Minimax Theory and its Applications

Volume 11 (2026), No. 5s, 450-463

Finite capacity two heterogeneous servers queueing model with encouraged arrivals, feedback, reneging and retention of reneged customers

P Kavitha[1], M. Thiagarajan[2]

[1*] *Research Scholar, Lecturer, Government Polytechnic College, Perambalur, Tamil Nadu, India. lecpkavithamaths@gmail.com, https://orcid.org/0009-0002-4702-2971*

[2]*Associate Professor and Head, PG & Research Department of Mathematics, St. Joseph's College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nādu, India. thiagarajan_ma1@mail.sjctni.edu, https://orcid.org/0009-0001-8990-115X*

Abstract:

Systems of service encouraging customer arrivals can be affected by increased congestion, especially where service capacities are varied and unhappy customers need to undergo service more than once. The situation is compounded by customer impatience because keeping customers can result in reduced abandonment and increased queue size. In this research paper, a model of finite capacity queuing system with two different servers, encouraged arrivals, feedback, reneging and retention of reneging customers developed. The arrival of customers is modeled using Poisson process with encouragement dependent arrival rate, service and impatience time are distributed exponentially. Numerical sensitivity analysis considers the impacts of the arrival rate parameter, reneging parameter, retention probability, and feedback probability. An increase in the arrival rate parameter from 3 to 10 results in an increase in expected system size from 8.2868 to 9.7438 and an increase in expected queue size from 6.2943 to 7.7438. An increase in the reneging parameter from 1 to 10 leads to a decrease in the expected system size from 8.3358 to 3.3901 even though the average reneging rate rises from 2.5355 to 6.0521. An increase in the retention probability from 0.10 to 1.00 leads to a reduction in the average reneging rate from 1.3461 to zero and an increase in queue size from 4.9851 to 5.9441. An increase in feedback probability from zero to 0.80 causes the system size to rise from 7.4841 to 9.7021. The model facilitates well-coordinated decisions related to customer motivation, customer retention, and service capabilities, whereas there is a need for validation and economic optimization before making recommendations for operational policies.

Keywords: Encouraged arrivals, heterogeneous service, finite capacity, feedback, reneging, retention of reneged customers.

1. Introduction

Queueing Theory plays a significant role in the evaluation and designing service systems in practical applications like telecommunication networks, banking system, health care services, computing systems, transport networks, and customer support center systems. Most of these systems are suffers with service congestion due to customer arrivals that are stochastic in nature, along with limited resources for offering services. It becomes an imperative to design realistic queueing models to represent service systems for improving the efficiency of the service system and ensuring customer satisfaction. Service facilities usually consist of more than one server that can offer different service capabilities depending upon the type of work to be done (Pravina et al., 2025). These queueing systems are known as heterogeneous servers queueing systems and each server offers different service rates. Heterogeneous servers are commonly applied for real-life environment. For example, hospitals in which doctors may possess different capabilities in providing services; call centers where agents differ in offering service; and manufacturing processes in which machines work at different rates.

Another crucial element of the modern service systems is the behavior of customers and customers may behave differently according to the environment of the service system. For example, service providers may use techniques such as encouraged arrival, whereby customers are motivated to come into the system using methods like incentives, advertisement, and better service assurance. Encouragement arrival techniques are producing efficient

result in an increase in the actual arrival rate and consequently impact the entire system's performance. Customers waiting in the queue may get intolerant and may exit from the system queue before getting service (Khan & Paramasivam, 2022; Som & Kumar, 2018). This occurrence is known as reneging. Reneging is common in many service providing systems, which includes call centers, hospitals, and ticket booking systems, wherein customers may renounce their place in queue due to waiting for a very long time. Service providing organizations are focused on to retain the customers who have left the queue by offering retention methods. Feedback is one of the important characteristics for improving the efficiency, which involves a customer who has received his service re-entering the system for more service. Feedback processes can be seen in systems that involve manufacturing, where rework has to be done; in technical support systems; and in systems that involve repairs, where a customer may need multiple service requests to be processed (Kumar, 2022). A number of models have been designed to capture these characteristics of service systems individually. However, very few studies have been done on the combined effect of all these characteristics on a service system with a limited capacity. Based on these pragmatic considerations, the current research formulates a model of a two-server finite capacity queueing process incorporating the aspects of encouraged arrivals, feedback, reneging, and retention of reneged customers. The proposed queueing process modeled using a continuous time Markov chain approach where steady state analysis done using the balance equations approach. Performance measures like mean system size, mean queue size, and reneging rates have been derived (Xu et al., 2026). The findings from the current study shed light on how customer behavior and heterogeneity impact queue efficiency, and prove helpful to managers for devising effective policies for managing the queue process.

1.1 Research objective

The primary objective of the proposed technique is to develop and analyze a finite-capacity for a two-server queueing model with heterogeneous service rates, encouraged arrivals, feedback, reneging, and retention of reneged customers. This research is focusing on to formulate the stochastic model and derive its steady-state probability distribution and key performance measures. This research has examined the effectiveness of arrival rates, service rates, encouragement factors, feedback probability, reneging rates, and retention probability on system performance. This research paper aims to evaluate the performance attributes such as server utilization, expected queue length, system size, waiting time, throughput, and customer retention. The results are intended to provide useful insights for improving resource utilization and service efficiency in practical finite-capacity service systems.

This research article is presenting a finite capacity two heterogeneous servers queueing model with encouraged arrivals, feedback, reneging and retention of reneged customers. The remaining section of the paper has been organized as follows; the section 2 is presenting related work for the heterogeneous server model and section 3 is presenting methodology for the proposed queueing model. The results and discussion of the proposed queueing model has discussed in the section 4. The conclusion of the research paper is discussed in section 5.

2. Related work

Queueing models have increasingly become more realistic through the inclusion of customer impatience, heterogeneous service rates, system interruptions and economic optimization. includes discouraged arrivals, customer impatience and self-switching server characteristics and connects demand behavior with varying service environment (Pravina et al., 2025). But due to the restriction of single server, the applicability of this queueing model is not applicable for systems that consist of multiple servers with heterogeneous capacity. On the other hand, focuses on encouraged arrivals, feedback, balking and reneging customers using an M/M/1/N approach (Khan & Paramasivam, 2022). Even though, this queueing model clearly distinguishes repeat customers from customers' impatience and retention, its restrictions of Markovian single server make the applicability of the model restricted to Markovian arrival and Markovian service process. Heterogeneous service rates are studied where the concept of reverse balking and reneging are introduced and focuses on transient queueing system under finite capacity and finite population as well as reneging (Som & Kumar, 2018; Kumar, 2022). The two queueing models differ in terms of assumptions. Therefore, the assumptions should be carefully analyzed before comparing their findings. The two-phase heterogeneous service rates introduce synchronous working vacation interruptions (Xu et al., 2026).

Nevertheless, this approach can be used only if the real system operates according to the service phases and interruptions stated above. In fact, M/M/c queueing approach in includes heterogeneous servers along with balking and reneging, providing more service capacity representation than approaches using one server; however, Poisson arrivals and exponentially distributed service times can be considered restrictive because of high variability in operational data (Sudhesh & Azhagappan, 2019). GI/M/2/N approach with heterogeneous servers, working vacations, feedback, and impatience and Bat Algorithm-based cost optimization is proposed to overcome limitations caused by the assumption of Poisson arrivals in (Guendouzi & Bouzebda, 2025). Exponential service times and two-server system are still assumptions inherent in this approach, while the optimization process

involves convergence and sensitivity analyses. The approach related to global financial institutions and regulation in economics provides more context, although it does not prove any contribution in the field of queueing models (Salave, 2026). Single-server approach in involves encouraged arrivals, impatient customers, and service switching (Ismailkhan et al., 2025). In comparison with discouraged arrivals model in its applicability depends on the ability of the real system to increase or decrease arrivals (Poomalai et al., 2026).

To integrate retrials, push-out, and feedback into analytical and metaheuristic methods, incorporating customer behaviours which are distinct from standard balking and reneging (Poomalai et al., 2026). But the use of these systems can apply only if customers do retrial, experience multiple services, or push-out. Its focus integrates encouraged arrivals and feedback together with the probabilistic detection of unqualified elements, using a detection criterion, the output of which cannot directly be generalized to service performance optimization (El-Hadidy & Alraddadi, 2026). To study open access service systems and information system design with respect to its context, despite the lack of an explicit comparison benchmark for heterogeneous queuing systems (Fallahhusein et al., 2025). The model of an electric vehicle charging system discussed in looks at finite capacity service provision (Varshney et al., 2026).

Its relevance to other service systems depends on whether their resource limitations and service-adjustment processes are similar. It develops a heterogeneous two-server model that has reverse balking, position-based reneging, retention, and cost minimization (Tamuli et al., 2024). It provides a better characterization of customer abandonment but needs to be validated in terms of the specified server system and customer behaviour before being used for other service systems. In this, it includes admission control, retrials, heterogeneous servers, finite queues, and maintenance where the focus is on the relationship between access and service processes (Saravanan et al., 2023). The maintenance process in this work is different from working vacations and unavailability due to failures; hence consistency in assumptions regarding interruptions is key in comparison. It analyses latency increases in mobile and satellite internet connections, but network latency itself cannot prove the relevance of a customer impatience model (Husseyn et al., 2025).

It takes into account customer impatience and vacation service for a GI/M(k)/1/N system, providing economic and performance perspectives while maintaining the one server system setup (Guendouzi & Saadaoui, 2026). It looks at heterogeneous servers, disasters, repairs, balking and reneging, moving on to reliability of services (Vimala, 2026). However, disasters cause interruptions that are different processes compared to scheduled maintenance or lower speed vacation service. This model takes into consideration competitive behavior of service providers and heterogeneous customers and servers, thus paying attention not to behavioural probabilities but to customer behavior within a competitive environment. This requires different set of assumptions than those in traditional congestion models (Hanukov & Hassin, 2026).

To provide an IoT and electrical systems application setting, but one whose importance in relation to queueing theory needs to be explicitly made in relation to task arrival, service resources, or congestion (Verma & Mehta, 2024). Lastly, considers heterogeneous information and Bayesian effects in strategic customers in queues, showing how information may affect decisions about joining the queue. Comparison with other conventional balking and reneging theories must therefore consider how customer information and decision making is modelled (Cao et al., 2026). Together, the above studies show complementary developments rather than sequential ones towards a better theoretical development. A worthwhile contribution in research would involve the identification of the exact mechanisms that need to be considered together, the validity of the assumptions and their performance under identical parameters. Of importance would be whether added complexity to the model contributes to significant differences in waiting time, throughput, customer retention, server utilization, and cost.

2.1 Research gap

The study of heterogeneous servers with queueing systems have been analyzed world widely. Because it represents reality like call centers, hospitals, banking systems, and customer support centers. A number of research studies have examined heterogeneous multi server queueing models integrate customer's behaviors like balking, reneging, feedback, and retrials.

In many research models, the incorporation of customers retention strategies are not given much space to reneging. But in real life, the service system institutions attempt to hold on impatient clients by the way of incentives and reminders. The study feedback mechanisms avoid the effect of customer encouragement polities which determine customers to retain in the system. In addition, many research works are come to conclude that homogeneous servers or infinite system capacity that may not precisely exemplify the real service situations where service cost may vary and the system cubage is restricted.

3. Methodology

This section is providing a detailed explanation for the proposed system model with model description.

3.1 Notations and parameters

Table 1: Notations and parameters.

Symbol / Notation	Definition / Description	Mathematical Expression & Conditions
λ	Baseline Customer Arrival Rate Before Encouragement	$\lambda > 0$
η	Proportional Increase in Arrivals Due To Encouragement (e.g., $\eta = 0.30$ represents a *30%* increase)	$\eta \geq 0$
$\lambda(1+\eta)$	Effective Poisson Arrival Rate After Encouragement	—
μ_1, μ_2	Exponential Service Rates of Servers 1 and 2, Respectively	$\mu_1, \mu_2 > 0$ (differ for heterogeneous servers)
σ_1, σ_2	Probabilities Of Assigning an Arriving Customer to Server 1 Or Server 2 When the System Is Empty	$\sigma_1 + \sigma_2 = 1$
N	Maximum Number of Customers Permitted in the System (Waiting and Receiving Service)	—
n	Number of Customers Currently in the System	$0 \leq n \leq N$
α	Probability that a Customer Returns to the Queue for Additional Service After Service Completion	$0 \leq \alpha \leq 1$
β	Probability that a Customer Departs After Service Completion	$\beta = 1 - \alpha$
ξ	Exponential Impatience Rate of Each Waiting Customer (Mean Time to an Impatience Event is $1/\xi$)	$\xi > 0$
q	Probability of Retaining a Customer Following an Impatience Event	$0 \leq q \leq 1$
p	Probability that an Impatient Customer Abandons the System	$p = 1 - q$
P_{00}	Steady-state Probability that the System Is Empty and Both Servers are Idle	$P_{00} \geq 0$
P_{10}	Steady-state Probability that only Server 1 is Serving a Customer and Server 2 is Idle	$P_{10} \geq 0$
P_{01}	Steady-State Probability that only Server 2 is Serving a Customer and Server 1 is idle	$P_{01} \geq 0$
P_{11}	Steady-State Probability that Both Servers are Busy and No Customers are Waiting	$P_{11} \geq 0$
P_n	Steady-state Probability of Having n Customers in the System ($P_0 = P_{00}$, $P_1 = P_{10} + P_{01}$, $P_2 = P_{11}$)	$P_n \geq 0$, $\sum_{n=0}^{N} P_n = 1$
k	Integer Index Used in the Products Appearing in the stationary-probability expressions	—
L_s	Expected number of Customers in the System, including those Receiving Service	—
L_q	Expected Number of Customers Waiting, Excluding Those Receiving Service	—

R_r	Average Rate of Customer Abandonment After Unsuccessful Retention	$R_r = \xi p L_q$
RR	Average Rate of Successful Retention Events	$RR = \xi q L_q$

The notation specifies the characteristics of the arrival process, service facility, customer dynamics, and performance metrics of the proposed queueing model in table 1. Customers arrive into the system at a rate of $\lambda(1+\eta)$, where service is provided by two heterogeneous servers with service rates μ_1 and μ_2, under the assumption that the total number of customers in the system is limited by a finite capacity N.Upon completion of service, a customer departs permanently from the system with probability $\beta = 1 - \alpha$, or reenters the queue for additional service with probability α. Waiting customers experience impatience at rate ξ, after which are retained in the system with probability q or abandon the system with probability $p = 1 - q$.

Steady-state probabilities characterize system occupancy and server states, enabling the derivation of key performance measures, including L_s (average system size), L_q (average queue length), R_r (average abandonment rate), and R_R (retention rate).

3.2 Model description

Following assumptions have been utilized in the proposed system model.

1. The term Encouraged arrival has been defined in terms of arrivals in the model. The customer's arrival is occurring one by one in accordance to the Poisson process with respect to the parameter of λ(1+η). The η represents the percentage of increasing the number of customer arrival rate from past or observed data. For an example, an organization is offering concession for the customers and the percentage is increasing with respect to 30 % or 120%, then η=0.3 or η=1.2 respectively.
2. Customers are handled with two servers with two different service rates known as μ_1 and μ_2 respectively. The service times is defined based on the exponential distribution. A new customer is arriving and joins in the system with the probability rate of either σ_1 or $\sigma_2 = 1 - \sigma_1$ in front of server 1 or in front of server 2 respectively.
3. According to the assumption, the capacity of the system is finite with the size of N and the queue is design with the working principle of First Come First Served.
4. The customer can either join at the end of the queue with the new probability value α or can leave the system with probability value of β $(\alpha + \beta = 1)$ based on the feedback.
5. If a customer is waiting in queue for a long time for acquiring the service and this may lead to impatient and leave the queue. The customer reneging times (ξ) are exponentially distributed.
6. The probability value for a customer retention rate of a reneged customer is defined as q and the probability value for a customer leaving without acquiring service is $p = 1 - q$.
7. The probability value for zero customer in the system is known as P_{00}. $P_{10}$is known as probability value with one customer being served by server 1 and P_{01} is the probability value with one customer being served by server 2. The probability value P_n is defined for the n number of customers in the system as follows, $n = 2,3, \dots, N$.

Also, $P_0 = P_{00}$

$P_1 = P_{10} + P_{01}$, and

$P_2 = P_{11}$.

3.3 Model description for customer flow structure using two server queueing model

Figure 1 illustrates the customer flow and analytical methodology of the proposed finite-capacity heterogeneous two-server queueing model. Customers arrive at an encouraged rate of λ(1 + η), enter a first-come, first-served queue, and receive service from servers operating at different rates, μ_1 and μ_2. Arrivals are blocked when the system is full. Impatient waiting customers are retained with probability τ or abandon the system with probability 1 − τ. After service completion, customers either return through feedback for additional service or depart. These transitions are formulated using a continuous-time Markov chain, and steady-state balance equations are solved under the normalization condition. The resulting stationary probabilities support the evaluation of expected system size, queue length, average reneging rate, and average retention rate, together with sensitivity analysis of the model parameters.

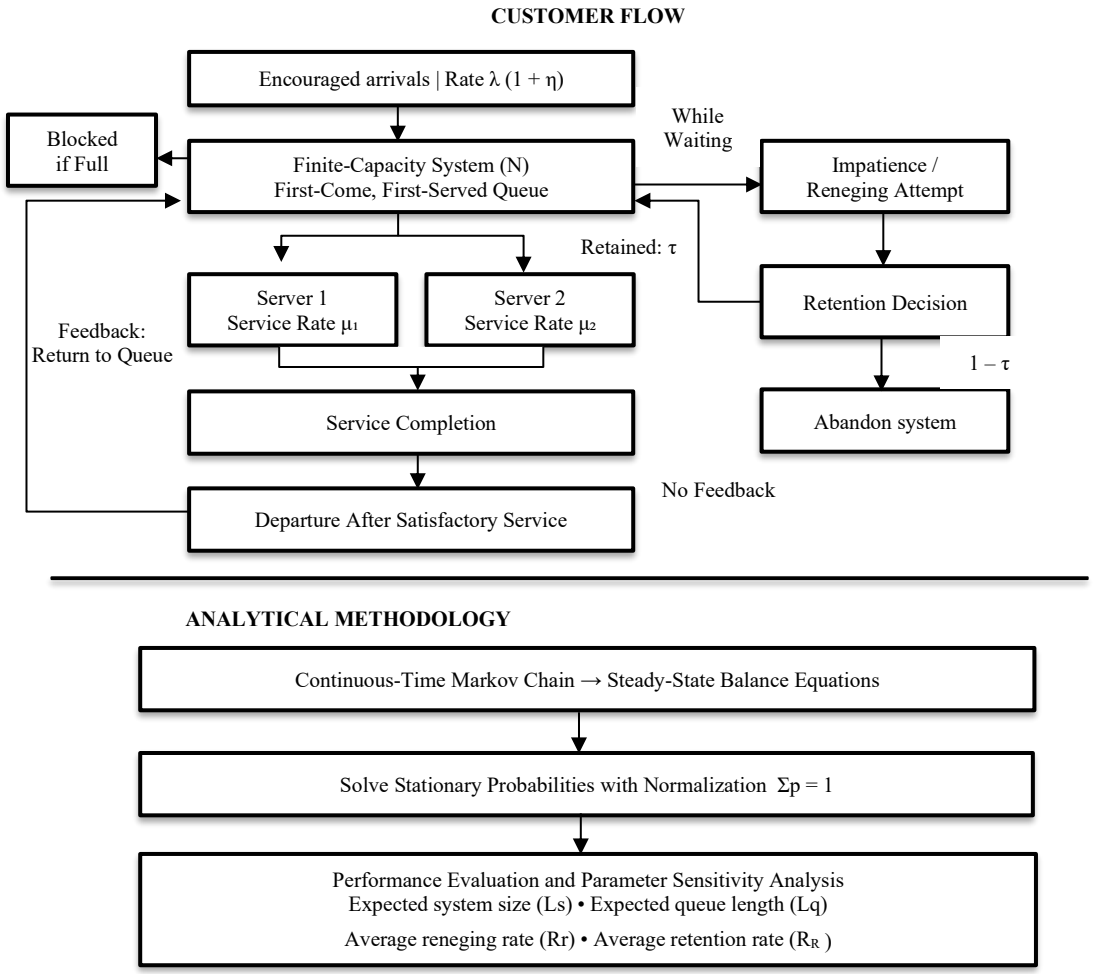

Figure 1: Customer-flow structure and analytical methodology of the proposed finite-capacity heterogeneous two-server queueing model.

4. Results and discussion

This section provides a detailed discussion about the steady-state equations and solutions for the proposed system model. The performance analysis for the proposed has been conducted based on the expected system size, expected queue length, average rate of reneging, and average rate of retention. Numerical illustration for the proposed system model has been discussed in the following section

4.1. Steady – state equations

The steady state equations are

Let $A = \lambda(1 + \eta)$ denote the effective arrival rate, and let $d_i = \beta\mu_i$ $(i = 1,2)$ denote the service-completion rates resulting in permanent departure. Feedback completions do not reduce system occupancy. Each waiting customer experiences an impatience event at rate ξ and abandons with probability $p = 1 - q$. Therefore, when $n \geq 2$, the total rate of downward transitions from occupancy n to $n - 1$ is given by in equation (1)

$$D_n = d_1 + d_2 + (n - 2)\xi p \qquad (1)$$

The retained customer does not leave the system and does not contribute to any transition among the states of occupancy. The arrivals contribute to occupancy at a rate of A as long as the system is not full; otherwise, are blocked.

The steady-state balance equations governing the system (assuming $N \geq 3$, with $P_1 = P_{10} + P_{01}$ and $P_2 = P_{11}$) are formulated as follows:

1. Empty State ($n = 0$): The departure flow is equal to the arrival flow due to services leaving in equation (2).

$$AP_{00} = d_1P_{10} + d_2P_{01} \quad (2)$$

2. Server 1 Busy (P_{10}): Entry happens either through an arrival to an empty system assigned to server 1 or through a departure from server 2 when both the servers are busy in equation (3).

$$(A + d_1)P_{10} = d_2P_2 + A\sigma_1P_{00} \quad (3)$$

3. Server 2 Busy (P_{01}): Similarly, for server 2 alone in equation (4)

$$(A + d_2)P_{01} = d_1P_2 + A\sigma_2P_{00} \quad (4)$$

4. Occupancy Two ($n = 2$): Both servers are busy and no customers are waiting in equation (5)

$$(A + D_2)P_2 = AP_1 + D_3P_3 \quad (5)$$

5. Intermediate Occupancy ($3 \le n \le N - 1$): Arrivals and departures generate the balance relation in equation (6)

$$(A + D_n)P_n = AP_{n-1} + D_{n+1}P_{n+1}, \quad 3 \le n \le N - 1 \quad (6)$$

6. System Capacity ($n = N$): Accepted arrivals are impossible at the boundary in equation (7)

$$AP_{N-1} = D_NP_N \quad (7)$$

$$\lambda(1 + \eta)P_{00} = \beta\mu_1P_{10} + \beta\mu_2P_{01}, \quad n = 0 \quad \rightarrow (8)$$

$$\big(\lambda(1 + \eta) + \beta\mu_1\big)P_{10} = \beta\mu_2P_{11} + \lambda(1 + \eta)\sigma_1P_{00}, \quad n = 1 \quad \rightarrow (9)$$

$$\big(\lambda(1 + \eta) + \beta\mu_2\big)P_{01} = \beta\mu_1P_{11} + \lambda(1 + \eta)\sigma_2P_{00}, \quad n = 1 \quad \rightarrow (10)$$

$$\big(\lambda(1 + \eta) + \beta(\mu_1 + \mu_2)\big)P_2 = \big(\beta(\mu_1 + \mu_2) + \xi p\big)P_3 + \lambda(1 + \eta)P_1, \quad n = 2 \quad \rightarrow (11)$$

$$\big(\lambda(1 + \eta) + \beta(\mu_1 + \mu_2) + (n - 2)\xi p\big)P_n = \big(\beta(\mu_1 + \mu_2) + (n - 1)\xi p\big)P_{n+1} + \lambda(1 + \eta)P_{n-1}, \quad 3 \le n \le N - 1 \quad \rightarrow (12)$$

$$\lambda(1 + \eta)P_{N-1} = \big(\beta(\mu_1 + \mu_2) + (N - 2)\xi p\big)P_N, \quad n = N \quad \rightarrow (13)$$

Equations (8) – (13) explain the equilibrium between incoming and outgoing probability flows for all the states of the system. In particular, equation (8) maintains the balance between the number of arrivals at an empty system and permanent departures from both servers. Equations (9) and (10) cover the states with one active server, taking into consideration the probability that the server gets its workload and the departure flows due to service completions. Equation (11) covers the state with two customers in service and zero waiting customers, considering the inflow from the state with three customers. Equation (12) generalizes this equilibrium for intermediate states ($3 \le n \le N - 1$), where $n - 2$ waiting customers generate a total abandonment rate of $(n - 2)\xi p$. Finally, equation (13) governs the full-capacity state ($n = N$), where new arrivals are blocked. Throughout these balance relations, $\lambda(1 + \eta)$ denotes the encouraged arrival rate, $\beta\mu_i$ is the constant departure rate for server i, and ξp represents the effective abandonment rate per customer.

4.2. Steady-state solutions

Summing the balance equations over all states below occupancy n cancels internal transitions. The remaining flows cross the boundary between occupancies $n - 1$ and n, yielding in equation (14)

$$AP_{n-1} = D_nP_n, \quad 2 \le n \le N \quad (14)$$

Consequently, the recursive relationship is given in equation (15)

$$P_n = \frac{A}{D_n}P_{n-1}, \quad 3 \le n \le N \quad (15)$$

Expanding this recursion expresses any state probability in terms of P_2 in equation (16)

$$P_n = P_2\prod_{k=3}^{n}\frac{A}{D_k}, \quad 3 \le n \le N \quad (16)$$

Boundary probabilities and coefficients

To determine the boundary probabilities, let us express them in terms of P_{00} in equation (17)

$$P_{10} = c_{10}P_{00}, \qquad P_{01} = c_{01}P_{00}, \qquad P_2 = c_2P_{00} \tag{17}$$

Using equations (2) and (3) from the balance relations, the coefficients c_{10} and c_{01} are given in EQUATION (18)

$$c_{10} = \frac{d_2c_2 + A\sigma_1}{A + d_1}, \qquad c_{01} = \frac{d_1c_2 + A\sigma_2}{A + d_2} \tag{18}$$

Substituting these expressions into equation (1), and applying $\sigma_1 + \sigma_2 = 1$, yields the coefficient for P_2 in equation (19)

$$c_2 = \frac{A^2(A + d_1\sigma_2 + d_2\sigma_1)}{d_1d_2(2A + d_1 + d_2)} \tag{19}$$

The occupancy-one probability is therefore:

$$P_1 = (c_{10} + c_{01})P_{00} \tag{20}$$

In equation (20) Substituting $D_k = d_1 + d_2 + (k-2)\xi p$, the remaining state probabilities become in equation (21)

$$P_n = c_2P_{00}\prod_{k=3}^{n}\frac{A}{d_1 + d_2 + (k-2)\xi p}, \qquad 3 \le n \le N \tag{21}$$

Normalization and existence

Applying the total probability normalization condition in equation (22)

$$P_{00} + P_{10} + P_{01} + \sum_{n=2}^{N} P_n = 1 \tag{22}$$

To obtain the exact closed-form expression for the empty-state probability P_{00}:

$$P_{00} = \left[1 + c_{10} + c_{01} + c_2\left(1 + \sum_{n=3}^{N}\prod_{k=3}^{n}\frac{A}{d_1 + d_2 + (k-2)\xi p}\right)\right]^{-1} \tag{23}$$

In equation (23) These expressions fully determine all stationary probabilities. For positive arrival and service rates ($\beta > 0$) and finite system capacity N, the finite irreducible Markov chain possesses a unique stationary distribution. Notably, the restrictive infinite-capacity stability condition $A < d_1 + d_2$ is not required here due to the finite capacity N.

4.3. Performance measures

In this section, to derive some important performance measures of the queueing model.

i. Expected system size (L_s):

$$L_s = \sum_{n=1}^{N} nP_n = P_1 + 2P_2 + \sum_{n=3}^{N} nP_n \qquad \rightarrow (24)$$

$$L_s = \left(\frac{\lambda(1+\eta)+\beta(\mu_1\sigma_2+\mu_2\sigma_1)}{2\lambda(1+\eta)+\beta(\mu_1+\mu_2)}\right)\frac{\lambda(1+\eta)}{\beta\mu_1\mu_2}(\mu_1+\mu_2)P_{00} + 2\left(\frac{\lambda(1+\eta)+\beta(\mu_1\sigma_2+\mu_2\sigma_1)}{2\lambda(1+\eta)+\beta(\mu_1+\mu_2)}\right)\frac{\lambda(1+\eta)}{\beta\mu_1}\frac{\lambda(1+\eta)}{\beta\mu_2}P_{00} + \sum_{n=3}^{N} n\left(\frac{\lambda(1+\eta)+\beta(\mu_1\sigma_2+\mu_2\sigma_1)}{2\lambda(1+\eta)+\beta(\mu_1+\mu_2)}\right)\frac{\lambda(1+\eta)}{\beta\mu_1}\frac{\lambda(1+\eta)}{\beta\mu_2}\prod_{k=3}^{N}\frac{\lambda(1+\eta)}{\beta(\mu_1+\mu_2)+(k-2)\xi p}P_{00}$$

ii. Expected queue length of the system (L_q)

$$L_q = \sum_{n=1}^{N}(n-2)P_n = \sum_{n=3}^{N}(n-2)P_n \qquad \rightarrow (25)$$

$$L_q = \sum_{n=3}^{N}(n-2)\left(\frac{\lambda(1+\eta)+\beta(\mu_1\sigma_2+\mu_2\sigma_1)}{2\lambda(1+\eta)+\beta(\mu_1+\mu_2)}\right)\frac{\lambda(1+\eta)}{\beta\mu_1}\frac{\lambda(1+\eta)}{\beta\mu_2}\prod_{k=3}^{N}\frac{\lambda(1+\eta)}{\beta(\mu_1+\mu_2)+(k-2)\xi p}P_{00}$$

Average rate of reneging (R_r)

$$R_r = \sum_{n=3}^{N}(n-2)\xi p P_n \qquad \rightarrow (26)$$

$$R_r = \sum_{n=3}^{N}(n-2)\xi p\left(\frac{\lambda(1+\eta)+\beta(\mu_1\sigma_2+\mu_2\sigma_1)}{2\lambda(1+\eta)+\beta(\mu_1+\mu_2)}\right)\frac{\lambda(1+\eta)}{\beta\mu_1}\frac{\lambda(1+\eta)}{\beta\mu_2}\prod_{k=3}^{N}(n-2)\xi p\left(\frac{\lambda(1+\eta)+\beta(\mu_1\sigma_2+\mu_2\sigma_1)}{2\lambda(1+\eta)+\beta(\mu_1+\mu_2)}\right)\frac{\lambda(1+\eta)}{\beta\mu_1}\frac{\lambda(1+\eta)}{\beta\mu_2}P_{00}$$

Average rate of retention (R_R)

$$R_R = \sum_{n=3}^{N}(n-2)\xi q P_n \qquad \rightarrow (27)$$

$$R_R = \sum_{n=3}^{N}(n-2)\xi q\left(\frac{\lambda(1+\eta)+\beta(\mu_1\sigma_2+\mu_2\sigma_1)}{2\lambda(1+\eta)+\beta(\mu_1+\mu_2)}\right)\frac{\lambda(1+\eta)}{\beta\mu_1}\frac{\lambda(1+\eta)}{\beta\mu_2}\prod_{k=3}^{N}\frac{\lambda(1+\eta)}{\beta(\mu_1+\mu_2)+(k-2)\xi p}P_{00}$$

Equations (24) – (27) quantify system occupancy, congestion, abandonment, and retention. Specifically, the expected system size L_s accounts for all customers currently receiving service and waiting in the queue, whereas the expected queue length L_q restricts focus strictly to waiting customers. Given two servers, any system state containing $n \geq 3$ customers feature $n-2$ waiting customers. Consequently, the average customer reneging (abandonment) rate is given by $R_r = \xi p L_q$, capturing customers who depart following unsuccessful retention efforts, while the average successful retention rate is expressed as $R_R = \xi q L_q$. Collectively, these performance metrics delineate the operational balance between maintaining active customer participation and mitigating system congestion.

4.4 Numerical illustration

This section presents detailed numerical illustrations of the model. The table 1 has been recorded from the computed values based on the following assumptions $\mu_1 = 3, \mu_2 = 4$, $\eta = 1.5, \sigma_1 = 0.3, \sigma_2 = 0.7$, $\xi = 0.3$, and $N = 10$

Table 2: Performance analysis with respect to arrival data.

Arrival Rate (λ)	Expected System Size (L_s)	Expected Queue Length (L_q)	Average Reneging Rate (R_r)	Average Retention Rate (R_R)
3	8.2868	6.2943	0.7553	1.1329
4	9.0279	7.0286	0.8434	1.2652
5	9.3327	7.3329	0.8799	1.3199
6	9.4939	7.4939	0.8993	1.3489
7	9.5929	7.5929	0.9111	1.3667
8	9.6596	7.6596	0.9192	1.3787
9	9.7076	7.7076	0.9249	1.3874
10	9.7438	7.7438	0.9293	1.3939

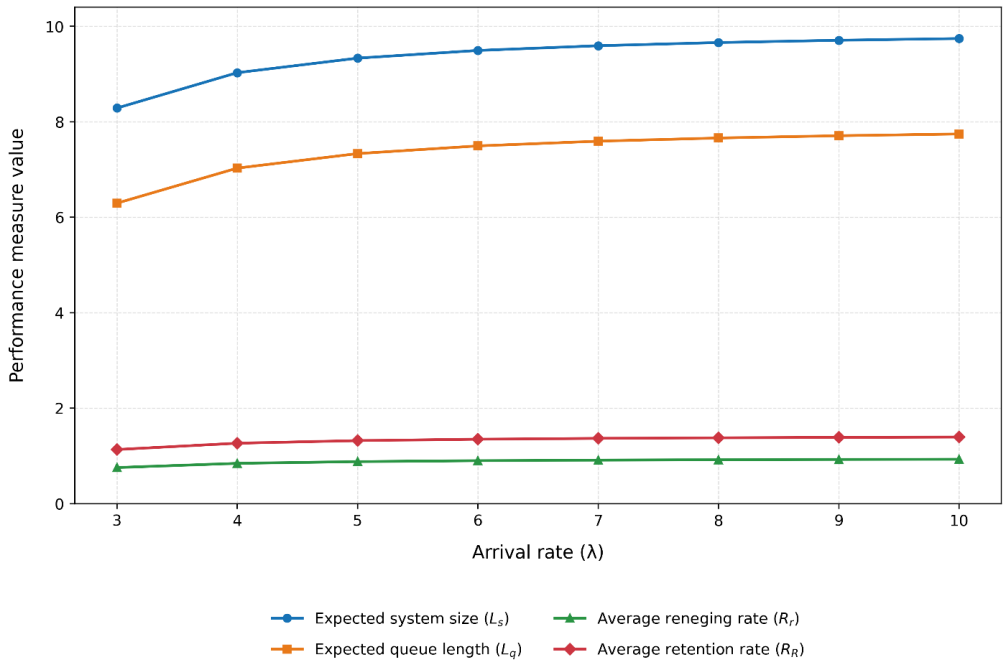

Figure 2: Performance Analysis with respect to arrival data.

A detailed performance analysis has been conducted for the following number of customers in the queue, expected queue length, reneging rate, and customer retention rate and it has been listed in the table 2 and figure 2. The expected system size is gradually increasing from 8.2868 to 9.7438 based on the increasing in arrival rate. The expected queue length (Lq) is increasing from 6.2943 to 7.7438. This increasing behavior is shows that more customers accumulated in the system based on the increases in the arrival intensity. The reneging rate (Rr) is increasing from 0.7553 to 0.9293 and it shows that a higher arrival rate leads to more abandon the queue due to increase in waiting time. The customer retention rate is gradually increasing from 1.1329 to 1.3939, it shows that a greater number of impatient customers are successfully retained through the retention mechanism. The overall observation in the result demonstrates that increasing the arrival rate improves customer inflow and increases the congestion rate, queue length, reneging, and the need for effective retention strategies

Practical application implies that increasing the number of customers requires service capacity to be considered. In the case of constant service costs, growth in demand leads to larger queue lengths and increased abandonments, thus promotional efforts do not guarantee better service performance on their own. The increased retention rate is caused by a larger number of retention efforts performed under conditions of heavy congestion and should not be perceived as a result of the better efficiency of retention with the same probability of retention.

The table 3 has been calculated based on the assumptions of $\mu_1 = 3,\ \mu_2 = 4,\ \eta = 1.5, \sigma_1 = 0.3, \sigma_2 = 0.7, \lambda = 4,$ and $N = 10$

Table 3: Performance Analysis with respect to reneging rate.

ξ	L_s	L_q	R_r	R_R
1	8.3358	6.3388	2.5355	6.3388
2	7.2258	5.2364	4.1891	10.4727
3	6.2129	4.2359	5.0831	12.7077
4	5.4145	3.4527	5.5244	13.8111
5	4.8179	2.8721	5.7442	14.3606
6	4.3730	2.4429	5.8629	14.6572
7	4.0349	2.1196	5.9349	14.8372
8	3.7716	1.8701	5.9842	14.9604
9	3.5615	1.6727	6.0216	15.0539
10	3.3901	1.5130	6.0521	15.1302

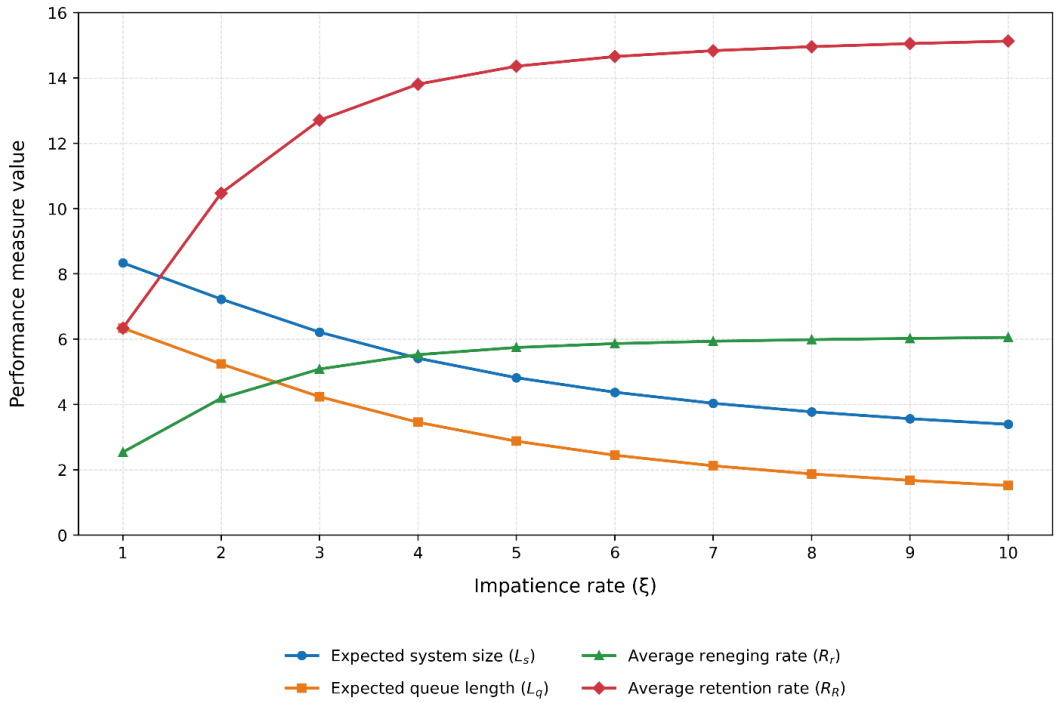

Figure 3: Performance Analysis with respect to reneging rate.

According to the performance measures in table 3 and figure 3, the impact of increasing the reneging parameter on the queueing system has been taken by increasing ξvalue from 1 to 10. The expected system size (Ls) is decreasing gradually from 8.3358 to 3.3901 and the expected queue length (Lq) is decreasing from 6.3388 to 1.5130. These values are indicating the customers are leaving the system more quickly and this reduce the congestion between the customers for acquiring service. More customers are abandoning in the queue due to impatience grows due to the increasing value in reneging rate (Rr) from 2.5355 to 6.0521. The retention rate (R_R) is increasing from 6.3388 to 15.1302 shows that retention policies actively participating to recover an increasing the number of reneged customers. These results are showing the stronger customer impatience reduce system occupancy

The decrease in length of the queue is the result of customers leaving prior to service being provided and not faster service provision. This means that a smaller queue could hide unfulfilled demand and customer loss. In a practical sense, this means that managers cannot mistake lower occupancy for efficiency. Higher retention rate also means higher numbers of retention incidents, which may or may not involve different customers but may be the result of multiple retention attempts with the same customer.

The table 4 has been recorded the performance measures with respect to the retention rate by assuming the following values, $\mu_1 = 3,\ \ \mu_2 = 4, \eta = 1.5, \sigma_1 = 0.3, \sigma_2 = 0.7$, $\lambda = 4, \xi = 0.3,$ and $N = 10$

Table 4: Performance Analysis with respect to the retention rate.

q	L_s	Lq	R_r	R_R
0.10	6.9416	4.9851	1.3461	0.1496
0.15	6.9948	5.0369	1.2844	0.2267
0.20	7.0483	5.0890	1.2214	0.3053
0.25	7.1021	5.1414	1.1568	0.3856
0.30	7.1562	5.1941	1.0908	0.4675
0.35	7.2104	5.2471	1.0232	0.5509
0.40	7.2649	5.3001	0.9540	0.6360
0.45	7.3195	5.3534	0.8833	0.7227
0.50	7.3742	5.4069	0.8111	0.8111
0.55	7.4291	5.4605	0.7372	0.9009
0.60	7.4841	5.5142	0.6617	0.9926
0.65	7.5391	5.5680	0.5846	1.0858

0.70	7.5941	5.6218	0.5061	1.1806
0.75	7.6491	5.6757	0.4257	1.2770
0.80	7.7041	5.7295	0.3438	1.3751
0.85	7.7590	5.7833	0.2603	1.4748
0.90	7.8139	5.8371	0.1751	1.5761
0.95	7.8685	5.8906	0.0884	1.6788
1.00	7.9230	5.9441	0.0000	1.7832

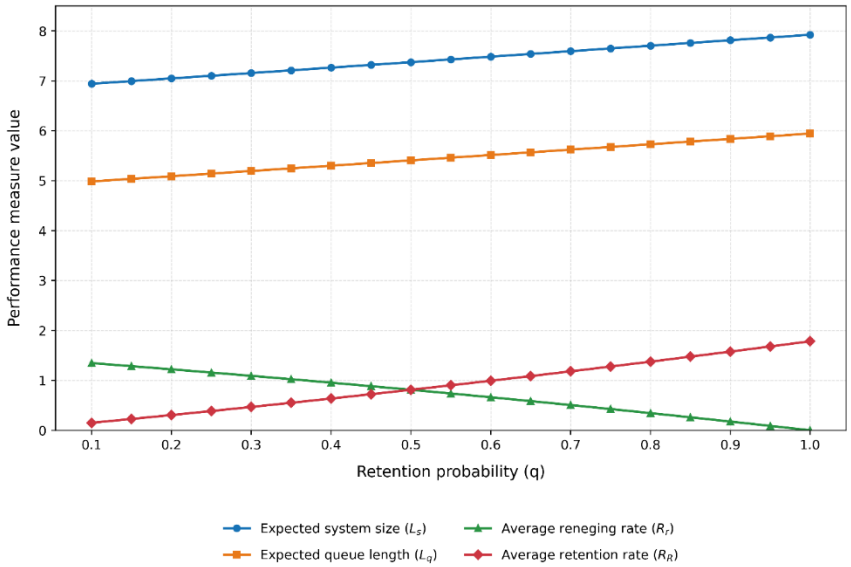

Figure 4: Performance Analysis with respect to retention rate.

The customer retention probability (τ) has been recorded in the table 4 and figure 4 by varying probability rate from 0.10 to 1.00. The expected system size (Ls) is getting increases from 6.9416 to 7.9230 and same way expected queue length (Lq) is increasing from 4.9851 to 5.9441. This result shows that the retaining more customer result in a larger number of customers retaining in the system. The reneging rate (R_R) is decreasing gradually from 1.3461 to 0; it shows that increases in the retention probability. The increases in the retention probability effectively eliminate customer abandonment. The retention rate (R_R) is increasing consistently from 0.1496 to 1.7832 and it shows that the improvement in retention strategy. This result is confirming that higher retention probability substantially reduces customer loss by maintaining greater customer participation in the service provider system.

A lower retention probability implies lower abandonments but results in more waiting customers in the finite-capacity queueing system. As such, there arises a compromise between ensuring customer participation and dealing with the created congestion. In that regard, retention programs should always be backed up with sufficient service capacity. The abandonments equaling zero for (q = 1) arise as a consequence of the fact that all impatient customers are retained; the former is not an indication that full retention is possible.

The table 5 shows the results in the performance analysis with respect to feedback by assuming $\mu_1 = 3$, $\mu_2 = 4$, $\eta = 1.5$,$\sigma_1 = 0.3$,$\sigma_2 = 0.7$, $\xi = 0.3$, and $N = 10$

Table 5: Performance Analysis based on the feedback.

α	L_s	L_q	R_r	R_R
0	7.4841	5.5142	0.6617	0.9926
0.2	8.4076	6.4139	0.7697	1.1545
0.4	9.0279	7.0287	0.8434	1.2652
0.6	9.4309	7.4311	0.8917	1.3376
0.8	9.7021	7.7031	0.9244	1.3865

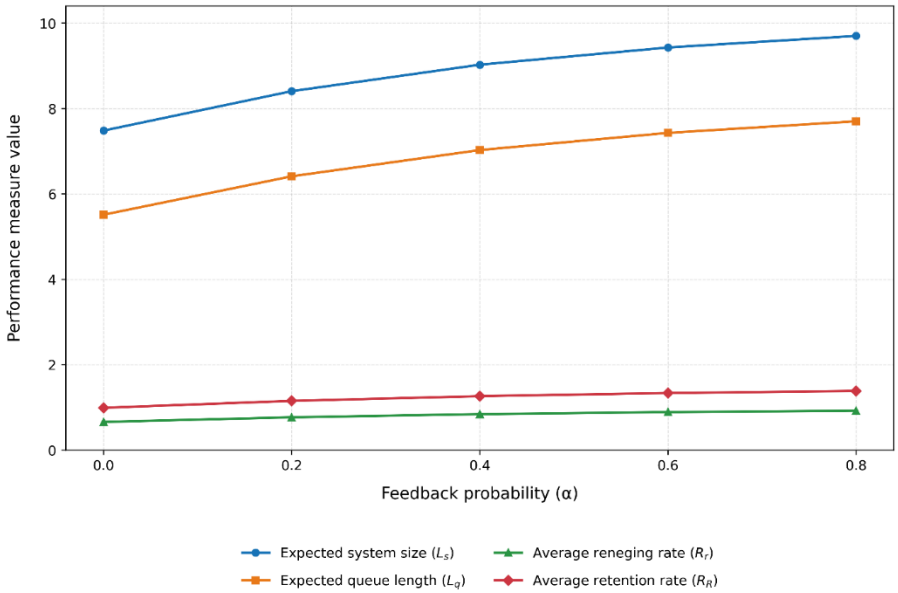

Figure 5: Performance Analysis with respect to feedback.

Table 5 and figure 5 presents a performance measures with the effect of the feedback probability. The expected probability is increasing from 7.4841 to 9.7021 and the expected queue length (Lq) is increasing from 5.5142 to 7.7031 based on the increases in the feedback probability from 0.0 to 0.8. This result shows the more customers re-enter in to the system after service completion and this result leads to higher system occupancy

The reneging rate (Rr) is increasing from 0.6617 to 0.9244 and this increasing value is indicating that longer queues due to repeated customer arrivals. The customer retention rate R_R is increasing from 0.9926 to 1.3865; it shows that the retention mechanism becomes increasingly important as system congestion grows. The higher feedback probability value is improving customer continuity and increasing congestion rate.

Higher chances of feedback imply higher demands for service from those customers who have already been serviced. The repeat service usage takes up resources that would otherwise be available to new customers. It is, thus, important to note that feedback does not necessarily reflect the improvement in customer satisfaction and service delivery. In case of rework and feedback, higher service completion quality can help reduce repeat usage. In the case where there is an increase in demand for service, proper planning for capacity is necessary.

5. Conclusion

A finite capacity queuing model with two heterogeneous servers, customer arrivals, feedback, reneging, and rejoining was proposed in this study. Steady state balance equations and stationary probabilities were used to obtain closed form expressions for expected system size, expected queue size, average reneging rate and average retention rate. From the numerical analysis carried out in this study, the effects of customer behavior and service restrictions on congestion and customer losses have been seen. Raising the arrival rate parameter from 3 to 10 caused an increase in expected system size from 8.2868 to 9.7438 and expected queue size from 6.2943 to 7.7438. Also, the average reneging rate increased from 0.7553 to 0.9293. On the other hand, increasing the reneging rate from 1 to 10 resulted in decrease in system size from 8.3358 to 3.3901 and queue size from 6.3388 to 1.5130 while the average reneging rate increased from 2.5355 to 6.0521. A lower occupancy when impatience is more pronounced simply indicates that customers were leaving the service system. An increase in the retention probability from 0.10 to 1.00 decreased the average number of reneging from 1.3461 to 0.0000, but increased the queue length from 4.9851 to 5.9441. Likewise, an increase in the feedback probability from 0.0000 to 0.80 increased the system size from 7.4841 to 9.7021. This suggests that there is always a balance between retaining the customers and repeated service, on the one hand, and requiring larger capacity, on the other. Numerical sensitivity analysis does not imply statistical significance and empirical generalization. The retention policy may include repeated actions on the same customers and there is no economic criterion that is maximized. The future research should confirm the validity of model assumptions based on real-world data on the system performance, develop the model for more than one server and non-exponential distributions, and include state-dependent arrivals, failures, repairs, and vacation. Economic optimality should consider the balance between the retention cost and loss from abandonment and congestion.

References

1. Cao, J., Guo, Y., Hou, X., & Fang, C. (2026). Heterogeneous information provision and Bayesian effect for queues with strategic customers. *Journal of the Operational Research Society*, 1-22. https://doi.org/10.1080/01605682.2026.2641673
2. El-Hadidy, M. A. A., & Alraddadi, R. (2026). Probabilistic detection model for unqualified units under encouraged arrivals in a feedback Markovian queue: a search-theoretic approach with discounted effort and waiting time reduction. *Applied Mathematics in Science and Engineering*, *34*(1), 2648958. https://doi.org/10.1080/27690911.2026.2648958
3. Fallahhusein, M., Fernandes, R. B., Naga Raju, M., ugli, I. S. S., Nijaguna, G. S., & Udayakumar, R. (2025). Open Access Service Models: Impact on Information System Design. *Indian Journal of Information Sources and Services, 15*(2), 148–153. https://doi.org/10.51983/ijiss-2025.IJISS.15.2.20
4. Guendouzi, A., & Bouzebda, S. (2025). Cost Optimization in a GI/M/2/N Queue with Heterogeneous Servers, Working Vacations, and Impatient Customers via the Bat Algorithm. *Mathematics*, *13*(21), 3559. https://doi.org/10.3390/math13213559
5. Guendouzi, A., & Saadaoui, A. (2026). Customer impatience in the GI/M (k)/1/N queue with working vacation: Economic and performance analysis. *Croatian Operational Research Review*, *17*(1), 17-27. https://doi.org/10.17535/crorr.2026.0002
6. Hanukov, G., & Hassin, R. (2026). Duopoly competition in a service market with strategic heterogeneous customers and servers. *Annals of Operations Research*, 1-39. https://doi.org/10.1007/s10479-026-07157-2
7. Husseyn, M., Saranya, G., Babu, K., Kishore, D., & Kiruthikadevi, K. (2025). Characterizing Latency Inflation in Mobile and Satellite Internet Connections. *Journal of Internet Services and Information Security*, *15*(3), 573-586. https://doi.org/10.58346/JISIS.2025.I3.039
8. Ismailkhan, E., Nagarajan, G., Sundar Raj, M., & Immaculate, S. (2025). Analysis of a Single-Server Queue with Encouraged Arrivals, Impatient Customers and Dynamic Service Switching. *Reliability: Theory & Applications*, *20*(4), 191-199. http://doi.org/10.24412/1932-2321-2025-489-191-199
9. Khan, I. E., & Paramasivam, R. (2022). Reduction in waiting time in an M/M/1/N encouraged arrival queue with feedback, balking and maintaining of reneged customers. *Symmetry*, *14*(8), 1743. https://doi.org/10.3390/sym14081743
10. Kumar, B. (2022). Transient analysis of M/M/C queuing model with reneging, finite capacity and population. *International Journal of Mathematics in Operational Research*, *21*(1), 1-25. https://doi.org/10.1504/IJMOR.2022.120339
11. Poomalai, S., Jafari, S., & Nanjappan, J. V. (2026). Analysis and Cost Optimization of a Retrial Queue with Push-Out and Feedback Using Analytical and Metaheuristic Approaches. *Axioms*, *15*(3), 204. https://doi.org/10.3390/axioms15030204
12. Pravina, C. D., Kamala, P., Sreelakshmi, S., & Damodaran, S. (2025). Analysis of a Queue Subject to Discouraged Arrivals, Customer Impatience, and Self-Switching Server Dynamics. *Contemporary Mathematics*, 2738-2751. https://doi.org/10.37256/cm.6320256285
13. Salave, A. P. (2026). Analysing the Role of Global Financial Institutions in Shaping Economic Policy and Regulation. *International Academic Journal of Economics, 13*(1), 9-15. https://doi.org/10.71086/IAJE/V13I1/IAJE1302
14. Saravanan, V., Poongothai, V., & Godhandaraman, P. (2023). Admission control policy of a two heterogeneous server finite capacity retrial queueing system with maintenance activity. *Opsearch*, *60*(4), 1902-1925. https://doi.org/10.1007/s12597-023-00669-6
15. Som, B. K., & Kumar, R. (2018). A heterogeneous queuing system with reverse balking and reneging. *Journal of Industrial and Production Engineering*, *35*(1), 1-5. https://doi.org/10.1080/21681015.2017.1297739
16. Sudhesh, R., & Azhagappan, A. (2019). Analysis of an M/M/c queue with heterogeneous servers, balking and reneging. *International Journal of Operational Research*, *36*(3), 293-309. https://doi.org/10.1504/IJOR.2019.103120
17. Tamuli, A., Das, D., & Choudhury, A. (2024). Optimizing the Performance of Multi-server Heterogeneous Queueing Systems with Dynamic Customer Behaviour: Tamuli et al. *Sankhya B*, *86*(2), 366-414. https://doi.org/10.1007/s13571-024-00340-0
18. Jun, L., Kim, L., & Xe, L. (2026). Utilization-Aware Chiplet Dataflow Engine for End-To-End Learned Image Compression with Entropy–Transform Pipeline Fusion. *Journal of Integrated VLSI and Signal Intelligence*, *1*(3), 24-31.
19. Verma, A., & Mehta, N. (2024). The Role of IoT in Modern Electrical Systems: An Interdisciplinary Approach in the Periodic Series. *Smart Grid Integration*, 13-18. Periodic Series in Multidisciplinary Studies.
20. Vimala, K. (2026). Analysis of the M/M/2 Queue with Heterogeneous Servers, System Disaster, Server Repair, Customers' Impatience, Balking and Reneging. *system*, *14*(1).
21. Xu, W., Li, L., & Fan, W. (2026). Optimization Analysis of a Multi-Server Queueing System with Two-Phase Heterogeneous Services and Synchronous Working Vacation Interruptions. *Mathematics*, *14*(5), 874. https://doi.org/10.3390/math14050874