Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 1259-1267
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.5636

Analysis of average waiting time and server utilization factor using queueing theory in cloud computing environment

Saurabh Adhikari[a,*], Maha A. Hutaihit[b], Moumita Chakraborty[a], Sawsan dheyaa Mahmood[c], Benjamin Durakovic[d], Souvik Pal[e], D. Akila[f], Ahmed J. Obaid[g]

[a] *Swami Vivekananda University, India*
[b] *Department of Communication Engineering Collage of Engineering, University of Diyala: Baqubah, Diyala, Iraq*
[c] *Department of Electricity Engineering College of Engineering, University of Tikrit, Tikrit, Iraq*
[d] *International University of Sarajevo, Bosnia.*
[e] *Department of Computer Science and Engineering Global Institute of Management and Technology, India*
[f] *Department of Information Technology, Vels Institute of Science, Technology and Advanced Studies, India*
[g] *Faculty of Computer Science and Mathematics, University of Kufa, Iraq*

(Communicated by Madjid Eshaghi Gordji)

Abstract

In industry-academy studies, the cloud computing model goes way above the ground. Cloud has emerged as a fantastic business model for service users and, depending on consumer requirements, can be used pay per usage base. Due to inadequate hardware or software resources, When the quantity of client requests for their high-demand service requirements is large, they prefer to wait in a server queue. As a result, in this study, Reduction in overall waiting time and server utilization factor has been focused on. Comparison has been made on average waiting time and analysis made on server utilization using the M/M/c queuing model.

Keywords: waiting time, queueing model, server utilization, Cloud computing.

*Corresponding author
Email addresses: saurabhadhikari@svu.ac.in (Saurabh Adhikari), mahaabbashutaihit@uodiyala.edu.iq (Maha A. Hutaihit), moumita.chakraborty@svu.ac.in (Moumita Chakraborty), sawsan.d.mahmood@tu.edu.iq (Sawsan dheyaa Mahmood), bdurakovic2@gmail.com (Benjamin Durakovic), souvikpal22@gmail.com (Souvik Pal), akiindia@yahoo.com (D. Akila), ahmedj.aljanaby@uokufa.edu.iq (Ahmed J. Obaid)

Received: July 2021 *Accepted:* September 2021

1. Introduction

The study of waiting lines is known as queuing theory. It is a field of applied mathematics that makes use of stochastic process concepts. A. K. Erlang, a Danish engineer, is credited with the development of queuing theory when he released his work on automated telephone exchanges in 1909. Essentially, it is a system in which clients need service and the service that can be offered is constrained. The service times are randomly selected. Some clients may have to wait if the number of servers is restricted [18]. Cloud computing, with its increasing applicability and popularity, not only provides huge potential, but also confronts several obstacles in its growth process [2].

The Theory of Queues has a unique attraction for mathematicians interested in stochastic processes because it gives a simple example that is both I stationary and (ii) not Markovian in general [7]. CCU's tasks, according to Luqun Li [10], were classified into various classes, each with a distinct priority. And they are on the server at some rate based on a distribution by Poisson, but a generic distribution follows the process time of the server for each task. As a result, a non-preemptive M/1 G queuing model was developed. With a non-preventive method, work planning in Cloud Computing is converted into a problem with queue planning for M / 1G.

Every day, we are irritated by the inconvenience of having to wait in line. In our increasingly packed and urbanized world, the problem is becoming more common. There are not only apparent lines at traffic, airport check-in stations and supermarkets, but also hidden delays in optical and wireless channels due to phone calls and data packets. Time, money, and resources are wasted for us all. Production and communication networks worldwide depend on the control of queues. Due to greater availability to high-speed Internet and advancements in virtualization and distributed computing, cloud computing garnered popular attention. The information technology is currently widely employed and evolving. When programs and data go onto the cloud, they not only change where, but also how computing takes place [8].

A. Objective of the Manuscript

In the previous part, we have discussed the queueing model, multi-server capacity, minimisation of time to wait and queue length. We focused on reducing the average wait time and evaluating the server use of a multi-server queueing approach. In this trial, the average server time and server use was compared by the M/M/c model.

2. Literature review

Souvik Pal et al. [15] They operate with the CloudSim simulation tool, which was used to figure out how many CPU cores and how long it takes to run. Reducing the amount of time spent waiting is also significant. When requesting a large number of jobs, it is necessary to wait for servers that might expand the queue and raise the waiting time. This research also looks at a queuing model with a large number of servers and limited capacity in order to reduce wait times and queue length.

Lizheng Guo et al. [4] They offer the mode, function, and method for synthesis optimization. Finally, simulation is performed using synthesis optimization methods. Simulation results are compared and evaluated with conventional optimization, This shows that the suggested methodology may minimise average expectation times, average queue time and customer number. The data centre is built as a service centre with a number of applications and an endless task application buffer which may be used as the M/M/m queue system.

Bashir Yusuf Bichi et al. [1] They concentrated on mathematical formulation utilizing a queuing system approach to demonstrate how a system's throughput and time delay might differ between a

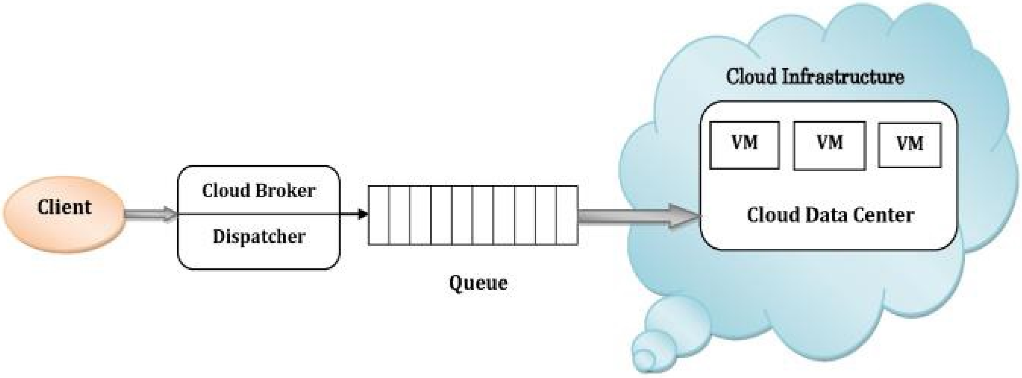


Figure 1: The skeleton Model of Cloud Server Queuing

single server system and a many server systems in a cloud-computing environment. They can easily observe that the throughput of the system with several serving units is greater than that of the system with a single serving unit. Furthermore, the waiting time for a specific request before it can be executed on a single server is longer than when a multiple server system is used.

Einollah Jafarnejad Ghomi et al. [5] Their research provided a taxonomy of cloud computing queuing models and The course of AQTMCC and the necessity for further study on it have been decided in the future. The topics of future study include: robustness, automatic Knowledge of QuoS and SLA-aware resource delivery, event-based predictions of workloads, provision and maintenance of cross cloud resources, model delivery, class-based and SLA-based service.

Khaled Salah et al. [3] proposed a paper that analyze the CDC performance with the developed stochastic model with the queuing theory. CDC platforms are developed in such a way that QOS of the cloud has been achieved. We give numeric examples of how the model predicts how many virtual machine instances a set of QoS-criteria needs to be met. We use a DES to evaluate our analysis model (Discrete Event Simulator). The recommended model can predict the required number of VMs meet the expected QoS objectives with variation of the arrival request rate, according to the results of their research and simulation.

Roxana MARCU et al. [11] The article proposed a cloud platform to create a fully integrated, industry-specific medical system. An analytical model has been proposed for achieving excellent performance with low cost for scheduling cloud. In terms of the number of requests, response time to wait, and demand drop rate, the performance of each established priority class is measured. The experimental results demonstrate that the recommended model can support several arrival requests and give a fast answer time depending on priority classes. S. Pal et al. [16, 14] have discussed scheduling algorithm and probabilistic algorithm to enhance the performance matrices in cloud computing environment.

3. System model

As shown in Figure 1, The cloud service queuing system's framework model has been defined.

With the stated requirement, the cloud user communicates with the cloud broker. Cloud brokers [15], who function as negotiators or mediators, help users in preserving all operations, the number of CPU-cores required and the time required for a user request to be handled, such as the special data center requested, execution time for the relevant requests. Users can also use the cloud broker to provide themselves with resources.

The cloud broker analyzes and submits the request to the data centre. In a short period of time,

Table 1: Initial Parameter ([10])

Λ	μ
20	40
60	70
120	122

when a great deal of requests is initiated, a queue result. The demands are carried out in conjunction with the Virtual Machines' VMM development. The data center distributes resources based on their availability. After the queries have been completed, Virtual machines have to be shut off correspondingly. We found the M/M/C model in that work to calculate the average server and duration of waiting.

M/M/c. M/M/c: The system of M/M/c queuing [13, 6, 16, 14, 17, 9, 12, 19] was named as the only model capable of analysing and closing the density function of the likelihood when a new system comes. The M/M/c queuing model in a multiple server system allows for the analysis and treatment of the optimization problem. The quantities of a service, the operational load of an application environment, the configuration of a multi-server system, the agreement on levels of service, customer pleasure etc, are the elements of your model pricing. The study also investigated two alternative models of server power and speed.

4. Results analysis

When the server's capacity is insufficient to handle the quantity of requests in a reasonable amount of time, waiting queues will be formed. Methods of arrive and service, and servers [18, 2], are necessary in respect of the fundamental queuing model. The M/M/c queuing model was described in this chapter. The rate of Poisson distribution and exponential distribution are used to represent the service request and process time, respectively. Kendal's notation [7] is correct, the average rate of arrival and the average rate of operation are denoted by λ and μ, respectively. The model M/M/c has been tested here and the server number has changed (denoted as c) and the usage factor being compared as well. To ensure that the system is stable, where the usage factor $[\rho = \frac{\lambda}{\mu} \leq 1]$ is considered, an equilibrium condition must be considered. As shown in Table 1, it took into consideration the average arrival and average rates of service. The number of servers was compared with the average system client numbers (Ls), the average number of queue customers (Lq), the average system waiting time (Ws), and the average queue waiting time (Wq) 1.

The comparison study was shown in figures 2, 3, 4, 5, with reference to the above table 1.

The preceding chart and research shows that expanding servers decrease the average waiting time and average queue and device users. The use factor (U) can now be set as follows:
$U = \frac{\rho}{c} = \frac{\lambda}{c\mu}$. Therefore, in the Table 2, Utilization factor has been compared.

The server utilization factor may be decreased by increasing the number of servers, according to this study. Figure 6 illustrates this. As seen in the figures from 7, 8, 9, 10, 11, 12, we have also evaluated the probability of customers in a graphical way in case of two different models.

Fig 7 to fig. 12 shows the performance of probability of customerd in M/M/2 and M/M/3 with a wide range of average rate of arrival. The probability of customers satisfaction without waiting for service has been increased with increase in number of system and decrease in average rate arival.

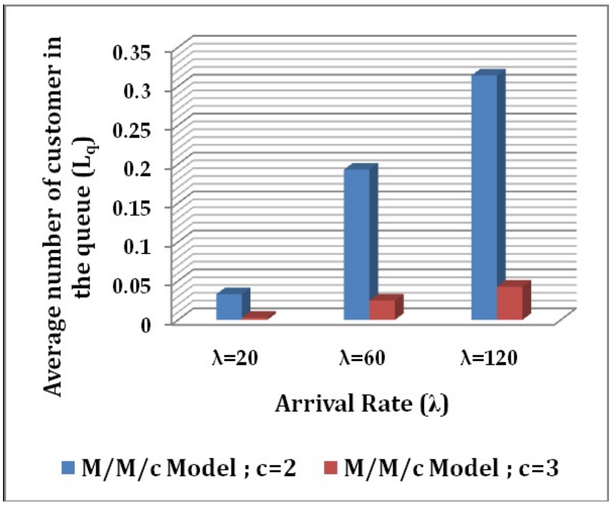


Figure 2: Outcome with Lq

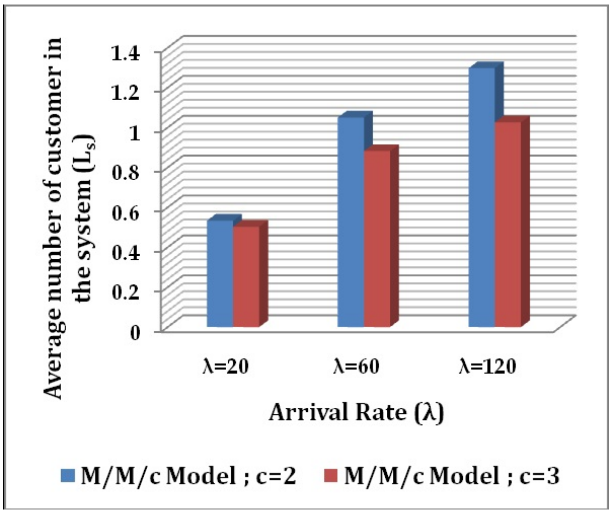


Figure 3: Outcomes with Ls

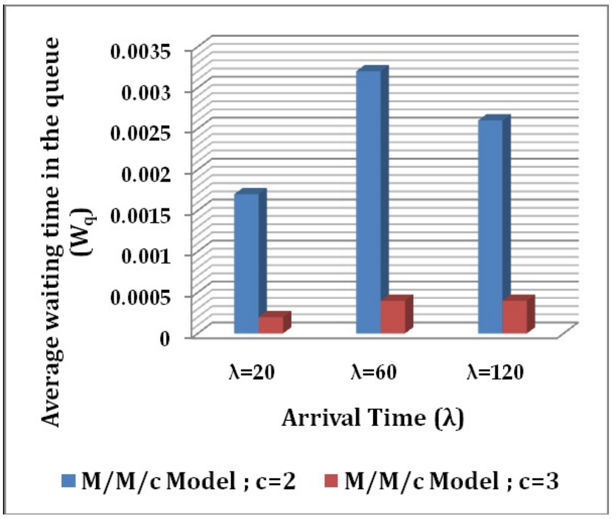


Figure 4: Outcomes with Wq

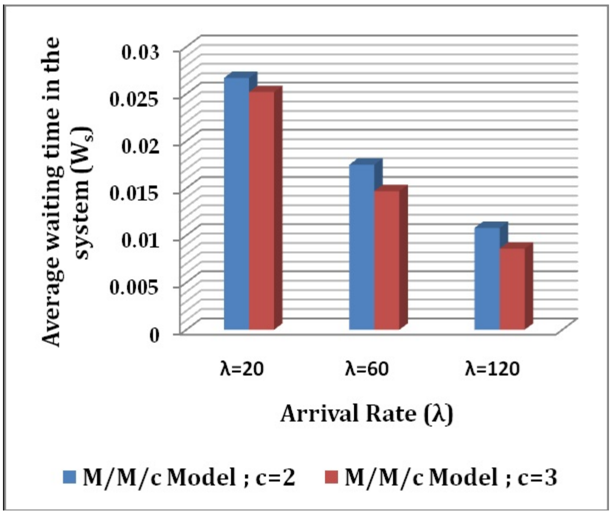
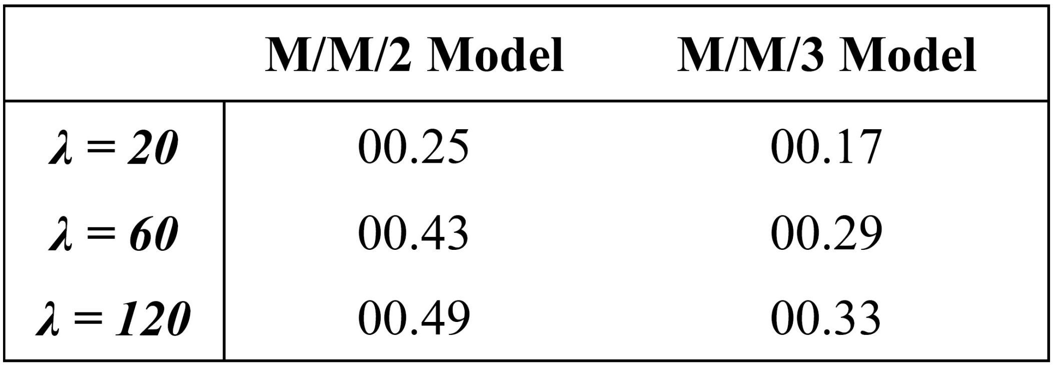


Figure 5: Outcomes with Ws

Table 2: OUTCOME ANALYSIS FOR SERVER UTILIZATION



	M/M/2 Model	M/M/3 Model
λ = 20	00.25	00.17
λ = 60	00.43	00.29
λ = 120	00.49	00.33

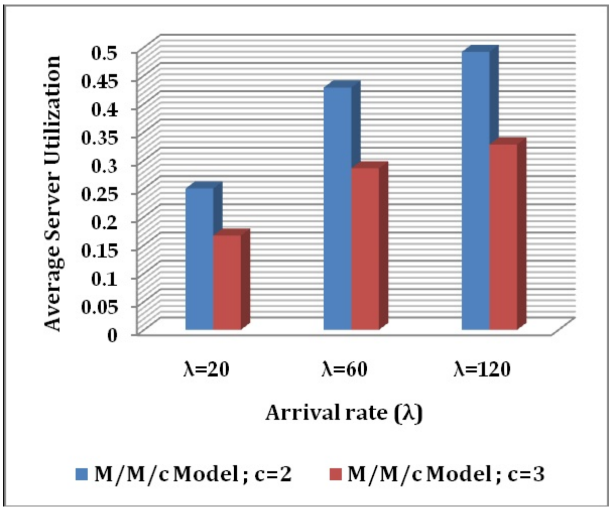


Figure 6: Comparison Outcomes with server utilization

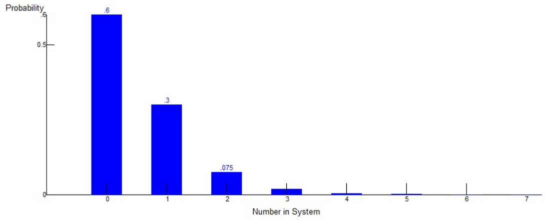


Figure 7: Probability of customers in M/M/2 Model ($\lambda = 20$)

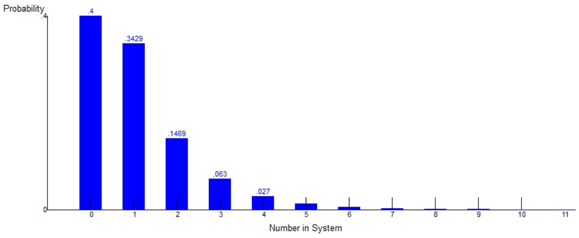


Figure 8: Probability of customers in M/M/2 Model ($\lambda = 60$)

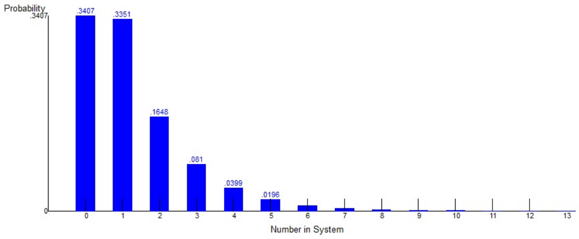


Figure 9: Probability of customers in M/M/2 Model ($\lambda = 120$)

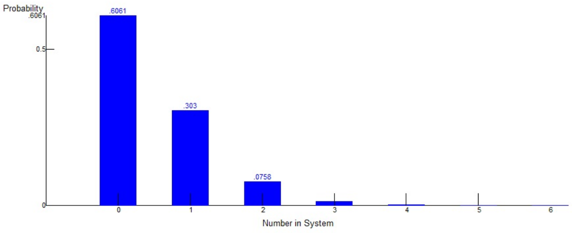


Figure 10: Probability of customers in M/M/3 Model ($\lambda = 20$)

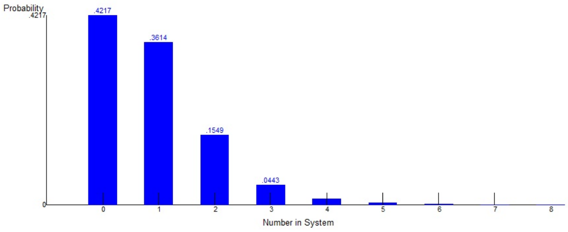


Figure 11: Probability of customers in M/M/3 Model ($\lambda = 60$)

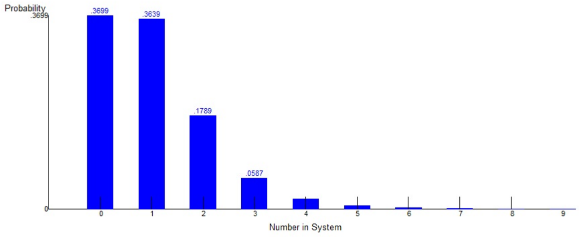


Figure 12: Probability of customers in M/M/3 Model ($\lambda = 120$)

5. Conclusion

This paper investigated the analysis of web application deployment on cloud environment for studying the cost point of view. This analysis is useful for estimating the future requirement of the cloud resources based on number of user requests. As a result, in this study, We concentrated on the overall reduction in waiting time and use of servers. We compared the average time of expectation and server use analysis using the M/M/C queuing model. The server utilization factor has been decreased by increasing the number of servers and expanding servers will decrease the average waiting time and average queue and device users.

References

[1] B.Y. Bichi and T. Ercan, *An efficient queuing model for resource sharing in cloud computing*, Int. J. Engin. Sci. 3(10) (2014) 36–43.

[2] C. Cheng, J. Li and Y. Wang, *An energy-saving task scheduling strategy based on vacation queuing theory in cloud computing*, Tsinghua Sci. Tech. 20(1) (2015) 28–39.

[3] S. El Kafhali and K. Salah, S*tochastic modelling and analysis of cloud computing data center*, 2017 20th Conf. Innov. Clouds Internet Networks (2017) 122–126.

[4] L. Guo, T. Yan, S. Zhao and C. Jiang, *Dynamic performance optimization for cloud computing using M/M/m queueing system*, J. appl. Math. 2014 (2014).

[5] E. Jafarnejad Ghomi, A.M. Rahmani and N.N. Qader, *Applying queue theory for modeling of cloud computing: A systematic review*, Concur. Comput. Pract. Exper. 31(17) (2019) e5186.

[6] S. Jeyalaksshmi, M.S. Nidhya, G. Suseendran, S. Pal and D. Akila, *Developing mapping and allotment in volunteer cloud systems using reliability profile algorithms in a virtual machine*, 2021 2nd Int. Conf. Comput. Autom. Knowledge Manag. (2021) 97–101.

[7] D.G. Kendall, *Some problems in theory of queues*, J. Roy. Stat. Soc. Series B 13(2) (1951) 151–185.

[8] A.D. Khomonenko, S.I. Gindin and K.M. Modher, *A cloud computing model using multi-channel queuing system with cooling*, In 2016 XIX IEEE Int. Conf. Soft Comput. Measur. (2016) 103–106.

[9] G. Lakshmi, M. Ghonge and A.J. Obaid, *Cloud based IoT smart healthcare system for remote patient monitoring*, EAI Endorsed Trans. Pervasive Health Tech. (2021).

[10] L. Li, *An optimistic differentiated service job scheduling system for cloud computing service users and providers*, 3rd IEEE Int. Conf. Multimedia Ubiquitous Engin.(MUE '09) (2009) 295–299.

[11] R. Marcu, I. Danila, D. Popescu, O. Chenaru and L. Ichim, *Message queuing model for a healthcare hybrid cloud computing platform*, Stud. Inf.Control 26(1) (2017) 95–104.

[12] A.S. Nori and A.O. Abdulmajeed, *Design and implementation of Threefish cipher algorithm in PNG file*, Sustain. Engin. Innov. 3(2) (2021) 79–91.

[13] A. Outamazirt, K. Barkaoui and D. Aïssani, *Maximizing profit in cloud computing using M/G/c/k queuing model*, 2018 Int. Symp. Prog. Syst. (2018) 1–6.

[14] S. Pal, R. Kumar, L.H. Son, K. Saravanan, M. Abdel-Basset, G. Manogaran and P.H. Thong, *Novel probabilistic resource migration algorithm for cross-cloud live migration of virtual machines in public cloud*, J. Supercomput.75 (2019) 5848–5865.

[15] S. Pal and P.K. Pattnaik, *A Simulation-based approach to optimize the execution time and minimization of average*

waiting time using queuing model in cloud computing environment, Int. J. Electrical & Comput. Engineering (2088-8708), 6(2) (2016) 743–750.

[16] S. Pal and P.K. Pattnaik, *Adaptation of Johnson sequencing algorithm for job scheduling to minimize the average waiting time in cloud computing environment*, J. Engin. Sci. Tech. 11(9) (2016) 1282-1295.

[17] R. Regin, A.J. Obaid, A. Alenezi, F. Arslan, A.K. Gupta and K.H. Kadhim, *Node replacement based energy optimization using enhanced salp swarm algorithm (Es2a) in wireless sensor networks*, J. Engin. Sci. Tech. 16(3) (2021) 2487–2501.

[18] L. Tadj, *Waiting in line*, Potential IEEE 14(5) (1996) 11–13.

[19] M. Tripathi, *Facial image denoising using AutoEncoder and UNET*, Heritage Sustain. Develop.3(2) (2021) 89–96.