ISSN 0304-9892 (Print) ISSN 2455-7463 (Online)
www.vijnanaparishadofindia.org/jnanabha

Jñānābha, **Vol. 55(2) (2025), 123-134**

ASSESSMENT OF ENHANCING COST EFFICIENCY IN CLOUD COMPUTING SYSTEMS THROUGH QUEUEING MODELS

Jitendra Kumar and Vikas Shinde
Department of Engineering Mathematics & Computing Madhav Institute of Technology & Science, Deemed University Gwalior, Madhya Pradesh, India-474005
Email: jkmuthele@mitsgwalior.in; vpshinde@mitsgwalior.in

(Received: March 07, 2024*; In format: March* 11, 2024*; Revised: November* 19, 2025*; Accepted: November* 24, 2025*)*

DOI: https://doi.org/10.58250/jnanabha.2025.55214

Abstract

This article deals with cost effectiveness in cloud computing systems using $M/G/c$ queueing model with primary focus on minimizing waiting time. Our object is to investigate and enhance the cost effectiveness of cloud computing systems which utilize diverse services such as infrastructure as a service, platform as a service, software as a service, monitoring as a service, and communication as a service and anything as a Service(*IaaS, PaaS, SaaS, MaaS, CaaS, XaaS*). We provide analytical results to compute the various performance measures and energy efficiency. Sensitivity analysis is also carried out with the emphasis on energy consumed and energy efficiency.The integration of cloud computing and virtualization services is an important aspect explored in this study.

2020 Mathematical Sciences Classification: 90B22, 37M21

Keywords and Phrases: Queueing Models, Performance Evaluation, Cloud Services, Total Energy Consumed, Total Cost, Energy Efficiency

1 Introduction

Cloud computing networking represents an emerging commercial infrastructure that pledges to obviate the necessity for maintaining expansive computer facilities. Positioned as the next phase in the development of on-demand infrastructure technology services, production, maintenance and other facilities. Cloud computing networking offers a computing base intended to replace the traditional computing network infrastructure. The purpose of this change is to control costs and streamline Providers of cloud computing networks deliver cloud services through one or more distribution systems, encompassing computing, storage, network resources and other functionalities. The performance evaluation requirements of this innovative model pave the way for the provisioning of networking resources on-demand. Cloud computing is feature planning in networking in which services occur at widely different stages favour of planning in which services are widely on different phases. Every day, we are facing by the inconvenience of having to wait in lines. 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. Therefore, the Production and communication networks worldwide depend on the control of queues.Particularly, in cloud computing we are also facing the problem of data centre receives data processing, customer order overloading and many other problems of cloud networking etc. These challenging issues can be resolved by queuing models.

2 Types of Cloud Computing

Cloud services, accessed over the internet, transform data, applications, and computing management for businesses and individuals. This includes Infrastructure as a Service *(IaaS)* for virtualized resources, Platform as a Service *(PaaS)* providing a development platform, and Software as a Service *(SaaS)* delivering pre-built software solutions. Monitoring as a Service *(MaaS)* is under the umbrella of anything as a Service *(XaaS)* in cloud computing. Cloud services enhance scalability, flexibility, and cost-effectiveness, enabling users to pay for utilized resources. Crucial for modern computing, cloud technology promotes innovation, collaboration, and agility across industries. In Figure 2.1, specific cloud computing service types were omitted as noted.

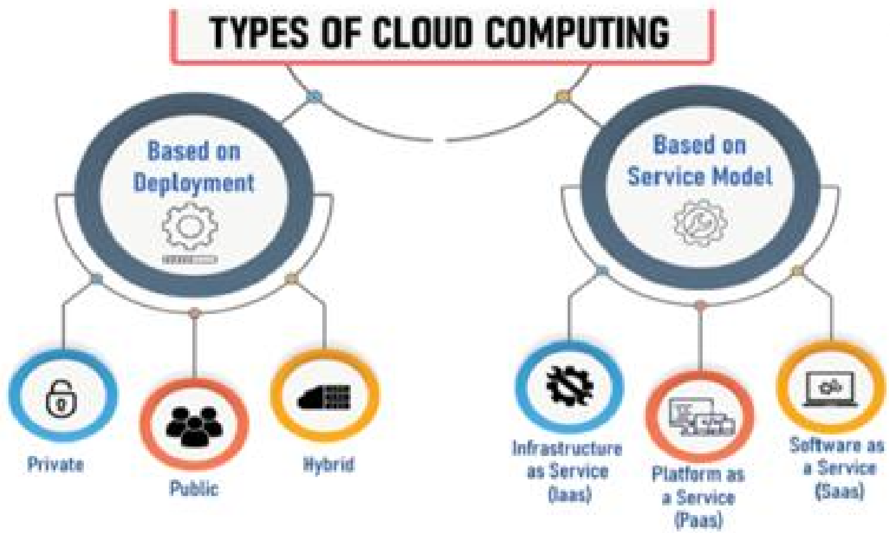

***Figure* 2.1:** *Classification of Cloud Computing Networks*

2.1 Infrastructure as a Service

Infrastructure as a Service, plays a crucial role in cloud computing, offering virtualized resources online. Users can flexibly access and oversee servers, storage, and networking through a pay-as-you-go model, enabling organizations to expand their *IT* capabilities without substantial investments in physical hardware. Infrastructure as a Service proves cost-effective by relieving enterprises from on-site hardware maintenance, allowing users to focus on application development and business innovation while delegating the intricacies of infrastructure management to the cloud.

2.2 Platform as a Service

Platform as a Service, a vital cloud computing model, streamlines application development and deployment. It offers a ready-to-use environment, enabling developers to code without handling infrastructure intricacies. This accelerates development, promotes collaboration, and simplifies deployment. Abstracting the infrastructure layer, *PaaS* eases scalability and maintenance. Featuring automatic updates and built-in services, it empowers businesses for efficient innovation, reducing application time-to-market and optimizing resource usage. *PaaS* stands as a valuable solution for agile and streamlined development.

2.3 Software as a Service

Software as a Service is a cloud computing cornerstone, delivers software over the internet, allowing users direct access without installation. It eliminates update and hardware management hassles, providing a cost-effective, scalable solution. With a subscription model, businesses deploy and use software efficiently, paying for consumed services. This enhances accessibility, collaboration and flexibility, integral to modern business operations. *SaaS* streamlines workflows, ensures seamless updates and offers a user-friendly experience globally.

2.4 Monitoring as a Service

Monitoring as a Service is a cloud-based solution offering extensive *IT* infrastructure monitoring. It enables remote tracking analysis of performance metrics and issue identification. *MaaS* provides real-time insights, proactive alerts and historical data to enhance system reliability. With a subscription model, users access tools without on-premises infrastructure. Crucial for maintaining a stable *IT* environment, *MaaS* supports proactive issue resolution, minimizing downtime. It is essential for businesses prioritizing performance with reliability in a streamlined and efficient manner.

2.5 Anything as a Service

Anything as a service is a broad cloud computing category encompassing diverse offerings beyond Infrastructure as a Service, Platform as a Service and Software as a Service. It offers a flexible framework, delivering software to infrastructure as services over the internet. This allows businesses to access on-demand resources, tailoring solutions to specific needs. Anything as a service promotes scalability, cost-efficiency and innovation, providing a versatile platform for diverse services. As a dynamic concept, anything as a service continually adapts to emerging technological demands, shaping the future of cloud computing.

2.6 Communication as a Service

Communication as a Service is a cloud-based model delivering various communication tools over the internet, including voice, video, messaging and collaboration applications. It enables organizations to integrate communication solutions without on-premises infrastructure. This scalable and flexible approach fosters efficient communication, allowing businesses to adapt to changing needs. Communication as a Service enhances collaboration, streamlines communication processes and improves overall connectivity, making it an essential component in the modern digital workplace.

3 Literature Review

This section explores enhancing cost efficiency in cloud computing through queueing models, utilizing diverse services (*IaaS, PaaS, SaaS, MaaS, CaaS, XaaS*). It assesses total energy consumption, costs, and energy efficiency. Key indicators such as queue length, waiting time in queue, and utilization factor optimize resource allocation. The review provides insights into how queueing models effectively manage energy consumption, costs, and operational efficiency across various cloud services, benefiting both practitioners and researchers. Boxma *et al.* [3] addresses the approximation of mean waiting time in an $M/G/s$ queueing system. This work likely delves into developing methods to estimate the average waiting time in queueing systems with general service time distributions and multiple servers (s). The findings contribute to queueing theory and system optimization. Cao *et al.* [5] focus on determining the optimal multi-server configuration for maximizing profits. Chaisiri *et al.* [6] focused on the optimization of resource provisioning cost in cloud computing. This study likely explores methods and strategies to enhance the efficiency of resource allocation in cloud environments, with a specific emphasis on minimizing associated costs. Eisa [7] enhances cloud computing scheduling based on queuing models, while Anupama [1] utilizes queuing theory to analyze the performance measures of clouds with infinite servers. The work by Wang *et al.* [25] and Furht [8] provides a perspective on cloud computing as a whole, offering a comprehensive overview of the field. Gross and Harris [9] on fundamentals of queueing theory serves as a reference for understanding queueing systems, adding depth to the theoretical framework. Ghose *et al.* [10] adopt an interacting stochastic models approach for end-to-end performability analysis in infrastructure-as-a-service cloud environments. Khazaei *et al.* [12] focused on the performance analysis of cloud computing centers. The study employs $M/G/m/m{+}r$ queueing systems to assess and analyze the efficiency and effectiveness of cloud computing environments. This likely involves evaluating resource utilization, response times, and overall system performance within cloud centers. Kumar and Shinde [13] focused on the performance evaluation of bulk arrival and bulk service with multiple servers using a queue model. This study likely delves into assessing the efficiency and outcomes of systems involving bulk arrivals and services with the utilization of multiple servers through a queueing model. Lakshmi and Bindhu [15] contribute a queuing model aimed at improving the quality of service by reducing waiting times in the cloud. Ma and Mark [16] provided an approximation method for estimating mean queue length in $M/G/c$ systems, improving analytical accuracy for complex service distributions and offering practical performance predictions for multi-server queueing environments. Mei *et al.* [17] explored a profit maximization scheme with guaranteed quality of service in cloud computing. This study likely delves into developing strategies that balance profit objectives for service providers with ensuring and maintaining a specified quality of service in cloud computing environments. Praveen *et al.* [19] examined the utilization of Queueing Theory in Cloud Computing for waiting time reduction; this study likely investigates the application of Queueing Theory principles to minimize waiting times in the realm of cloud computing services. Resing [20] and Whitt [24] providing essential theoretical underpinnings for understanding queueing systems and their applications in cloud environments. Suakant *et al.* [22] addressed the performance measurement of cloud computing services. This study likely involves evaluating and assessing the effectiveness and efficiency of various aspects of cloud services to provide insights into their overall performance. akahashi's [23] focused on an approximate formula for the mean waiting time in an $M/G/c$ queue, this study is presumed to offer a mathematical approach for estimating the average waiting time in a queueing system with variable service times and a fixed number of servers. Xong and Perros [26] address service performance and analysis in cloud computing, providing valuable insights. Xia *et al.* [27] explores stochastic modeling and performance analysis related to migration-enabled and error-prone cloud environments. The study likely delves into understanding the impact of migration processes and error occurrences on the overall performance of cloud systems. Zhuravlev *et al.* [28] present a survey of energy-cognizant scheduling techniques, contributing to the broader context of resource optimization in cloud computing. Ghomi *et al.* [11] gave systematic literature review to evaluate the related literature, and 71 articles were selected as primary studies that were classified the modeling techniques of cloud computing using the queuing theory in seven categories based on their focus area: (i) performance, (ii) quality of service, (iii) workflow scheduling, (iv) energy savings, (v) resource management, (vi) priority-based servicing, and (vii) reliability. Adhikari *et al.* [2] employed reduction in overall waiting time and server utilization factor along with comparison made on average waiting time and analyze server utilization using the $M/M/c$ queuing model. Outamazirt *et al.* [18] formulated $M/M/c/K$ queuing model with impatient customers for maximizing the profit in cloud computing. Kafhali [14] presented a stochastic model based on queuing theory to aid in studying and analyzing performance in Cloud data centers (CDC). CDC platforms are modeled with an open queuing system that can be used to estimate the expected Quality of Service (QoS) guarantees the cloud can offer. Berl *et al.* [4] cloud computing has recently received considerable attention, as a promising approach for delivering information and communication technologies (ICT) services by improving the utilization of data center resource more or less its play an important role of cost effectiveness. Finally, we described encapsulates a diverse range of research efforts, spanning stochastic modeling, performance analysis, queueing theory applications, and profit maximization schemes in cloud computing. The findings collectively contribute to a deeper understanding of the challenges, optimizations, and theoretical foundations within the dynamic landscape of cloud environments. Roch *et al.* [21] investigated cloud-edge-sensors infrastructure using $M/M/c/K$ queuing theory to analyse agricultural data systems performance which focuses on optimizing data handling and evaluates the system configuration impacts on performance. The model significantly enhances cost efficiency and scalability in cloud computing.

4 Optimization of the Cloud Computing

We described cloud computing networking there are a lot of customers who access the facility. This model comprises of a cloud structure in Figure 4.1 which can be a facility inside. The service hub is multi-points of admission to all types of clienteles all over the world. The service hub is an assortment of administration assets that are given by the supplier to have all applications for can apply to utilize administration as indicated by the various types of mentioning and pays some cash to the supplier of the assistance. Cloud computing supplier constructs the service center to be used by clienteles, like that Flipkart, Amazon, and others which providers several kinds of behaviors. In this paper, we use on-request cases. For examples let you pay for registering limit continuously/month/day/hour with no extended responsibilities. This liberates the expenses and intricacies of preparation, buying and keeping up with equipment, programming and changes what are normally huge fixed expenses into a lot more modest variable expenses.

4.1 Study of Cloud Computing Systems

We explore the dynamics of a cloud computing system. Typically, an influx of numerous user service requests occurs within the information processing system, as depicted in Figure 4.1. Various hubs act as gathering places and when client service requirements arise, the cloud computing community responds by offering diverse services tailored to users' adaptive needs. This includes considerations for pricing across different services and the operational structure of the cloud computing network. The service model of cloud computing networks, as depicted in Figure 4.1, can be conceptualized as a queueing model illustrated in Figure 4.2. Let's consider a scenario where k users are utilizing n independent service stations. The continuous arrival of demands, originating from two or more clients, follows an exponential random variable pattern in the cloud computing system. This leads to customer arrivals adhering to a Poisson process with an arrival rate $lambda_i$, distributing requests in the scheduler queue to different computing servers with the arrival rate contingent on the scheduler. In a data center, computing services denoted as S1, S2, S3,..., and Sn, and a multi-server system with a service rate of $\alpha\mu_i$ for each service, the total arrival rate is

$$\lambda = \sum_{i=1}^{n} \lambda_i$$

and the total service rate is

$$\mu = \sum_{1}^{n} \alpha\mu_i$$

. The system's suitability for various types of services to users is confirmed by the observed concept that the service rate follows a Poisson process. Therefore, the queueing model is well-suited for the cloud computing networking system, accommodating different service types for our users.

***Figure* 4.1:** *Classification of Cloud Computing Networks*

4.2 Roll of Queue System in Cloud Computing Networking

Queue systems play a pivotal role in optimizing performance and resource allocation within the realm of cloud computing networking. In cloud environments, where numerous users simultaneously access services, efficient task management becomes crucial. Queue systems effectively manage and prioritize incoming requests, orchestrating the orderly execution of tasks. They contribute to load balancing, ensuring that resources are allocated judiciously among various tasks and users. Queue systems enhance the responsiveness and reliability of cloud services, preventing bottlenecks and minimizing waiting times. By regulating the flow of tasks, these systems enable cloud providers to maintain optimal service levels, even during peak periods. The strategic implementation of queue systems in cloud computing networking not only streamlines operations but also enhances user satisfaction by promoting a seamless and responsive computing experience. Their role is integral in aligning resource utilization with demand, ultimately fostering the efficiency and scalability of cloud computing networks.

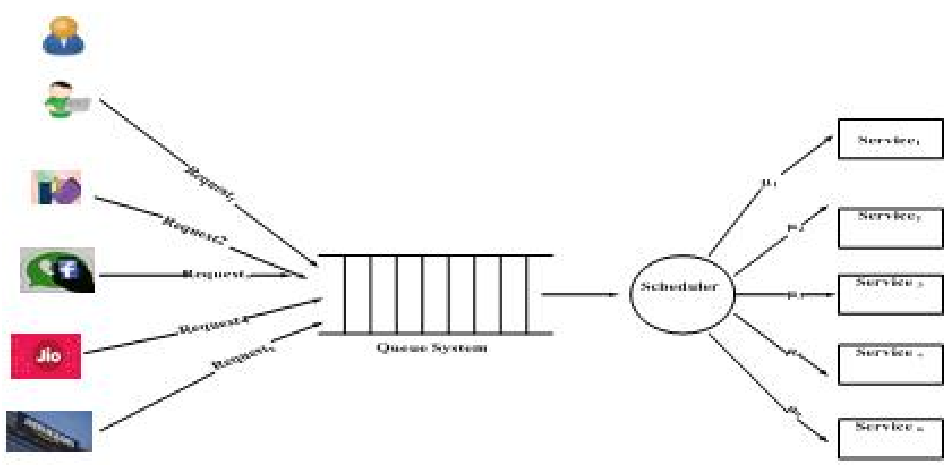

***Figure* 4.2:** *A queueing model various service in cloud computing networking*

5 Mathematical Model

We have constructed a mathematical model utilizing the $M/G/c$ queueing model for cloud computing networking. The model accommodates the global surge in requests and the extensive array of services offered by cloud computing networks, with unrestricted sources of clients and queueing model capacity. The mathematical model's steady-state equation is represented as S = 0, 1, 2...; these equilibrium conditions can be defined from the steady-state transition probability diagram of the $M/G/c$ model, as shown in Figure 5.1. From the state- transition diagram to arise some possibility as when the state is $k\,(0 \leq n \leq c)$. Currently, n

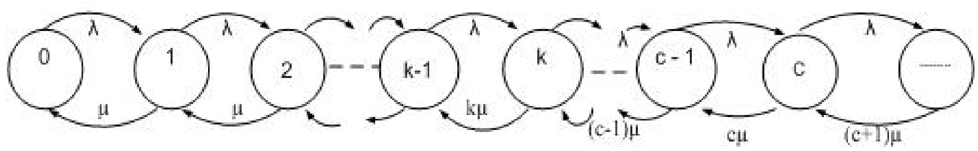

Figure* 5.1*: *Steady-state Probability Diagram*

services are active, while the remaining $c - n$ services are idle. In this scenario, where n is greater than c,c services are occupied, leaving $n - c$ users in a waiting state, contributing to the size of the queue.We assume $\rho = \dfrac{\lambda}{\mu\, c} < 1$, the queue model $M/G/c$ following steady-state equations:

$$\left.\begin{aligned} p_1 &= c\,\rho\,p_0 \\ p_2 &= \frac{c^2}{2!}\rho^2 p_0 \\ p_3 &= \frac{c^3}{3!}\rho^3 p_0 \\ &\vdots \\ &\vdots \\ p_m &= \frac{c^n}{n!}\rho^n p_0 \end{aligned}\right\} \tag{5.1}$$

In general,

$$p_m = \left.\begin{matrix} \dfrac{c^n}{c!}\,\rho^n\,p_0, & \text{when } 0 \leq n \leq c, \\ \dfrac{n^c}{c!}\,\rho^n\,p_0, & \text{when } n \geq c \end{matrix}\right\}. \tag{5.2}$$

We obtain P_0 from equation (5.2) and use the normalization condition

$$\sum_{n=0}^{\infty} p_n = 1,$$

in which its solution is

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!}\right)^{-1}. \tag{5.3}$$

We demonstrate the importance of cloud computing system performance in enhancing operational efficiency. This encompasses factors such as queue length, waiting time in the system, and queue, along with resource utilization, total energy consumption, total cost, and energy efficiency considerations. Through the evaluation and optimization of these aspects, we guarantee a smooth service delivery, address user demands, and uphold overall system reliability within the dynamic cloud environment. The analysis is supported by the use of specific notations and the derivation of mathematical formulas, employing a respected $M/G/c$ queuing model. where:

L_s	=	Queue length in the system,
L_q	=	Queue length in the queue,
W_s	=	Waiting time in the system,
W_q	=	Waiting time in the queue,
TEC	=	Total energy consumed,
EE	=	Energy efficiency,
TC	=	Total cost,
λ	=	Arrival rate in this system,
μ	=	Service rate for each user in this system,
c	=	Number of servers,
ρ	=	Utilization factor.

Expected Number of Customers in the System (E[L_s]):

$$L = \frac{\lambda}{\mu(1-\rho)}.$$

Expected Number of Customers in the Queue (E[L_q]):

$$L_q = \frac{\rho^c P_0 \lambda^c}{c!(1-\rho)^2}.$$

Expected Time of Customers in the System (E[W_s]):

$$W = \frac{1}{\mu - \frac{\lambda}{c}}.$$

Expected Time of Customers in the Queue (E[W_q]):

$$W_q = \frac{\rho P_0 \lambda}{c(1-\rho)}.$$

Total Energy Consumed (TEC):

$$TEC = c \times (\text{Power consumption per server}) \times (\text{Average server utilization}).$$

Energy Efficiency (EE):

$$EE = \frac{\lambda \times (\text{Average service time per customer})}{c \times (\text{Power consumption per server}) \times (\text{Average server utilization})}.$$

Total Cost (TC):

$$TC = C_s + C_w.$$

Here, C_s represents the cost per server multiplied by the expected server utilization (ρ), where ρ is the fraction of time the servers are busy. Additionally, C_w denotes the cost per customer waiting multiplied by the expected number of customers waiting (L_q).

6 Sensitivity Analysis

In this section, we study cloud computing networks with various types of service platforms. The optimization process to find the optimal balance among the total energy consumed total cost and energy efficiency with utilization factor with service cost and waiting cost per server to achieve efficient system operation with various types parameters as for need to our object. We examine specific cases to assess performance metrics in cloud computing networks, alongside considerations for energy consumption.

Case I

In this case, we investigate scenarios where the arrival rate is increased. Our focus is on analysing performance metrics such as queue length, waiting time, utilization factor, as well as total energy consumed, energy efficiency and total cost within cloud computing networks. We utilize the $M/G/c$ queue model, taking into account parameters such as arrival rate (λ), service rate (μ), and the number of servers (c), with predetermined values for λ, μ, and c. Let us assume that the service rate $\mu = 5.5$ and the number of servers $c = 5$ are fixed in this case.

Table 6.1: Performance Analysis of Cloud Computing Networks (Case 1)

Scenarios	c	μ	λ	TEC	TC	EE
When Arrival Rate Increases	5	5.5	2	36.36	2200.00	0.12
	5	5.5	3	54.55	2190.05	0.19
	5	5.5	4	72.73	2150.78	0.26
	5	5.5	5	90.91	2136.38	0.34
	5	5.5	6	109.09	2136.03	0.42
	5	5.5	7	127.27	2238.71	0.52
	5	5.5	8	145.45	2583.39	0.62
	5	5.5	9	163.64	3860.86	0.74
	5	5.5	10	181.82	7296.41	0.86
	5	5.5	11	200.00	16162.59	1.01
	5	5.5	12	218.18	37297.98	1.17
	5	5.5	13	236.36	85093.21	1.36
	5	5.5	14	254.55	188977.83	1.57
	5	5.5	15	272.73	407973.79	1.81
	5	5.5	16	290.91	859603.31	2.10
	5	5.5	17	309.09	1777184.83	2.45
	5	5.5	18	327.27	3628722.51	2.86
	5	5.5	19	345.45	7369605.77	3.37
	5	5.5	20	363.64	15011301.80	4.02
	5	5.5	21	381.82	30986992.83	4.87

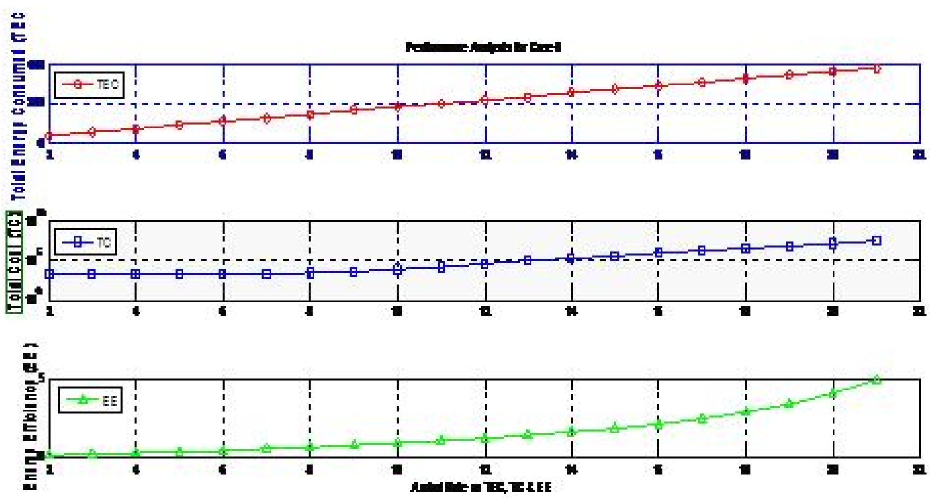

***Figure* 6.1:** *Classification of Cloud Computing Networks*

Case- II

In this case, we examine scenarios where the number of servers is augmented to analyze performance metrics such as queue length, waiting time, utilization factor, as well as total energy consumption, energy efficiency and overall cost within cloud computing networks. This investigation employs the M/G/c queue model, incorporating parameters like arrival rate (λ), service rate (μ), batch size of service, and the number of servers (c), with predetermined values for λ, μ, and c. Let us assume that the service rate $\mu = 5.5$ and the arrival rate $\lambda = 5$ are fixed in this case.

*Table **6.2**: **Performance Analysis of Cloud Computing Networks** (Case* **II***)*

Scenarios	**c**	λ	μ	**TEC**	**TC**	**EE**
When Service Rate Increases	5	5	5.2	90.91	2206.38	0.34
	5	5	5.4	89.29	2195.89	0.34
	5	5	5.6	87.72	2155.44	0.35
	5	5	5.8	86.21	2135.04	0.35
	5	5	6.0	84.75	2104.67	0.36
	5	5	6.2	83.33	2084.34	0.36
	5	5	6.4	81.97	2004.04	0.36
	5	5	6.6	80.65	1993.76	0.37
	5	5	6.8	79.37	1903.50	0.37
	5	5	7.0	78.13	1853.27	0.38
	5	5	7.2	76.92	1813.06	0.38
	5	5	7.4	75.76	1762.86	0.39
	5	5	7.6	74.63	1702.68	0.39
	5	5	7.8	73.53	1602.51	0.40
	5	5	8.0	76.92	1553.06	0.38
	5	5	8.2	78.13	1453.27	0.38
	5	5	8.4	79.37	1403.50	0.37
	5	5	8.6	74.63	1342.68	0.39
	5	5	8.8	73.53	1302.51	0.40
	5	5	9.0	71.43	1202.21	0.41

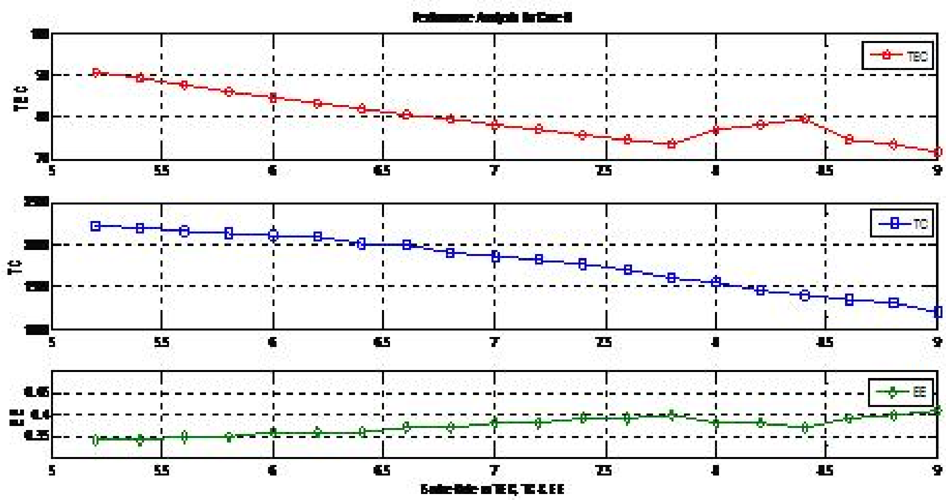

***Figure* 6.2:** *Classification of Cloud Computing Networks*

Case- III

In this case, we examine scenarios where the number of servers is augmented to analyze performance metrics such as queue length, waiting time, utilization factor, as well as total energy consumption, energy efficiency and overall cost within cloud computing networks. This investigation employs the $M/G/c$ queue model, incorporating parameters like arrival rate, service rate, batch size of service rate and the number of servers, with predetermined values for λ, μ and c. Let we assume the service rate $= 5.5$ and number of arrival rate $= 5$ are fixed in this case.

Table **6.3***: Performance Analysis of Cloud Computing Networks (Case* **III***)*

Scenarios	λ	μ	**c**	**TEC**	**TC**	**EE**
When Servers Increase	10	6.5	2	153.85	80440.74	2.03
	10	6.5	3	153.85	28727.66	1.28
	10	6.5	4	153.85	11517.82	1.05
	10	6.5	5	153.85	4467.47	0.94
	10	6.5	6	153.85	2517.61	0.87
	10	6.5	7	153.85	2135.10	0.83
	10	6.5	8	153.85	2005.55	0.80
	10	6.5	9	153.85	1990.44	0.78
	10	6.5	10	153.85	1900.03	0.77
	10	6.5	11	153.85	1850.00	0.76
	10	6.5	12	153.85	1810.00	0.75
	10	6.5	13	153.85	1760.00	0.74
	10	6.5	14	153.85	1700.00	0.73
	10	6.5	15	153.85	1600.00	0.72
	10	6.5	16	153.85	1550.00	0.72
	10	6.5	17	153.85	1450.00	0.71
	10	6.5	18	153.85	1400.00	0.71
	10	6.5	19	153.85	1340.00	0.71
	10	6.5	20	153.85	1300.00	0.70
	10	6.5	21	153.85	1200.00	0.70

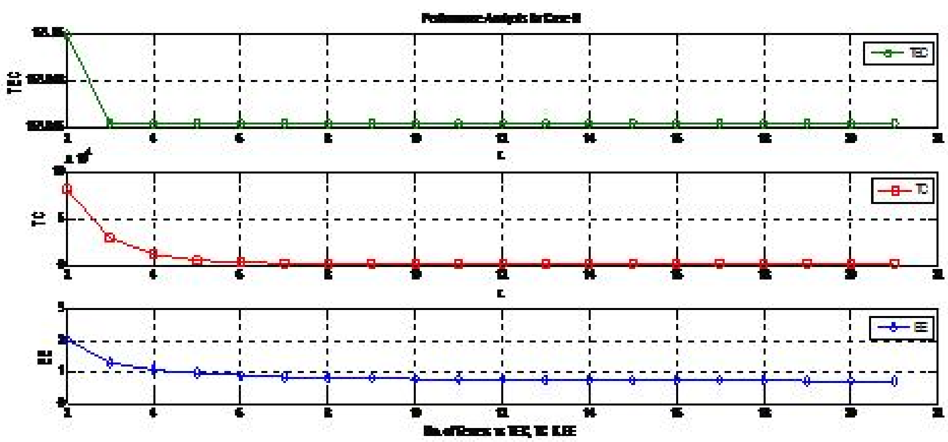

Figure* 6.3: Classification of Cloud Computing Networks*

7 Result Discussion

The proposed model delivers the most optimal cost by considering the number of servers, arrival rate, and service rate, leveraging cloud services by numerical illustrations. This optimization extends to both total energy consumption and energy efficiency, catering to diverse service requirements across various scenarios. Consequently, it is imperative to architect and assess the optimal cost, ensuring a seamless transition into the computer design for future prospects. In this context, we delineate the details of the three aforementioned cases for a comprehensive understanding.

For case 1, when the arrival rate is 2, the total energy consumption is 36.36, the total cost is 2200.00, and the energy efficiency is 0.12.As the arrival rate progressively increases, there is a concurrent escalation in both energy consumption and cost. This dynamic relationship among the arrival rate, energy consumption cost and energy efficiency provides valuable insights into the system's behavior under diverse workloads. Such insights serve as a crucial foundation for informed decision-making, guiding strategies for system optimization and resource allocation to enhance overall efficiency and cost-effectiveness.

For case II, when the service rate is 5.5, the system exhibits a total energy of 90.91, a corresponding total cost of 2206.38, and an energy efficiency of 0.34. Notably, as the service rate undergoes to increase both the total energy consumed and the associated costs display fluctuations, thereby influencing changes in energy efficiency. This dataset offers valuable insights into the system's performance across a spectrum of service rates. Such insights serve as a crucial foundation for informing decision-making, facilitating optimization and resource allocation strategies that consider the efficiency and cost factors. This comprehensive understanding of the system's behavior under varying service rates contributes to the development of well-informed strategies at enhancing overall system performance.

For case III, when the number of servers is 2, the total energy consumed is 153.85, the total cost is 80440.74, and the energy efficiency is 2.03.As the number of servers increases, then dynamic interplay unfolds: the total energy consumed and associated costs exhibit fluctuations, leading to shifts in energy efficiency. This dataset not only sheds light on the system's behavior during the scaling of servers but also offers crucial insights. In turn of server as a valuable guide for decision-making processes related to resource allocation and system optimization. By the efficiency and cost considerations, this information aids in formulating strategies that align with the system's optimal performance under varying server configurations.

8 Conclusion

This study delves into the evaluation of costs in cloud computing systems, with a specific emphasis on minimizing waiting times through the utilization of the $M/G/c$ queueing model. Our investigation centres on the dynamic landscape of cloud computing, covering *IaaS, PaaS, SaaS, MaaS, CaaS, and XaaS*. The primary objective is to scrutinize and optimize the cost dynamics inherent in these systems. A central theme emerges in the form of the integration of cloud computing with virtualization services, and highlighting its pivotal role in bolstering efficiency and cost reduction. Through a detailed examination of numerical illustrations, we explore the influence of varying parameters on cloud services, providing valuable insights into the intricate relationships among factors such as the number of servers, arrival rate and service rate. This research significantly contributes to the on-going discourse on optimizing costs in cloud computing, laying the groundwork for future developments and enhancements in the field. As the demand for cloud services continues to grow and underscore the critical importance of refining cost considerations efficient and economically viable cloud computing environment. There are lot of real world applications will be performed by the queueing models. In this investigation, we analyse the $M/G/C$ queueing model by scaling and tuning the various constraints and consequently the observations has been depicted in figures 6.1 to 6.3. The proposed model emphasis for various services such as (*IaaS, PaaS, SaaS, MaaS, CaaS, and XaaS*)to manage incoming requests of clients over loading of the systems for reducing their waiting time. Further this work can be extended for customer order and data processing on large scale which is the most realistic application of queueing models in cloud computing.

References

[1] A. Anupama, Using Queuing Theory the Performance Measures of Cloud with Infinite Servers, *International Journal of Computer Science & Engineering Technology (IJCSET)*, **5**(1) (2014), 18-21.

[2] S. Adhikari, M. A. Hutaihit, M. Chakrabortya, S. Mahmood, B. Durakovic, S. Pal, D. Akila, Ahmed J. Obaid, Analysis of average waiting time and server utilization factor using queueing theory in cloud computing environment, *Int. J. Nonlinear Anal. Appl.* 12, Special Issue, winter and spring (2021), 1259-1267. DOI:10.22075/ijnaa.2021.5636.

[3] O. J. Boxma, J. W. Cohen and N. Huffels, Approximations of the mean waiting time in an *M/G/s queueing system, Operations Research*, **27**(6) (1979), 1115-1127.

[4] A. Berl, ErolGelenbe, M. D. Girolamo, G. Giuliani, H. de Meer1, M. Q. Dang and K. Pentikousis, Energy-Efficient Cloud Computing, T*he Computer Journal,* **53**, (2009), 1-7.

[5] J. Cao, K Hwang, K. Li, Optimal multi-server configuration for profit maximization in cloud computing, *IEEE Transactions on Parallel and Distributed Systems*, **24**(6) (2012), 1087-1096.

[6] S. Chaisiri, B. S. Lee and D. Niyato, Optimization of Resource Provisioning Cost in Cloud Computing, *IEEE on Transactions on Services Computing*, **5**(2) (2012), 164-177.

[7] M. Eisa, Enhancing Cloud Computing Scheduling Based on Queuing Models, *International Journal of Computer Applications*, **85**(2) (2014), 17-23.

[8] Furht, *Cloud Computing Fundaments, Handbook of Cloud Computing,* Springer New York Dordrecht Heidelberg London, 2010.
[9] Gross and Harris, *Fundamental of Queueing Theory,* John Wiley & Sons, 4th edition, 2014.
[10] R. Ghose, K.S. Trivedi, V.K. Naik and D.S. Kim, End-to-End Perform ability Analysis for Infrastructure-as-a-Service Cloud: *An Interacting Stochastic Models Approach, Conference paper,* DOI: 10.1109/PRDC.2010.30, (2010), 125-231.
[11] E. J. Ghomi, A. M. Rahmani and N. N. Qader, Applying queue theory for modeling of cloud computing: *A systematic review, Concurrency Computat PractExper*, **31**(17) (2019), 1-31, DOI: 10.1002/cpe. 5186.
[12] H. Khazaei, J. Misic and V. B. Misic, Performance analysis of cloud computing centres using *M/G/m/m+r* queueing systems, *IEEE Transactions on Parallel and Distributed Systems*, **23**(5) (2012), 936-943.
[13] J. Kumar and V. Shinde, Performance Evaluation Bulk Arrival and Bulk Service with Multi-Server using Queue Model, *International Journal of Research in Advent Technology*, **6**(11) (2018), 3069-3076.
[14] S. El Kafhali and K. Salah, Stochastic modelling and analysis of cloud computing data center, 2017 20*th Conf. Innov. Clouds Internet Networks* (2017) 122-126.
[15] G. V. Lakshmi and C. S. Bindhu, A Queuing Model To Improve Quality of Service by Reducing Waiting Time in Cloud, International Journal of Soft Computing and Engineering (*IJSCE*), **4**(5) (2014), 1-3.
[16] B. N. W. Ma and J. W. Mark, Approximation of the mean queue length of an *M/G/c* queueing system, *Operations Research,* **43**(1) (1995), 158-165.
[17] J. Mei, K. Li and A. Ouyang, A profit maximization scheme with guaranteed quality of service in cloud computing, *IEEE Transactions on Computers*, 64(11) (2015), 3064-3078.
[18] A. Outamazirt, K. Barkaoui and D. Assani, Maximizing profit in cloud computing using *M/G/c/k* queuing model, *International Symposium on Programming and Systems* ,(*ISPS*) (2018), 1−6.
[19] T. S. D. Praveen, K. Satish and A. Rahiman, The Queueing Theory in Cloud Computing to Reduce the Waiting Time, *International Journal of Computer Science & Engineering Technology (IJCSET)*, **1** (2011), 110-112.
[20] J. A. Resing, *Queueing Theory,* Eindhoven, the Netherlands: Eindhoven University of Technology, Press, 2002.
[21] J. W. Roch, E. Gomes, V. Barbosa, A. Sabino, L. N. Lima, G. Callou, F. A. Silva, E. Choi, T. A. Nguyen, D. Min and J. W. Lee, Enhancing data harvesting systems: Performance quantification of Cloud-Edge-sensor networks using queueing theory, *ICT Express*, **11**(4) (2025), 597-602.
[22] S. Suakanto, S. H. Supangkat, Suhardi and R. Saragih, Performance Measurement of Cloud Computing Services, *International Journal on Cloud Computing: Services and Architecture(IJCCSA)*, **2**(2) (2012), 9-20.
[23] Y. Takahashi, An Approximation Formula for the Mean Waiting time of an *M/G/c* Queue, *Journal Operational Research Society*, **20** (1977), 150-163.
[24] W Whitt, Approximations for the *GI/G/m* queue, *Production and Operations Management*, **2**(2) (1993), 114-161.
[25] L. Wang, G. V. Laszewski, A. Younge, X. He, M. Kunze, J. Tao and C. Fu, Cloud computing a perspective study, *New Generation Computing*, **28** (2010), 137-146.
[26] K. Xong and H. Perros, Service Performance and Analysis in Cloud Computing, *Proceedings of the 2009 Congress on Services - I,* Los Alamitos, CA, USA, (2009), 693-700.
[27] Y. N. Xia, M. C. Zhou, X Luo, Stochastic modeling and performance analysis of migration-enabled and error-prone clouds, *IEEE Transactions on Industrial Informatics*, **11**(2)(2015), 495-504.
[28] S. Zhuravlev, J. C. Saez, Blagodurov Setal, Survey of energy-cognizant scheduling techniques, *IEEE Transactions on Parallel and Distributed Systems*, **24**(7) (2012), 1447-1464.