*, Geneva, Switzerland, 2017, pp. 1-6, doi: 10.1109/GIOTS.2017.8016213.

[15] P. Tedeschi, and S. Sciancalepore, "Edge and fog computing in critical infrastructures: Analysis, security threats, and research challenges," *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, Stockholm, Sweden, 2019, pp. 1-10, doi: 10.1109/EuroSPW.2019.00007.

[16] J. Zhang, B. Chen, Y. Zhao, X. Cheng, and F. Hu, "Data security and privacy-preserving in edge computing paradigm: Survey and open issues," *IEEE Access*, vol. 6, pp. 18209-18237, 2018, doi: 10.1109/ACCESS.2018.2820162.

[17] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30-39, Jan. 2017, doi: 10.1109/MC.2017.9.

[18] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud computing and grid computing 360-degree compared," *2008 Grid Computing Environments Workshop*, Austin, TX, 2008, pp. 1-10, doi: 10.1109/GCE.2008.4738445.

International Journal of Distributed and Cloud Computing
9 (1) 2021, 18-27
http://www.publishingindia.com/ijdcc

Amazon EC2 Locations for Spot Pricing - A Hierarchical Clustering Approach

Veena Khandelwal[1*] and Shantanu Khandelwal[2]
[1]Department of Computer Science & Engineeing, SRM Institute of Science and Technology, India.
Email: vn.khandelwal@gmail.com
[2]Manager, KPMG Services Pte. Ltd., Singapore. Email: shantanukhandelwal@protonmail.com
*Corresponding Author

Abstract: EC2 provides cloud services through several regions and availability zones. Spot instances are offered by Amazon EC2 through spot pricing. Amazon launched it simplified spot pricing model at Reinvent: 2017. The prices change less frequently now and are more predictable. To ensure availability, spot instances can be acquired from diversified locations. The biggest concern to the user is the identification of regions and availability zones to choose for optimizing cost and maximizing availability. The study attempts to find similarity in cost per instance among different EC2 locations according to Ward's Hierarchical Clustering Algorithm. The analysis uses past 60 days history traces from last week of September to November 2019 of four different compute instance types across all Amazon EC2 regions and availability zones. Results suggest a significant amount of similarity in terms of cost of instance in spot pricing across different locations. This raises user's confidence in adopting spot market.

Keywords: Amazon EC2 regions, Availability zones, Hierarchical clustering, Spot instances.

I. Introduction

Cloud computing delivers different types of services in terms of resource capacities, irrespective of their physical location. These services are governed by an important agreement called Service Level Agreements. This agreement governs both the parties including cloud computing service provider and cloud end user at different pricing models. Amazon Web Services spot instances, Microsoft Azure's low-priority virtual machines and Google Cloud Engine offer pre-emptible virtual machines without any availability guarantee.

Amazon Elastic Compute Cloud (EC2) offers three different pricing models, namely Reserved Pricing, On-demand Pricing and Spot Pricing. Customers find spot pricing attractive as spot instances are offered at a very cheap price. Excess computing capacity is sold by Amazon as spot instances. Spot instances are offered by Amazon EC2 through spot pricing at a substantially low price without any availability guarantee, as spot instances can be interrupted by the cloud service provider at any time.

Amazon provides Infrastructure as a Service (IaaS) from various diversified locations comprising several regions and availability zones. The entire spot pricing model is built around bidding on excess resources at deep cost savings for executing a varied workload in a specified region and availability zone.

Earlier spot prices spiked higher than the on-demand price multiple times a day on every weekday. Spot prices were non uniformly distributed and exhibited seasonal nature. There were price variations up to 77% during different weekdays and hours [1].

In November 2017, Amazon simplified its spot pricing policy [2]. Spot prices are now stable and have lower volatility. After the change, spot prices hover between a small range without much volatility. The new spot instance market features significantly more stable with far fewer price events and longer times between price events. They are now predictable for longer period of time with lesser price events. The seasonal nature of the spot market has changed. The prices in the new model are updated far less frequently and are uniformly distributed throughout the week.

Amazon has also changed its spot instance termination policy. Spot instance market price is dynamically decided by Amazon and does not depend on the user submitted bids. With the new spot pricing model, when the demand exceeds the capacity, the price remains relatively the same, but some instances are reclaimed automatically by Amazon EC2, not for price, but for the capacity-oversubscribed. Spot instances can now be terminated when there is an increase in demand or decrease in supply due to more demand of reliable on-demand instances. Instances cannot be purchased at the published price because that does not reflect the actual supply and demand curve for the excessive resources. However, in order to control costs, maximum spot price can be set by the users depending upon their budget and time of executing jobs on spot instances.

In order to address this, Amazon provides a 2 minute termination notice. Depending on the needs, user can specify whether they want to hibernate, stop, or terminate spot instances when they are interrupted during the notice period [3]. Spot instances with bid prices higher than market price may be terminated as well. Bidding higher prices no more guarantees higher reliability or execution duration. Users can no longer control termination. Users no more bid exorbitantly for the spot instances i.e., even more than the on-demand prices when prices are jumping highly. Instead, they pay the current spot price. Users no more now get free lunch also which was earlier utilized for executing small jobs of duration less than one hour at no cost.

New spot pricing policy degrades the reliability of spot instances as well as it excludes the use of spot price prediction techniques. With the new pricing model, we have essentially traded one problem for another: predictability of getting and keeping an instance for the predictability of price. With the new spot pricing it has become much easier to get started with as the prices are now steady and there are fewer price adjustments than earlier. Also Hibernation feature introduced with the new spot pricing policy prevents data loss and makes execution of high performance computing jobs much easier. The new pricing policy has led to generally higher prices [4].

With this new spot pricing model, a couple of counter-intuitive things can and do occur:

- Bidding for spot instances at the current price does not result in acquiring an instance.
- An existing instance may be reclaimed even though the bid price was higher than the current price.

With the introduction of new spot pricing model, spot instances are suitable for executing applications such as high performance computing jobs, short-span workloads, development environments, large scale simulations such as IPL-like traffic pattern required by research community, high concurrency events, financial modelling and analysis, big data analytics, paid web services, distributed applications, testing and developing, web scraping and many others that do not require Reserved Instances. For use cases such as scientific research and computing, high concurrency events, large scale simulations where cost is an important aspect, there is a great amount of flexibility required for the adoption of spot instances. Finance folks are interested in same number of servers at lesser cost and developers are interested in same cost but more performance. To optimize cost and reduce the impact of spot instance volatility, customers can execute their jobs at diversified locations across multiple regions and availability zones. For acquiring spot instances at diversified locations and building spot price aware scheduling algorithms, the biggest concern to the user is the identification of regions/availability zones to choose to optimize cost and availability.

The objective of this study is to help developers design high availability solutions with failover at low cost using spot instances for applications that have significantly long execution time. The study identifies EC2 locations that have most similar hourly spot price variations to reduce the impact of volatility. By computing similarity between hourly spot prices of several compute instance types across different locations, this work illustrates the locations that have most similar price changes. We find that prices across availability zones within a region i.e., availability zones and prices in some regions also are very similar.

With the earlier spot pricing policy, spot instances were only suitable for execution of short jobs due to highly volatile nature of spot instances. We study new spot pricing policy across 16 different Amazon EC2 regions and 64 availability zones to design high availability solutions with failover using spot prices for long running sudden tasks at a low cost.

The rest of this paper is structured as follows. Section II of this paper examines related work. Section III examines the spot history data, Amazon infrastructure and different compute instances. Section IV focuses of Hierarchical Clustering technique used in this paper. Section V shows hierarchical clustering results of our analysis. Section VI will perform discussions on our results to glean any interesting insights. Finally Section VII will summarise and give our overall conclusions.

II. Related Work

Very few works are based on this new spot pricing model. Although there are several papers published since the new spot pricing model was introduced, most of them analyze spot history traces before 2017. The extensive work done earlier is now no longer applicable as spot price dynamics have changed significantly. Very few authors study new spot pricing model. We discuss all papers based on the old and new spot pricing model in the related work.

Prior to November 2017, it was believed that spot prices were market driven, but actually it was not so. Very high peaks which correspond to spot prices much higher than on-demand prices were observed. An unreliable instance cannot be acquired by any user by bidding at bid price higher than on-demand price. According to [5] there must be some hidden underlying mechanism governing spot price by Amazon EC2. When the spot prices were low during off-peak hours/off-peak days, spot instances were suitable for executing small and medium sized jobs of length 4 to 12 hours while optimizing execution cost. With frequent price changes, acquiring spot instances required forecasting spot prices, determining bidding strategy suitable for the job type, integration of check-pointing and fault-tolerance policies to complete job execution before the deadline. Much work has been done in these areas to optimize cost, reduce execution time and increase reliability of spot instances.

Authors of [6] model spot instance lifetime using a Markov Chain. [7] employ normal approximation to model the spot price distribution. [8] assume spot prices are normally distributed and focus on achieving availability. [9] make use of discrete semi-Markovian chain to model the spot price variations. [10]

perform analysis of spot price predictability. [11] make use of predictive model based on artificial neural networks. [5] perform reverse engineering on how prices are set. [12] show that Gaussian distribution shows better fit. Authors perform statistical analysis of spot instances using a mixture of Gaussian distribution to model inter-price time of each spot instance and determine the time correlation in terms of hour-in-day and day-of-week. Authors of [13] use Gradient Descent algorithm to predict one-day-ahead and one-week-ahead spot prices. Authors of [14] make use of several moving average statistical methods to predict the next hour spot prices. [15] use Random Regression Forest to predict one-day and one-week-ahead spot prices.

Authors of [16] study Amazon EC2 new spot pricing policy. Authors study spot prices in 4 different regions and conclude that location plays a critical role in spot instance pricing. Spot prices still differ depending on the granularity of that location i.e., AWS region and availability zone. According to [17] with the new spot pricing policy, frequency of price events has reduced substantially. Hourly and weekly price trends have also normalized. According to [4] since November 2017, when market driven pricing policy changed to retail pricing policy, spot prices are continually increasing.

However, some recent works in cloud computing are also based on security issues in adopting cloud computing, Big Data security and improving performance. Authors of [18] discuss various security issues and challenges associated with Cloud computing environment and current solution space. Authors of [19] choose optimized service broker routing policy based on different parameters to increase the overall performance of the cloud computing systems. Authors of [20] use cloud computing to operate with the Big Data to establish an architecture relaying on the security of the network in order to improve the security issues. Authors of [21] improve the multi-user data sharing mechanism in hybrid cloud using public cloud's cost savings and elasticity with the private cloud's security and customisation. Authors of [22] propose an efficient algorithm for advanced scalable Media-based Smart Big Data (3D, Ultra HD) on Intelligent Cloud Computing systems.

III. Spot History Data

The study involves 60 days spot price traces beginning from 24-09-2019 to 23-11-2019. Spot prices history traces were obtained using AWS Command Line Interface from AWS website directly and stored in csv file.

Table I: Amazon EC2 Regions, Availability Zones and Instance Types used for Study

Region Asia Pacific		*10*	*eu-west-2 (London)*
1	*ap-northeast-1 (Tokyo)*	AZ26	eu-west-2a
AZ0	ap-northeast- 1a	AZ27	eu-west-2b
AZ1	ap-northeast- 1c	AZ28	eu-west-2c
AZ2	ap-northeast- 1d	*11*	*eu-west-3 (Paris)*
2	*ap-northeast-2 (Seoul)*	AZ29	eu-west-3a
AZ3	ap-northeast- 2a	AZ30	eu-west-3b
AZ4	ap-northeast- 2b	AZ31	eu-west-3c
AZ5	ap-northeast- 2c	*Region South America*	
3	*ap-south-1 (Mumbai)*	*12*	*sa-east-1 (Sao Paulo)*
AZ6	ap-south-1a	AZ32	sa-east-1a
AZ7	ap-south-1b	AZ33	sa-east-1b
AZ8	ap-south-1c	AZ34	sa-east-1c
4	*ap-southeast-1 (Singapore)*	*Region US-East*	
AZ9	ap-southeast-1a	*13*	*us-east-1(N. Virginia)*
AZ10	ap-southeast-1b	AZ35	us-east-1a
AZ11	ap-southeast-1c	AZ36	us-east-1b
5	*ap-southeast-2 (Sydney)*	AZ37	us-east-1c
AZ12	ap-southeast-2a	AZ38	us-east-1d
AZ13	ap-southeast-2a	AZ39	us-east-1f
AZ14	ap-southeast-2c	*14*	*us-east-2 (Ohio)*
Region Canada		AZ40	us-east-2a
6	*ca-central-1 (Central)*	AZ41	us-east-2b

AZ15	ca-central-1a	AZ42	us-east-2c
AZ16	ca-central-1b	*Region US-West*	
Region EU		*15*	*us-west-1 (Oregaon)*
7	*eu-central-1 (Frankfurt)*	AZ43	us-west-1a
AZ17	eu-central-1a	AZ44	us-west-1c
AZ18	eu-central-1b	*16*	*us-west-2 (N. California)*
AZ19	eu-central-1c	AZ45	us-west-2a
8	*eu-north-1 (Stockholm)*	AZ46	us-west-2b
AZ20	eu-north-1a	AZ47	us-west-2c
AZ21	eu-north-1b	AZ48	us-west-2d
AZ22	eu-north-1c	*Instance Types*	
9	*eu-west-1 (Ireland)*	1	c5.2xlarge
AZ23	eu-west-1a	2	c5.4xlarge
AZ24	eu-west-1b	3	c5.9xlarge
AZ25	eu-west-1c	4	c5.18xlarge

Spot prices history consists of spot prices from 16 Amazon EC2 regions shown as bold and numbered, consisting of 49 availability zones numbered AZ0 to AZ48 and 4 compute instance types, namely c5.2xlarge, c5.4xlarge, c5.9xlarge and c5.18xlarge as shown in Table I. Spot instances in availability zone sa-east-1b started on 16-10-2019. As spot price are more predictable and change less frequently in the new spot pricing model, the study considers same spot price for sa-east-1b availability zone for time period beginning from 24-09-2019 to 15-10-2019. The study uses python 3.7.3 to implement hierarchical clustering using Ward's Method by importing scipy.cluster.hierarchy module.

IV. Methodology

A. Hierarchical Cluster Analysis

Cluster analysis is a statistical method that refers to the methods used to group different data objects into groups (clusters) based on similarity/dissimilarity among the data such that the intra-cluster similarity is high and inter-cluster similarity is low. The resulting partitioning into different sub classes reveals internal structures if any, and aids in improving our understanding of the data. Hierarchical clustering begins by considering each point as an individual cluster and at each step; clusters that are similar to each other are merged. Similarity at each step is determined based on the proximity metrics of the clusters. Split and merge process of the clusters is recorded in a tree like structure called Dendrogram. In Spot Pricing, hierarchical clustering is used to identify distinct groups of potential regions and availability zones so that, for example, bidding decision can be appropriately targeted.

B. Hierarchical Agglomerative Methods

1) Single Linkage

Single Linkage finds distance between closest clusters. Distance between two clusters *x* and *y* is given as:

$$D(x,y) = \sum_{i=1}^{S_x} \sum_{j=1}^{S_y} Min\{d\,(x_i, y_i)\} \tag{1}$$

where, S_x and S_y denote cluster sizes. Proximity between two clusters is the proximity between their two closest objects [23]. Cluster merging of the two closest clusters takes place in each step. The method produces clusters of irregular shapes. We avoid single linkage for spot pricing cluster analysis as it does not take into account cluster structure and results in forming long chains of the data. Single Linkage can sometimes form clusters with objects in different cluster when objects in other cluster are nearer than to the objects within its own cluster. The edges in the clusters start connecting and very soon all the data falls into one cluster. We need to avoid this and find different clusters.

2) Complete Linkage

Complete Linkage defines distances as the maximum instance distance over all possible pairs in two given clusters [24], which is a non local measure. Complete linkage has the opposite effect of single linkage. It can keep similar objects in separate clusters.

$$D(x,y) = \sum_{i=1}^{S_x} \sum_{j=1}^{S_y} Max\{d\,(x_i, y_i)\} \tag{2}$$

Complete Linkage finds distance between the farthest elements in the clusters. It forces well balanced, tighter spherical clusters than single linkage with consistent diameter i.e., equal size and also does not take into account cluster structure. This method is suitable for compact regular clusters. It is extremely sensitive to outliers.

3) Average Linkage Method

Average linkage method is a compromise between nearest neighbor distance computation of Single linkage and farthest neighbor computation of complete linkage method [25]. It tends to maximize the coherency of the joined clusters. This approach to calculate dissimilarity between two clusters takes all observations from each groups, computes their dissimilarity and calculates the average of the dissimilarity [26]. As all instance distances influence the distance calculation, this algorithm is robust towards outliers than complete linkage. It forms uniform sized, convex shaped clusters. It can also break large sized clusters into smaller ones.

$$D(x,y) = \frac{1}{S_x . S_y} \sum_{i=1}^{S_x} \sum_{j=1}^{S_y} d\,(x_i, y_i) \tag{3}$$

Mathematically, similarity can be written as:

$$sim(C1, C2) = \sum D(x,y)/|C1| \times |C2| \tag{4}$$

where, $C1$&$C2$ are clusters and $x_i \in C1$& $x_j \in C2$.

4) Ward's Method

This approach of calculating the similarity between two clusters is an alternative approach to single linkage clustering. It is robust against outliers but biased towards formation of spherical clusters. It is computationally expensive but performs fewer computations. The clusters formed are of fairly equal size, are less than optimal clusters [27]. At each step Ward's method joins the two clusters to make one cluster that minimizes variance measured by Sum of Squared Error (SSE) at each step. SSE is a measure of the distance between a cluster's instance attribute value and the cluster mean value. Instead of looking at the diameter of the result, it looks for the deviation of the result. It then merges the results with the smallest deviation.

$$SSE = \sum_{i}^{S_x} \sum_{j}^{S_y} |X_{ij} - Y_i|^2 \tag{5}$$

$$D(x,y) = SSE(x,y) - \big(SSE(x) + SSE(y)\big) \tag{6}$$

$$sim(C1, C2) = \sum_{1}^{p} (D(x,y)^2\,)/|C1| \times |C2| \tag{7}$$

where, p=61 days for which spot prices were collected in this study, number of points to be merged into clusters is 49 which refer to the different availability zones and $C1$ & $C2$ are clusters and $x_i \in C1$ & $x_j \in C2$. Hierarchical clustering performed for spot pricing uses Ward's method for calculating similarity between two clusters. For spot pricing cluster analysis, our basic requirement is that the price variations within a cluster should beminimum. This would help in identifying EC2 locations that have similar spot prices with minimum spot price variations. Similarity is calculated by finding the sum of the squared distances between any two points. Hierarchical clustering algorithm extracts data from spot history traces. Dendrograms are used to represent cluster merging process.

- On the x axis there are labels that specify indices of price traces in data.
- On the y axis we see the distances (of the 'Ward' method).

V. Hierarchical Clustering Results

After performing Hierarchical clustering using Ward's method, we show dendrograms of spot prices for four different compute instance types as mentioned in Table I to illustrate which clusters join at each level and the distance between clusters formation. Cluster formation is done using Ward's method based on the spot pricing history traces of 4 compute instance types in all 49 availability zones. When there is a large difference in the distance as calculated using between clusters at subsequent levels, it suggests that at lower level, clusters that are relatively close together were joined whereas, at the following level, clusters that were joined were distant. This is shown in dendrograms using three different horizontal lines at y=16, y=26 and y=45. The optimum number of clusters in spot pricing is the number present just before that large jump in distance shown by different horizontal lines in the dendrograms.

A. c5.2xlarge Instance Type

From 16 Amazon EC2 regions, initially 6 clusters are formed based upon their spot price similarity calculated using Ward's method. This is shown by line y=15 in Fig. 1. Cluster 1 is formed by AZ0 to AZ8 which belong to regions ap-northeast-1, ap-northeast-2 and ap-south-1. Cluster 2 is formed by AZ9 to AZ16 which belong to regions ap-southeast-1, ap-southeast-2 and ca-central-1. Cluster 3 formed by AZ17 to AZ25 which belong to regions eu-central-1, eu-north-1 and eu-west-1. Cluster 4 is formed by AZ26 to AZ31 which belong to regions eu-west-2 and eu-west-3. Cluster 5 is formed by AZ32 to AZ39 which belong to regions sa-east-1 and us-east-1. Cluster 6 is formed by AZ40 to AZ48 which belong to regions us-east-2, us-west-1, us-west-2. These 6 clusters are formed at the least vertical distance, i.e., have the lowest variations and hence highest similarity in prices. Hierarchical merging of clusters for different regions at line y=16 are shown in Table II and Table III.

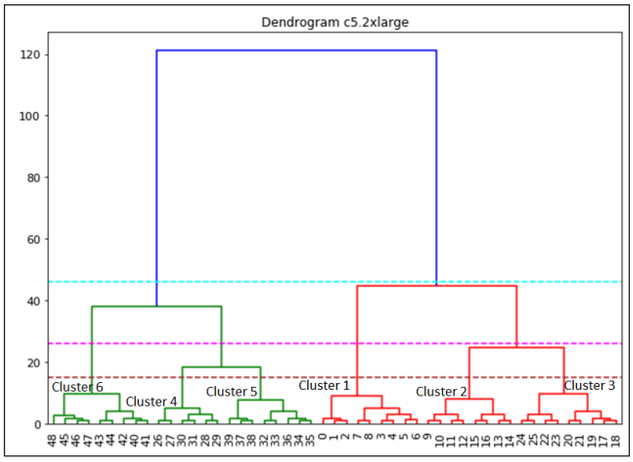


Fig. 1: Cluster Formation using Ward's Method for c5.2xlarge Instance Type

B. c5.4xlarge Instance Type

From 16 Amazon EC2 regions, initially 6 clusters are formed based upon their spot price similarity calculated using Ward's method. This is shown by line y=15 in Fig. 2. Cluster 1 is formed by AZ0 to AZ5 which belong to regions ap-northeast-1 and ap-northeast-2. Cluster 2 is formed by AZ9 to AZ14 which belong to regions ap-south-1, ap-southeast-1 and ap-southeast-2. Cluster 3 formed by AZ15 to AZ25 which belong to regions ca-central-1, eu-central-1, eu-north-1 and eu-west-1. Cluster 4 is formed by AZ26 to AZ36 which belong to regions eu-west-2, eu-west-3 and sa-east-1 and availability zones 1a and 1b of region us-east 1. Cluster 5 formed by AZ37 to AZ39 which belong to regions availability zones 1c, 1d, 1f of region us-east-1, us-east-2, us-west-1 and us-west-2. These 5 clusters are formed at the least distance, i.e., have the lowest variations and hence highest similarity in prices. At line y=26, clusters 1, 2 merge. These clusters are less similar to each other as they are formed at a distance of 26. At line y=46, clusters 3, 4 merge to form one cluster. These clusters are most dissimilar to each other as they have joined far apart.

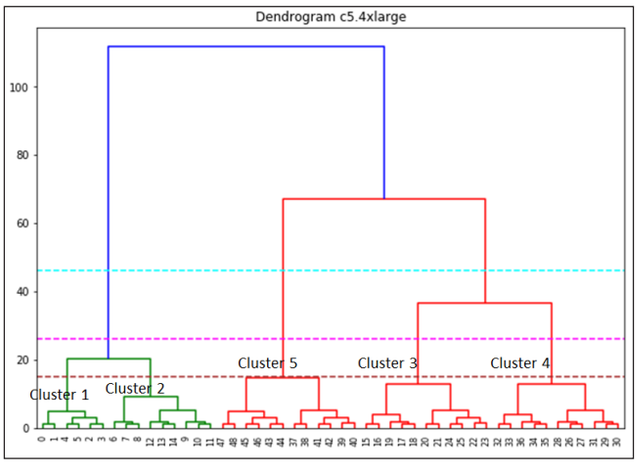


Fig. 2: Cluster Formation using Ward's Method for c5.4xlarge Instance Type

C. c5.9xlarge Instance Type

From 16 Amazon EC2 regions, initially 6 clusters are formed based upon their spot price similarity calculated using Ward's method. This is shown by line y=15 in Fig. 3. Cluster 1 is formed by AZ0 to AZ8 which belong to regions ap-northeast-1, ap-northeast-2 and ap-south-1. Cluster 2 is formed by AZ9 to AZ16 which belong to regions ap-southeast-1, ap-southeast-2 and ca-central-1. Cluster 3 formed by AZ17 to AZ25 which belong to regions eu-central-1, eu-north-1 and eu-west-1. These 6 clusters are formed at the least distance, i.e., have the lowest variations and hence highest similarity in prices.

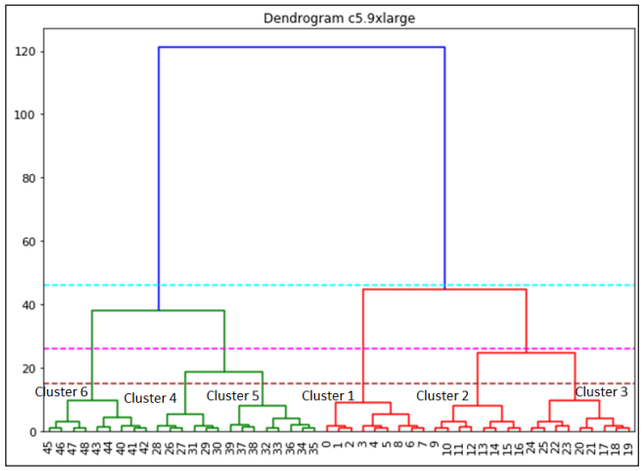


Fig. 3: Cluster Formation using Ward's Method for c5.9xlarge Instance Type

The cluster formation in c5.9xlarge instance types at line y=15 is exactly as c5.2xlarge. This suggests that spot price changes of both instance types are very similar in all regions and availability zones. At line y=26, clusters 2 and 3 merge. These clusters are less similar to each other as they are formed at a distance of 26. At line y=46, clusters 1, 2, 3 merge to form one cluster. These clusters are most dissimilar to each other as they have joined far apart. A slight variation in merging of clusters is observed for these regions with respect to c5.2xlarge instance type. But, these does not make any significant impact to our study as we are interested in most similar regions, i.e., lowest level merge and not most dissimilar regions, corresponding to highest level merge.

D. c5.18xlarge Instance Type

From 16 Amazon EC2 regions, initially 5 clusters are formed based upon their spot price similarity calculated using Ward's method. This is shown by line y=15 in Fig. 4. Cluster 1 is formed by AZ0 to AZ11 which belong to regions ap-northeast-1, ap-northeast-2, ap-south-1 and ap-southeast-1. Cluster 2 is formed by AZ9 to AZ19 which belong to regions ap-southeast-2 and ca-central-1 and eu-central-1. Cluster 3 formed by AZ20 to AZ31which belong to regions eu-north-1, eu-west-1, eu-west-2 and eu-west3. Cluster 4 is formed by AZ32 to AZ38 which belong to regions sa-east-1 and us-east-1. Cluster 5 is formed by AZ39 to AZ48 which belong to availability zone us-east-1a

of region us-east 1 and regions us-east-2, us-west-1, us-west-2. These 5 clusters are formed at the least distance, i.e., have the lowest variations and hence highest similarity in prices. The cluster formation in c5.18xlarge instance types is similar to c5.2xlarge to a great extent. This suggests that spot price changes of both instance types are very similar in all regions and availability zones.

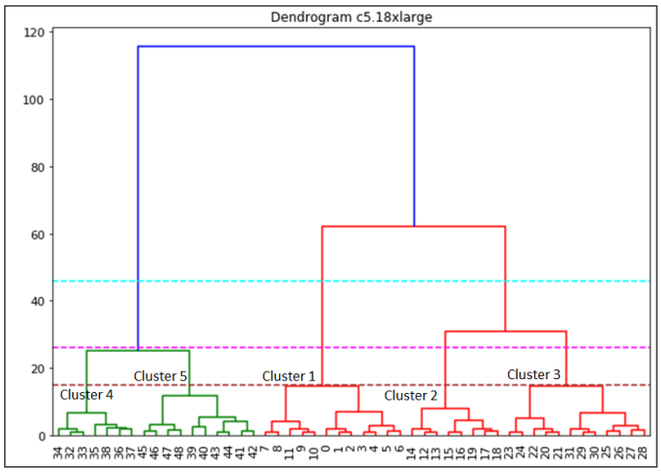


Fig. 4: Cluster Formation using Ward's Method for c5.18xlarge Instance Type

At line y=26, clusters 2, 3 merge and clusters 4, 5 merge. These clusters are less similar to each other as they are formed at a distance of 26. At line y=46, clusters 1, 2, 3 merge to form one cluster. These clusters are most dissimilar to each other as they have joined far apart. A slight variation in merging of clusters is observed for these regions with respect to c5.2xlarge instance type. Again, this does not make any significant impact to our study as we are interested in most similar regions, i.e., lowest level merge and not most dissimilar regions corresponding to highest level merge.

Hierarchical merging of clusters for different availability zones is shown in Table IV. The table shows formation of clusters 1 to 6 in different availability zones. From Cluster 1 results, we find that zones AZ0 to AZ8 form one cluster for most of the instance types. From cluster 2 results we find that zones AZ19 to AZ14 form one cluster. From cluster 3 results, we find that zones AZ17 to AZ25 form one cluster for most of the instance types. From cluster 4 results, we find that zones AZ26 to AZ31 form one cluster for most of the instance types. From cluster 5 results, we find that zones AZ32 to AZ39 form one cluster for most of the instance types. From cluster 6 results, we find that zones AZ40 to AZ48 form one cluster for most of the instance types.

Table II: Hierarchical Clustering for Cluster 1, Cluster 2 and Cluster 3 for Different Compute Instance Types Showing Different Regions that Merge due to Similarity in Spot Price for Different Compute Instance Types at Line y=15

Instance Type	Cluster 1 (Regions)	Cluster 2 (Regions)	Cluster 3 (Regions)
c5.2xlarge	ap-northeast-1	ap-southeast-1	eu-central-1
	ap-northeast-2	ap-southeast-2	eu-north-1
	ap-south-1	ca-central-1	eu-west-1
c5.4xlarge	ap-northeast-1	ap-south-1	ca-central-1
	ap-northeast-2	ap-southeast-1	eu-central-1
		ap-southeast-2	eu-north-1
			eu-west-1
c5.9xlarge	ap-northeast-1	ap-southeast-1	eu-central-1
	ap-northeast-2	ap-southeast-2	eu-north-1
	ap-south-1	ca-central-1	eu-west-1
c5.18xlarge	ap-northeast-1	ap-south-1	ap-southeast-2
	ap-northeast-2	ap-southeast-1	ca-central-1
	ap-south-1		eu-central-1

Table III: Hierarchical Clustering for Cluster 4, Cluster 5 and Cluster 6 for Different Compute Instance Types Showing Different Regions that Merge due to Similarity in Spot Price for Different Compute Instance Types at Line y=15

Instance Type	Cluster 4 (Regions)	Cluster 5 (Regions)	Cluster 6 (Regions)
c5.2xlarge	eu-west-2	sa-east-1	us-east-2
	eu-west-3	us-east-2	us-west-1
			us-west-2
c5.4xlarge	eu-west-2		us-east-1
	eu-west-3		us-east-2
	sa-east-1		us-west-1
			us-west-2
c5.9xlarge	eu-west-2	sa-east-1	us-east-2
	eu-west-3	us-east-1	us-west-1
			us-west-2
c5.18xlarge	eu-north-1	sa-east-1	us-est-2
	eu-west-1	us-east-1	us-west-1
	eu-west-2		us-west-2
	eu-west-3		

Table IV: Hierarchical Clustering for Cluster 1, Cluster 2, Cluster 3, Cluster 4, Cluster 5 and Cluster 6 for Different Compute Instance Types Showing Different Availability Zones that Merge due to Similarity in Spot Price for Different Compute Instance Types at Line y=15

Instance Type	Cluster 1	Cluster 2	Cluster 3
c5.2xlarge	AZ0 to AZ8	AZ9 to AZ16	AZ17 to AZ25
c5.4xlarge	AZ0 to AZ5	AZ6 TO AZ14	AZ15 to AZ25
c5.9xlarge	AZ0 to AZ8	AZ9 TO AZ16	AZ17 to AZ25
c5.18xlarge	AX0 to AZ11	AZ12 TO AZ19	AZ20 to AZ31
Instance Type	Cluster 4	Cluster 5	Cluster 6
c5.2xlarge	AZ26 to AZ31	AZ32 to AZ39	AZ40 to AZ48
c5.4xlarge	AZ26 to AZ36		AZ37 to AZ48
c5.9xlarge	AZ26 to AZ31	AZ32 to AZ39	AZ40 to AZ48
c5.18xlarge	AZ32 to AZ38	AZ39 to AZ48	

Authors of [28] in their study confirm the lack of practical and deployable market-driven mechanism for using spot instances. The result of clustering analysis cannot be compared with any other study as this is the first study since Amazon changed it new spot pricing policy in November 2017.

VI. Conclusion

Academic community strongly advocated spot market whey it was introduced, but it is still not reflected by consumer behaviour. Consumers still doubt about Amazon's spot-pricing scheme. This is due to the lack of published information from Amazon EC2 regarding termination of spot instances. At some point of time even over-subscribed instances are terminated and at some other time even under-subscribed instances sustain. In order to play a promising role in the sustainability of EC2 IaaS exploitation at low cost, some availability guarantee is required that the job will not be interrupted frequently and will be completed within its deadline. Spot consumers prefer simple mechanisms without using complex procedures at the backend

such as spot price prediction, check pointing and migration which are a major obstacle in spot market adoption. This analysis identifies regions/availability zones with similar price variations which would help user in spot bidding decisions and designing high availability solutions at diversified EC2 locations with failover at low cost without falling back to on-demand instances when spot instances are not available. This study would fulfil consumer's real needs and raise their confidence in exploiting spot market and design practical and easily deployable spot mechanisms. The study explores spot prices of only 4 compute instance types. It could be extended to other instance types also. In future we would like to study on outliers analysis of the spot prices aiming to enhance customer confidence in the adoption of spot instances for lowering costs.

References

[1] K. Veena, C. Anand, and G. C. Prakash, "Temporal and spatial trend analysis of cloud spot instance pricing in Amazon EC2," *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech),* Aug. 2016.

[2] R. Pary, New Amazon EC2 Spot Pricing Model: Simplified Purchasing without Bidding and Fewer Interruptions | Amazon Web Services. Accessed: Jan. 16, 2020. [Online]. Available: https://aws.amazon.com/blogs/compute/new-amazon-ec2-spot-pricing/

[3] Docs.aws.amazon.com, Spot Instance Interruptions - Amazon Elastic Compute Cloud. Accessed: Jan. 26, 2020. [Online]. Available: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-interruptions.html

[4] G. George, R. Wolski, C. Krintz, and J. Brevik, "Analyzing AWS spot instance pricing," *2019 IEEE International Conference on Cloud Engineering (IC2E),* Jun. 2019.

[5] O. A. Ben-Yehuda, M. Ben-Yehuda, A. Schuster, and D. Tsafrir, "Deconstructing Amazon EC2 spot instance pricing," *ACM Transactions on Economics and Computation*, vol. 1, no. 3, pp. 1-20, Sep. 2013.

[6] N. Chohan, C. Castillo, M. Spreitzer, M. Steinder, A. N. Tantawi, and C. Krintz, "See spot run: Using spot instances for mapreduce workflows," *HotCloud*. 2010, pp. 7-7.

[7] A. Andrzejak, D. Kondo, and S. Yi, "Decision model for cloud computing under SLA constraints," *2010 IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, Aug. 2010.

[8] M. Mazzucco, and M. Dumas, "Achieving performance and availability guarantees with spot instances," *2011 IEEE International Conference on High Performance Computing and Communications*, Sep. 2011

[9] M. Zafer, Y. Song, and K.-W. Lee, "Optimal bids for spot VMs in a cloud for deadline constrained jobs," *2012 IEEE 5th International Conference on Cloud Computing*, Jun. 2012.

[10] H. Zhao, M. Pan, X. Liu, X. Li, and Y. Fang, "Optimal resource rental planning for elastic applications in cloud market," *2012 IEEE 26th International Parallel and Distributed Processing Symposium*, May 2012.

[11] R. M. Wallace, V. Turchenko, M. Sheikhalishahi, I. Turchenko, V. Shults, J. L. Vazquez-Poletti, and L. Grandinetti, "Applications of neural-based spot market prediction for cloud computing," *2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*, Sep. 2013.

[12] B. Javadi, R. K. Thulasiram, and R. Buyya, "Characterizing spot price dynamics in public cloud environments," *Future Generation Computer Systems*, vol. 29, no. 4, pp. 988-999, Jun. 2013.

[13] V. K. Singh, and K. Dutta, "Dynamic price prediction for amazon spot instances," *2015 48th Hawaii International Conference on System Sciences*, Jan. 2015.

[14] J. L. Lucas-Simarro, R. Moreno-Vozmediano, R. S. Montero, and I. M. Llorente, "Cost optimization of virtual infrastructures in dynamic multi-cloud scenarios," *Concurrency and Computation: Practice and Experience*, vol. 27, no. 9, pp. 2260-2277, Dec. 2012.

[15] V. Khandelwal, A. K. Chaturvedi, and C. P. Gupta, "Amazon EC2 spot price prediction using regression random forests," *IEEE Transactions on Cloud Computing*, vol. 8, no. 1, pp. 59-72, Jan. 2020.

[16] N. Ekwe-Ekwe, and A. Barker, "Location, location, location: Exploring Amazon EC2 spot instance pricing across geographical regions," *2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, May 2018.

[17] M. Baughman, S. Caton, C. Haas, R. Chard, R. Wolski, I. Foster, and K. Chard, "Deconstructing the 2017 changes to AWS spot market pricing," *Proceedings of the 10th Workshop on Scientific Cloud Computing - ScienceCloud'19*, 2019.

[18] Z. Gou, S. Yamaguchi, and B. B. Gupta, "Analysis of various security issues and challenges in cloud computing environment," *Identity Theft*, pp. 221-247.

[19] A. M. Manasrah, A. Aldomi, and B. B. Gupta, "An optimized service broker routing policy based

on differential evolution algorithm in fog/cloud environment," *Cluster Computing*, vol. 22, no. S1, pp. 1639-1653, Dec. 2017.

[20] K. E. Psannis, C. Stergiou, and B. B. Gupta, "Advanced media-based smart big data on intelligent cloud systems," *IEEE Transactions on Sustainable Computing*, vol. 4, no. 1, pp. 77-87, Jan. 2019.

[21] G. Gao, L. Wu, and Y. Yan, "A secure storage scheme with key-updating in hybrid cloud," *International Journal of High Performance Computing and Networking*, vol. 13, no. 2, p. 175, 2019.

[22] C. Stergiou, K. E. Psannis, B. B. Gupta, and Y. Ishibashi, "Security, privacy & efficiency of sustainable cloud computing for big data & IoT," *Sustainable Computing: Informatics and Systems*, vol. 19, pp. 174-184, Sep. 2018.

[23] F. Murtagh, and P. Contreras, "Algorithms for hierarchical clustering: An overview," *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86-97, Dec. 2011.

[24] R. J. Jensen, G. Dunn, and B. S. Everitt, "An introduction to mathematical taxonomy," *Systematic Botany*, vol. 8, no. 1, p. 103, Jan. 1983.

[25] D. P. Berrar, W. Dubitzky, and M. Granzow (Eds.), "A practical approach to microarray data analysis," 2003.

[26] M. R. Anderberg, "Conceptual problems in cluster analysis," *Cluster Analysis for Applications*, pp. 10-24, 1973.

[27] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236-244, Mar. 1963.

[28] Z. Li, H. Zhang, L. O'Brien, S. Jiang, Y. Zhou, M. Kihl, and R. Ranjan, "Spot pricing in the cloud ecosystem: A comparative investigation," *Journal of Systems and Software*, vol. 114, pp. 1-19, Apr. 2016.

International Journal of Distributed and Cloud Computing
9 (1) 2021, 28-35
http://www.publishingindia.com/ijdcc

Internet-of-Things (IoT): An In-Depth Examination of the Issues and Challenges for the Purpose of Ensuring Long-Term Viability

Pritam Kumar
Lecturer, Digital Business Management, Martin de Tours School of Management and Economics,
Assumption University, Thailand. Email: pkumar@msme.au.edu

Abstract: The primary objective of this research is to identify and analyze all the issues and challenges associated with the Internet-of-Things in the context of business management. The paper develops an exploratory research model to identify and analyze all the issues and challenges associated with the Internet-of-Things in business administration. The researcher gathered all the relevant data and information with the help of a previous literature review recently published on the Internet-of-Things and business management. To compile relevant content on the subject, the researcher used secondary resources such as Internet sites, the official website of organizations that have adopted the Internet-of-Things for management, published research papers, peer-reviewed journals, magazines, periodical articles, and books. This research study will aid researchers and management in developing a unique security framework to minimize and eliminate security and privacy risks and the adoption of the Internet-of-Things using new communication tools and technologies in business management practices. According to the findings of this study, the Internet-of-Things not only assist in business processes and activities but also improve the efficiency and productivity of the available things.

Keywords: Business management, Internet-of-Things, Sustainable development.

I. Introduction

The adoption and implementation of new technology introduce risk and uncertainty into the business environment. For any organization to compete effectively with global competitors in this highly competitive era, intelligent and innovative tools and techniques were required. The Internet-of-Things (IoT) aids in enhancing the capacity and ability of various business activities to achieve predetermined organizational objectives by providing support Loai [1]. When an organization adopts and implements the Internet-of-Things in its operations for the first time, it requires basic hardware and software assets. As a result, it requires a large amount of financial support and technical knowledge about the utilization of all the applications and technologies available. There is a variety of building blocks available to get the process started and ready to be used.

Over the past few years, the Internet-of-Things has had a significant impact on many aspects of human life, including the following areas: residences, educational institutions, workplaces, industrial factories, hospitals, and health-care facilities, as well as agricultural environments, Parvaneh Asghari [2]. As a result of technological advancements and innovation, as well as rapid changes and convergence in the fields of wireless and digital communication systems, micro-electro-mechanical systems, and digital electronics technology, the Internet-of-Things has been introduced. According to Sagar [3], large objects connected to the Internet have improved the organization's overall performance. Because of technological innovation and advancement, the modern generation is faced with a variety of challenges. Because of global competition and new technologies, the Internet-of-Things plays a significant role in environmental protection, social welfare, economic prosperity, and business activities. Every organization strives to reach many customers in the shortest amount of time and at the lowest possible cost; this is only possible when organizations use the Internet and advanced technology Abubaker [4].

As a result, to maintain global competitiveness, every organization must conduct accurate strategic planning and extensive use of advanced technology in all aspects of its operations. The results of these efforts will be beneficial in the long run Ali [5]. As a result, organizations must focus on developing innovative technologies that will help them to become more competitive in the future. The Internet-of-Things is a necessity for the organization in terms of policy formulation. It has created several opportunities that can assist in cost reduction, revenue enhancement, and operational efficiencies.

II. Literature Review

Researchers examine a variety of previous literature in this section of their research. 1st, the researcher looks at other recent publications in the field of study. Many books and articles are available about issues and challenges associated with Internet-of-Things applications in business management. During this

section, the researcher summarises the most important and relevant findings from previous research and explains how this research differs from previous research findings.

According to Falguni Jindal [6], new innovative growing technology aids in advancing science and technology as quickly as possible. The Internet-of-Things makes use of robotics and virtual reality to its advantage. It connects and supports many living and non-living objects, which can then transfer and share information within a community network. It also develops automated processes to assist humans and make their lives more convenient. It is proposed in this research paper that the Internet-of-Things will face several challenges in the future. These include compatibility and long-term viability, security of the technical standards and connectivity, investment in the business, intelligent analysis and actions, and legal challenges, among other things. According to the findings of this study, data security was the most crucial factor in its adaptation.

According to Zainab [7], the significant challenges and issues of the Internet-of-Things environment are discussed in detail, emphasizing the future research directions in this study area. Current developments have brought about the emergence of the Internet-of-Things as a new and innovative technology that incorporates a modern and cutting-edge wireless telecommunications network. The Internet-of-Things can be defined as interoperability and intelligent node for interconnected devices in an infrastructure network. This study aims to identify the issues and challenges associated with the adoption and implementation of a connectivity concept on the Internet-of-Things. The Internet-of-Things environment played a significant role in minimizing the various challenges and their impact on organizational performance, classified into 2 categories: General challenges and Unique challenges.

According to Abubaker Haddud [4], the impact of the adoption and implementation of the Internet-of-Things in the supply chains of a specific organization is investigated. This research reveals the most significant advantages, benefits, and challenges that have been previously documented in the literature. This study aims to determine the impact of the Internet-of-Things on organizational performance and the supply chain in general. The researcher creates a research model that will make use of an online survey to collect data. To accomplish this, the researcher selected 87 respondents as a sample to collect the necessary information for completing the survey. University scholars from various countries across 6 continents were chosen to participate in the survey as respondents. Respondents were authors and co-authors of academic papers published in the Decision Science Institute's annual conference proceedings in 2015 and 2016, respectively. The adoption of the Internet-of-Things could provide some significant benefits to specific organizations and their supply chains, according to the findings of this paper.

A study conducted by Ali Kamali [5] examines the fuzzy analytic network process (FANP), which was used to determine the most effective Internet-of-Things tools and technology challenges in Iran and prioritize the factors mentioned previously. This paper focuses on the approach and challenges that were implemented, including privacy, business-related factors, technological factors, security issues, legal and regulatory challenges, and cultural elements, among other things, in information security. All these factors had an impact on the development of Internet-of-Things technology. According to the findings, security, privacy, and technological challenges were the most critical factors that directly impacted the Internet-of-Things' adoption rate.

An investigation by Parvaneh Asghari [2] into various Internet-of-Things application domains is carried out to determine the various approaches used in its applications. The researcher used the Systematic Literature Review (SLR) method to examine the previous literature in various fields of study. The purpose of this study was to conduct a statistical and analytical analysis of the most recent research techniques used in Internet-of-Things applications. The researchers chose research articles that were published between 2011 and 2018.This study used a technical taxonomy, which revealed the Internet-of-Things applications approaches according to the ones that were selected through the SLR process. The Internet-of-Things applications covered in this study included smart cities, environmental monitoring, health care, commercial, and industrial aspects. There were several disadvantages to different studies, and the weaknesses of the previous paper were discussed in this study. The study also identified potential future research challenges and issues in Internet-of-Things applications.

It has been suggested by Habib [8] that the Internet-of-Things converts the digital world into a physical world. This research study has made human life more accessible, secure, happier, and more comfortable than previously. This paper examines the impact of the Internet-of-Things on the growth of the economy and its various applications. It is proposed in this paper to identify applications that cover all aspects of human life and make connectivity possible from anywhere, at any time, and to anything in the future, and to do so in a cost-effective manner. This research provides some recommendations for implementing this research of connectivity, which opens several new research questions. Finally, this paper examines the various aspects of future Internet-of-Things applications and the research challenges.

According to Sagar Bhat [3], there are several challenges on the Internet-of-Things. This research looked at the future of the Internet, specifically the part that would assist intelligent communicating things to improve life and business operations. This study aims to identify future challenges associated with the Internet-of-Things, which will be comprised of heterogeneously connected devices in the future. According to the findings of this study, Internet-of-Things will become a reality within the next 5 years. It is demonstrated in this study that the Internet-of-Things has a wide range of potential as a powerful technology in everyday easy and fast life. All the opportunities provided by the Internet-of-Things have resulted in numerous different applications based on connectivity and its use. This study provides some guidance for future research into intelligent

applications for more innovative offices, enterprises, factories, and homes and more innovative hospitals, more convenient and easy-to-use transportation systems, among other things.

Khanna [9], a researcher in applications, evaluates the various contributions of researchers in various fields of application. In this research paper, the researcher compiles several research studies relevant to the current state of the Internet-of-Things. This review of the literature was carried out on various aspects of the Internet-of-Things, including applications, technologies, and challenges, among other things. The papers examined various parameters to search for each application in turn. This research also discovered that there are currently existing challenges in this field, which were highlighted. Finally, this study identifies some future research directions in Internet-of-Things that were identified during the research process to prepare novel researchers in this field to search for and improve upon current standings in Internet-of-Things.

According to N. Angelova [10], the Internet-of-Things was a new paradigm that had a significant impact on a wide range of industries and sectors, including agriculture and farming, health; education; and industry, among many others. This study investigates the various changes in the adoption of advanced technology and the Internet-of-Things that have occurred to support work. This research aims to identify the most critical functions that make human life more efficient, easier, and more comfortable. Consumers, governments, and businesses all benefited from incorporating the Internet-of-Things into any business model, regardless of the industry. It identified several challenges and threats, including those related to security, privacy, and standards.

III. The Research Objectives

A. Significant Goal

The primary goal of this research is to identify and analyze all the issues and challenges associated with the Internet-of-Things in the context of corporate management.

Secondary Objectives of the Study

- To ascertain the relationship between the use of the Internet of Things and the performance of business management personnel.
- To provide some recommendations to business managers on how to recognize and define the challenges and issues associated with the Internet-of-Things to ensure the long-term viability of their organizations.

Query for Research

This research aims to identify and analyze all the issues and challenges associated with the Internet-of-Things in the context of business management. In addition, the research uncovers information that helps to answer the following research questions more specifically:

RQ.1. How do business managers deal with the issues and challenges of adopting and implementing the Internet-of-Things?

RQ.2. What are the potential advantages of the Internet-of-Things, as well as their implications for business management?

RQ.3. What strategies do business managers use to overcome all these obstacles to the long-term development of their organizations?

B. Research Design/Data Analysis Process

The paper develops an exploratory research model to identify and analyze all the issues and challenges associated with the Internet-of-Things in business administration. The researcher gathered all the relevant data and information with the help of a previous literature review recently published on the Internet-of-Things and business management. To compile relevant content on the subject, the researcher used secondary resources such as Internet sites, the official website of organizations that have adopted the Internet-of-Things for management, published research papers, peer-review journals, magazines, periodicals articles, and books.

C. The Internet-of-Things as a Concept

The phrase "Internet-of-Things" is comprised of 2 words, namely, the Internet and things. Using the Internet as a connector, we can connect 2 things: living things such as humans and other animals or birds and non-living things such as sensors and other technologies, electronic devices, and other resources used in business activities.

Definition of Internet-of-Things

"An open and comprehensive network of intelligent objects that can self-organize, share information, data, and resources, and react and act in response to situations and changes in their environments."

D. According to Oracle India

"IoT (Internet-of-Things) refers to a network of physical objects (also known as "things") that have been embedded with sensors, software, and other technologies to communicate and exchange data with other devices and systems; over the internet."

IV. Issues and Challenges

When it comes to adopting and implementing the Internet-of-Things to improve the performance of various business

management functions, every business organization faces several issues and challenges. In this section, the researcher identifies the difficulties associated with the Internet-of-Things in business administration. In addition, the new development in the IoT has spawned a slew of issues, challenges, and questions from a legal and regulatory standpoint, among other things.

- *Security*: Because of the insufficient and improper use of the Internet-of-Things, organizations are experiencing significant security issues. The organization required a large amount of data to make various managerial decisions, so management implemented an extensive data analysis system, cloud computing, data warehousing, and data mining, in addition to several new hubs to the web and the systems. If an organization does not prioritize security, it will become a target for hackers and attackers. Many businesses are suffering because of security flaws in their systems. According to the study's findings, researchers discovered that malware was being distributed against their servers by many Internet-of-Things devices used in basic applications, such as closed-circuit cameras and smart home devices.
- The Internet-of-Things provides enormous benefits to its users. However, it also brings with it several issues and challenges that must be addressed. Organizational managers, users, researchers, and security specialists recognize the importance of privacy risks and cyber security Parvaneh Asghari and colleagues [2]. Unfortunately, users may not be aware of or acknowledge the security consequences in some cases. However, the security and privacy of databases are the most important factors to consider when using an internet-based system because if a security breach occurs, it can result in massive data loss and damage.
- *Network Security*: Network security is a powerful mechanism that can be exerted within a network to ensure that the Internet-of-Things operates in a trustworthy manner. Network security is the most critical element. Every organization devotes significant resources to security architecture, which is designed from a communication perspective. This is not suitable for the Internet-of-Things system; therefore, reliable strategies for improving the security of the Internet-of-Things network must be developed.
- *Virtualization*: Virtualization is the ability to share hardware resources across multiple operating systems and is a virtualization capability. Using virtualization technology, it is possible to run multiple operating systems and software programs on the same server, such as applications or services, by creating multiple virtual machines within a physical machine. Every organization faces challenges on the Internet-of-Things environment, divided into 3 categories: the absence of a registry mechanism, the discovery and heterogeneity of resources, and the absence of interaction between services and events on the Internet-of-Things environment.
- *Connectivity*: The connection of multiple devices is one of the most significant challenges of the Internet-of-Things; this will either stop or eliminate the currently existing technologies and the associated structure. The existing technologies are being used to connect, authorize, and authenticate several terminals in a specific network using a client/server architecture and centralized administration. As a result, this model is only valid and appropriate for the current situation. It is not suitable for future requirements because organizations will need to connect billions of devices into a single network. In addition, high speed and new innovative technologies are required by organizations to compete in the global marketplace, and that require many devices to connect many employees, suppliers, distributors, consumers, and other stakeholders to improve performance and generate significant profits. Consequently, proper connectivity with others in a single network is required to function correctly.
- *Compatibility and Longevity*: The concept of the Internet-of-Things is advancing at a rapid pace. It includes a variety of tools and technologies and a focus on a standard that will improve the compatibility and long-term viability of internet-based technologies. However, it will present some difficulties and challenges in establishing additional innovative hardware and software to develop communication with many devices simultaneously. Non-unification of cloud services, unavailability of standardized protocols and operating systems, and variations in firmware among Internet-of-Things tools and technologies are just a few of the issues concerning the Internet-of-Things' compatibility. Furthermore, in the future, various technologies and devices currently in use in organizations will become obsolete or require updating because innovations in the technology environment will cause all these technologies to become obsolete or require updating.
- *Standards*: It collects activities that store, process, and handle information collected from various databases and sensors, collectively referred to as "the Internet-of-Things." Technology conventions are a combination of communication protocols, networks, and data aggregation, all of which help improve the data by increasing the scope, scale, and frequency of data available for investigation. The standard is the most critical element in a data analysis system; by employing the proper standard, every organization can obtain accurate results and information that can be used for various managerial decisions. If the organization does not have proper standards in place for network, protocol, and data aggregation, it will create some difficulties in managing the organization.
- *Privacy*: When it comes to the implementation of the Internet-of-Things in business management, privacy is

a significant challenge. The ability of the Internet-of-Things to function appropriately is dependent on the privacy preferences of users and organizations. So many times, organizations are confronted with issues of privacy, as well as the potential risks and harms that come with the Internet-of-Things and other cutting-edge technological advancements. Because of data loss and privacy concerns, small and medium-sized businesses are delaying adopting and implementing Internet-of-Things technology. As a result, organizations must be aware of and respect users' privacy rights, which are fundamental concepts in ensuring users' education, self-assurance, and confidence in connected devices, the Internet-of-Things, and related services. In addition, many users and the achievement of predetermined goals and outcomes require internet-connected devices to share and exchange data over the internet. As a result, it is a location where hackers can congregate in large numbers. In the event of a data breach and improper use of private information, an organization can be severely disrupted, resulting in significant financial and other risks.

- *Scalability*: Scalability refers to the capacity and ability of a network and system to handle increased traffic or to take advantage of the growing scale of a business environment to achieve higher performance. When it comes to Internet-of-Things, scalability is one of the most significant challenges because it describes the organization's efficiency and ability to deal with higher performance and the organization's long-term growth. The increasing number of internet-connected devices is having an impact on the performance of the network as well as the performance of the organization. As a result, many organizations are now rethinking their existing Internet-of-Things strategies by developing and updating the internet's foundational technologies. However, to achieve global scale, massive additional investments are required.
- *Intuitive Analysis and Decision-Making*: Finally, data collection and analysis are 2 steps that must be completed before an IoT implementation can be considered complete. The process of data analysis is dependent on cognitive models and technological advancements. Specific parameters in the data analysis process cause intelligent actions to be implemented in the context of the Internet-of-Things.

V. Theoretical Implication

This study examines the significant issues and challenges associated with the Internet-of-Things in business management and their implications for business operations and performance. When it comes to perceptions of the Internet-of-Things' impact on business management and the considerations that organizations should consider when adopting and implementing Internet-of-Things solutions, the findings of this study can be considered a valuable source of information. For the implementation of the Internet-of-Things in business management, this study provides fundamental knowledge and theoretical background. Furthermore, it can be used as a starting point for future research that aims to understand basic concepts and aspects related to various benefits resulting from adopting the Internet-of-Things in business operations and management. The findings and results of this research work identify some previously unidentified gaps in the literature regarding the adoption and utilization of the Internet-of-Things for business management.

VI. Managerial and Practical Implications

According to this research, the issues and challenges associated with the Internet-of-Things in business management are being investigated. Specifically, the findings of this study indicate that when the management allows unauthorized and insufficiently protected access to users' private data, it exposes the systems to risks and harmful attacks that compromise the system's security and safety, as well as the personal data's security and safety. As a result, it is necessary to develop an appropriate security system to protect personal information and data, or else users will not be able to use the internet-based system. Because this study provides a thorough understanding of data security and privacy, managers and organizations should adopt and implement Internet-of-Things while taking appropriate security precautions. This research study will aid researchers and management in developing a unique security framework to minimize and eliminate security and privacy risks, as well as the adoption of the Internet-of-Things using new communication tools and technologies, as well as the application of Internet-of-Things technologies in business management practices. A significant aspect of future research that provides practical aspects related to the adoption of Internet-of-Things in business management and its advantages can be derived from this study, which provides fundamental knowledge of the challenges that arise in adopting and implementing Internet-of-Things for various business activities and management functions. According to the findings of this study, various managerial considerations should be given to the selection of Internet-of-Things technology, which has the potential to have technological implications for organizations. To improve the capabilities of management functions, this study provides important and relevant information to managers and future users to define, accept, develop, and promote the Internet-of-Things systems to enhance their capabilities.

TABLE I: SUMMARY OF SOME OF THE RELATED WORKS

The Researcher and The Research Title	The Aim of the Study	Techniques	Solutions	The Results
Falguni Jindal, Rishabh Jamar, Prathamesh Churi (2018), Future and Challenges of Internet-of-Things	Identify the future challenges of Internet-of-Things	Data from Internet-of-Things sensors.	The authors have utilized the sensors data for secure service infrastructure with remote control options.	Security and data privacy will have a significant role in reducing the future challenges of the Internet-of-Things.
Zainab, Hesham A. Ali, Mahmoud M. Badawy (2015), Internet-of-Things (IoT): Definitions, Challenges and Recent Research Directions	Reviewing the main challenges of the Internet-of-Things environment by concentrating on the current research directions in this topic	Wireless telecommunication network.	The researcher has reviewed the main idea to design the structure of the Internet-of-Things that considers three dimensions: independent network, information items, and intelligent applications.	The Internet-of-things environment has various challenges mainly divided into 2 parts, i.e., General challenges and Unique challenges, which was affected the performance of The Internet-of-Things.
Ali Kamali (2018), A Fuzzy Analytic Network Process (FANP) Approach for Prioritizing Internet-of-Things Challenges in Iran	Identify and prioritize the Internet-of-Things technology development challenges in Iran.	Fuzzy analytic network process (FANP).	The researcher utilized the fuzzy analytic network process (FANP), which was applied to identify Iran's essential Internet-of-Things technology development challenges.	Technological, privacy, and security challenges are the crucial factors that affect the Internet-of-Things.
Parvaneh Asghari (2018), Internet-of-Things Applications: A Systematic Review	Categorize analytically and statistically, and analyze the current research techniques on Internet-of-Things applications.	Systematic Literature Review (SLR).	The researcher has used systematic literature review method to identify the advantages and limitations of every study review and identify the future challenges of Internet-of-Things.	This study identifies the various advantages and weaknesses of the previous paper, it also finds out the future research challenges and issues in internet of things applications.
Habib Ur Rehman (2017), Future Applications and Research Challenges of IOT	Examines the impact of the Internet-of-Things on the growth of the economy and the various applications of the Internet-of-Things.	Internet-of-Things applications domains.	The researcher has used the internet of things application to determine the impact of various Internet-of-Things applications on the economy.	Internet-of-Things will give humans comfort and improve the efficiency of things and make them intelligent.
Abubaker Haddud, A, DeSouza, A. R. Khare, A and Lee, H, (2017), Examining Potential Benefits and Challenges Associated with the Internet-of-Things Integration in Supply Chains	Examine the impact of adopting the Internet-of-Things in organizational supply chains to verify potential vital benefits and challenges in the literature.	Online survey and researcher select 87 participants as a sample, The IBM Statistical Package for the Social Sciences (SPSS) version 21 for analysis of data.	The researcher has utilized a valid and reliable research instrument that included 65 scale items.	Internet-of-Things can give significant advantages for achieving competitive advantages for organizations and their supply chains.

VII. Observation and Conclusion

This paper aims to identify and analyze the various issues and challenges associated with the Internet-of-Things in the context of business management. Every organization undergoes changes in business activities and the surrounding environment, such as technological innovation, to improve organizational functions. By utilizing the Internet-of-Things, businesses can reduce future uncertainty while also improving organizational management. The business collected data from sensors, mobile networks, official websites, and consumer databases. Because the business collected essential data from various devices, networks, and data centres in this competitive environment, the organization required a secure and safe service infrastructure that ensured the security and privacy of data. The Internet-of-Things, a combination of the Internet and physical objects, improves business management and business activities and the quality of all human life. According to the findings of this study, the Internet and things assist in business processes and activities. However, they also improve the efficiency and productivity of the available things. Technology and internet-based activities are constantly being updated, and the changing nature of the Internet-of-Things will make it the most powerful tool for improving the capacity of business management to ensure the long-term success of an organization. Researchers have discussed various issues and challenges related to the Internet-of-Things (IoT) in business management in this research study. It also helps to improve overall management functions and activities such as production, supply chain management, customer relationship management, tracking product usage, and monitoring marketing activities and campaigns.

Additionally, the Internet-of-Things improves customer service by enabling real-time communication and focusing on meeting the expectations and needs of consumers. The Internet-of-Things can enable managers to forecast each customer's demand to reduce the challenges of providing consumer services. As a result, managers can improve customer satisfaction by utilizing the Internet-of-Things. According to the findings of this study, the Internet-of-Things can also assist in cost reduction, time savings, and risk avoidance. Marketing managers should make use of the Internet-of-Things to educate consumers. Consumer personal data is collected and stored by businesses using internet-based technology. All this data is used solely for commercial purposes. Businesses should identify the challenges and issues associated with the adoption and implementation of Internet-of-Things tools and technologies in business management to improve the capabilities and efficiencies of their business activities and management. According to the findings of this study, the Internet-of-Things provides a more intelligent way to combine digital and physical things to improve the capabilities of management processes while maintaining a safe environment.

VIII. Recommendation

According to the findings of this study, organizations should identify their existing technological and organizational environments and focus on finding something different to minimize the issues and challenges associated with the adoption and implementation of the Internet-of-Things that should be done differently.

- Organizations should concentrate on overcoming issues and challenges, and risks to reduce the problems associated with adopting and implementing new innovative internet-based technology, such as the Internet-of-Things.
- To make business management more accessible and more effective, developing some standard systems for the Internet-of-Things is necessary. For this purpose, organizations should concentrate on security architectures that meet all the requirements for security and privacy issues in the context of Internet-of-Things.
- Privacy policies for all applications, devices, and protocols that the organization should use should be developed and enforced by the organization.
- With the help of new technology, there should be a requirement for global accountability of legal aspects to minimize privacy and security challenges.
- The organization should emphasize extensive data analysis, personalization, dynamic analysis, and search and exploration.
- Constraints and Prospective Orientation.

This research examines various previous research papers and the researcher's perceptions of the issues and challenges associated with the Internet-of-Things in the context of business organizations. Future research may examine the advantages and drawbacks of Internet-of-Things when adopting and implementing Internet-of-Things within organizations, using primary data sources such as questionnaires, surveys, and observational data. This may aid in the identification of potential benefits and limitations that were not addressed in this study, which is beneficial.

The study's primary limitation was that it examined the issues and challenges associated with the Internet-of-Things in business management but did not explain how organizations should adopt the Internet-of-Things in business management. As a result, future research may investigate this topic.

Various studies should be carried out to examine the security, privacy, and transparency of the technology being used. It will assist the manager, users, and future researchers resolve the various challenges that have arisen because of this research.

The author draws their conclusions based on secondary data, which is limited. It would be preferable to collect primary data, opinions, and viewpoints from experts and users who have the first-hand experience with the challenges and issues associated with the Internet-of-Things in the management of businesses. This paper will also assist practitioners and researchers in understanding the various issues and research challenges associated with the Internet-of-Things in management, which will help develop research trends for the future.

References

[1] L. Tawalbeh, F. Muheidat, M. Tawalbeh, and M. Quwaider, "IoT privacy and security: challenges and solutions," *Applied Sciences*, vol. 10, no. 12, p. 4102, 2020, doi: 10.3390/app10124102. [Online]. Available: www.mdpi.com/journal/applsci

[2] P. Asghari, A. M. Rahmani, and H. H. Seyyed Javadi, "Internet-of-things applications: A systematic review," *Computer Networks*, vol. 148, pp. 241-261, 2019.

[3] S. Bhat, O. Bhat, and P. Gokhale, "Applications of IoT and IoT: Vision 2020," *International Advanced Research Journal in Science, Engineering and Technology* (ISO 3297:2007 Certified), vol. 5, no. 1, Jan. 2018, ISSN [Online] 2393-8021, ISSN [Print] 2394-1588.

[4] A. Haddud, A. DeSouza, A. Khare, and H. Lee, "Examining potential benefits andchallenges associated with the Internet-of Things integration in supply chains," *Journal of Manufacturing Technology Management*, vol. 28, no. 8, pp. 1055-1085, 2017. [Online]. Available: http://dx.doi.org/10.1108/jmtm 0520170094

[5] A. K. Mohammadzadeh, S. Ghafoori, A. Mohammadian, R. Mohammadkazemi, B. Mahbanooei, and R. Ghasemi, "A Fuzzy Analytic Network Process [FANP] approach for prioritizing Internet-of-Things challenges in Iran," *Technology in Society*, vol. 53, pp. 124-134, May 2018. [Online]. Available: https://doi.org/10.1016/j.techsoc.2018.01.007

[6] F. Jindal, R. Jamar, and P. Churi, "Future and challenges of Internet-of-Things," *International Journal of Computer Science & Information Technology [IJCSIT]*, vol. 10, no. 2, pp. 13-25, 2018, doi: 10.5121/ijcsit.2018.10202.

[7] Z. H. Ali, H. A. Ali, and M. M. Badawy, "Internet-of-Things [IoT]: Definitions, challenges and recent research directions," *International Journal of Computer Applications*, vol. 128, no. 1, pp. 37-47, 2015, ISSN: 0975-8887.

[8] H. Ur Rehman, Mohd. Asif, and M. Ahmad, "Future applications and research challenges of IOT," Conference Paper, Dec. 2017, doi: 10.1109/ICICT.2017.8320166.

[9] A. Khanna, and S. Kaur, "Internet-of-Things [IoT], applications and challenges: A comprehensive review," *Wireless Personal Communications*, vol. 114, pp. 1687-1762, 2020. [Online]. Available: https://doi.org/10.1007/s11277-020-07446-4

[10] N. Angelova, G. Kiryakova, and L. Yordanova, "The great impact of Internet-of-Things on business," *Trakia Journal of Sciences*, vol. 15, no. suppl. 1, pp. 406-412, 2017, doi: 10.15547/tjs.2017.s.01.068, ISSN [Print] 1313-7069, ISSN [Online] 1313-3551. [Online]. Available: http://www.uni-sz.bg

International Journal of Distributed and Cloud Computing
9 (1) 2021, 36-48
http://www.publishingindia.com/ijdcc

Processor Allocation Two-Dimensional Mesh-Based Multi-Computer Systems: Protecting against Over-Partitioning and Maintaining Certain System Utilization Level

Sulieman A. Bani-Ahmad
Professor of Data Science, Department of Intelligent Systems, Faculty of Artificial Intelligence Al-Balqa Applied University, Jordan. Email: sulieman@bau.edu.jo

Abstract: Gradual Request-Partitioning-Based (GRPB) processor allocation strategies try to remedy the problem internal fragmentation by having requests get be allocated non-contiguously in the form of gradually produced blocks in case contiguous allocation is not possible. In this article, we experimentally show that GRPB techniques suffer from the problem of over-partitioning, and thus, negatively affects key performance indicators of multi-computer systems. This paper also proposes a framework for using fuzzy logic to control specific key performance levels in two-dimensional mesh-based multi-computer systems. We demonstrated that it is possible to implement a fuzzy-based feedback control system to control a number of performance indicators in the parallel system. The proposed allocation scheme prevents over-splitting of parallel jobs by forcefully limit to the maximum number of blocks that can be assigned to any parallel job. This maximum number is referred to as the partitioning-bound. Preventing parallel processes from getting over-partitioned is vitally important as this (i) reduces the distance of inter-processor source-to-destination communication and (ii) helps in avoiding the "inter-process interference"; the main cause of communication contention. Fuzzy-based allocation provides a mechanism to dynamically control the partitioning-bound. Instead of pre-setting and having a fixed partitioning-bound level, we provide a way to set this value at run time taking into consideration the current status to multi-computer system.

Keywords: Fuzzy-control, Gradual request-partitioning, Internal fragmentation, Multi-computer systems, Non-contiguous processor allocation.

I. Introduction

Two-Dimensional Mesh-based Multi-computer consists of a set of processing elements connected through a 2D mesh-like "interconnection network". In a typical mesh interconnection network the processing elements are arranged as the nodes of a two-dimensional grid. In this grid, direct point-to-point links exist connecting each node to its adjacent nodes. Mesh multi-computers are suitable for computer problems that require parallel computing. This involves subdividing large computer problems into smaller ones, each executing in a different process. In such case, the execution of processes is carried out simultaneously each on a dedicated set of processing elements. Examples of such computer applications are matrix computations, image processing and problems [1, 2].

Processor allocation in a multi-computer system is concerned with the dedication of the requested frame of processing elements for a specific computer job in hand. Three categories of processor allocation schemes are studies in literature, those are; contiguous, non-contiguous and mixed.

One issue with contiguous processor allocation strategies in 2D mesh-based multi-computer systems is its low system utilization as they fail to eliminate the effect of external and internal fragmentation. Hence, these strategies provide noticeably low and limited performance, mainly; system utilization and response time [3, 4, 5].

In non-contiguous processor allocation, jobs that fail to be allocated in a contiguous manner are subdivided into smaller sub-blocks that are individually allocated in non-contiguous manner instead of waiting until a submesh of the requested size and orientation is available in the system. In practice, lifting the condition of contiguity helps reducing processor fragmentation, and pre-execution waiting delay, and significantly increases system utilization. However, this non-contiguity results in increasing the overall communication overhead due to message contention increase. The Contention is the situation where multiple processes attempt to transmit a message across one shared link simultaneously. The extra communication overhead load depends on how dispersed is a partitioned parallel request (referred to as the degree of non-contiguity) [5, 1].

Gradual Request-Partitioning-Based (GRPB) processor allocation strategies try to tackle this problem on non-contiguity by gradually partitioning job requests and non-contiguously allocating them in the form of blocks (in case contiguous allocation is not possible). Studies demonstrate that GRPB allocation strategies demonstrate having the advantages of both contiguous allocation (e.g., low communication latency) and non-contiguous allocation strategies (e.g., high system utilization and throughput). This is achieved through preserving a *relatively acceptable* level of contiguity between allocated blocks of a parallel job.

Two groups of GRPB allocation strategies are studies in literature; (i) the "Adaptive Non-Contiguous Allocation" (ANCA) and (ii) the "Bounded-Gradual-Partitioning" (BGP) allocation. Those allocation schemes attempt to tackle the problem of internal and external fragmentation through allowing non-contiguity allocated submeshes [6, 7]. In ANCA, for instance, this is done by splitting the requested frame into two sub-frames of *equal sizes* (if possible). Splitting in this scheme is best applied to the longest dimension of the two dimensions of the request. In BGP, however, this is achieved by recursive partitioning the requested frame into two sub-frames: one is *large* and the other is *small*. BGP scheme prevents over-splitting of parallel jobs by limiting the maximum job dispersal represented by preventing the number of non-contiguous blocks assigned to any served parallel job from exceeding specific maximum value. This maximum value is referred to as the *partitioning-bound*.

Preventing parallel processes from getting over-partitioned is vitally important in shortening communication source-to-destination path and in avoiding the inter-process interference. In this work, we aim at providing mechanisms to control the partitioning-bound. Instead of pre-setting and having a fixed partitioning-bound level, we intend to provide a way to set this value at run time taking into consideration the current status to multi-computer system. For example, one way is to consider the current (or cumulative) system utilization to dynamically and automatically decide this value.

Consider the following 2D mesh multi-computer in Fig. 1. In this figure, the multi-computer consists of a mesh of size 5 x 6; a total of 30 processing units. Black nodes represent busy elements (already allocated and occupied). White nodes are available ones. Assume further that a process of size 2 x 2 has arrived and is waiting to be served. Under this scenario, contiguous processor allocation strategies fail to allocate this job despite the fact that enough free-processes are available in the system. BGP allocation strategies lift the contiguity condition and allocate multiple blocks. In this case, two blocks of size 1 x 2 each can be successfully allocated.

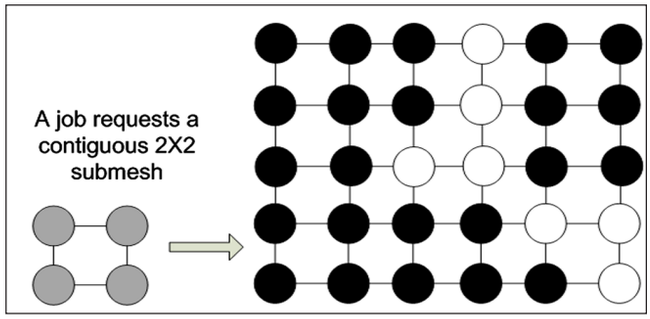


Fig. 1: A Multi-Computer System with 5 x 6 2D Mesh

One issue with BGP allocation strategies is that they may aggressively partition or over-partition parallel jobs by assigning them many sub-blocks. Preventing this phenomenon is important as it reduces the degree of non-contiguity. As a result, the distance of inter-processor communication path becomes shorter and the inter-process interference that creates communication contention reduces.

Our goal in this study is to provide mechanisms to control the partitioning-bound. Instead of pre-setting the partitioning-bound level, we would like to decide this value at run time. The proposed mechanisms take into consideration the current (or cumulative) system utilization and probably other performance metrics to decide this value.

The rest of this paper is organised as follows: Literature review is presented in Section II. The proposed allocation strategy is explained in Section III. In Section IV and V the experimental settings and results are presented. Section V compares the proposed approach to the already presented work in literature. Section VI is the conclusion.

II. Literature Review

A. Contiguous Processor Allocation

Contiguous allocation strategies work by allocating a single and contiguous block of idle processing elements for the execution of the parallel job in hand [8, 3, 4, 5, 9, 10, 11]. In fact, contiguity of processing elements minimizes the source-to-destination distance of inter-processor communication paths and, thus, in tackles the phenomenon of inter-process interference that causes "communication contention". Further, contiguity help reduce the power consumption of multiprocessor architectures, e.g., 2D mesh multi-computers [12, 13].

Other advantages of this category of allocation strategies are:

- Allocated subframes preserve the same exact shape (i.e., topology) as the underlying multi-computer system.

- The number of multi-computers allocated to a particular parallel job exactly matches that parallel job request with no internal fragmentation [14]. Thus, in mesh-connected multi-computers, parallel jobs are allocated to submeshes [14, 3, 4, 5, 9, 10].

Next is a list of well-known contiguous processor allocation strategies:

Two-Dimensional Buddy: In this allocation mechanism, the parallel system is assumed to be square-shaped with the two side-lengths be equal to a power of two (e.g., 2^i). In this case, the size of incoming parallel requests is rounded-up to a square-shape with side-lengths as the nearest power of two. Consequently, this allocation mechanism suffers from the problem of internal fragmentation. Rounding up the sides of parallel requests results in allocating submeshes larger than the actually requested submeshes [9].

In [10], the Frame Sliding (FS) allocation mechanism is proposed to solve the problem of internal fragmentation. This is achieved by allowing parallel requests of any arbitrary shape and size to be allocated. That is; parallel requests are viewed as frames of rectangular shapes. The Frame Sliding algorithm works by sliding each of the requested frames across the system's 2-D or 3-D mesh to examine for an unallocated submesh in the system [10].

In [11], the author has proposed The Best-Fit (BF) and First-Fit (FF) the allocation schemes in order to guarantee the recognition of a unallocated sub-meshes taking into consideration the orientation factor. Each of the 2 approaches works by scanning the entire system's mesh. Identified and allocated submeshes follow the same size, shape, and orientation of the requested frame. In case a frame with the same shape and size exists in the parallel system but the orientation is different, these two strategies fail. To solve this issue, the Adaptive-Scan (AS) scheme is proposed in [15]. The adaptive-scan changes solve this problem by rotating and changing the orientation of the requested submesh being searched for in case the required sub-mesh with the requested orientation is not available.

B. Non-Contiguous Processor Allocation

Studies show that allocation schemes with high *recognition ability* for idle sub-meshes can significantly improve the probability of successfully assigning parallel requests into the system and, consequently, reduce the job pre-execution waiting delay [16].

Further performance-related studies show that performance of already existing contiguous allocation strategies a significant improvement cannot be improved [4, 14, 17]. In practice, the system utilisation performance indicator can significantly degrade because of external fragmentation problem [4, 14]. External fragmentation occurs when parallel jobs fail to be served while enough unallocated processing elements in the parallel system exist.

Hardware advances in 2D-mesh interconnection network, such as the use of wormhole routing technique and the availability of high-performance switching techniques, have resulted in decreasing the communication latency and in making this parameter be less dependent on the distance between the communicating nodes [4, 7, 14]). This has resulted in making non-contiguous allocation plausible in parallel systems. By elimination of the restriction of contiguity in processor allocation, the pre-execution waiting time becomes significantly less. The "Multiple Buddy System" (MBS) and the Paging allocation strategies are examples of non-contiguous allocation algorithms proposed in the literature.

In the MBS strategy introduced in [18], the 2-D mesh of the parallel system is divided into a set of disjoint sub-meshes of square-shape with widths and heights are of powers of two. The MBS algorithm works as follows: the number of required processors is factorized into base-4blocks. In case the requested frame is not available, MBS recursively breaks-down large requests into 4blocks until it produces sub-blocks of the desired overall original size of the parallel job.

In the Paging algorithm that is introduced in [19], the entire system's mesh is sub-divided into "virtual" pages in the form of square frames of side-lengths of 2^i, where i is a parameter referred to as the index parameter. The index parameter is a pre-assigned positive integer.

C. Request-Partitioning-Based (RPB) Allocation Schemes

Partially non-contiguous allocation strategies out-perform totally non-contiguous strategies in terms of jobs dispersal [6, 19]. In Paging algorithm, for example, a reasonable degree of contiguity is gained depending on the indexing scheme used. To illustrate, job dispersal is less when decreasing the index parameter. Increasing the index parameter, however, produces higher internal processor fragmentation [19].

Request-Partitioning-Based allocation reduces the problems of (i) internal and external fragmentation, (ii) low average system utilization (ASU), and (ii) low system throughput. This is achieved by permitting parallel requests to be gradually partitioned and allocated in a non-contiguous manner in case contiguous allocation is not successful [20]. Notice that, in general, the probability of successfully allocating small sub-frames is high.

In literature, two groups of RPB allocation schemes can be found; (i) the "Adaptive Non-Contiguous Allocation" (ANCA) [21] and the (ii) Bounded-Gradual-Partitioning (BGP) allocation [16, 22].

The ANCA algorithm works as follows: First, the algorithm attempts to do contiguous allocation, if this step fails, allocation is achieved by recursively splitting the requested frames into two sub-frames of equal sizes (roughly). Splitting is performed at the longest dimension of the two dimensions; the width and the height [14, 6].

In BGP, however, allocation of parallel jobs is achieved by gradually splitting the original frame into two sub-frames; one large and another small sub-frame of processing elements ([20]0a; [20]0b).

In [2], the authors proposes the Minimum Interference Paging (MIP) non-contiguous processor allocation strategy which attempts to reduce the distances among processors allocated using a paging variant that chooses a set of processors with the lowest distance between the first and last allocated processors, where the distance is the number of processors between the first allocated node and last allocated node. MIP is comparatively evaluated with Paging (0), Multiple Buddy System (MBS) and Adaptive Non-contiguous Allocation (ANCA). The results show that MIP exhibits superior performance in terms of average turnaround time of jobs.

In [21], the author suggests a new noncontiguous allocation strategy called All Request Shapes Greedy Available Busy List (ARSGABL for short). ARSGABL considers all possible request shapes when attempting allocation for a job request experimental simulation-based results confirm that the ARSGABL strategy improves system performance in terms of the average turnaround times of jobs.

D. Unbounded vs Bounded Gradual Request Partitioning (GRP)

GRP-based allocation schemes partitions requests at the longest dimension. This approach works as follows: first, it attempts to find a single and idle submesh using any contiguous allocation strategy (First-Fit or Best-Fit for instance). If contiguous allocation fails and enough number of free processing elements is available in the system, the request is splitted into two sub-frames that are allocated based on the same contiguous allocation approach mentioned above. These steps are repeated until the job request is satisfied. The two new sub-requests are reclusively served until success (the best-fit algorithm is used in the experimental part of this research). This recursive procedure continues until the current job's request is fulfilled. Notice that this approach places *no limit to* the job dispersal.

A pseudo code for the "unbounded GRP-based allocation strategy" is sketched in Fig. 2. The key idea in here is to try to allocate the largest possible contiguous sub-meshes.

```
Procedure GRP-BF(a, b):
Begin
    JobSize = a × b
    If (number of free processors < JobSize) Return failure
    List AllocatedPIDs={}; // PIDs allocate to the job
    Return GRP-BFAllocate (a, b, &AllocatedPIDs);
End

Procedure GRP-BFAllocate (a, b, AllocatedPIDs)
Begin
    S(x, y) = FIND_BF (S(a, b);
    If (S(x, y)  != null)
    {
        Add the PIDs of S to the list AllocatedPIDs;
        Return success;
    }
    Else
    {
        If(a>=b)
            α1= a-1;  β1=b; α2= 1;  β2=b;
        else
            α1= a;  β1=b-1; α2= a;  β2=1;
            Return GRP-BFAllocate (α1, β1, &AllocatedPIDs);
            Return GRP-BFAllocate (α2, β2, &AllocatedPIDs);
    }
End
```

Fig. 2: Pseudo Code for the Unbounded GRP-BF Allocation Strategy

Experiments show that "unbounded GRP-based" allocation strategies may cause jobs get highly dispersed due to *over-partitioning* depending on the temporal state of the multi-computer system mesh. Theoretically, a request for n processors can be partitioned into n blocks of size one processing element each! The negative effect of this phenomenon appears more at heavy intra-process communication loads [VE03]. To remedy this drawback, the *Bounded* GRP-BF is proposed in [20, 22]. This processor allocation algorithm is sketched in Fig. 3. The key difference between Fig. 2 and Fig. 3 is that we put an upper bound on the "number of blocks" into which requests can be subdivided into. This helps prevention of over-partitioning requests.

```
BlockCount; // block count for the current job request
MaxBlockCount; // maximum allowed blocks per job
Procedure BGRP-BF(a, b):
Begin
    JobSize = a × b
    If (number of free processors < JobSize) Return failure
    List AllocatedPIDs={}; // PIDs allocate to the job
    BlockCount=0;
    Return BGRP-BFAllocate(a, b, &AllocatedPIDs);
End

Procedure BGRP-BFAllocate (a, b, AllocatedPIDs)
Begin
    S(x, y) = FIND_BF (S(a, b);
    If (S(x, y)  != null)
    {
    Add the PIDs of S to the list AllocatedPIDs;
    BlockCount= BlockCount+1;
    Return success;
    }
    Else
    {
        If (BlockCount > MaxBlockCount) Return failure
        If(a>=b)
            α1= a-1;  β1=b; α2= 1;  β2=b;
        else
            α1= a;  β1=b-1; α2= a;  β2=1;
            Return BGRP-BFAllocate (α1, β1, &AllocatedPIDs);
            Return BGRP-BFAllocate (α2, β2, &AllocatedPIDs);
    }
End
```

Fig. 3: Pseudo Code for the Bounded GRP-BF Allocation Strategy

III. Proposed Allocation Strategy: Bounded Gradual Request Partitioning with Controlled Job Dispersal

One observation on the pseudo code presented in Fig. 3 is that the upper bound *MaxBlockCount* is preset in the code. The user decides upon this value and the value remains unchanged regardless of performance indicators of the multi-computer system. In this work, we propose mechanisms that help dynamically decide upon this parameter in order to maintain specific targeted system performance level(s). For that, a new procedure is added to the pseudo code of Fig. 3. Fig. 4 represents the pseudo code with one more procedure added to the strategy, that is the Refresh MaxBlockCount() procedure. This procedure is responsible for calculating the new value of the *MaxBlockCount* variable. The new value of this variable should take into consideration the current status of the multi-computer system by calculating some key performance indicators that will be presented shortly. This procedure should be invoked after each successful allocation decision.

```
1  BlockCount; // block count for the current job
2  MaxBlockCount; // maximum allowed blocks per job
3
4  Procedure RefreshMaxBlockCount()
5  Begin
6      . . .
7  End
8
9  Procedure BGRP-BF(a, b):
10 Begin
11     JobSize = a × b
12     If (number of free processors < JobSize) Return failure
13     List AllocatedPIDs={}; // PIDs allocate to the job
14     BlockCount=0;
15     Return BGRP-BFAllocate(a, b, &AllocatedPIDs);
16 End
17
18 Procedure BGRP-BFAllocate (a, b, AllocatedPIDs)
19 Begin
20     S(x, y) = FIND_BF (S(a, b);
21     If (S(x, y)  != null)
22     {
23         Add the PIDs of S to the list AllocatedPIDs;
24         BlockCount= BlockCount+1;
25         RefreshMaxBlockCount();
26         Return success;
27     }
28     Else
29     {
30         If (BlockCount > MaxBlockCount) Return failure
31         If(a>=b)
32             α1= a-1;  β1=b; α2= 1;  β2=b;
33         else
34             α1= a;  β1=b-1; α2= a;  β2=1;
35             Return BGRP-BFAllocate (α1, β1, &AllocatedPIDs);
36             Return BGRP-BFAllocate (α2, β2, &AllocatedPIDs);
37     }
38 End
```

Fig. 4: Pseudo Code for the Bounded GRP-BF Allocation Strategy with the MaxBlockCount Parameter Automatically Determined and Calculated at Run Time

A. Multi-Computer System's Key Performance Indicators

The following is a list of performance indicators that we may intend to control in the multi-computer system. Those indicators should contribute to our decision of determining the value of the MaxBlockCount parameter in the allocation strategy code.

- *"Mean Response Time" (MRT):* RT is the time elapsed from the arrival of a job until the first real output produced from that job.
- *"Mean Job Service Time" (MJST):* The ST of a job is the time elapsed from the successful allocation of the job until the moment the job releases the system resources after it finishes.
- *(Mean or Temporal) "Percent-System-Utilization" (PSU):* Percent-system-utilization of a multi-computer system is defined as the percentage of the busy or allocated processors within a system, this value between 0 and 1 (normalized by dividing by the system's mesh size).
- *"Mean-Packet-Blocking-Time" (MPBT):* For a specific message, this performance indicator is defined as the total time the head of the communication messages (using wormhole routing) is blocked at each and every station over the source-to-destination path.
- *"Mean Packet Latency" (MPL):* In wormhole routing, each message is subdivided into a number of packets. MPL for a specific message is defined as the average of the time that all packets within a parallel job will be sent between processors (From source processor to distention processor).
- *Mean-Number-of-Blocks-Per-Job (MBPJ):* MBPJ is defined as the number of disjoint sub-frames assigned to the parallel jobs that were served since the system is put into service.

Next is the nature of relationship between the *MaxBlockCount* parameter in Fig. 3 and each of the six performance indicators listed before.

Table I: The Nature of Relationship between the MaxBlockCount Parameter and Keyparallel System Performance Indicators

Performance Indicator	Relationship with the *MaxBlockCount* Parameter
MRT	Indirect
MJST	Indirect
PSU	Direct
MPBT	Direct
MPL	Direct
MBPJ	Direct

Notice that, in general, as the *MaxBlockCount* parameter increases the response time and service time of parallel jobs decrease. The reason is that the pre-execution waiting time of jobs significantly decreases because they get allocated early due to lifting the condition of contiguity.

Comparatively, increasing the *MaxBlockCount* parameter increases job dispersal, this causes the number of Blocks Per Job, packet latency time, and packet blocking time all

increase as well. This creates a direct relationship between the *MaxBlockCount* parameter value and those three performance indicators. Notice that, increasing the *MaxBlockCount* parameter value causes the Percent-system-utilization performance indicator to increase as well.

B. Fuzzy Control Systems: Basics

Fuzzy logic is a technique used to handle lexically imprecise propositions with elastic constraints [23]. Practically, fuzzy logic provides a framework for handling sets of rules expressed in an imprecise form. One very basic concept of fuzzy logic is the linguistic variables. Intrinsically, fuzzy logic uses the whole interval of real numbers between zero (pure False) and one (pure True). For example, expressing temperature is possible through a linguistic variable that takes values such as high, low and quite low. Such linguistic variables are naturally embedded in the fuzzy rules of a fuzzy controller and in our nature as humans. In fact, with fuzzy logic processing expressions from natural language, such as 'small', 'large', and 'approximately equal' becomes possible in computers [23].

C. Definitions: Fuzzy Logic Basic Terms

i) Fuzzy Sets and Degree of Memberships

A fuzzy set is defined as a set whose member elements have degrees of membership. In classical sets, the membership of member elements in a given set is assessed in binary terms (either an object is a member of a set or not). Fuzzy sets are different, in the set of "young" people for example, a new born baby will is definitely a member of this fuzzy set, similarly, a 100 years old human is not a member of this set, humans of ages 20, 30, or 40 years partially members of this set with decreasing levels of membership. The grade of membership for all members of a fuzzy set describes that fuzzy set.

An object's grade of membership is a real number in the interval between 0 and 1 (Fig. 5). The Grade of membership is often denoted by the Greek letter μ. Higher number implies higher membership.

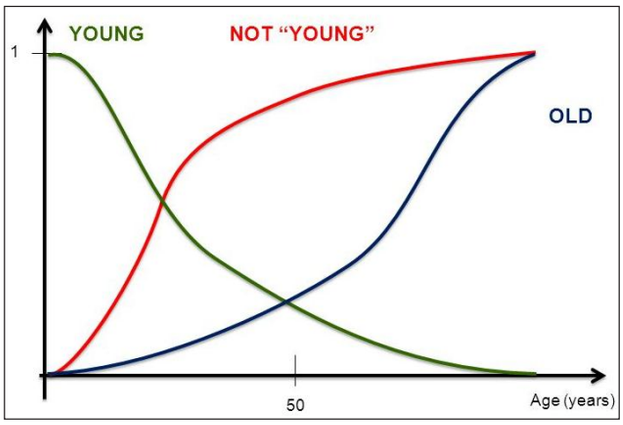


Fig. 5: The Sets More or "Young", "Not Young", and "Old"

Objects within a fuzzy set are taken from a 'universe of discourse' or 'universe' for short. A universe contains all objects that can come into consideration which, depends on the context of course. In describing 'taste' for instance, the elements of this term are taken from a universe such as {bitter, sweet, sour, …}.

Every object or element in the universe of discourse belongs to fuzzy set to some grade as mentioned before. The function that relates a number to each element x of the universe of discourse is called the membership function and is denoted by μ(x).

ii) Continuous or Discrete Membership Functions

Two types of membership functions can be used: continuous or discrete. In the continuous form, the membership function is basically a mathematical function. A membership function is for example bell-shaped (also called a π-curve), s-shaped (called s-curve), or z-shaped (called z-curve or reverse s-curve). All membership functions in Fig. 5 are of the second type. A π-curve membership function in the same context of Fig. 5 would be the "middle age" function. In the discrete type of membership function and the universe are discrete members or points in a list (i.e., vector). Depending on the application, sometimes it can be more convenient to use discrete membership function with sampled (discrete) representation [24].

iii) Representing Continuous Membership Functions

Cosine functions can be used to represent a variety of s-shaped membership functions as follows:

$$s(x_l, x_r, x) = \left\{ \begin{array}{ll} 0 & , x < x_l \\ \frac{1}{2} + \frac{1}{2}\cos\left(\frac{x - x_r}{x_r - x_l}\pi\right) & , x_l \le x \le x_r \\ 1 & , x > x_r \end{array} \right\}$$

Where x_l is the left breakpoint, and x_r is the right breakpoint. The z-shaped curve is just a vertical reflection of this curve as follows:

$$z(x_l, x_r, x) = \left\{ \begin{array}{ll} 1 & , x < x_l \\ \frac{1}{2} + \frac{1}{2}\cos\left(\frac{x - x_l}{x_r - x_l}\pi\right) & , x_l \le x \le x_r \\ 0 & , x > x_r \end{array} \right\}$$

The π-shaped cure can be constructed as a combination of the s-shaped and z-shaped curves such that the peak of the new curve is flat in the interval $[x_3, x_4]$ as follows:

$$\pi(x_1, x_2, x_3, x_4, x) = \min(s(x_1, x_2, x), z(x_3, x_4, x))$$

Where x_1 is the left breakpoint, and x_2 is the right breakpoint.

A fuzzy set is normalized if any membership value in it is less than or equals to 1. To normalize a fuzzy set membership function (or vector), we divide each membership value by the largest membership in the complete set.

D. Fuzzy Control

The process of how fuzzy logic control systems work is shown in Fig. 6A and 6B. Firstly, a crisp set of input data are gathered

from the system to be controlled. In our case, the crisp input is the value of the performance indicator to be measured. After that, this crisp input is converted to a proper fuzzy set through the fuzzification step. Fuzzification implies using fuzzy linguistic variables, fuzzy linguistic terms and membership functions. Inference rules are then used to produce fuzzy outputs that are mapped to a crisp output, again using the proper membership functions, through the defuzzification step. Those steps are presented in Fig. 6 that shows the block diagram of a typical fuzzy control system. Notice that Fuzzy control systems use statements rather than equations. Those statements are represented as empirical rules [27, 28].

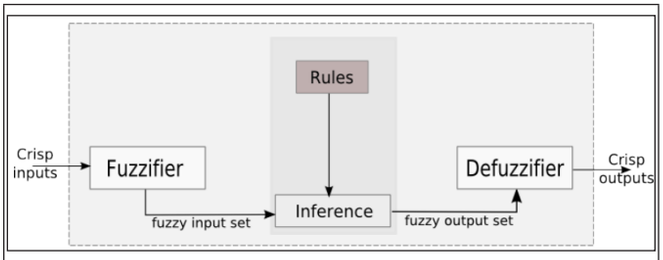


Fig. 6A: A Simple Fuzzy Control System

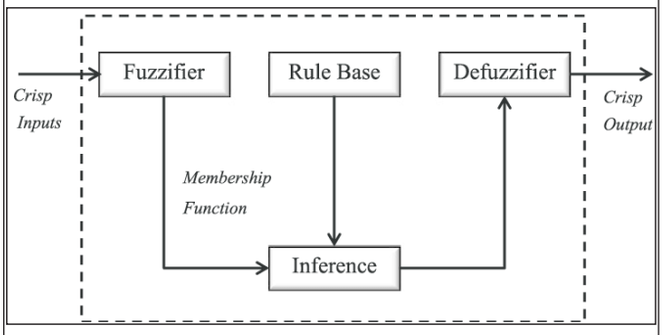


Fig. 6B: A Block Diagram of a Fuzzy Control System

Fig. 6

The fuzzy control system basically compares the detected crisp input values from the system environment and compares it to the targeted key performance indicators to compute the "steady-state error". The computed error (Fig. 7) is then used to modify specific system parameters that have a direct or indirect impact on the targeted performance indicators in the system in hand. The parameter of interest in our context is the maximum number of partitions allowed (*MaxBlockCount* in Fig. 3 above) [27, 28].

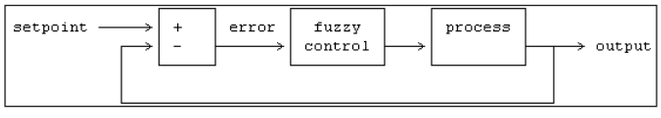


Fig. 7: Steady State Error for Closed-Loop Feedback Control System

IV. Experimental Setting

Every processor allocation mechanisms studied in this research is implemented in C programming language and included into the unix-based parallel simulation tool termed ProcSimity and introduced in [25, 26]. Table II lists the basic experimental setting of the simulated system:

Table II: The Nature of Relationship between the *MaxBlockCount* Parameter and Keyparallel System Performance Indicators

Simulator Setting	Value
Number of simulated jobs per run	1000
Number of runs	10
Job run-mode	Run to completion (no time-sharing)
Confidence level of calculated performance indicators	95%
Relative error of calculated performance indicators	5%
Simulated parallel jobs communication pattern	All-to-all communication pattern [18]
Scheduling mechanism used	First-Come First-Serve, FCFS
Simulated parallel system size	32 x 32 computing units (a total of 1024 computing elements).
Simulated parallel job size distribution	The uniform distribution with 1 to 10 units for each dimension
The set of performance factors considered	MRT, JST, the Simulation "Finish Time" (FT), PSU, MPBT, MPL, and MBPJ

V. Experimental Results and Observations

In this section, we report the experimental results of the proposed allocation strategies after we implemented them. This section is of three parts. In the first part, we comparatively evaluate the proposed techniques with the already available allocation strategies from literature. In this part, we do our experiments in a loaded system (with mean inter-arrival time between jobs be 1). In the second part, we present our observations on the impact of the temporal system load on the key performance indicators of the proposed and the present strategies from literature. In the third part, we present our observations on the impact of the communication load (i.e., # of messages/job) on the key performance indicators of the proposed and the present strategies from literature.

A. Part I: Performance in Loaded System

In this section, we comparatively evaluate the proposed techniques with the already available allocation strategies from literature. To this end, we do our experiments in a loaded system (with mean inter-arrival time between jobs be 1). In all experiments of this part, the number of messages per parallel job is fixed at 30 messages.

In the following table, we present some clarifications about the allocation strategies under study and their abbreviated names in the figures.

TABLE III: THE SET OF ALLOCATION STRATEGIES UNDER STUDY AND THEIR ABBREVIATED NAMES IN DEPICTED GRAPHS

Allocation Strategy	Information about the Strategy
"BGP(1,0)"	The Best-Fit allocation strategy [11]
"BGP(2,0)"	The Bounded Gradual Partitioning allocation strategy [6, 20, 22] MaxBlockCount is fixed at 2
"BGP(4,0)"	The Bounded Gradual Partitioning allocation strategy [6, 20, 22] MaxBlockCount is fixed at 4
"BGP(8,0)"	The Bounded Gradual Partitioning allocation strategy [6, 20, 22] MaxBlockCount is fixed at 8
"BGP(20,0)"	The Bounded Gradual Partitioning allocation strategy [6, 20, 22] MaxBlockCount is fixed at 20
"BGP(60,1)"	The proposed Bounded Gradual Partitioning allocation strategy MaxBlockCount is dynamically decided to make the temporal Percent-system-utilization be 60%.
"BGP(80,1)"	The proposed Bounded Gradual Partitioning allocation strategy MaxBlockCount is dynamically decided to make the temporal Percent-system-utilization be 80%.

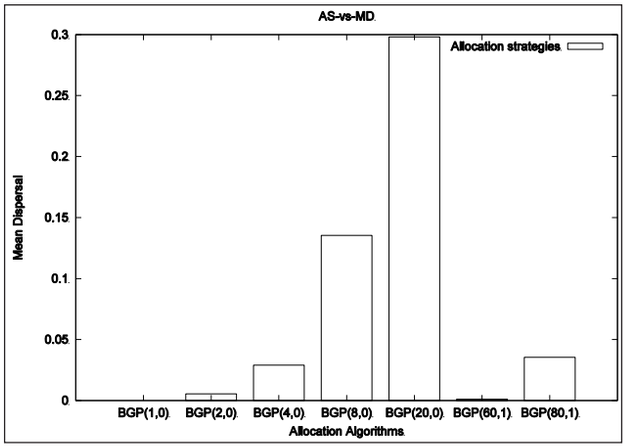


Fig. 8: Allocation Strategies versus Mean Dispersal

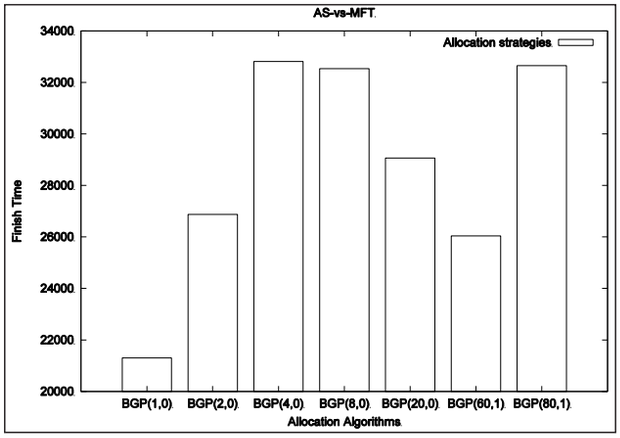


Fig. 9: Allocation Strategies versus Mean Finish Time

Fig. 8 depicts allocation strategies versus mean dispersal in a loaded system. The Bounded Gradual Partitioning allocation strategy where *MaxBlockCount* is fixed at 20 has shown the worst (highest) job dispersal. While the proposed allocation strategies (where *MaxBlockCount* is dynamically decided to make the temporal Percent-system-utilization be 60% and 80%) showed relatively good performance compared to the best allocation strategy in terms of this performance indicator; namely, the Best-Fit allocation strategy. Notice that the job dispersal of the Best-Fit strategy is zero as it assigns one single block for the parallel job.

Fig. 9 depicts allocation strategies versus mean simulation finish time in a loaded system. The Bounded Gradual Partitioning allocation strategy where *MaxBlockCount* is fixed at 20 has shown high mean simulation finish time. The same observation applies to all BGP allocation strategies. This is expected because the simulator simulates 1000 parallel jobs per run for a total of ten runs. Since BGP strategies increases Percent-system-utilization and since the job scheduler uses the FCFS scheduling mechanism, the simulator takes long finish time as relatively large parallel jobs has to remain in the system's waiting queue before they can be successfully allocated and served. This causes the RT and the ST of individual parallel jobs become relatively high. This is clear from Fig. 10 and 11 (shown next). Fig. 10 and 11 shows MJRT and MJST depicted against allocation strategies.

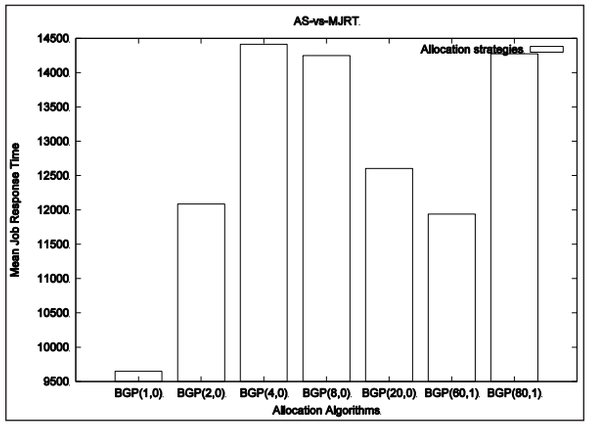


Fig. 10: Allocation Strategies versus MJRT

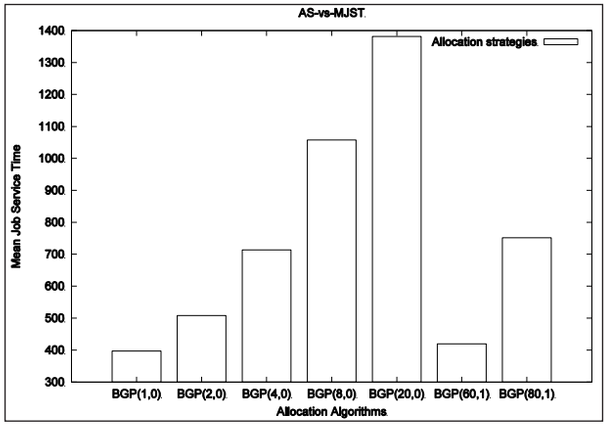


Fig. 11: Allocation Strategies versus Mean-Job-Service-Time

Fig. 12 and 13 depicts the allocation strategies under study versus MPBT and MPL in a loaded system. For a specific message, MPBT is computed as the total time the head of the communication messages (using wormhole routing) is blocked at each and every station over the source-to-destination path. Notice that in wormhole routing, each message is subdivided into a number of packets. MPL for a specific message is computed as the average of the time that all packets within a parallel job will be sent between processors (From source processor to distention processor). The Best Fist allocation strategy has shown the best (lowest) packet blocking time and packet latency of all allocation strategies. This is expected as this allocation strategy does not disperse parallel jobs. The worst allocation strategies according to these two performance indicators are the Bounded Gradual Partitioning strategies where *MaxBlockCount* is fixed at 8 and 20. The proposed allocation strategies have showed better performance and in the case of having the *MaxBlockCount* be dynamically determined to maintain 60% Percent-system-utilization has shown comparable performance to the Best-Fit allocation strategy.

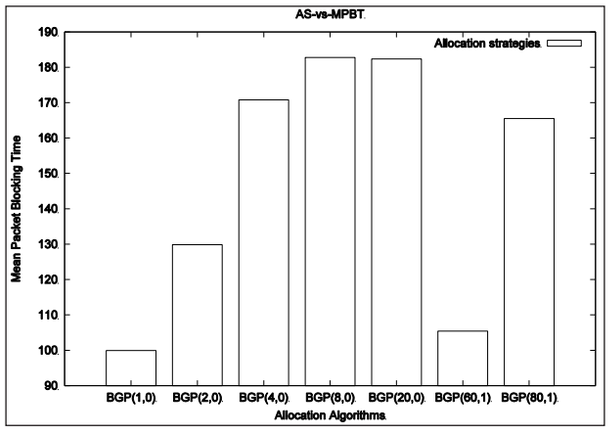


Fig. 12: Allocation Strategies versus MPBT

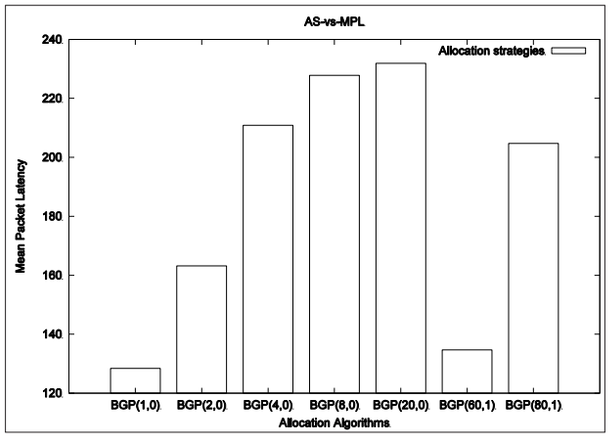


Fig. 13: Allocation Strategies versus Mean Packet Latency

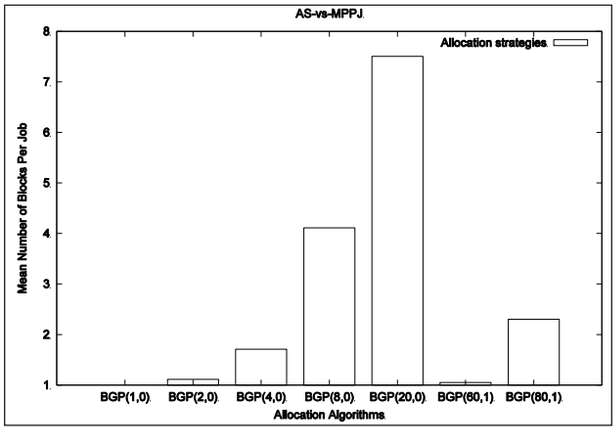


Fig. 14: Allocation Strategies versus MBPJ

Fig. 14 depicts allocation strategies versus MBPJ. MBPJ is computed as the number of *disjoint sub-frames* assigned to the parallel jobs that were served since the system is put into service. As discussed earlier, the Best-Fit allocation strategy do not disperse allocated parallel jobs, each job is assigned one single block. However, the worst allocation strategies according to this factor are the Bounded Gradual Partitioning strategies where *MaxBlockCount* is fixed at 8 and 20. The proposed allocation strategies, on the other hand, it has shown better performance and in the case of having the *MaxBlockCount* be dynamically determined to maintain 60% PSU has shown comparable performance to the Best-Fit allocation strategy.

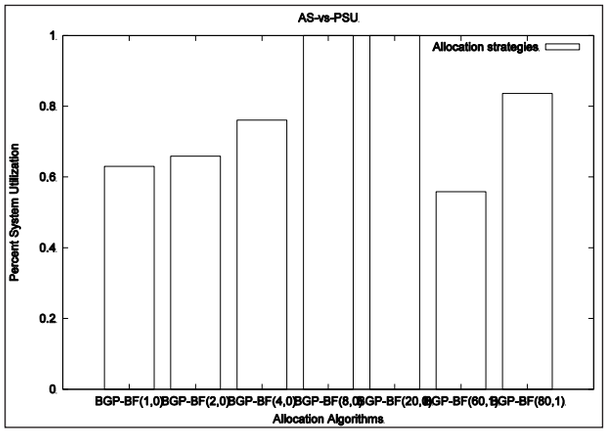


Fig. 15: Allocation Strategies versus Percent-System-Utilization

B. Part II: Performance over System Load

Fig. 16 depicts the system load versus mean number of blocs per job. Notice that as the system gets more loaded, the number of blocks assigned to served parallel jobs increases as the temporal system utilization of the multi-computer system increases. This makes it difficult to successfully allocate parallel jobs without dispersing them. This is clear in Fig. 17 that depicts System load versus mean dispersal. This, in turns, causes the ST of the job to increase. This observation can be noticed in Fig. 18 that depicts system load versus MJST.

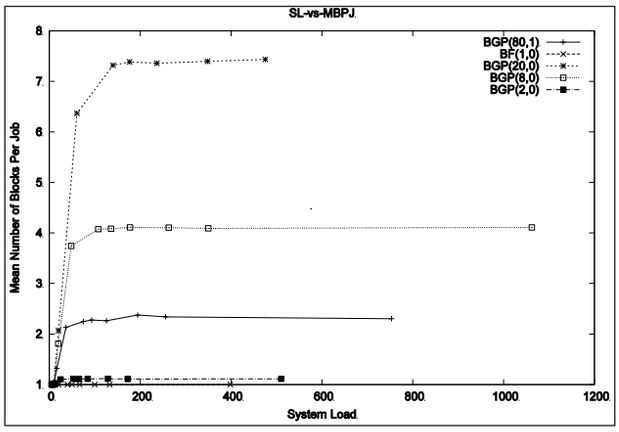


Fig. 16: System Load versus MBPJ

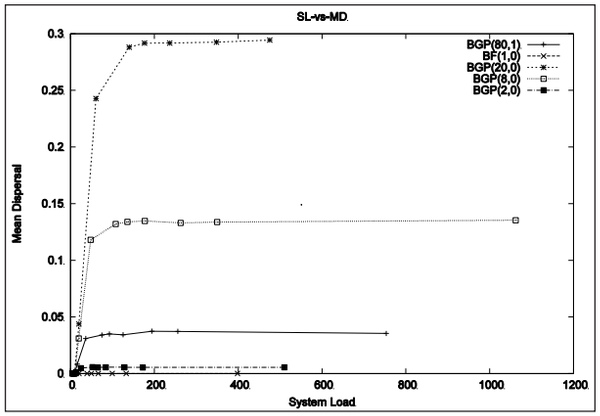


Fig. 17: System Load versus Mean Dispersal

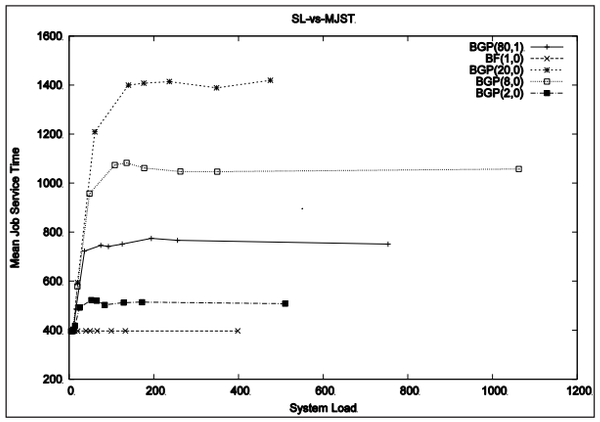


Fig. 18: System Load versus Mean-Job-Service-Time

C. Part III: Performance over Communication Load

In this section, we study the communication load impact on the proposed and examined allocation strategies in this study. In all experiments in this section, we use the well-known all-to-all communication pattern. This communication pattern is used as it produces high message contention. In all-to-all communication or broadcasting, each processing element serves as a source as well as a destination. A parallel process sends the same m-word-size message to every other process allocated to that particular parallel job.

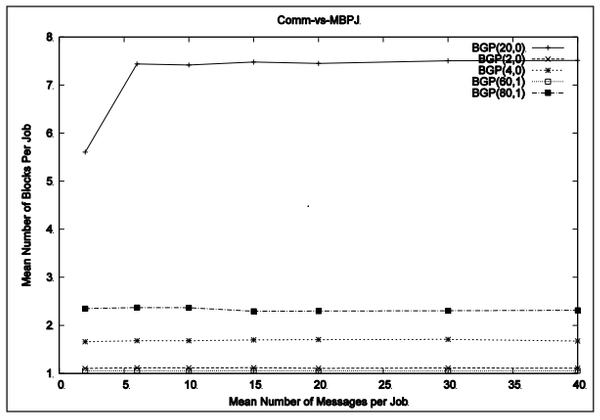


Fig. 19: Communication Load versus MBPJ

Fig. 19 depicts the communication load (represented by the number of messages per parallel job) versus MBPJ. And Fig. 20 depicts the communication load versus mean dispersal level of parallel jobs. We observe that the proposed allocation strategies have shown relatively good and comparable performance compared to the Bounded Gradual Partitioning strategies where *MaxBlockCount* is fixed at 8 and 20.

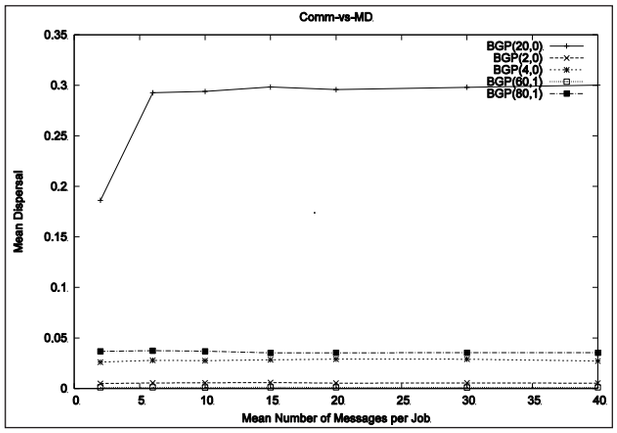


Fig. 20: Communication Load versus Mean Dispersal

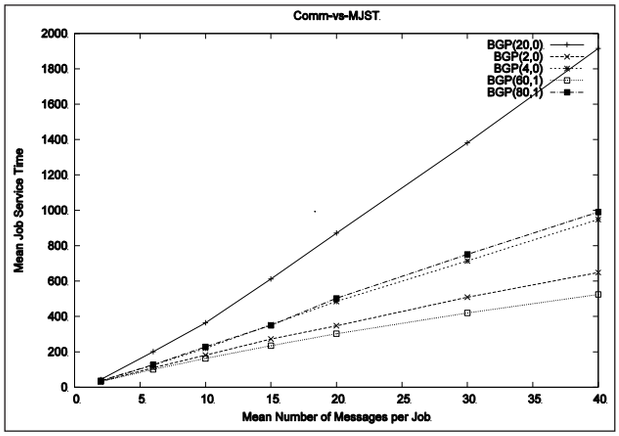


Fig. 21: Communication Load versus MJST

Fig. 19 depicts the communication load versus MPBT. Again, we notice that the proposed allocation strategies have shown relatively good and comparable performance compared to the Bounded Gradual Partitioning strategies where *MaxBlockCount* is fixed at 8 and 20. Those later strategies disperse parallel jobs and, thus, increase the packet latency and the packet blocking times due to message contention. This can also be observed in Fig. 20 that depicts communication load versus MPL.

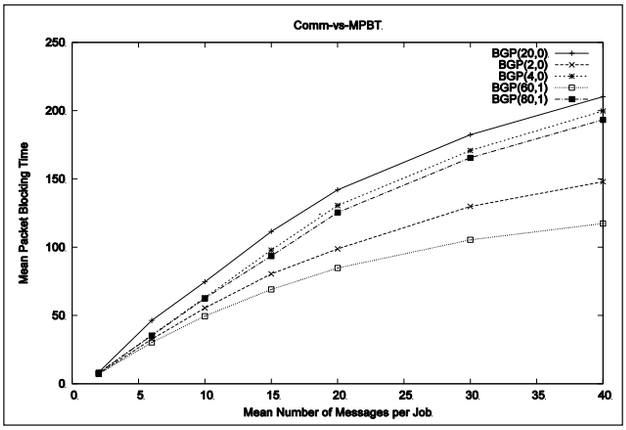


Fig. 22: Communication Load versus MPBT

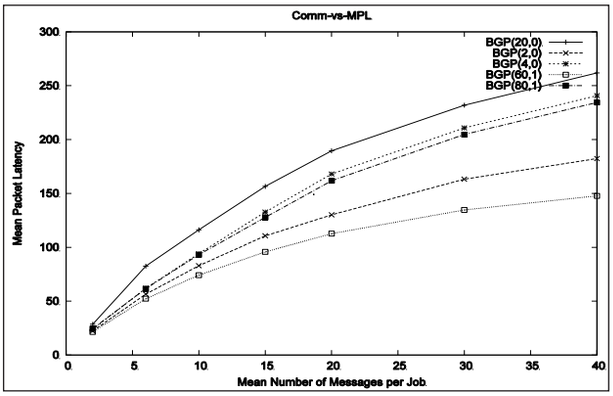


Fig. 23: Communication versus MPL

VI. Conclusion

In this paper, we proposed a framework for using fuzzy logic to control specific key performance levels in two-dimensional mesh-based multi-computer systems. We demonstrated that it is possible to implement a fuzzy-based feedback control system to control a number of performance indicators in the parallel system. We showed that those indicators can be controlled through increasing and decreasing the *MaxBlockCount* parameter in the newly proposed Bounded GRP-BF processor allocation strategy. In this context, the *MaxBlockCount* parameter automatically determined and calculated at run time through a fuzzy-based feedback control system.

The proposed allocation scheme prevents over-splitting of parallel jobs by preventing jobs get over-splitted by controlling the *partitioning-bound*. Fuzzy-based allocation provides a

mechanism to dynamically control the partitioning-bound. Instead of pre-setting and having a fixed partitioning-bound level, we provide a way to set this value at run time taking into consideration the current status to multi-computer system.

As a future work, we will be working on adding a tuning parameter(s) to the proposed approach that grants the operating system (and the user) to control how different performance indicators contribute to the decision of the partitioning-bound value. To this end, a number of tuning methods of controller can be utilized. Thus, we can comparatively evaluate them.

References

[1] S. Bani-Mohammad, I. Ababneh, and M. Yassen, "A new compacting non-contiguous processor allocation algorithm for 2D mesh multicomputers," *JITR*, vol. 8, no. 4, pp. 57-75, 2015.

[2] S. Bani-Mohammad, and I. Ababneh, "Improving system performance in non-contiguous processor allocation for mesh interconnection networks," *Simulation Modelling Practice and Theory*, vol. 80, pp. 19-31, ISSN: 1569-190X, 2018.

[3] G. M. Chiu, and S. K. Chen, "An efficient submesh allocation scheme for two-dimensional meshes with little overhead," *IEEE Transactions on Parallel and Distributed Systems*, vol. 10, no. 5, pp. 471-486, May 1999.

[4] Ababneh, "An efficient free-list submesh allocation scheme for two-dimensional mesh-connected multicomputers," *Journal of Systems and Software*, vol. 79, no. 8, pp. 1168-1179, Aug. 2006.

[5] Ababneh, and J. Davis, "Program-based static allocation policies for highly parallel computers," *Proc. IPCCC 95, IEEE Computer Society Press,* Scottsdale, AZ, USA, 28-31 Mar. 1995, pp. 61-68.

[6] S. Bani-Ahmad, "Processor allocation with reduced internal and external fragmentation in 2D mesh-based multi-computers," *Journal of Applied Sciences*, vol. 11, no. 6, pp. 943-952, 2011, doi: 10.3923/jas.2011.943.952.

[7] S. Bani-Mohammad, M. Ould-Khaoua, I. Ababneh, and L. Machenzie, "A fast and efficient processor allocation strategy which combines a contiguous and non-contiguous processor allocation algorithms," Technical Report: TR-2007-229, DCS Technical Report Series, Department of Computing Science, University of Glasgow, Jan. 2007.

[8] B. S. Yoo, and C. R. Das, "A fast and efficient processor allocation scheme for mesh-connected multi-computers," *IEEE Transactions on Parallel and Distributed Systems*, vol. 51, no. 1, pp. 46-60, Jan. 2002.

[9] K. Li, and K. H. Cheng, "A two-dimensional buddy system for dynamic resource allocation in a partitionable mesh connected system," *Journal of Parallel and Distributed Computing*, vol. 12, no. 1, pp. 79-83, May 1991.

[10] P. J. Chuang, and N. F. Tzeng, "Allocating precise submesh in mesh-connected systems," *IEEE Transactions Parallel and Distributed Systems*, pp. 211-217, Feb. 1994.

[11] Y. Zhu, "Efficient processor allocation strategies for mesh-connected parallel computers," *Journal of Parallel and Distributed Computing*, vol. 16, no. 4, pp. 328-337, 1992.

[12] D. Rupanetti, and H. Salamy, "Task allocation, migration and scheduling for energy-efficient real-time multiprocessor architectures," *Journal of Systems Architecture*, ISSN: 1383-7621, vol. 98, pp. 17-26, 2019.

[13] L. E. Rubio-Anguiano, A. C. Trabanco, J. L. B. Velasco, and A. Ramírez-Treviño, "Maximizing utilization and minimizing migration in thermal-aware energy-efficient real-time multiprocessor scheduling," *IEEE Access*, vol. 9, pp. 83309-83328, 2021, doi: 10.1109/ACCESS.2021.3086698.

[14] C. Y. Chang, and P. Mohapatra, "Performance improvement of allocation schemes for mesh-connected computers," *Journal of Parallel and Distributed Computing*, vol. 52, no. 1, pp. 40-68, Jul. 1998.

[15] J. Ding, and L. N. Bhuyan, "An adaptive submesh allocation strategy for two-dimensional mesh connected systems," *Proc. Int. Conf. Parallel Process. II*, Aug. 1993, pp. 193-200.

[16] S. Bani-Mohammad, I. M. Ababneh, and M. Yassen, "A new compacting non-contiguous processor allocation algorithm for 2D mesh multi-computers," *Journal of Information Technology Research (JITR)*, vol. 8, no. 4, 2015.

[17] P. Krueger, T. Lai, and V. A. Radiya, "Job scheduling is more important than processor allocation for hypercube computers," *IEEE Transactions on Parallel and Distributed Systems*, vol. 5, no. 5, pp. 488-497, May 1994.

[18] V. Lo, K. Windisch, W. Liu, and B. Nitzberg, "Non-contiguous processor allocation algorithms for mesh-connected multi-computers," *IEEE Transactions on Parallel and Distributed Systems*, vol. 8, no. 7, pp. 712-726, Jul. 1997.

[19] T. Liu, K. W. Huang, F. Lombardi, and L. N. Bhuyan, "A submesh allocation scheme for mesh-connected multiprocessor systems," *Proc. Int. Conf. Parallel Process. II*, Aug. 1995, pp. 159-163.

[20] S. Bani-Ahmad, "Submesh allocation in 2D-mesh multicomputers: Partitioning at the longest dimension of requests," *The International Arab Journal of Information Technology (IAJIT)*, vol. 10, no. 3, May 2013.

[21] S. Bani-Mohammad, "An efficient all shapes busy list processor allocation algorithm for 3D mesh multicomputers," *IJCAC*, vol. 7, no. 2, pp. 10-26, 2017.

[22] S. Bani-Ahmad, "Intra-job communication contention and request-partitioning-based allocation strategies in 2D-mesh multi-computers," *Third International Symposium on Parallel Architectures, Algorithms and Programming (PAAP'10),* Dalian, LiaoNing, P. R. China, 18-20 Dec. 2010.

[23] R. Stobart, "Tutorial on fuzzy control," IEE Colloquium on Two Decades of Fuzzy Control - Part 1, London, 1993, pp. 1/1-1/6.

[24] H. J. Zimmermann, *Fuzzy Set Theory and its Applications*. Kluwer Academic, Dordrecht, 1991.

[25] K. Windisch, J. V. Miller, and V. Lo, "ProcSimity: An experimental tool for processor allocation and scheduling in highly parallel systems," *Proceedings of the Fifth Symposium on the Frontiers of Massively Parallel Computation (Frontiers'95)*, Washington, USA, 6-9 Feb. 1995, pp. 414-421.

[26] ProcSimity V4.3 User's Manual, University of Oregon, 1997.

[27] J. Jantzen, Design of Fuzzy Controllers. Tech. Report No. 98-E 864 (design), Technical University of Denmark, Department of Automation, 30 Sep. 1999.

[28] J. Jantzen, Tutorial on Fuzzy Logic. Tech. Report No. 98-E 868, Technical University of Denmark, Department of Automation, 19 Aug. 1998.