International Journal of Mathematical, Engineering and Management Sciences
Vol. 9, No. 5, 1198-1209, 2024
https://doi.org/10.33889/IJMEMS.2024.9.5.063

Ram Arti Publishers

Cost Analysis of a Finite Capacity Queue with Server Failures, Balking, and Threshold-Driven Recovery Policy

Amit Kumar
Department of Mathematics,
Chandigarh University, Mohali, 140413, Punjab, India.
E-mail: amitk251@gmail.com

Savita
Department of Mathematics,
Chandigarh University, Mohali, 140413, Punjab, India.
Corresponding author: savitamathematics@gmail.com

Chandra Shekhar
Department of Mathematics,
Birla Institute of Technology and Science, Pilani, 333031, Rajasthan, India.
E-mail: Chandrashekhar@pilani.bits-pilani.ac.in

(Received on March 4, 2024; Revised on May 10, 2024 & July 1, 2024; Accepted on July 19, 2024)

Abstract
The study is on the cost dynamics which is concerned with finite queues in particular to susceptibility to server outages, threshold-driven recovery and Catastrophic events, but main focus is on balking behaviour of the customers. A thorough investigation of the financial implications of interruptions and efficacy to recovery strategies is done and to alleviate the impact of breakdowns, disasters and customer refusals the queue system designs through empirical research is conducted. Various recovery policies have been compared with the existing thresholds, and the study further endeavors to categorize the cost-effective approach to mitigate the economic concerns of server failures and catastrophic incidents. The study brings discernments on the expected development in the management of the queuing system along with the assistance in complete decision-making in cost efficiency and enhancement of the complete system performance.

Keywords- Customer behavior, Balking, Breakdown, Catastrophic, Threshold-based recovery policy.

1. Introduction

Queueing systems are pivotal in various manufacturing and service contexts, profoundly impacting operational efficiency, customer satisfaction, and overall financial performance. As organizations strive to optimize their processes, understanding and mitigating potential disruptions within finite capacity queues becomes imperative. This study investigates the intricate dynamics of such queues, particularly focusing on customer balking behavior and incorporating factors like server breakdowns, catastrophic events, and a threshold-driven recovery policy. In contemporary environments, server breakdowns pose significant challenges, leading to disruptions in service delivery and subsequent financial implications. These breakdowns can be caused by various issues, ranging from technical malfunctions to human errors, each requiring a different approach for effective resolution. Implementing robust maintenance protocols and real-time monitoring systems can help pre-emptively address potential breakdowns, thereby minimizing service disruptions and maintaining smooth operational flow.

The queueing systems gets disrupted through Catastrophic events like natural disasters, Power outages and cybersecurity fissures. This even adds to major financial and reputational damages to the queueing systems.

To diminish the impressions, comprehensive disaster recoveries and business continuity plans are developed. The plans have to bear the detailed response strategies, regular drills and advanced technologies which ensure the quick recovery and the operations gets sustained during the disruptions. These plans should include detailed response strategies, regular drills, and the integration of advanced technologies to ensure quick recovery and sustained operations during disruptions. Additionally, customer balking, where customers abandon queues due to observed disruptions, complicates financial impact analysis. Understanding the psychological factors behind balking-such as customers' perceptions of waiting times and service reliability-is crucial. By analysing customer behaviour and feedback, organizations can better anticipate balking tendencies and implement measures like improved communication and customer engagement strategies to retain customers during disruptions.

In the previous studies, many queueing systems aspects are discussed like impatience, adaptive rates of arrival, strategies for controlling and the server dynamics. The studies reveal that there is a gap and comprehensive study is required to consider the combined impact on the server failures, catastrophic events and also on the customer behaviours in concern to the queueing systems. The study targets to fill the existing cavity by analysing single- server queues and examine its effects of breakdowns, disasters and customer behaviours on the performance of queueing systems. There is a dire need to understand the fiscal systems to come up with the decision-making so as to develop the robust strategies for efficient resource allocation and effective queue management.

Numerous studies on understanding the queueing systems with impatience, adaptive arrival rates and control strategies. Boxma and Prabhu (2011) delves into the control of congestion and queue management through internet systems by exploring M/G/1 queue with impatience and adaptive arrival rates. Bu and Sun (2024) exert on client threshold tactics and on the balking habits of customers in queueing with variable service capabilities. Antline et al. (2023) insists on controlling strategies, on manageable arrival rates, and reverse balking in a single-server queueing model. Tian et al. (2024) reveals the equilibrium and socially optimal joining strategies in a Markovian queueing System with single server and with variable service charges under observable or unobservable queuing scenarios. Yanting and Na (2023) probed on customer's behaviour such as balking, recoiling, interjection and reneging, with the highlights on the need for comprehensive study on the behaviours in connection to queueing systems. Jain and Raychaudhuru (2023) incorporates an asynchronous vacation policy (AVP) to optimize the performance in multi-server queueing systems with assorted heterogeneous servers and Poisson arrivals. Ayyappan and Meena (2023) have assessed matrix analytic methods to check performance analysis and also check on the cost evaluation in single-server phase- type queueing model of complex server dynamics.

Many studies have identified that breakdowns are crucial for queueing systems (cf. Ayyappan and Meena, 2023; Karthick and Suvitha, 2023; Kumar, 2023; Shekhar et al., 2017; Shekhar et al., 2020a). The study focuses on the vital roles of queueing, particularly on the allocation of resources and effective management of customers, tasks or entities awaiting service. Adhering to FIFO (First-In and First-Out) principle the queues safeguard and ensure an orderly sequence of customers, which assists to prevent bottlenecks and ensures equitable treatment. Queues facilitate smooth flow and permit the servers to manage the customers and allocate resources effectively and systematically. This gives an overview into service status and progress, which is essential for tracking, observing and monitoring the issues and resolving the same on time. In addition, the breakdowns are informed about the capacity planning strategies after the analyses of the queues during the breakdowns. This helps in organization to identify and address weaknesses in system capacity or resource availability. This guarantees better readiness and resilience during the disruptions.

Chang et al. (2018), Ojobar and Ogini (2022), Rani et al. (2023), and Saxena et al. (2020) in their research work have discussed on the queueing theory and its threshold-based recovery policy as an important factor as it empowers the proactive management of disruptions and disturbances which result in failures. With the establishment of predefined thresholds and key indicators the queue length, waiting timings or the service charges; the organizations may identify the deteriorating conditions in the queueing systems in a particular time frame. Ahuja and Jain (2023) investigated on finite buffer and finite source models along with the ways of retrial policy to delay threshold recovery. They insist on using recursive technique for steady-state solutions and parametric non-linear programming to gauge on the system characteristics in fuzzy environments. Kumari and Sharma (2023) probed on F- policy as queueing system with single working vacation, has also analysed server breakdowns with the use of threshold recovery, has done transient analysis and has evaluated cost through numerical methods. Sanga and Antala (2024) has premeditated on the single unreliable server that is state-dependent and is the finite queueing model with F-policy admission control. The researcher has also addressed on the server breakdowns and repair facilities. It has focussed on the optimization of the total cost function for the extended service delivery with efficiency seeking profitability with the use of genetic algorithms and quasi-Newton methods. Ghanbarian and Mazumdar (2024) also explored on the fluctuations of empirical occupation procedures in the mean-field limits of large systems heterogeneous processor- sharing servers. They have investigated on the sensitivity to arrival parameters in the diffusion scale. Functional Central Limit Theorems (FCLTs) have been designed in the study to calculate the mean-field approximation faults and error functioning.

Kocer and Ozkar (2023) overviewed on queueing inventory system and server breakdowns as two customer types and as interruptions. They have used (LDQBD) level-dependent quasi-birth-death process, and have analysed steady-state probabilities with optimal inventory policies in a numerical way. Dublin et al. (2024) has observed single queueing system with finite buffering capacity, breakdowns and repairs. The customers decide individually and independently to stay or to leave in the breakdowns. This GI/M/1- type system is studied in its stead- state with its numerical samples and optimization problems. Yiming and Guo (2023) reconnoitred on the asymptotic behaviour of M/G/I retrial queueing system and server breakdowns while highlighting the stability of time-dependent solutions and system spectrum features.

Ammar (2021), Mishra and Banerjea (2020), Shekhar et al. (2020b), and Seenivasan and Abinaya (2022) studied the Catastrophic analysis, major causes of system failures and performance issues. They suggested to create a robust queueing model which can guarantee system reliability and resilience. Chakravarthy (2017) scrutinized on queueing model with catastrophes and checks on the delayed actions. They have implied matrix-analytic method for steady-state analysis. Raj and Jain (2024) in their study developed a multi- server catastrophic retrial queueing model to control pre-emptive prior scheduling and deployment. Kumar and Bura (2024) has defined a capacity Markovian queueing system along with the two identical servers, in consideration to the environmental conditions and random disasters and has deployed probability generating function techniques to seek transient state solutions. Wang et al. (2023) has directed his study on new methods for other calculating sojourn times in *MAP/M/1* queueing systems and has described three types of negative customers.

This paper unfolds as follows: Section 2 introduces the elementary assumptions underlying our queueing model. In Section 3, we present clear and concise analysis solutions. Section 4 focuses on key system characteristics. Sections 5 and 6 delve into building an expected cost function per unit of time, leveraging Newton's method for optimizing system capacity, threshold value, and service rate. Practical application examples illustrating the model's real-life utility are discussed in Sections 7 and 8. Finally, conclusions drawn from our findings are presented in Section 9.

2. Model Description and Stationary Probabilities

Threshold based recovery policy and catastrophic events are designed to check and analyse the limited-capacity queue model within the server breakdowns. The finite Markov model is designed to check birth-death process principles in queueing theory. The Markovian has permitted the system behaviour in different conditions and has ensured a complete comprehensive thinking on the model's performance and reliability.

2.1 Assumptions and Notations

M/M/1/K queue is studied along with server breakdowns, catastrophic events and threshold-based recovery policies. The model has finite capacity as K (which is less than infinity) and longs for customer balking or policies. The customers under such situations leave the queue as they cannot take the probability of $1-\delta$ and they cannot wait for long time. The key assumptions are outlined below:

- The model assumes customer arrivals follow a Poisson process with parameter λ. This enables us to understand the average arrival rate and anticipate arrival patterns in the investigated context.
- Customers advents as FIFO (First-come, first served basis) with the assurance to be served on their arrival. It helps customer's experience as it reflects a fair and organized service quality.
- During busy periods, service times follow to an exponential distribution with a mean of $\frac{1}{\mu}$. This statistical model defines the average duration of customer service, aiding in the assessment of system performance during peak demand.
- During peak hours, when the server is engaged with a customer, others must wait in a queue for their turn. Since the server can only serve one customer at a time, there may be delays in attending to subsequent arrivals.
- The server can experience breakdowns during busy period. The probability of a breakdown depends on the breakdown rate, time-to-breakdown assumed to follow an exponential distribution with a rate of α.
- When a breakdown occurs, the server remains out of service until the number of customers in the system reaches a specified threshold value, T $(1 \leq T \leq K)$. The repair times are exponentially distributed with a mean of $\frac{1}{\beta}$.
- The server after repairing, resumes the services till the outcome of next breakdown.
- Time-to-catastrophes tracks an exponential distribution with γ as average rate. The frequency and timings of catastrophic events is accessed through the statistical models.
- The stochastic processes are assumed to be independent it is assumed that they are analyzed separately. They are focused on complete understanding so as to avoid the influence of other processes.

2.2 Notations

The study delves into the mathematical representation with a current state of governing model at any time as *t*. Let

$$N(t) = \text{ the number of customers in the system at time } t \quad (1)$$

$$Y(t) = \text{ state of the server at time } t \quad (2)$$

where,

$$Y(t) = \begin{cases} 0, & \text{when the server is in a busy period} \\ 1, & \text{when the server is in a breakdown period} \end{cases}$$

Then, the system $\{Y(t), N(t); t \geq 0\}$ is a continuous time Markov process with a state space $S = \{(0,n) \mid n \in I_1\} \cup \{(1,n) \mid n \in I_2\}; I_1 = 0,1,2,\dots,K, I_2 = 1,2,\dots,K$. Furthermore, the steady-state probabilities of the system are defined as follows:

$$\pi_0(n) = \lim_{t\to\infty}\{Y(t) = 0, N(t) = n \mid n = 0, 1, \dots, K\}$$
$$\pi_1(n) = \lim_{t\to\infty}\{Y(t) = 1, N(t) = n \mid n = 1,2, \dots, K\}$$

2.3 Chapman-Kolmogorov Equations

The equations of Chapman- Kolmogorov acts as an instrument to check on the transitions in stochastics processes. In the Markov Chains it is crucial to understand the changings in systems in particular periods. In the study the equations are added to describe and focus on probabilistic transitions between different states of the model.

***Server is in busy state*:** The governing system of equations for the state $Y(t) = 0$ are is follows:

$$-\lambda\pi_{0,0} + \mu\pi_{0,1} + \gamma\left(\sum_{i=1}^{K}\pi_{0,n} + \sum_{i=1}^{K}\pi_{1,n}\right) = 0 \quad (3)$$
$$-(\lambda\delta + \mu + \alpha + \gamma)\pi_{0,1} + \lambda\pi_{0,0} + \mu\pi_{0,2} = 0 \quad (4)$$
$$-(\lambda\delta + \mu + \alpha + \gamma)\pi_{0,n} + \lambda\delta\pi_{0,n-1} + \mu\pi_{0,n+1} = 0; \quad n = 2,3, \dots, T-1 \quad (5)$$
$$-(\lambda\delta + \mu + \alpha + \gamma)\pi_{0,n} + \lambda\pi_{0,n-1} + \mu\pi_{0,n+1} + \beta\pi_{1,n} = 0; \quad n = T, T+1, \dots, K-1 \quad (6)$$
$$-(\mu + \alpha + \gamma)\pi_{0,K} + \lambda\delta\pi_{0,K-1} + \beta\pi_{1,K} = 0 \quad (7)$$

***Server is in breakdown state*:** The governing system of equations for the state $Y(t) = 1$ is as follows:

$$-(\lambda\delta + \gamma)\pi_{1,1} + \alpha\pi_{0,1} = 0 \quad (8)$$
$$-(\lambda\delta + \gamma)\pi_{1,n} + \lambda\delta\pi_{1,n-1} + \alpha\pi_{0,n} = 0; \quad n = 2, 3, \dots, T-1 \quad (9)$$
$$-(\lambda\delta + \beta + \gamma)\pi_{1,n} + \lambda\delta\pi_{1,n-1} + \alpha\pi_{0,n} = 0; \quad n = T, \dots, K-1 \quad (10)$$
$$-(\beta + \gamma)\pi_{1,K} + \lambda\delta\pi_{1,K-1} + \alpha\pi_{0,K} = 0 \quad (11)$$

The state probabilities mentioned earlier conform to a specific normalizing condition, which is:

$$\sum_{n=0}^{K}\boldsymbol{\pi_{0,n}} + \sum_{n=0}^{K}\boldsymbol{\pi_{1,n}} = \mathbf{1} \quad (12)$$

3. The Steady-State Solution

The system of simultaneous linear equations (Equations (3) to (11)) can be represented in matrix form as follows:

$$AX = 0 \quad (13)$$

where, **A** be a square matrix of size $(K+1)$, with the coefficients of state probabilities $\pi_{0,0}, \pi_{0,1}, \dots, \pi_{0,K}, \pi_{1,0}, \pi_{1,1}, \dots, \pi_{1,K}$ as its element and **X** be a column vector of unknown state probabilities of size $(K+1) \times 1$, and **0** be a null vector of the appropriate dimension. Using the normalizing condition in Equation (12), we get:

$$\bar{\mathbf{A}}\mathbf{X} = \mathbf{B} \quad (14)$$

where, matrix $\bar{\mathbf{A}}$ remains the same as the previously described matrix **A**, except that each element in the last row is replaced by 1. **B** is a column vector with $\left[\mathbf{B}_{(K+1)\times 1}\right] = 1$. The system of linear equations in matrix form can be represented in augmented matrix form as:

$$[\mathbf{A} \mid \mathbf{B}]$$

The non-homogeneous system of linear equations can be solved using numerical methods such as Gauss elimination extended (GEE) or the Successive Over Relaxation (SOR) method, with an over-relaxation parameter set to 1.25 in MATLAB (R2019b) software. In the consequent section, the classification of the

queueing system is determined based on the stationary probabilities obtained from solving this equation system.

4. Performance Characteristics

Studying the steady-state probabilities helps in understanding the system structures, leading to the derivation of performance metrics. These metrics are crucial for optimizing system performance and are invaluable for system managers and industrial engineers. They are critical in enhancing the Grade of Service (GoS) by predicting preventive requirements and queueing indices for a specific finite queueing system. The following classification highlights the most important performance metrics.

- The expected number of customers in the system

$$L_S = \sum_{n=1}^{K} n\pi_{0,n} + \sum_{n=1}^{K} n\pi_{1,n} \quad (15)$$

- The probability of the server being busy

$$P_B = \sum_{n=1}^{K} \pi_{0,n} \quad (16)$$

- The probability of the server being in a state of breakdown

$$P_D = \sum_{n=1}^{K} \pi_{1,n} \quad (17)$$

- The probability of the server being in an idle state

$$P_I = \pi_{0,0} \quad (18)$$

- Probability that the system is blocked

$$P_{SB} = \pi_{0,K} + \pi_{1,K} \quad (19)$$

- Average balking rate

$$R_B = \sum_{n=1}^{K} (1-\delta)\lambda(\pi_{0,n} + \pi_{1,n}) \quad (20)$$

- Expected waiting time in the system

$$W_S = \frac{L_S}{\lambda_{eff}} \quad (21)$$

where, λ_{eff} is the effective arrival rate and can be written as

$$\lambda_{eff} = \lambda\pi_{0,0} + \lambda\delta\left(\sum_{n=1}^{K-1} \pi_{0,n} + \sum_{n=1}^{K-1} \pi_{1,n}\right) \quad (22)$$

- Throughput of the system

$$Th = \mu \sum_{n=1}^{K} \pi_{0,n} \quad (23)$$

5. Cost Analysis

This research aims to determine the optimal parameters for minimizing the expected total cost. Our approach involves defining a suitable cost function designed for this purpose. In particular, this function considers the repair facility rates (μ) as decision variables. Additionally, we will consider various cost components associated with different activities. We assume that these cost factors related to each activity follow a linear relationship. Below is a breakdown of the cost elements associated with individual activities:

C_h ≡ The amount charged for holding every customer in the system per unit of time.

$C_i \equiv$ The unit time cost during server idle periods.
$C_b \equiv$ The unit cost incurred during periods of server activity
$C_d \equiv$ The unit time cost in the event of server breakdown.
$C_{sb} \equiv$ The fixed cost incurred for each lost customer due to system blockage.
$C_w \equiv$ The cost of customer wait time per unit of time.
$C_m \equiv$ The rate at which a service is provided can be expressed as the cost per unit time.
$C_{bb} \equiv$ The cost associated to each balked unit per unit time

The expected cost function is given by,

$$TC(\mu) = C_h L_S + C_i P_I + C_b P_B + C_d P_D + C_{sb} \lambda P_{SB} + C_w W_S + C_m \mu + C_{bb} R_B \quad (24)$$

6. Newton-Quasi Method

The Newton-Quasi method is employed to solve complex functions in cases where using Newton's method is impractical due to the challenging computation of the Hessian matrix. It detects extreme points of a function by seeking gradient zeroes, eliminating the requirement to compute the Hessian matrix. This method is particularly valuable in optimization problems, such as minimizing expected total cost, where computing derivatives is challenging. The process involves initializing trial solutions, computing derivatives, and refining solutions iteratively until convergence.

***Algorithm*:** Newton-Quasi Method
***Require*:** Tolerance $\in$
***Ensure*:** Minimum value $\Omega(\mu^*)$
(i): $i \leftarrow 0$
(ii): $\mu_i \leftarrow \mu \triangleright$ Initial trial solution
(iii): **Repeat**
(iv): $\Omega_{\mu_i} \leftarrow$ Compute expected total cost $TC(\mu_i)$
(v): $\Omega'_{\mu_i} \leftarrow$ Compute first derivative $\frac{d\,TC(\mu_i)}{d\mu_i}$
(vi): $\Omega''_{\mu_i} \leftarrow$ Compute second derivative $\frac{d^2\,TC(\mu_i)}{d\mu_i^2}$
(vii): $\mu_{i+1} \leftarrow \mu_i - \frac{\Omega'_{\mu_i}}{\Omega''_{\mu_i}}$
(viii): $i \leftarrow i + 1$
(ix): $\mu^* \leftarrow \mu_i$
(x): $\mu_i \leftarrow \mu_{i+1}$
(xi): **until** $\left|\frac{\partial \Omega}{\partial \mu}\right| < \in$

7. Numerical Results

The analytical outcomes of the finite capacity system's performance metrics alone are inadequate to validate the model's effectiveness. To assess the practical applicability of the proposed finite Markov model with balking, breakdowns, and a threshold-driven recovery policy, numerous numerical experiments are conducted in MATLAB.

For the default parameters $K = 10$; $T = 5$; $\lambda = 1$; $\mu = 5$; $\delta = 0.5$; $\alpha = 0.6$; $\beta = 3$; $\gamma = 0.3$; $C_h = 14$; $C_i = 295$; $C_b = 360$; $C_d = 135$; $C_{sb} = 11$; $C_r = 209$; $C_m = 7$; $C_{bb} = 20$, we performed a series of numerical experiments, and the results are presented in **Table 1** and **Table 2**, which include various performance metrics discussed in the previous section. Additionally, **Figure 1** illustrates a surface plot for

$TC(\mu)$, providing a visual representation of the total cost as a function of the service rate (μ) wrt K.

Table 1. Numerical simulation regarding different system characteristics w.r.t. T, μ, λ.

(T, μ, λ)	L_s	P_B	P_D	P_I	P_{SB}	λ_{eff}	W_S	Th	$TC(\mu)$
(4.0, 3.0, 1.0)	0.7557	0.1841	0.2582	0.5577	1.0000	0.7557	2.2253	0.2211	301.6505
(5.0, 3.0,1.0)	0.8478	0.1733	0.2783	0.5484	1.0000	0.8478	2.1652	0.2258	299.1240
(6.0, 3.0 ,1.0)	0.9121	0.1661	0.2897	0.5441	1.0000	0.9121	2.1308	0.2279	297.7721
(7.0 ,3.0, 1.0)	0.9554	0.1615	0.2964	0.5421	1.0000	0.9555	2.1109	0.2289	297.0295
(4.0, 3.0 ,1.0)	0.7557	0.1841	0.2582	0.5577	1.0000	0.7557	2.2253	0.2211	301.6505
(4.0,3.5,1.0)	0.6854	0.1686	0.2400	0.5914	1.0000	0.6854	2.6601	0.2043	305.7430
(4.0, 4.0,1.0)	0.6273	0.1555	0.2238	0.6206	1.0000	0.6273	3.1046	0.1897	309.8705
(4.0, 4.5, 1.0)	0.5784	0.1443	0.2096	0.6461	1.0000	0.5784	3.5570	0.1769	313.9908
(4.0, 5.0, 1.0)	0.5366	0.1347	0.1969	0.6685	1.0000	0.5366	4.0157	0.1658	318.0842
(4.0, 3.0 ,1.0)	0.7557	0.1841	0.2582	0.5577	1.0000	0.7557	2.2253	0.2211	301.6505
(4.0, 3.0, 1.5)	1.0482	0.2582	0.2896	0.4522	1.5000	0.6988	2.1313	0.4108	309.3454
(4.0, 3.0, 2.0)	1.3230	0.3260	0.2991	0.3750	1.9996	0.6616	2.1028	0.6250	320.3649
(4.0,3.0, 2.5)	1.5998	0.3876	0.2973	0.3151	2.4980	0.6404	2.1081	0.8562	333.1695

Table 1 demonstrates that as we delay by increasing the threshold value, the expected number of customers in the system also increases. This increase in the expected number of customers in the system directly influences other performance metrics. Changes in service and arrival rates also result in noticeable changes in performance metrics. Moreover, changes in parameters directly impact the expected total cost, which highlights the need to optimize the expected total cost for the optimal service rate. Similar illustrations corresponding to parameters α, β, and δ are summarized in **Table 2**. A higher failure rate leads to inefficiency in system behavior, which can be addressed by enhancing the recovery rate.

Table 2. Numerical simulation regarding different system characteristics w.r.t. α, β, δ.

(α, β, δ)	L_s	P_B	P_D	P_I	P_{SB}	λ_{eff}	W_S	Th	$TC(\mu)$
(0.4, 3.0, 0.3)	0.614	0.179	0.221	0.601	1.000	0.614	2.338	0.120	303.310
(0.5, 3.0, 0.3)	0.671	0.167	0.257	0.576	1.000	0.671	2.228	0.127	297.655
(0.6, 3.0, 0.3)	0.721	0.157	0.289	0.554	1.000	0.721	2.132	0.134	292.718
(0.7, 3.0 ,0.3)	0.765	0.148	0.317	0.534	1.000	0.765	2.048	0.140	288.369
(0.4,3.0, 0.3)	0.614	0.179	0.221	0.601	1.000	0.614	2.338	0.120	303.310
(0.4, 3.5, 0.3)	0.613	0.179	0.220	0.601	1.000	0.613	2.339	0.120	303.328
(0.4, 4.0, 0.3)	0.613	0.179	0.220	0.601	1.000	0.613	2.339	0.120	303.343
(0.4, 4.5, 0.3)	0.612	0.179	0.220	0.601	1.000	0.612	2.339	0.120	303.354
(0.4, 5.0, 0.3)	0.612	0.179	0.220	0.601	1.000	0.612	2.339	0.120	303.364
(0.4,3.0,0.3)	0.614	0.179	0.221	0.601	1.000	0.614	2.338	0.120	303.310
(0.4, 3.0 ,0.4)	0.667	0.186	0.216	0.598	1.000	0.667	2.351	0.161	306.034
(0.4, 3.0, 0.5)	0.715	0.194	0.211	0.594	1.000	0.715	2.366	0.203	308.868
(0.4, 3.0, 0.6)	0.759	0.203	0.206	0.591	1.000	0.759	2.382	0.245	311.756

The surface plot in **Figure 1** illustrates TC(μ) for different values of system capacity *K* and service rate μ. It highlights the issue of higher costs associated with a larger system capacity and service rate. It also inspires us to formulate an optimization problem to determine the optimal design parameter value for the service rate.

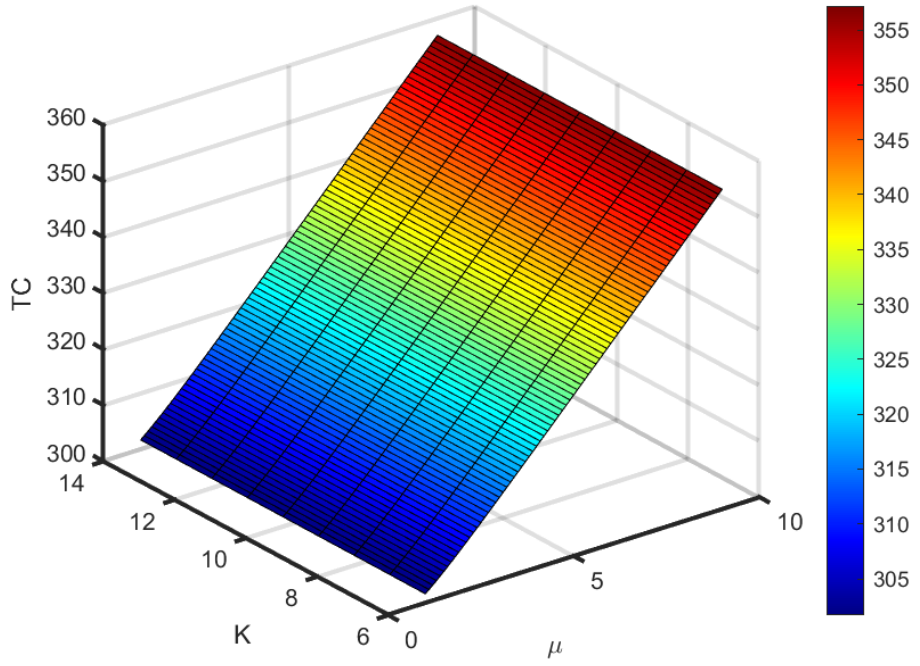

Figure 1. Total cost function w.r.t. K and μ.

For optimal analysis, we employ the Newton Quasi method to determine the optimal value for the service facility. To this end, numerous experiments are carried out using following default values for the system parameters. $K = 15$; $T = 6$; $\lambda = 1$; $\delta = 0.7$; $\alpha = 0.6$; $\beta = 3$; $\gamma = 0.3$; $C_h = 14$; $C_i = 295$; $C_b = 360$; $C_d = 135$; $C_{sb} = 11$; $C_r = 209$; $C_m = 7$; $C_{bb} = 20$. The outcomes are detailed in **Table 3** and **Table 4**.

Table 3. The optimal service rates and expected total cost for different system parameters.

(K, T, λ, δ)	μ^*	$TC(\mu^*)$	Number of iterations	$\frac{\partial TC}{\partial \mu}$
(9, 4, 1.0, 0.5)	0.9567	289.6029	4	3.8147×10^{-6}
(11, 4, 1.0, 0.5)	0.9584	289.6102	4	1.01094×10^{-6}
(13, 4, 1.0, 0.5)	0.9587	289.6112	4	1.5259×10^{-6}
(15, 4, 1.0, 0.5)	0.9588	289.6113	5	0
(15, 5, 1.0, 0.5)	0.5990	283.4116	6	1.1444
(15, 6, 1.0, 0.5)	0.2138	278.2001	4	0
(15, 6, 1.2, 0.5)	0.6634	285.7090	5	7.6294×10^{-6}
(15, 6, 1.4, 0.5)	1.0696	292.6407	4	3.5666×10^{-6}
(15, 6, 1.6, 0.5)	1.4431	299.1592	4	2.6434×10^{-6}
(15, 6, 1.0, 0.6)	0.5925	286.4102	6	7.6294×10^{-6}
(15, 6, 1.0, 0.7)	0.9432	293.5740	4	3.8147×10^{-6}

Table 3 presents a summary of the optimal service rate values obtained for minimizing the expected total cost using the Newton-Quasi method. A higher service rate, indicating a more efficient service facility, is essential to accommodate a larger number of customers, especially with increasing system capacity. An increased threshold value indicates a deteriorating service system, highlighting the need for further research. A higher service rate is typically associated with a higher arrival rate of customers. Similar optimal analyses are presented in **Table 4**, corresponding to breakdown and recovery rates.

Table 4. The optimal service rates and expected total cost for different system parameter.

(α, β, γ)	μ^*	$TC(\mu^*)$	Number of iterations	$\frac{\partial TC}{\partial \mu}$
(0.3,3.0,0.4)	0.7916	308.6701	6	3.8147×10^{-6}
(0.4,3.0,0.4)	0.5134	299.8588	6	1.0147×10^{-6}
(0.5,3.0,0.4)	0.2948	292.1842	6	3.8147×10^{-6}
(0.5,4.0,0.4)	0.2102	292.6581	3	0.3458×10^{-6}
(0.5,4.5,0.4)	0.3638	292.8215	3	0×10^{-6}
(0.5,5.0,0.4)	0.3784	292.9540	4	0×10^{-6}
(0.5,3.0,0.5)	0.0111	295.0019	3	1.1444

In summary, we recommend implementing measures to minimize breakdowns for as long as possible. Efficient service facilities are essential to prevent customer balking, even at increased costs. Prompt recovery is essential.

8. Real-Life Application

In the context of physician supervision, managing an emergency department in an urban hospital setting can be highly challenging. The limited number of treatment rooms available and the staff's systematic handling of an influx of patients seeking urgent medical care contribute to the complexity. Patient arrivals in such environments often follow a Poisson process, mirroring the random and sudden nature of medical emergencies. The patients keep arriving randomly and result to "server breakdowns". This interrupts the medical workflows and trigger the automatic responses based on preset thresholds as the count of patient numbers. "Patient balking" results due to unattended response and they deicide immediately to leave the hospital without seeking treatment. Therefore, hospitals need plans for server breakdowns and must be equipped in handling the patient balking so as to accommodate quality service and care to the patients.

9. Conclusions

The study targets to analyze the $M/M/1/K$ queue and the server breakdowns, on threshold-based recovery policy and on customer balking system. The study imposes on Markov process theory and has revealed steady-state equations with the solution via Gaussian elimination so as to detect the customer distribution. The key systems are analyzed and the expected cost system is developed as per unit time with an insight into economic implications on the queueing systems. The study implied Newton's method to identify the optimal service price (μ^*) and minimizes the estimated cost. The study targets on cost-effective guidance for decision-making. With numerous numerical experiments and illustrations, the parameter interactions are systematically observed and the optimal values (K^*, μ^*) and the minimal estimated price (F (K^*, μ^*)) is determined. Importantly, we illustrated these findings with a real-world example, highlighting their relevance for decision-makers dealing with server breakdowns, balking thresholds, and customer behavior. This study enhances our theoretical understanding of queueing systems and provides actionable insights for strategic decision-making in various operational contexts.

Conflict of Interest

The authors confirm that this article's contents have no conflict of interest.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors would like to thank the editor and anonymous reviewers for their comments that helped improve the quality of this work.

References

Ahuja, A., & Jain, A. (2023). Fuzzy analysis of a queueing system featuring an unreliable service provider and geometric arrivals by incorporating constant retrial policy and delayed threshold recovery. *Journal of Ambient Intelligence and Humanized Computing, 14*(6), 7499-7518. https://doi.org/10.1007/s12652-022-04455-y.

Ammar, S.I. (2021). Fluid M/M/1 catastrophic queue in a random environment. *RAIRO-Operations Research, 55*(2021), S2677-S2690. https://doi.org/10.1051/ro/2020100.

Antline, N.B., Balaji, R., & Saradha, M. (2023). Control policy, manageable arrival rates, and reverse balking in mutually dependent queueing model. *Contemporary Mathematics, 5*(1), 17-25. https://doi.org/10.37256/cm.5120243407.

Ayyappan, G., & Meena, S. (2023). Phase type queueing model of server vacation, repair and degrading service with breakdown, starting failure and close-down. *Reliability: Theory & Applications, 18*(1), 464-483.

Boxma, O.J., & Prabhu, B.J. (2011). Analysis of an M/G/1 queue with customer impatience and adaptive arrival process. In *International Conference on NETwork Games, Control and Optimization* (pp .1-4). IEEE. Paris, France.

Bu, Q., & Sun, Y. (2024). Stationary analyses for a double-ended queueing system with random service capacity and balking customers. *Journal of Computational and Applied Mathematics, 437*, 115473. https://doi.org/10.1016/j.cam.2023.115473.

Chakravarthy, S.R. (2017). A catastrophic queueing model with delayed action. *Applied Mathematical Modelling, 46*, 631-649. https://doi.org/10.1016/j.apm.2017.01.089.

Chang, C.J., Ke, J.C., & Chang, F.M. (2018). Unreliable retrial queue with loss and feedback under threshold-based policy. *International Journal of Industrial and Systems Engineering, 30*(1), 1-20. https://doi.org/10.1504/ijise.2018.094608.

Dudin, A.N., Chakravarthy, S.R., Dudin, S.A., & Dudina, O.S. (2024). Queueing system with server breakdowns and individual customer abandonment. *Quality Technology & Quantitative Management, 21*(4), 441-460. https://doi.org/10.1080/16843703.2023.2215630.

Ghanbarian, S., & Mazumdar, R.R. (2024). Mean-field fluctuations at diffusion scale in threshold-based randomized routing for processor sharing systems and applications. *Stochastic Models, 40*(2), 296-339. https://doi.org/10.1080/15326349.2023.2250418.

Jain, A., & Raychaudhuri, C. (2023). Cost optimization in multi-server queueing model with balking. *AIP Conference Proceedings, 2699*(1), 020008. https://doi.org/10.1063/5.0139432.

Karthick, V., & Suvitha, V. (2023). An analysis of three servers Markovian multiple vacation queueing system with servers breakdown. *IAENG International Journal of Applied Mathematics, 53*(2), 1-8.

Kocer, U.U., & Ozkar, S. (2023). A production queueing-inventory system with two-customer and a server subject to breakdown. *Annals of Operations Research, 331*, 1089-1117. https://doi.org/10.1007/s10479-023-05275-9.

Kumar, A. (2023). Single server multiple vacation queue with discouragement solved by confluent hypergeometric function. *Journal of Ambient Intelligence and Humanized Computing, 14*(5), 6411-6422. https://doi.org/10.1007/s12652-020-02467-0.

Kumar, D., & Bura, G.S. (2024). Transient solution of a two homogeneous servers Markovian queueing system with environmental, catastrophic and restoration effects. *International Journal of Mathematics and Statistics Studies, 12*(1), 45-53. https://doi.org/10.37745/ijmss.13/vol12n14553.

Kumari, U., & Sharma, D.C. (2023). Transient analysis of an unreliable system with working vacation and threshold recovery. *Mathematics in Engineering, Science & Aerospace, 14*(3), 761.

Mishra, R., & Banerjea, A.C. (2020). Neurological damage by coronaviruses: A catastrophe in the queue! *Frontiers in Immunology, 11*, 1-13. https://doi.org/10.3389/fimmu.2020.565521.

Ojobor, S.A., & Ogini, N.O. (2022). Threshold recovery policy for the machine interference repair problem with server vacations. *Journal of Physics: Conference Series, 2199*, 012020. https://doi.org/10.1088/1742-6596/2199/1/012020.

Raj, R., & Jain, V. (2024). Resource optimization for MMAP[c]/PH[c]/S catastrophic queueing model with PH retrial times. *OPSEARCH*. https://doi.org/10.1007/s12597-023-00731-3. (In press).

Rani, S., Jain, M., & Meena, R.K. (2023). Queueing modeling and optimization of a fault-tolerant system with reboot, recovery, and vacationing server operating under admission control policy. *Mathematics and Computers in Simulation, 209*, 408-425. https://doi.org/10.1016/j.matcom.2023.02.015.

Sanga, S.S., & Antala, K.S. (2024). Performance analysis and cost investigations for state-dependent single unreliable server finite queue under F-policy using GA and QNM. *Journal of Computational and Applied Mathematics, 441*, 115679.

Saxena, A., Claeys, D., Zhang, B., & Walraevens, J. (2020). Cloud data storage: A queueing model with thresholds. *Annals of Operations Research, 293(1)*, 295-315. https://doi.org/10.1007/s10479-019-03279-y.

Seenivasan, M., & Abinaya, R. (2022). Markovian queueing model with single working vacation and catastrophic. *Materials Today: Proceedings, 51*(8), 2348-2354. https://doi.org/10.1016/j.matpr.2021.11.572.

Shekhar, C., Kumar, A., & Varshney, S. (2020a). Parametric nonlinear programming for fuzzified queuing systems with catastrophe. *International Journal of Process Management and Benchmarking, 10*(1), 69-98. https://doi.org/10.1504/ijpmb.2020.104233.

Shekhar, C., Kumar, N., Gupta, A., Kumar, A., & Varshney, S. (2020b). Warm-spare provisioning computing network with switching failure, common cause failure, vacation interruption, and synchronized reneging. *Reliability Engineering & System Safety, 199*, 106910. https://doi.org/10.1016/j.ress.2020.106910.

Shekhar, C., Raina, A.A., Kumar, A., & Iqbal, J. (2017). A survey on queues in machining system: Progress from 2010 to 2017. *Yugoslav Journal of Operations Research, 27*(4), 391-413. https://doi.org/10.2298/yjor161117006r.

Tian, R., Su, S., & Zhang, Z.G. (2024). Equilibrium and social optimality in queues with service rate and customers' joining decisions. *Quality Technology & Quantitative Management, 21*(1), 1-34. https://doi.org/10.1080/16843703.2022.2157695.

Wang, Y., Wang, J., & Zhang, G. (2023). The effect of information on the strategic behavior in a Markovian queue with catastrophes and working vacations. *Quality Technology & Quantitative Management, 21*(5), 688-721. https://doi.org/10.1080/16843703.2023.2243199.

Yanting, C., & Na, Y. (2023). A multi-server priority queueing system with customer balking, interjections and reneging. *Operations Research and Management Science, 32*(5), 49-55. https://doi.org/10.12005/orms.2023.0148.

Yiming, N., & Guo, B.Z. (2023). Asymptotic behavior of a retrial queueing system with server breakdowns. *Journal of Mathematical Analysis and Applications, 520*(1), 126867. https://doi.org/10.1016/j.jmaa.2022.126867.