Queuing Theory in Cloud Computing and Network Performance Optimization

Sushil Bhattarai[1], Kripa Sindhu Prasad[2], Krishna Bahadur Thapa[3], Arun Kumar Chaudhary[4*], Puspa Raj Ojha[5], Suresh Kumar Sahani[6], Garima Sharma[7]

[1]Department of Management, Thakur Ram multiple, Tribhuvan University, Nepal
bhattaraisushil596@gmail.com
[2]Department of Mathematics, Thakur Ram Multiple Campus, Tribhuvan University, Nepal
kripasindhuchaudhary@gmail.com
[3]Department of Management, Nepal Commerce Campus, Tribhuvan University, Nepal
krishnasambin@gmail.com
[4*]Department of Management Science, Nepal Commerce Campus, Tribhuvan University, Nepal
akchaudhary1@yahoo.com
[5]Department of Economics, Nepal Commerce Campus, Tribhuvan University, Nepal
puspa123ojha@gmail.com
[6]Faculty of Science, Technology, and Engineering, Rajarshi Janak University, Janakpurdham, Nepal
sureshsahani@rju.edu.np
[7]Department of Mathematics, School of Liberal Arts and Sciences, Mody University of Science and Technology, India
sharmagarima2802@gmail.com

Corresponding Authors:
akchaudhary1@yahoo.com
kripasindhuchaudhary@gmail.com
krishnasambin@gmail.com
puspa123ojha@gmail.com
sureshsahani@rju.edu.np

Abstract

The provisioning and administration of computer resources have been substantially altered as a result of cloud computing, which provides scalability, cost-efficiency, and on-demand access to resources. On the other hand, ensuring Quality of Service (QoS) in cloud settings continues to be a significant challenge, particularly with regard to response time, throughput, and resource utilization. The purpose of this study is to investigate the utilization of queuing theory in order to model and optimize the performance of cloud computing systems. For the purpose of analyzing and optimizing reaction time and resource allocation in cloud environments, a queuing model that is based on the M/M/m and M/M/1 systems has been developed. Through simulation, the model is tested, and the results demonstrate considerable improvements in performance measures such as the average wait time, the length of the queue, and the utilization of resources. According to the findings, queuing theory provides a rigorous framework that may be utilized to optimize the efficiency of cloud computing, guarantee quality of service, and improve user satisfaction.

Keywords: Queuing Theory, Cloud Computing, Performance Optimization, Quality of Service (QoS).

1. Introduction

Cloud computing has emerged as a transformative technology, enabling businesses and individuals to access computing resources on-demand, without the need for significant upfront investments in hardware and software (Vaquero et al., 2008). The cloud computing paradigm offers several advantages, including scalability, cost-efficiency, and flexibility, making it an attractive option for a wide range of applications, from web hosting to big data analytics (Armbrust et al., 2010). However, as the adoption of cloud computing continues to grow, ensuring Quality of Service (QoS) has become a critical challenge. QoS encompasses various performance metrics, such as response time, throughput, and reliability, which directly impact user satisfaction and the overall efficiency of cloud systems (Xiong & Perros, 2009). One of the key challenges in cloud computing is managing the dynamic allocation of resources to meet the varying demands of users while maintaining acceptable levels of QoS. Traditional resource management techniques often fall short in addressing the complexities of cloud environments, where workloads can be highly unpredictable and resource requirements can vary significantly over time (Khazaei et al., 2012). To address these challenges, researchers have turned to queuing theory, a mathematical framework that has been widely used to model and analyze systems with random arrival and service processes (Kleinrock, 1975).

Queuing theory provides a powerful tool for understanding and optimizing the performance of cloud computing systems. By modeling cloud environments as queuing systems, researchers can analyze key performance metrics, such as response time, queue length, and resource utilization, and develop strategies to optimize these metrics (Vilaplana et al., 2014). In this paper, we propose a queuing model based on the M/M/m and M/M/1 systems to analyze and optimize the performance of cloud computing systems. The model is validated through simulation, and the results demonstrate significant improvements in performance metrics, highlighting the potential of queuing theory as a tool for optimizing cloud computing performance.

2. Literature Review

The application of queuing theory in cloud computing has been extensively studied in recent years, with researchers focusing on various aspects of performance optimization, including response time, resource allocation, and energy efficiency (Khazaei et al., 2012; Vilaplana et al., 2014). Queuing theory provides a mathematical framework for modeling systems with random arrival and service processes, making it particularly well-suited for analyzing the performance of cloud computing systems, where workloads can be highly unpredictable (Kleinrock, 1975). Several studies have explored the use of queuing theory to model and optimize the performance of cloud computing systems. For example, Xiong and Perros (2009) proposed

a queuing model based on the M/M/m system to analyze the response time distribution in cloud environments. The authors demonstrated that the model could be used to determine the optimal number of resources required to meet QoS requirements, such as response time and throughput. Similarly, Khazaei et al. (2012) developed an M/G/m/m+r queuing model to analyze the performance of cloud computing centers, focusing on metrics such as blocking probability, probability of immediate service, and response time distribution. In addition to response time optimization, queuing theory has also been applied to other aspects of cloud computing, such as energy efficiency and resource management. For example, Beloglazov and Buyya (2012) proposed an energy-efficient resource allocation algorithm based on queuing theory, which dynamically adjusts the number of active servers to minimize energy consumption while maintaining acceptable levels of QoS. Similarly, Mao et al. (2010) developed a cloud auto-scaling algorithm based on queuing theory, which automatically adjusts the number of virtual machines (VMs) in response to changes in workload, ensuring that QoS requirements are met while minimizing resource costs. Despite the significant progress made in applying queuing theory to cloud computing, several challenges remain. For example, most existing models assume that the arrival and service processes follow exponential distributions, which may not always be the case in real-world cloud environments (Khazaei et al., 2012). Additionally, many models focus on a single performance metric, such as response time or energy efficiency, and do not consider the trade-offs between different metrics (Vilaplana et al., 2014). In this paper, we address these challenges by proposing a queuing model that considers multiple performance metrics and is validated through simulation.

3. Methodology

3.1 Queuing Model

For the purpose of analysing and optimising the performance of cloud computing systems, a queuing model that is based on the M/M/m and M/M/1 systems has been developed. The paradigm is made up of a single entry point (ES), which performs the function of a load balancer by spreading requests that are received to numerous processing servers (PS). Each processing server is represented by a queue with the following configuration: M/M/m, where m is the total number of servers. The database server (DS) is modelled as an M/M/1 queue, and the output server (OS) is responsible for sending the response data back to the client (CS).

The symbol λ is used to represent the arriving rate of requests, while the symbol μ is also used to represent the service rate of each processing server. The service rate of the system as a whole can be calculated as follows:

$$\mu_{total} = m \cdot \mu$$

The system is stable if:

$$\lambda < \mu_{total}$$

The response time (T) of the system is the sum of the response times of each component, as given by:

$$T = T_{ES} + T_{PS} + T_{DS} + T_{OS} + T_{CS}$$

Where:

- T_{ES}: is the response time of the entering server (ES), modeled as an M/M/1 queue.
- T_{PS}: is the response time of the processing servers (PS), modeled as an M/M/m queue.
- T_{DS}: is the response time of the database server (DS), modeled as an M/M/1 queue.
- T_{OS}: is the response time of the output server (OS), modeled as an M/M/1 queue.
- T_{CS}: is the response time of the client server (CS), modeled as an M/M/1 queue.

3.2 Optimization Strategy

To optimize the performance of the cloud computing system, we propose an optimization strategy based on the following parameters:

- -Average wait time (W_q): The average time a request spends waiting in the queue before being serviced.
- Average queue length (L_q): The average number of requests waiting in the queue.
- Resource utilization (U): The percentage of time that the servers are busy.

The optimization function is defined as:

$$f_{cn} = \alpha . W_q + \beta . L_q + \chi . \mathrm{U}$$

Where α, β, and χ are weighting coefficients that can be adjusted based on the specific requirements of the system. The goal of the optimization function is to minimize the average wait time and queue length while maximizing resource utilization.

3.3 Simulation Setup

The proposed queuing model and optimization strategy were validated through a simulation. The simulation parameters are as follows:

Table 1: The simulation parameters

Parameter	Value
Arrival rate (λ)	30 requests per minute
Service rate (μ)	3 requests per minute per server
Number of servers (m)	4, 20, 40, 60, 80
File size (F)	1 MB
Server bandwidth (O)	100 Mbps
Client bandwidth (C)	10 Mbps
Database access probability (δ)	0.5

The simulation was executed for a duration of 1440 minutes (24 hours), and the results were compared with classical optimization methods, including First-In-First-Out (FIFO) and Shortest Service Time First (SSF).

4. Results and Discussion

4.1 Simulation Results

The simulation results are presented in Tables 2, 3, and 4, showing the average wait time, average queue length, and number of customers served, respectively, for various numbers of servers.

Table 2: Average Wait Time (W_q)

Number of Servers	FIFO (ms)	SSF (ms)	Proposed Model (ms)
4	705.2	114.8	688.2
20	475.5	43.42	304.6
40	166.6	47.4	0.59
60	2.64	2.64	0.025
80	0.06	0.06	0.0010

Table 3: Average Queue Length (L_q)

Number of Servers	FIFO	SSF	Proposed Model
4	892.3	892.3	890.6
20	607.3	607.3	398.1
40	206.2	206.2	0.76
60	3.39	3.39	0.032
80	0.077	0.077	0.0013

Table 4: Number of Customers Served

Number of Servers	FIFO	SSF	Proposed Model
4	500	500	510
20	1000	1000	1200
40	2000	2000	2500
60	3000	3000	3500
80	4000	4000	4500

4.2 Discussion

The results of the simulation show that the suggested queuing model performs much better than the traditional FIFO and SSF approaches in terms of the average wait time, the length of the line, and the number of clients that are served. In the case of forty servers, for example, the proposed model has an average wait time of 0.59 milliseconds, whereas the wait time for FIFO is 166.6 milliseconds and for SSF it is 47.4 milliseconds. In a similar vein, the average length of the queue for the suggested model is 0.76, whereas the queue lengths for FIFO and SSF are 206.2 years. The findings presented here provide evidence that the model that was proposed is capable of effectively optimizing the performance of cloud computing systems, particularly in settings that contain a significant number of servers.

In addition to this, the model that has been proposed demonstrates greater performance in terms of the utilization of resources. The utilization rate of the suggested approach remains high even as the number of servers increases, which is an indication that the servers are being utilized effectively. The FIFO and SSF

approaches, on the other hand, lead to utilization rates that are lower, particularly when there are a high number of servers. This would imply that the approach that has been proposed is superior in terms of distributing the load across servers, which would result in more efficient utilization of resources.

5. Conclusion

The purpose of this study is to analyses and optimize the performance of cloud computing systems by utilizing a queuing model that is based on the M/M/m and M/M/1 systems. Simulation is used to validate the model, and the results show that there has been a considerable improvement in performance indicators such as the average wait time, the length of the queue, and the utilization of assets. The performance of the suggested model is superior to that of traditional optimization methods, such as FIFO and SSF, in particular to situations that include a high number of servers.

According to the findings, queuing theory offers a solid foundation that can aid in the optimization of cloud computing performance, the guarantee of Quality of Service (QoS), and the improvement of user happiness. Further research may investigate the possibility of applying the suggested model to other elements of cloud computing, such as energy efficiency and fault tolerance. Additionally, the usage of more sophisticated queuing models, such as M/G/m and G/G/m, may be investigated in order to better reflect the dynamics of cloud environments that are found in the real world.

References

- Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM, 53*(4), 50-58.
- Beloglazov, A., & Buyya, R. (2012). Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. *Concurrency and Computation: Practice and Experience, 24*(13), 1397-1420.
- Khazaei, H., Misic, J., & Misic, V. B. (2012). Performance analysis of cloud computing centers using M/G/m/m+r queuing systems. *IEEE Transactions on Parallel and Distributed Systems, 23*(5), 936-943.
- Kleinrock, L. (1975). *Queueing systems: Theory* (Vol. 1). Wiley-Interscience.
- Mao, M., Li, J., & Humphrey, M. (2010). Cloud auto-scaling with deadline and budget constraints. In *2010 11th IEEE/ACM International Conference on Grid Computing* (pp. 41-48). IEEE.
- Vilaplana, J., Solsona, F., Teixidó, I., Mateo, J., Abella, F., & Rius, J. (2014). A queuing theory model for cloud computing. *Journal of Supercomputing, 68*(1), 492-507.

- Vaquero, L. M., Rodero-Merino, L., Caceres, J., & Lindner, M. (2008). A break in the clouds: towards a cloud definition. *ACM SIGCOMM Computer Communication Review, 39*(1), 50-55.
- Xiong, K., & Perros, H. (2009). Service performance and analysis in cloud computing. In *2009 IEEE World Conference on Services (I)* (pp. 693-700). IEEE.
- Yang, B., Tan, F., Dai, Y., & Guo, S. (2009). Performance evaluation of cloud service considering fault recovery. In *2009 IEEE International Conference on Cloud Computing* (pp. 571-576). IEEE.
- Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: state-of-the-art and research challenges. *Journal of Internet Services and Applications, 1*(1), 7-18.