ҚАЗАҚСТАН РЕСПУБЛИКАСЫНЫҢ ҒЫЛЫМ ЖӘНЕ ЖОҒАРЫ БІЛІМ МИНИСТРЛІГІ
МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН
MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE REPUBLIC OF KAZAKHSTAN

ХАЛЫҚАРАЛЫҚ УНИВЕРСИТЕТІ INTERNATIONAL UNIVERSITY SINCE 2009 МЕЖДУНАРОДНЫЙ УНИВЕРСИТЕТ

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

2025 (21) 1

ақпан - наурыз

ISSN 2708–2032 (print)
ISSN 2708–2040 (online)

Халықаралық ақпараттық және коммуникациялық технологиялар журналы
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Меншіктенуші: «Халықаралық ақпараттық технологиялар университеті» АҚ (Алматы қ.).
Қазақстан Республикасы Ақпарат және әлеуметтік даму министрлігінің Ақпарат комитетінде – **20.02.2020** жылы берілген.
№ KZ82VPY00020475 мерзімдік басылым тіркеуіне қойылу туралы куәлік.
Тақырыптық бағыты: ақпараттық технологиялар, әлеуметтік-экономикалық жүйелерді дамытудағы цифрлық технологиялар, ақпараттық қауіпсіздік және коммуникациялық технологияларға арналған.
Мерзімділігі: жылына 4 рет.
Тиражы: 100 дана
Редакцияның мекенжайы: 050040, Алматы қ-сы, Манас к-сі, 34/1, 709-кабинет, тел: +7 (727) 244-51-09.
E-mail: ijict@iitu.edu.kz Журнал сайты: https://journal.iitu.edu.kz

---

Международный журнал информационных и коммуникационных технологий
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Собственник: АО «Международный университет информационных технологий» (г. Алматы).
Свидетельство о постановке на учет периодического печатного издания в Министерство информации и общественного развития Республики Казахстан **№ KZ82VPY00020475,** выданное от **20.02.2020 г.**
Тематическая направленность: информационные технологии, информационная безопасность и коммуникационные технологии, цифровые технологии в развитии социо-экономических систем**.**
Периодичность: 4 раза в год.
Тираж: 100 экземпляров.
Адрес редакции: 050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09.
E-mail: ijict@iitu.edu.kz Сайт журнала: https://journal.iitu.edu.kz

«International Journal of Information and Communication Technologies»
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Owner: International Information Technology University JSC (Almaty).
The certificate of registration of a periodical printed publication in the Ministry of Information and Social Development of the Republic of Kazakhstan, Information Committee No. KZ82VPY00020475, issued on 20.02.2020.
Thematic focus: information technology, digital technologies in the development of socio-economic systems, information security and communication technologies
Periodicity: 4 times a year.
Circulation: 100 copies.
Editorial address: 050040. Manas st. 34/1, Almaty. +7 (727) 244-51-09. E-mail: ijict@iitu.edu.kz
Journal website: https://journal.iitu.edu.kz

МАЗМҰНЫ

ӘЛЕУМЕТТІК-ЭКОНОМИКАЛЫҚ ЖҮЙЕЛЕРДІ ДАМЫТУДАҒЫ ЦИФРЛЫҚ ТЕХНОЛОГИЯЛАР

АҚПАРАТТЫҚ ТЕХНОЛОГИЯЛАР

АҚПАРАТТЫҚ ҚАУІПСІЗДІК ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАРҒА АРНАЛҒАН

СОДЕРЖАНИЕ

ЦИФРОВЫЕ ТЕХНОЛОГИИ В РАЗВИТИИ СОЦИО-ЭКОНОМИЧЕСКИХ СИСТЕМ

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ И КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ

CONTENT

DIGITAL TECHNOLOGIES IN THE DEVELOPMENT OF SOCIO-ECONOMIC SYSTEMS

INFORMATION TECHNOLOGY

INFORMATION SECURITY AND COMMUNICATION TECHNOLOGIES

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 1. Number 21 (2025). Pp. 113–126
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.21.1.008

CROSS-SYSTEM ANALYSIS OF QUEUEING SYSTEMS INTERACTIONS IN DISTRIBUTED NETWORKS

S.B. Mukhanov*, D.M. Amrin, S.Zh. Zhakybpekov
International Information Technology University, Almaty, Kazakhstan.
E-mail: s.mukhanov@edu.iitu.kz

Samat B. Mukhanov — PhD in Computer Systems and Software Engineering, Head of «Computer Engineering» Department, assistant-professor, International Information Technology University, Almaty, Kazakhstan E-mail: s.mukhanov@iitu.edu.kz, https://orcid.org/0000-0001-8761-4272;
Daryn M. Amrin — master student of Software Engineering, «Computer Engineering» Department, International Information Technology University, Almaty, Kazakhstan
E-mail: 41376@iitu.edu.kz, https://orcid.org/0009-0003-0684-6947;
Syrym Zh. Zhakypbekov — Doctoral (PhD) student in «Computer Engineering», senior lecturer, «Computer Engineering» Department, International Information Technology University, Almaty, Kazakhstan E-mail: syrymzhakypbekov@gmail.com, https://orcid.org/0000-0001-9112-5922.

Abstract. The increasing reliance on cloud computing has brought both remarkable opportunities and complex challenges. Distributed networks, such as cloud computing environments, are inherently dynamic and require robust systems to handle varying workloads efficiently. This study explores the intricate interactions between multiple queueing models within these systems, providing a fresh perspective on their impact on performance. By using queueing theory, we delve into key metrics like response time, system throughput, and resource allocation strategies. The authors' analysis highlights the advantages of horizontal and vertical scaling, container orchestration, and the critical role of network bandwidth and latency in optimizing cloud infrastructure performance. The researchers also examines how queueing behaviors affect service-level agreements (SLAs), system availability, fault tolerance, and energy efficiency. The integration of analytical and simulation-based approaches enables to evaluate real-world scenarios and identify performance bottlenecks. The findings not only demonstrate the practical applications of queueing theory but also emphasize its relevance in managing resource contention and improving elasticity in multi-tenant environments. This research lays the groundwork for future development of predictive and adaptive models, contributing to the design of smarter, more scalable, and resilient cloud-based systems.

Keywords: distributed networks, cloud computing, queueing theory, response time, workload management, server utilization

For citation: Amrin D.M., Mukhanov S.B., Zhakypbekov S.Zh. CROSS-SYSTEM ANALYSIS OF QUEUEING SYSTEMS INTERACTIONS IN DISTRIBUTED NETWORKS//INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES. 2025. Vol. 6. No. 21. Pp. 113–126 (In Eng.). https://doi.org/10.54309/IJICT.2025.21.1.008.

МЕЖСИСТЕМНЫЙ АНАЛИЗ ВЗАИМОДЕЙСТВИЙ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ В РАСПРЕДЕЛЕННЫХ СЕТЯХ

С.Б. Муханов*, Д.М. Амрин, С.Ж. Жакыпбеков
International Information Technology University, Almaty, Kazakhstan.
E-mail: s.mukhanov@edu.iitu.kz

Самат Б. Муханов — PhD, ассистент-профессор, кафедра Компьютерной инженерии, Международный университет информационных технологий, Алматы, Казахстан
E-mail: s.mukhanov@iitu.edu.kz, https://orcid.org/0000-0001-8761-4272;
Дарын М. Амрин — магистрант кафедры «Компьютерная инженерия», специальность «Программная инженерия», Международный университет информационных технологий, Алматы, Казахстан
E-mail: 41376@iitu.edu.kz, https://orcid.org/0009-0003-0684-6947;
Сырым Ж. Жакыпбеков — докторант (PhD) специальности «Компьютерная инженерия», сеньор-лектор, кафедра Компьютерной инженерии, Международный университет информационных технологий, Алматы, Казахстан
E-mail: syrymzhakypbekov@gmail.com, https://orcid.org/0000-0001-9112-5922.

Аннотация. Растущая зависимость от облачных вычислений принесла как замечательные возможности, так и сложные проблемы. Распределенные сети, такие как среды облачных вычислений, по своей сути динамичны и требуют надежных систем для эффективной обработки изменяющихся рабочих нагрузок. В данном исследовании изучаются сложные взаимодействия между несколькими моделями массового обслуживания в изучаемых системах, что дает свежий взгляд на их влияние на производительность. Используя теорию массового обслуживания, авторы углубляются в ключевые показатели, такие как время отклика и стратегии распределения ресурсов. Проведенный анализ подчеркивает преимущества горизонтального масштабирования, вертикального масштабирования и критическую роль пропускной способности сети в оптимизации производительности облачной инфраструктуры. Результаты не только демонстрируют практическое применение теории массового обслуживания, но и под-черкивают ее ценность в решении реальных проблем в распределенных сетях. Это ис-следование закладывает основу для будущих исследований в области прогностических моделей и гибридных подходов, прокладывая путь для более умных, более масштаби-руемых облачных решений.

Ключевые слова: распределенные сети, облачные вычисления, теория массового обслуживания, время отклика, управление рабочей нагрузкой, использова-ние сервера.

Для цитирования: Амрин Д.М., Муханов С.Б., Жакыпбеков С.Ж. Межсистемный анализ взаимодействий систем массового обслуживания в распределенных сетях // МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОН-НЫХ ТЕХНОЛОГИЙ. 2025. Т. 6. No. 21. Стр. 113–126. (На анг.). https://doi.org/10.54309/IJICT.2025.21.1.008.

Introduction

Nowadays, cloud computing has become a popular and efficient tool for industrial and academic applications. Traditional on-premises computing is becoming less efficient for businesses with the evolution of servers. Cloud computing provides capabilities of distributed networks, in which resources are dynamically allocated for tasks on demand and scalability (Katal et al., 2023). Additionally, it provides different service models: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Cloud computing enables scaling vertically, like increasing performance and resources within one virtual machine, and horizontally, like adding more virtual machines to improve performance and balance the load (Chun & Choi, 2014).

However, despite these advantages, cloud environments come with several challenges. One major issue arises when there is a high flow of requests to the servers. In such cases, the waiting time increases, server utilization may exceed optimal thresholds, and system responsiveness deteriorates, potentially leading to increased operational costs and reduced quality of service. Another critical problem is the possibility of task rejection or dropping due to queue overflow or time constraints. Tasks may be discarded before execution because of users' impatience, deadline expiration, or system failures, which can negatively affect user satisfaction and SLA compliance.

To address these problems and minimize the impact of unexpected outcomes, cloud computing systems have widely adopted queueing theory for performance analysis and modeling. Queueing theory allows system architects and researchers to analyze the behavior of cloud infrastructures under different load conditions by studying metrics such as arrival rate, service rate, number of servers, queue length, and system utilization. By modeling and simulating these parameters, decision-makers can identify bottlenecks, predict future behavior under varying workloads, and improve system design.

Furthermore, by applying queueing models, organizations can fine-tune cloud configurations to achieve a balance between performance and cost. For instance, resource allocation strategies can be optimized to dynamically provision or de-provision virtual machines based on demand, thus ensuring efficient use of computational resources. This helps in maximizing resource utilization while minimizing idle time and unnecessary expenditures.

The application of queueing models also supports the design of more reliable and scalable cloud systems. By predicting queue lengths and response times, it is possible to enhance elasticity, reduce latency, and ensure high availability, even under bursty or unpredictable traffic conditions. Moreover, queueing theory can be extended to consider priority scheduling, load balancing algorithms, and multi-class task handling, making it a versatile tool for modern cloud computing environments.

This study aims to analyze the interactions between multiple queueing models in distributed networks as well as their impact on performance of the system.

The main objectives of research:

- Analyze how the queueing interactions in cloud computing affect performance;
- Develop mathematical models for each queueing system within the distributed network;
- Identify key parameters for accurate modeling;
- Compare model in different scenarios.

The relevance of this research is that the cloud computing is increasingly becoming the preferred choice for both industrial and academic applications, surpassing traditional computing in efficiency and scalability. Despite its advantages, cloud computing faces challenges such as high request flow that leads to increased waiting times and server utilization, potentially incurring unexpected costs. Queueing theory provides a robust framework for analyzing and addressing these challenges to improve efficiency of resource utilization. Cloud environments are inherently dynamic and must handle unpredictable workloads, fluctuating user demands, and varying task execution times. When the inflow of service requests exceeds the processing capability of the system, delays increase, resources may become overutilized, and quality of service deteriorates. In some cases, tasks may even be dropped due to queue overflow, user impatience, or service deadlines, affecting system reliability and violating service-level agreements (SLAs).

To address these concerns, queueing theory is employed to model the behavior of cloud systems under different load conditions. It enables the analysis of key parameters such as arrival rate, service rate, queue length, and system utilization. By simulating these interactions, it is possible to design more responsive and scalable cloud infrastructures.

Moreover, queueing models support decision-making related to horizontal and vertical scaling strategies, efficient load balancing, and dynamic resource provisioning. This leads to improved elasticity, cost optimization, and better management of computing resources. By predicting potential bottlenecks and evaluating system performance in advance, organizations can prevent service degradation and maintain high availability.

Therefore, integrating queueing theory into the design and operation of cloud platforms not only enhances system efficiency but also contributes to building resilient, adaptive, and cost-effective cloud computing environments that meet the evolving demands of modern users and applications.

Literature Review

Cloud infrastructure may have various problems, like unexpected internet connection failure, hardware failures, security breaches, resource overload, latency issues, data loss or corruption, which results in disconnection of users and server maintenance, thus leading to performance issues and customer dropout. All these events can happen randomly. Literature review of current study reveals multiple queueing theory approaches for distributed networks, like cloud. Authors use different queueing theory approaches in their works for performance evaluation and system modeling.

Adhikari used M/M/c queueing model for comparison of average waiting time and its analysis, and they focused on reduction of these parameters (Adhikari et al., 2021). They conducted an analysis of web application development on cloud environment at cost point of view and came to conclusion that by increasing the number of severs the average time decreases the server utilization factor, average waiting time and average queue.

Kumar in his paper with the help of queueing theory modelled a cloud computing infrastructure in which some requests being dropped (Kumar et al., 2021). Load balancing performance of cloud computing systems were evaluated using M/M/1/N queueing model with related drops in queues. They used a single server Markovian queueing model with reneging, customers arrive by one based on Poisson process, with the finite capacity.

Ghazali proposed that effective task scheduling algorithms can balance the load, optimize resource usage and performance (Ghazali & Ben Tahar, 2024). They conducted analysis based on queueing theory. They also evaluated generalized processor sharing, in which the server handles all jobs simultaneously.

Arul Freeda assessed the efficiency of cloud service and also tried to improve it by trying to make waiting time smaller (Arul Freeda Vinodhini, 2020). By their observations, transient solution shows better results rather than one in steady state solution. In real time cloud, high end queue concepts are important to consider.

Jordi Vilaplana explores queuing models to analyze and optimize resource allocation in cloud computing environments (Vilaplana et al., 2014). The authors investigate how virtual machines share resources, focusing on CPU (Central Processing Unit), memory, and network utilization, and use queuing theory to address performance and efficiency challenges in cloud systems. Our study integrates some of their proposed techniques to further analyze system performance.

Jordi Vilaplana in his other study explores a queuing-based framework for analyzing fog computing systems (Mas et al., 2022). The study models fog architectures using a closed Jackson network to evaluate interactions between fog, cloud, and client devices. The proposed model focuses on resource allocation, data processing, and performance optimization in fog environments. It provides insights into task o oading, load balancing, and the probability-based routing of jobs in distributed systems.

Queueing theory in practice is used for performance optimization and automation. In an article by Tolosana-Calasanz et al., autonomic controller for streaming applications were proposed (Tolosana-Calasanz et al., 2017). Their system is based on queueing theory and feedback control.

In the work of Guo et al. strategies for optimization of scheduling of virtual machines in cloud computing environments were investigated (Guo et al., 2018), in which authors modeled the system using queueing theory for buffering virtual machines jobs requests into different virtual queues. For intra-queue buffering, a shortest-job-first (SJF) policy is applied, arranging job requests in ascending order based on their lengths. This method aims to reduce the average job completion time by prioritizing shorter tasks. However, SJF can lead to job starvation for longer-duration tasks. To mitigate this issue, the authors propose shortest-job-first buffering and reinforcement-learning-based scheduling (SJF-RL) to provide a low delay and throughput performances.

In their other work they investigated efficient virtual machine scheduling in cloud environments in which workloads vary in size and resource demands (Guo et al., 2022). By employing queueing theory, the authors proposed two scheduling strategies: shortest-job-first with min-min best fit and shortest-job-first with queue-length-based maxweight to optimize job completion time and resource allocation. Their approach addresses limitations of traditional methods like first-come-first serve by dynamically prioritizing jobs based on system conditions. Their simulation results demonstrated significant improvements in reducing delay and preventing job starvation, particularly with the shortest job-first and queue-length-based maxweight strategy.

Atmaca et al. proposed G/G/c model for the evaluation of performance with general interarrivals and service time distributions with multiple servers to meet service level agreement for several quality of service performance metrics like blocking probability and task response, because, according to their assumptions, the model with Poisson arrivals not a perfect choice for a system with arrivals' coefficient of variations that different from value 1, and their model best fit to realistic representation of cloud service dynamics (Atmaca et al., 2016). They consider variability in which model Poisson can have problems with under-provisioning or over-provisioning of resources in a cloud environment.

El Kafhali and Salah model cloud data centers as open queuing systems to estimate quality of service parameters, such as response time, drop rate, and CPU utilization, under varying request arrival rates and different numbers of virtual machine instances (El Kafhali & Salah, 2017). The study provides numerical examples illustrating how the model estimates the required number of virtual machine instances to meet specific quality of service targets. The authors cross-validate their analytical model using a discrete event simulator, and the results demonstrated that the proposed model effectively estimates the number of virtual machines needed to achieve quality of service objectives as arrival request rates change.

The paper of Abd Eldjalil and Lyes for electric vehicle charging-discharging service based on cloud computing proposes a priority-queuing model integrated with cloud computing to optimize electric vehicle charging and discharging processes (Abd Eldjalil & Lyes, 2017). By considering factors such as battery state, user preferences, and station availability, their model dynamically schedules requests to minimize waiting times and enhance infrastructure utilization.

Bai et al. in their study introduce a queuing-based approach to evaluate the performance of cloud data centers with heterogeneous servers in which they develop a mathematical model that accounts for varying service rates, job sizes, and resource allocation strategies to better reflect real-world cloud environments (Bai et al., 2015).

Study of Zuo et al. proposes a scheduling approach to enhance resource utilization and load balancing in cloud computing environments (Zuo et al., 2018). The authors categorize tasks based on their resource demands—such as CPU, I/O (Input/Output), and memory—and assign them to corresponding queues.

Li in his study proposes quantitative models that treats cloud as queueing system enable the assessment of a system's ability to adapt to workload fluctuations by provisioning and de-provisioning resources dynamically, which results in costs reduction and performance improvements (Li, 2020).

Gupta et al. in their study compared multiple scheduling algorithms, like First-Come-First-Serve, Shortest Job Next, Round Robin, and Priority Scheduling (Gupta et al., 2015). According to their results, no single algorithm universally outperforms others across all metrics; instead, each has distinct advantages and limitations depending on specific workload characteristics and system requirements

While previous studies have explored different queuing mechanisms, there remains a need for further analysis in specific cloud environments.

Building on this, our study focuses on analyzing a queueing theory model in a cloud computing environment, providing insights into system performance and efficiency based on theoretical and analytical evaluation.

Materials and methods

There are lots of queueing theory models that can be considered for cloud computing. In this study a combination of couple models will be used.

Table 1 – queueing theory models

Regular models	
M/M/1	It is a single server queueing model in both arrivals and service times. It follows exponential distribution and is usually used for initial performance estimation and small non-scalable applications, and basic analysis.
M/M/c	Multi-server queue model, in which "c" means number of servers. Arrival and service times follows exponential distribution. It is more useful for real world applications, especially in load balancing.
M/M/1/K	The same as M/M/c model, but with finite capacity "K", in which arrivals are blocked when the system is full. Useful for applications that have limited resources and analyzing dropped requests in real systems.
M/M/c/K	Like M/M/1/K model, but with multiple servers that serve requests. Useful for analyzing systems with limited capacity and processes that work in parallel.
M/M/c/N	Like M/M/C, but with a finite N number of people that should be served. Useful for services in which we know number of people that should be served.
Advanced models that are used in complex real-world applications	
G/G/1	Like M/M/1 model, but with non-exponential arrival and service time distributions
G/G/c	Like M/M/c model, but with non-exponential arrival and service time distributions
G/G/1/K	Like M/M/1/K model, but with non-exponential arrival and service time distributions

In this study the M/M/1 queueing system represents a load balancer in network. In M/M/1 model system consists of a single server with both arrivals and service times following exponential distributions (Atefi et al., 2016). It is a single-entry point, which then forwards requests to processing servers that are represented by computational resource in cloud, like core of processor or entire CPU, GPU, which are implemented by M/M/c system. M/M/c queueing system includes multiple servers with exponential distributions and Poisson arrivals (Rathod & Chowdhary, 2019). Then the processing server node accesses database server with certain probability and represented by M/M/1 system to avoid inconsistences in data.

Client server represented by M/M/1 as well because it needs to receive responses from server.

Selecting the right model is crucial for cloud-based servers' evaluation of performance. Distributed Network is represented by simulation of cloud computing models (e.g., public cloud, hybrid cloud) and their challenges related to distributed queueing systems.

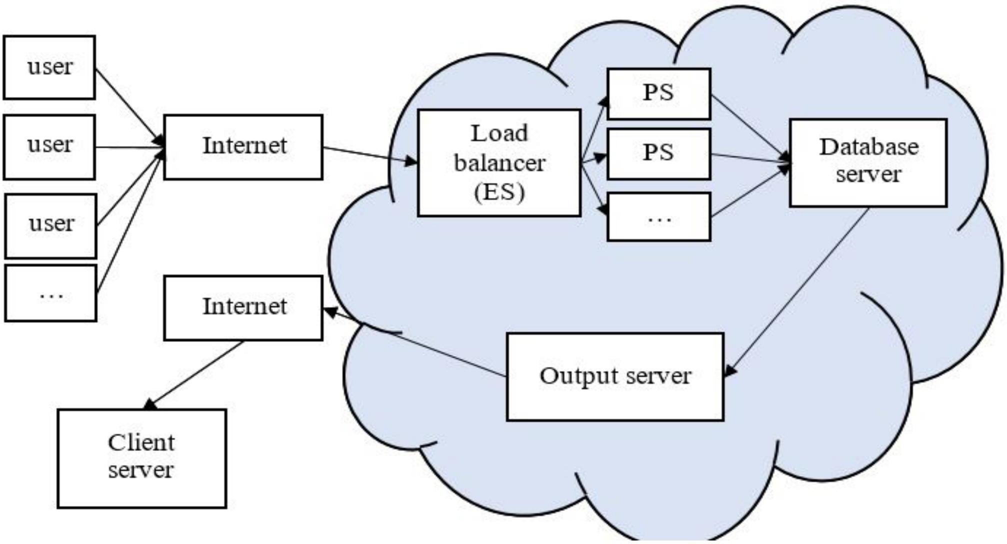

Figure 1 – queueing theory model in cloud computing

Markov chains and stochastic models are used to represent queue states.

Programming Language: Python. Libraries that are used: numpy, matplotlib.pyplot.

Kolmogorov equations represent continuous-time Markov processes.

ES - Entering Server MM1 Calculates the waiting time in the entering server.

$$W_{ES} = \frac{1}{\mu_{ES} - \lambda} \quad (1)$$

Where λ is the arrival rate,   is the service rate of the Entering Server.

PS: Processing Servers MMC Calculates the waiting time in the processing servers using the MMC model.

$$W_{PS} = W_{q,PS} + \frac{1}{\mu} \quad (2)$$

$$W_{q,PS} = \frac{L_{q,PS}}{\lambda} \quad (3)$$

$$L_{q,PS} = \frac{P_0 \cdot (\lambda/\mu)^c \cdot \rho_{PS}}{c! \cdot (1-\rho_{PS})^2} \quad (4)$$

$$P_0 = \left(\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c! \cdot (1-\rho_{PS})} \right)^{-1} \quad (5)$$

Where   is the service rate of processing servers, and c is the number of servers.

Database Server MM1 Calculates the waiting time in the database server using the adjusted arrival rate. Scaling Factor for Arrival Rate Modifies the effective arrival rate for the database server DS and adjusts the impact of γ on the system’s service capacity. Effective

Arrival Rate considers the utilization factor, or time-dependent factor .

$$W_{DS} = \frac{1}{\mu_{DS} - \delta \cdot \gamma} \quad (6)$$

$$\gamma = \frac{\lambda}{1-\tau} \quad (7)$$

Where is a scaling factor for arrival rate (probability of processing node accessing the database server), is the service rate of the database server, is utilization factor.

OS: Output Server (MM1) with Bandwidth Calculates the waiting time in the output server.

$$W_{OS} = \frac{1}{\mu_{OS} - \gamma} \quad (8)$$

Where γ is the arrival rate, is the service rate of the Output Server.

Client Server (MM1) with Bandwidth Calculates the waiting time in the client server.

$$W_{CS} = \frac{1}{\mu_{CS} - \gamma} \quad (9)$$

Where γ is the arrival rate, is the service rate of the Client Server.

$$W = W_{ES} + W_{PS} + W_{DS} + W_{OS} + W_{CS} \quad (10)$$

The total waiting time W is the sum of the waiting times for each server component in the system. Using this method the performance of servers will be modeled and evaluated.

Discussions and results

In this section multiple queueing systems in a distributed cloud computing environment are analyzed to assess interactions between these systems and how they affect performance metrics.

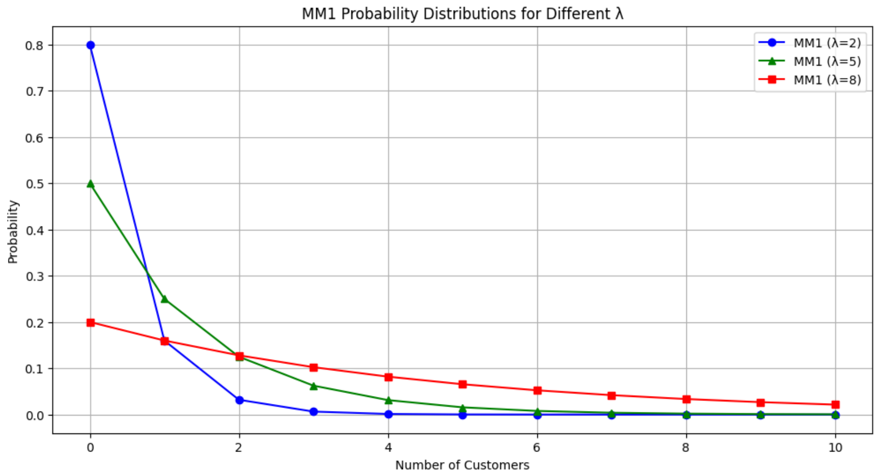

Figure 2 – Probability distributions for M/M/1 model with different arrival rates

Figure 2 illustrates the probability distributions for an M/M/1 queueing model to visualize different arrival rates. The x-axis represents the number of customers, while the y-axis denotes the probability. Higher arrival rates lead to an increased probability of having more customers in the system, which is crucial for performance and capacity planning. Arrival rates with a value of 8 have higher probabilities across the graph than those with values 2 and 5.

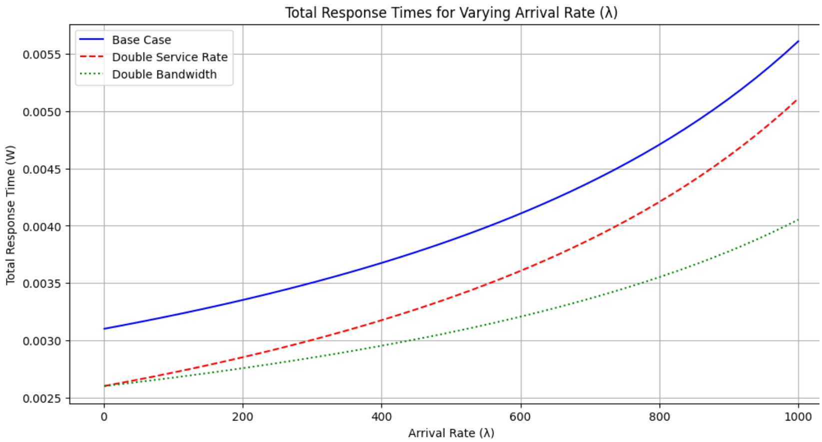

Figure 3 – Probability distributions for M/M/c model with different arrival rates

Figure 3 illustrates the probability distributions for an M/M/C queueing model with different arrival rates. Like the previous graph, the x-axis represents the number of customers, and the y-axis indicates the probability of having that specific number of customers in the system at any given time. It shows that higher arrival rates result in a greater likelihood of more customers in the system, like the M/M/1 model but with multiple servers.

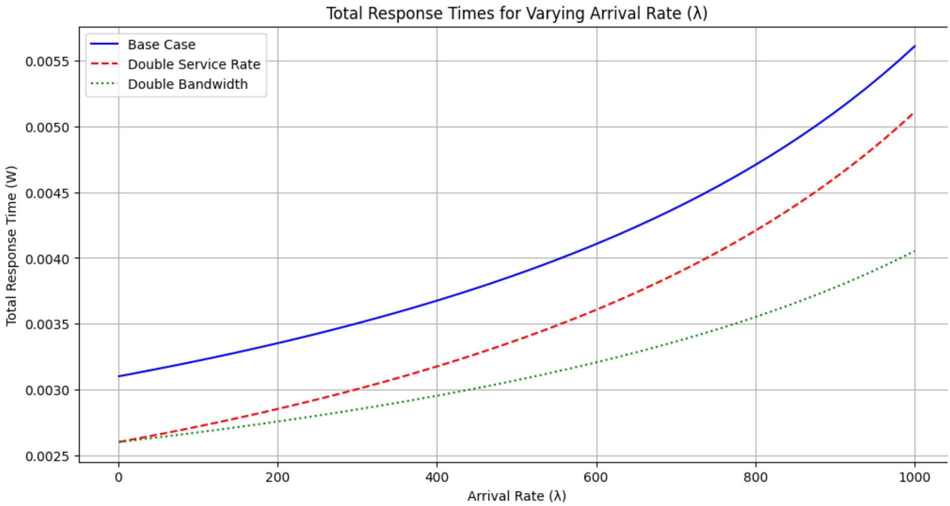

Figure 4 – Total response time (W) with varying arrival rates for M/M/c model

The figure 4 illustrates the relationship between the arrival rate and the response time W in an M/M/c queueing system with varying numbers of servers. The response time increases with the arrival rate for all configurations of servers. This is due to the increased overload of the system as more customers arrive. A notable observation is the effect of the number of servers on the response time. For lower arrival rates, the response time remains low and is not significantly affected by the number of servers. However, as the arrival rate increases, the response time becomes more sensitive to the number of servers. Increasing the number of servers can help mitigate the impact of higher arrival rates on the response time.

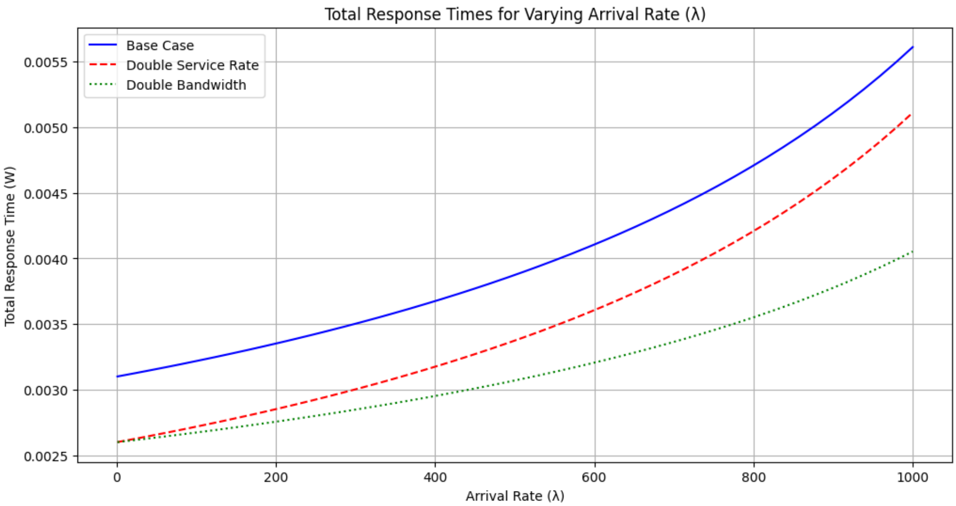

Figure 5 – Total response time (W) with varying arrival rates, base case, double service rate and double bandwidth scenario (= 1000, ES = 10000, c = 10, = 0.5, DS = 1000, OS = 2000)

As from figure 5, there is three cases for total response time with respect to arrival rates: base case, double service rate for processing servers, and double bandwidth. In case if we double both service rate and bandwidth size, both improve in performance, but bandwidth has more effect.

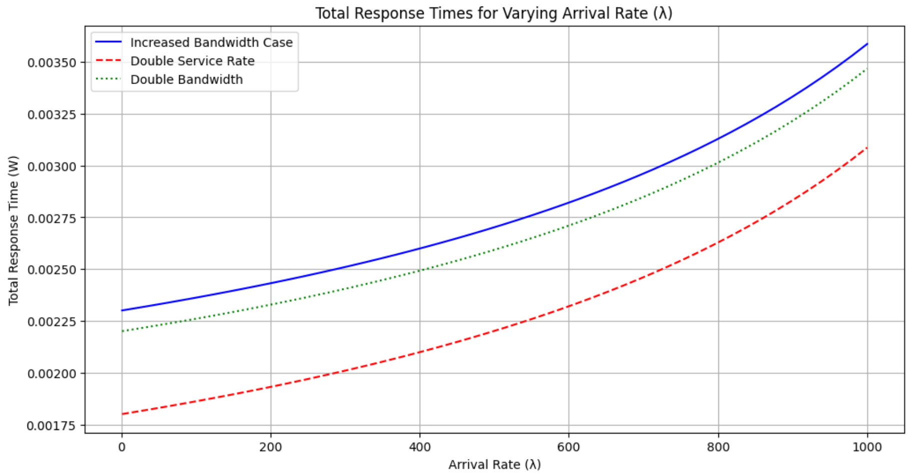

Figure 6 – Total response time (W) with varying arrival rates, base case (increased bandwidth), double service rate and double bandwidth scenario (= 1000, ES = 10000, c = 10, = 0.5, DS = 1000, OS = 10000)

From figure 6 we take the same parameters, but with increased base case scenario. In this case, improving service rate in processing servers have more effect.

Higher arrival rates (λ) increase the system's workload, leading to longer response times and higher server utilization. This trend was evident across all queueing models, particularly in the M/M/c model, where the presence of multiple servers mitigates, but does not eliminate performance degradation.

Adding more servers (c) in M/M/c models significantly reduced waiting times and response times for high arrival rates, demonstrating the benefits of horizontal scaling in cloud infrastructures. However, increasing the number of servers also decreased server utilization, indicating a tradeoff between performance and cost efficiency.

Bandwidth increase showed significant improvement in reducing total response time in the case of lower base bandwidth. This emphasizes the role of network infrastructure in cloud performance.

Doubling the service rate of processing servers had a notable effect on reducing response times in case of increased base bandwidth.

Properly modeling and analyzing queueing systems can guide resource allocation, improve load balancing, and reduce operational costs.

Conclusion

This study demonstrated how queueing theory models can be applied to analyze and optimize the performance of distributed cloud computing systems. Key considerations include modeling various queueing systems, including advanced models, to evaluate their effectiveness in different cloud scenarios. Identifying critical performance metrics, such as response time, and showing its dependence on arrival rates, service rates, and system configurations. This study demonstrated the impact of horizontal scaling, like adding servers, and vertical scaling, like increasing service rates, on overall performance. Highlighting the importance of network bandwidth in distributed environments, which significantly affects the total system response time. The study opens opportunities for further research, such as incorporating machine learning techniques to predict queue states and optimize resource allocation dynamically. Exploring hybrid models that combine analytical and simulation-based approaches could also enhance the robustness of performance evaluations in complex cloud environments. This research provides valuable insights for designing efficient distributed systems, ensuring scalability, and maintaining high-quality service delivery in modern cloud infrastructures.

REFERENCES

Abd Eldjalil C.D. & Lyes K. (2017). Optimal priority-queuing for EV charging-discharging service based on cloud computing. 2017. — IEEE International Conference on Communications (ICC). — 1–6. https://doi.org/10.1109/ICC.2017.7996494

Adhikari S., Hutaihit M., Chakraborty M., Mahmood S., Durakovic B., Pal S., Akila D. & Obaid A. (2021). Analysis of average waiting time and server utilization factor using queueing theory in cloud computing environment. — International Journal of Nonlinear Analysis and Applications. — 12(Special Issue). https://doi.org/10.22075/ijnaa.2021.5636

Arul Freeda Vinodhini G. (2020). CLOUD COMPUTING AS A QUEUE MODEL WITH SERVER BREAKDOWN. — Advances in Mathematics: Scientific Journal. — 9(10). — 8217–8225. https://doi.org/10.37418/amsj.9.10.51

Atefi K., Yahya S., Rezaei A. & Erfanian A. (2016). Traffic behavior of Local Area Network based on M/M/1 queuing model using poisson and exponential distribution. — 2016 IEEE Region 10 Symposium (TENSYMP). — 19–23. https://doi.org/10.1109/TENCONSpring.2016.7519371

Atmaca T., Begin T., Brandwajn A. & Castel-Taleb H. (2016). Performance Evaluation of Cloud Computing Centers with General Arrivals and Service. — IEEE Transactions on Parallel and Distributed Systems. — 27(8). — 2341–2348. https://doi.org/10.1109/TPDS.2015.2499749

Bai W.-H., Xi J.-Q., Zhu J.-X. & Huang S.-W. (2015). Performance Analysis of Heterogeneous Data Centers in Cloud Computing Using a Complex Queuing Model. — Mathematical Problems in Engineering. — 2015. — 1–15. https://doi.org/10.1155/2015/980945

Chun S.-H. & Choi B.-S. (2014). Service models and pricing schemes for cloud computing. — Cluster Computing. — 17(2). — 529–535. https://doi.org/10.1007/s10586-013-0296-1

El Kafhali S. & Salah K. (2017). Stochastic modelling and analysis of cloud computing data center. — 2017 20th Conference on Innovations in Clouds, Internet and Networks (ICIN). — 122–126. https://doi.org/10.1109/ICIN.2017.7899401

Ghazali M. & Ben Tahar A. (2024). A Queuing Theory Approach to Task Scheduling in Cloud Computing with Generalized Processor Sharing Queue Model and Heavy Traffic Approximation. — IAENG International Journal of Computer Science. — 51(10). — 1604–1611. Scopus.

Guo M., Guan Q., Chen W., Ji F. & Peng Z. (2022). Delay-Optimal Scheduling of VMs in a Queueing Cloud Computing System with Heterogeneous Workloads. — IEEE Transactions on Services Computing. — 15(1). — 110–123. https://doi.org/10.1109/TSC.2019.2920954

Guo, M., Guan, Q., & Ke, W. (2018). Optimal Scheduling of VMs in Queueing Cloud Computing Systems With a Heterogeneous Workload. — IEEE Access. — 6. — 15178–15191. https://doi.org/10.1109/ACCESS.2018.2801319

Gupta A., Bhadauria H.S., Singh A. & Patni J.C. (2015). A theoretical comparison of job scheduling algorithms in cloud computing environment. — 2015 1st International Conference on Next Generation Computing Technologies (NGCT). — 16–20. https://doi.org/10.1109/NGCT.2015.7375074

Katal A., Dahiya S. & Choudhury T. (2023). Energy efficiency in cloud computing data centers: A survey on software technologies. Cluster Computing. — 26(3). — 1845–1875. https://doi.org/10.1007/s10586-022-03713-0

Kumar R., Soodan B.S., Kuaban G.S., Czekalski P. & Sharma S. (2021). Performance Analysis of a Cloud Computing System using Queuing Model with Correlated Task Reneging. — Journal of Physics: Conference Series. — 2091(1). — 012003. https://doi.org/10.1088/1742-6596/2091/1/012003

Li K. (2020). Quantitative Modeling and Analytical Calculation of Elasticity in Cloud Computing. — IEEE Transactions on Cloud Computing. — 8(4). — 1135–1148. https://doi.org/10.1109/TCC.2017.2665549

Mas L., Vilaplana J., Mateo J. & Solsona F. (2022). A queuing theory model for fog computing. — The Journal of Supercomputing. — 78(8). — 11138–11155. https://doi.org/10.1007/s11227-022-04328-3

Rathod D. & Chowdhary Dr.G. (2019). Scalability of M/M/c Queue based Cloud-Fog Distributed Internet of Things Middleware. — International Journal of Advanced Networking and Applications. — 11(01). — 4162–

4170. https://doi.org/10.35444/IJANA.2019.11015

Tolosana-Calasanz, R., Diaz-Montes, J., Rana, O. F., & Parashar, M. (2017). Feedback-Control & Queueing Theory-Based Resource Management for Streaming Applications. IEEE Transactions on Parallel and Distributed Systems, 28(4), 1061–1075. https://doi.org/10.1109/TPDS.2016.2603510

Vilaplana, J., Solsona, F., Teixidó, I., Mateo Fornes, J., Abella, F., & Rius, J. (2014). A queuing theory model for cloud computing. The Journal of Supercomputing, 69, 1–16. https://doi.org/10.1007/s11227-014-1177-y

Zuo, L., Dong, S., Shu, L., Zhu, C., & Han, G. (2018). A Multiqueue Interlacing Peak Scheduling Method Based on Tasks' Classification in Cloud Computing. IEEE Systems Journal, 12(2), 1518–1530. https://doi.org/10.1109/JSYST.2016.2542251

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Правила оформления статьи для публикации в журнале на сайте:
https://journal.iitu.edu.kz
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Собственник: АО «Международный университет информационных технологий» (Казахстан, Алматы)

ОТВЕТСТВЕННЫЙ РЕДАКТОР
Мрзабаева Раушан Жалиқызы

КОМПЬЮТЕРНАЯ ВЕРСТКА
Асанова Жадыра

Подписано в печать 15.03.2025.
Формат 60х881/8. Бумага офсетная. Печать - ризограф. 9,0 п.л. Тираж 100
050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09).

Издание Международного университета информационных технологий
Издательский центр КБТУ, Алматы, ул. Толе би, 59