Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 54 (2) (2025), 710 – 737
DOI : 10.15672/hujms.1485216

Research Article

Performance and economic analysis of an unreliable single-server queue with general retrial times and varied customer patience levels

Nasreddine Dehamnia[1], Mohamed Boualem*[2], Djamil Aïssani[1]

[1] *University of Bejaia, Faculty of Exact Sciences, Research Unit LaMOS, 06000 Bejaia, Algeria*
[2] *University of Bejaia, Faculty of Technology, Research Unit LaMOS, 06000 Bejaia, Algeria*

Abstract

This paper presents a comprehensive mathematical analysis of an unreliable single-server retrial queue with general retrial times, serving two types of customer arrivals: high-patience and low-patience customers. Customers arrive in the system following two Poisson processes with different service rates. In addition, the model incorporates essential features such as service times, reserved times, and repair times, all following general distributions. The proposed model has practical applications in diverse domains, including healthcare systems, web traffic management, and call centers. Using the supplementary variable technique, we carry out an extensive analysis of the model. This approach allows us to derive the ergodicity condition for this Markov chain and compute its stationary distribution. The main performance measures of the system are expressed through the stationary state probabilities. Numerical illustrations are presented. Finally, we conduct an economic study to assess the impact of various system parameters on performance measures and total cost, offering a visual overview of the system's effectiveness and profitability. A comparative analysis with existing models shows how our approach generalizes traditional retrial queue models, which typically consider a single type of customer arrival, by considering two distinct customer classes. This contributes to the advancement of queueing theory and provides insight into optimizing real-world systems.

Mathematics Subject Classification (2020). 60K25, 90B22, 68M20

Keywords. Breakdowns and repairs, cost model, retrial queues, performance measures, server state probabilities, service orbit

*Corresponding Author.

Email addresses: nasreddine.dehamnia@univ-bejaia.dz (N. Dehamnia), mohammed.boualem@univ-bejaia.dz (M. Boualem), lamos_bejaia@hotmail.com (D. Aïssani)

Received: 18.05.2024; Accepted: 03.03.2025

1. Introduction

In a queueing scenario, such as a telecommunications system, a notable characteristic is that, when all servers are occupied, an incoming customer must leave the service area and return to the retrial group (orbit) after a specific period. Retrial queues provide an effective solution to these situations [4,16,37]. For example, if a server is unavailable when a customer arrives, they will join the orbit to attempt their requests in a random order and at random intervals [7]. Retrial queues are widely used to model stochastic phenomena in real-world systems, including wireless communications, IT, telephone networks, and healthcare, facilitating access to central processing services [14,19,43].

In numerous practical scenarios, servers are susceptible to unpredictable breakdowns, significantly impacting system performance [32, 40]. Consequently, the study of retrial queues with unreliable servers has received considerable attention. Li and Zhang [29] examined an $M/G/1$ retrial G-queue with general retrial times where the server continues to operate at a reduced service rate during breakdowns. They established stability conditions for the system and developed generating functions for the number of customers in the orbit. Gao et al. [18] analyzed an $M/G/1$ retrial queue with two types of breakdowns occurring during idle and busy periods, presenting stability conditions and utilizing the supplementary variable method to determine the steady-state probabilities. Tian and Zhang [39] investigated an unreliable $M/M/1$ queue with negative customers, where the arrival of a negative customer causes server failure, prompting immediate repair attempts. They provided steady-state probabilities and performance measures and discussed strategic customer behavior. Ayyappan and Udayageetha [6] studied a retrial queue with priority services, incorporating server breakdowns, startup/closedown times, and Bernoulli vacations. They derived the joint distribution of the server state and the number of customers in the system using the supplementary variable technique. Jagannathan and Sivasubramaniam [22] studied an unreliable retrial queue with batch arrivals, Bernoulli vacations, and impatient customer behavior. They assumed a delay before repair initiation after a breakdown and used the supplementary variable technique to derive steady-state results. Kumar et al. [27] addressed the Markovian machine interference problem and random switching failure, considering working vacations under a threshold policy. They account for synchronized impatience behavior and obtained steady-state probabilities and performance measures using the Successive Over-Relaxation method. Kumar et al. [26] analyzed a Markovian retrial queue where the server is subject to two types of breakdowns and repairs. They provided explicit expressions for the partial probability-generating functions of the server status and the number of customers in the orbit, along with key performance measures. Dudin et al. [15] investigated a single-server, non-preemptive priority queueing system with finite capacity and multiple customer types, considering batch arrivals and dynamic priority changes. They analyzed the stationary behavior of the system using a finite-state, multi-dimensional continuous-time Markov chain and computed key system characteristics.

In recent years, research on retrial queues with impatient customers has gained more attention. Impatience is the prominent characteristic, as customers often feel anxious waiting for services in real-life situations [8–11, 13]. Customer behavior, such as queueing and reverting, plays a crucial role in real-world queueing systems, where arrivals can be discouraged by long queues [38]. Balking occurs when customers decide not to enter the system upon arrival if they find the server unavailable [1, 30, 41, 44]. On the other hand, reneging occurs when customers join the system but leave before being served [23]. Recently, researchers have shown significant interest in exploring various aspects of customer flows entering the system, including different classes of customer arrivals [17,24,33], two classes of batch arrivals [5], priority customers [25, 28, 42], negative customers [39], impatient and persistent customers [38], two-way communication [2, 3] and batch arrival

[12, 44]. Although these studies have provided valuable insights into specific aspects of retrial queueing models with multiple types of customer arrivals, further investigations are needed to explore other aspects of customer flow entering the system, notably the impact of high-patience customers (those who are prepared to wait for longer periods of time in orbit before going into service) and low-patience customers (those who have a lower tolerance of waiting and are more inclined to become impatient and leave the queue more quickly). The significance of considering these two customer flows (high-patience and low-patience customers) is crucial in queueing modeling, as they influence customer behavior and can have an impact on the performance and efficiency of the retrial queueing system.

In the literature on retrial queues, several studies have addressed the management of systems where customer patience varies and service interruptions play a key role in the overall performance of the system. For instance, Hariom et al. [21] studied an inventory model with linear demand and stock outs in a three-level production system, highlighting the importance of managing replenishments based on customer patience levels. Similarly, Singh [34] proposed an inventory model for deteriorated items with holding and selling costs, where stockout management is also a critical factor. While their model is relevant, it does not account for the specific dynamics of impatient customers who leave the system. Singh et al. [35] then extended this analysis by proposing a model with quadratic demand, incorporating multi-level production processes, which is crucial for studying systems where customers return to the orbit after a service failure. Finally, Singh et al. [36] advanced the topic by introducing a supply chain approach, addressing the demand for both finished products and raw materials, while analyzing inventory management in production systems with degradation rates and service interruptions. These studies highlight the evolution of inventory and queueing models in complex environments, but they do not fully incorporate the consideration of retrial queues with high-patience and low-patience customers, which we address in this study.

In this paper, we examine a single-server retrial queue that incorporates two types of customers (high-patience and low-patience), server breakdowns, corrective repairs, retrial and reservation times, and distinct service rates. Such a model has practical applications in communication networks, healthcare systems, and website management, where server reliability and customer behavior significantly impact system performance. By jointly considering these factors, we provide a comprehensive mathematical analysis of the retrial queue's dynamics and performance in real-world scenarios.

Building on previous research, this study presents a comparative review that highlights the key contributions of our work. As shown in Table 1, we provide a structured comparison of retrial queue models, particularly those incorporating unreliable servers, customer impatience, and retrial behavior. Our work extends these models by proposing a comprehensive framework that includes (1) two customer types with distinct patience levels; (2) server failures and repairs modeled using a corrective maintenance mechanism; (3) a reservation process for low-patience customers; (4) a general retrial distribution that extends beyond the commonly assumed exponential retrial rate; and (5) an economic analysis of system costs. This framework addresses several limitations of previous studies and offers a more realistic approach to retrial queues. To obtain the stability condition and the explicit analytical solutions of this retrial queue, we employ the supplementary variables technique. This technique enables us to derive mathematical equations for the steady-state probability distributions that describe the behavior of the retrial queueing model as well as its performance metrics. Furthermore, it allows us to evaluate the impact of key parameters on system performance and conduct an economic study to analyze the total cost. As summarized in Table 1, our model contributes to advancing queueing theory by providing a more practical and versatile framework for analyzing retrial systems.

Table 1. Comparative literature review of retrial queue models with unreliable servers and impatient customers.

Authors	**Queue model**	**Server failures and repairs**	**Re-trial cus-tomers**	**Balk-ing**	**Reneg-ing**	**Reserv-ation process**	**Two types of customer arrivals**	**Solution method**	**Eco-nomic analysis**
Taleb & Aissani (2016) [38]	$M/G/1$ queue	✓	✓	✓	✓	×	✓	Generating functions	×
Zirem et al. (2019) [44]	$M^{[X]}/G/1$ queue with batch arrivals	✓	✓	✓	✓	✓	×	Supplementary variable	✓
Ayyappan & Udayageetha (2020) [6]	$M^{[X_1]}, M^{[X_2]}/ G_1, G_2/1$ queue with batch arrivals	✓	✓	×	×	×	✓	Supplementary variable	×
Sztrik et al. (2021) [37]	$M/G/1$ queue	✓	✓	✓	✓	×	×	Simulation	×
Kumar et al. (2022) [27]	$M/G/1$ queue with feedback	✓	✓	×	×	×	×	Matrix-geometric	✓
Mahanta et al. (2024) [31]	$M/G/1$ queue with feedback	×	✓	×	×	×	×	Supplementary variable	✓
Hamadouche et al. (2024) [20]	$M/G/1$ queue	✓	✓	✓	✓	×	✓	Embedded Markov chain	×
Proposed model	$M_1, M_2/G_1, G_2$ /1 queue	✓	✓	✓	✓	✓	✓	Supplementary variable	✓

The sections of the study are as follows: In Section 2, we provide the dynamics of the retrial queue, emphasizing its practical applications in real-world scenarios. In Section 3, we derive the stability condition and the steady-state distributions of the server states. In Section 4, we discuss various significant system performance metrics. In Section 5, we present numerical findings that demonstrate how certain parameters influence the performance metrics of the system. In Section 6, we dive into the detailed analysis of the total cost. In Section 7, we conclude the work done.

2. Formulation of the model

We consider an unreliable retrial queue denoted $M_1, M_2/G_1, G_2/1$. This system features a repairable server that efficiently handles arrivals from high-patience and low-patience customers. Furthermore, the system incorporates two distinct orbit structures. To improve comprehension of the system's dynamics, we provide a visual representation of this model in Figure 1. The model is characterized by the following description:

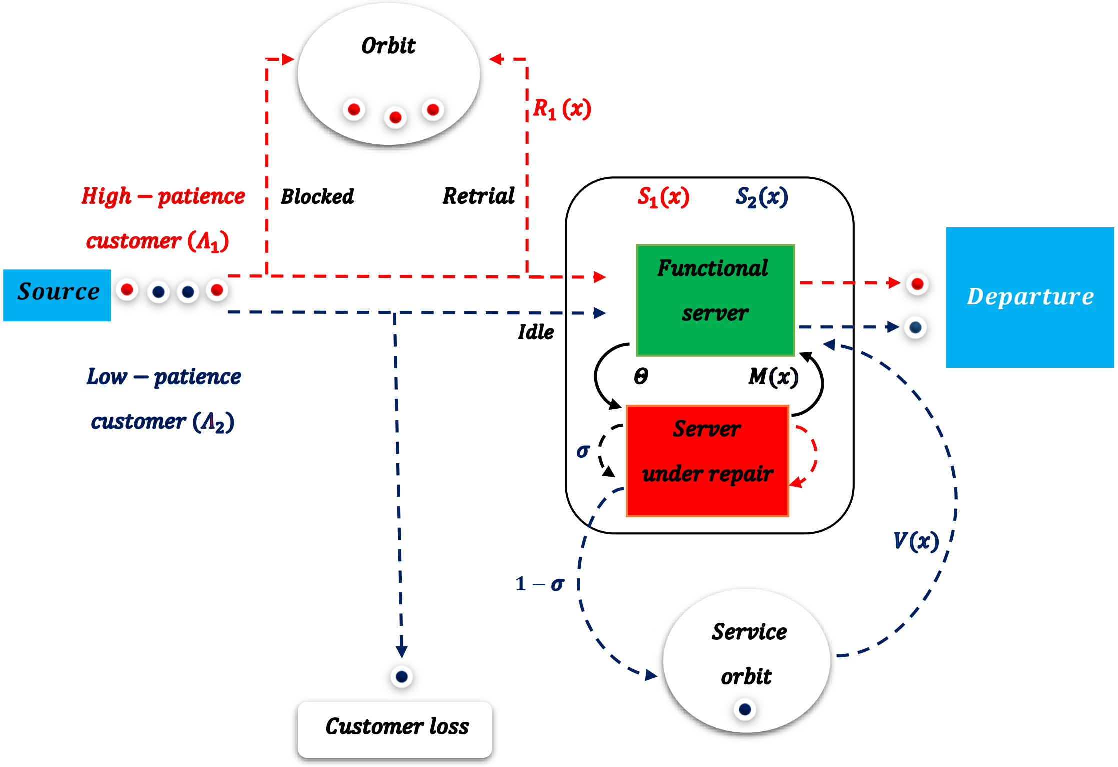


Figure 1. Comprehensive diagram of our system

i) Arrival process: The two types of customers independently arrive in the system following a Poisson process with different rates. The arrival rate for high-patience customers is denoted by $\Lambda_1 > 0$, while the arrival rate for low-patience customers is denoted by $\Lambda_2 > 0$.

ii) Service process: When a customer arrives and finds the server idle, the service for that customer begins immediately. However, if a high-patience customer encounters a blocked server (busy, down, or reserved), they leave the service area and enter into the orbit, becoming a source of repeated calls (retrial customers). In contrast, a customer with low patience who encounters a blocked server permanently leaves the system. We assume that the service policy follows the FCFS (First-Come, First-Served) discipline. The service time distributions for high-patience and low-patience customers are expressed by $S_1(x)$ and $S_2(x)$ respectively. The Laplace-Stieltjes transform (LST) of the service times distribution for high-patience customers is indicated by $L_{S_1}[z]$, while for low-patience customers, it is indicated by $L_{S_2}[z]$. Furthermore, we can compute the moments of the service-time distributions as follows: for high-patience customers, $\beta_{1j} = (-1)^j L^j_{S_1}[0]$, for low-patience customers, $\beta_{2j} = (-1)^j L^j_{S_2}[0]$, where j represents the order of the moment.

iii) Retrial process: The orbit is exclusively reserved for high-patience customers. When a primary customer (a customer outside the system) arrives first in the system, the high-patience secondary customer (retrial customers) opts to exit the service area and enter into the orbit and attempt service again after a random interval of time. The retrial time for high-patience customers follows a general distribution, denoted as $R_1(w)$. To analyze this distribution, we use the LST, represented as $L_{R_1}[z]$. Moreover, we define $\alpha_{1j} = (-1)^j L^j_{R_1}[0]$ as the jth moment of the retrial time distribution for customers with high patience.

iv) Breakdown process: The server undergoes active breakdowns, which means that it fails only while providing service. The duration of a failure follows an exponential distribution with a mean of $1/\Theta$.

v) Repair process: When a failure occurs, repairs are started immediately. Repair times follow a general distribution denoted by $M(y)$ and the LST is denoted by $L_M[z]$. The jth moment of the repair time distribution is given by the expression $\Gamma_j = (-1)^j L^j_M[0]$.

vi) Reservation process: When a service interruption occurs while a low-patience customer is being served, there are two options: either the customer remains in the service area with a probability σ, or the customer enters a service orbit with a complementary probability $(1-\sigma)$. On the other hand, when a service interruption occurs while a high-patience customer is being served, the latter remains in the service area. However, if a customer with low patience enters the service orbit due to a server failure, the server must wait for the customer to return after the repair. This period of time is referred to as the reserved time. The reserved time follows a general distribution, characterized by the function $V(v)$ and its LST, $L_V[z]$. The jth moment of the reservation time distribution is indicated by $\Delta_j = (-1)^j L^j_V[0]$.

We assume the following properties for various functions related to the system: $R_1(0) = 0$, $R_1(\infty) = 1$, $S_1(0) = 0$, $S_1(\infty) = 1$, $S_2(0) = 0$, $S_2(\infty) = 1$, $M(0) = 0$, $M(\infty) = 1$, $V(0) = 0$, $V(\infty) = 1$. $\overline{F}(x) = 1 - F(x)$ is the complementary cumulative distribution function of the probability distribution function (*pdf*) $F(x)$ on $[0,1]$. $L_F[s] = \int_0^\infty e^{-sx} dF(x)$ is the LST of *pdf*. $\overline{L}_F[s] = \int_0^\infty e^{-sx}(1 - F(x))\,dx = \frac{1-L_F[s]}{s}$ is the complementary LST. We also assume the mutual independence of all the introduced variables. The notation used to represent the conditional completion rates is as follows: $\alpha_1(w)dw$ is the conditional completion rate for repeated attempts by high-patience customers, where w represents the waiting time, calculated as $\frac{R_1(w)dw}{1-R_1(w)}$, considering the survival function $R_1(w)$. $\beta_1(x)dx$ is the conditional completion rate of the service for high-patience customers, with x representing the service time, calculated as $\frac{S_1(x)dx}{1-S_1(x)}$, considering the survival function $S_1(x)$. $\beta_2(x)dx$ is the conditional completion rate of the service for low-patience customers, with x representing the service time, calculated as $\frac{S_2(x)dx}{1-S_2(x)}$, considering the survival function $S_2(x)$. $\Gamma(y)dy$ is the conditional completion rate for repair, where y represents the repair time, calculated as $\frac{M(y)dy}{1-M(y)}$, considering the survival function $M(y)$. $\Delta(v)dv$ is the conditional completion rate for the reserved time, where v represents the reserved time, calculated as $\frac{V(v)dv}{1-V(v)}$, considering the survival function $V(v)$.

2.1. Practical applications of the proposed model

Example 2.1. Effective patient flow management is essential for maintaining the efficiency of the healthcare system, ensuring high-quality patient care, and optimizing resource utilization. Our $M_1, M_2/G_1, G_2/1$ retrial queue provides a robust analytical framework that enhances patient access to medical services while ensuring system stability and operational resilience. By integrating retrial mechanisms and system failures, this model accurately represents the complexities of real-world healthcare operations, particularly under fluctuating demand and resource constraints.

One of the key strengths of our model is its dynamic patient flow management. The healthcare facility operates as a server where patients are classified according to their willingness to wait. Patients with high patience continue to seek medical attention by retrying service after an initial denial, while patients with low patience are more likely to leave the system and seek alternative care if immediate service is unavailable. Patient arrivals are modeled as a Poisson process, reflecting the inherent randomness of medical emergencies and scheduled visits. The duration of service follows a general distribution that accounts for the variability in the complexity of treatment. The model also incorporates facility failures due to equipment malfunctions, resource shortages, or other operational problems. These failures follow an exponential distribution for time until failure, whereas repair durations are modeled using a general distribution.

A fundamental indicator of the effectiveness of the model is its adaptive response to peak-hour congestion and facility downtimes. When system capacity is exceeded or temporarily reduced, low-patience patients may permanently exit the system, while high-patience patients enter a retrial orbit, attempting to access medical care after a randomized delay. Furthermore, if a breakdown occurs during ongoing treatment, two possible outcomes are considered. With probability σ, the affected patient remains in the system and resumes treatment immediately after recovery from the system. In contrast, with probability $\overline{\sigma}$, the patient exits the facility but joins the service orbit, seeking treatment at a later stage. Upon system restoration, patients in the service area receive priority access to medical care, minimizing service interruptions and ensuring continuity of treatment. In addition, a reservation mechanism is implemented, allowing low-patience patients who entered the service orbit to resume their treatment immediately after repair. This improves system efficiency, promotes equitable resource allocation, and ensures a fair and effective patient service process.

Example 2.2. Efficient website traffic management is essential to ensure a seamless user experience, particularly during peak hours, when a surge in visitor requests can lead to server congestion. To analyze and optimize this traffic flow, we use the $M_1, M_2/G_1, G_2/1$ retrial queue, which provides a robust framework for modeling user interactions and server performance. In this model, the server represents the website's processing unit, while visitors are categorized into two distinct groups based on their patience levels. Visitor arrivals, regardless of their patience category, follow a Poisson process, accurately capturing the randomness of user traffic. Service times for high-patience and low-patience visitors are assumed to follow general distributions, reflecting the variability in processing requirements for different request types.

An important aspect of the model is its ability to incorporate reliability factors from the server. The server may experience failures after a certain operational period, which is modeled using an exponential distribution. When the server is operational but under high demand, low-patience visitors may choose to leave the system, whereas high-patience visitors enter a retrial orbit, attempting to reconnect after a randomly determined waiting period. This retrial mechanism allows the system to regulate congestion effectively while prioritizing users who are willing to wait for service.

In the event of server failure and subsequent recovery, a priority-based service mechanism is implemented to accommodate low-patience visitors. This system designates a reserved time period modeled as a random variable during which low-patience visitors receive preferential access to the server. The primary objective of this priority mechanism is to minimize service abandonment, ensuring that low-patience users can complete their transactions efficiently upon server restoration.

3. Analysis of the steady-state probability distribution

The aim of this section is to establish the steady-state distributions of the retrial queue using the supplementary variable method and generating functions. To achieve this goal, we constructed a mathematical model for the retrial queueing system, incorporating the notation and assumptions outlined in the preceding section. This model allows us to depict the system's behavior as a Markov process, where transitions between different states are determined by transition probabilities. Figure 2 depicts the state transition diagram. To describe the stochastic behavior of the retrial queue at time t, we define a set of variables, denoted as $\{X(t), t \geq 0\}$, as follows:

$$\{X(t)\}_{t\geq 0} = \{\Upsilon(t), \Psi(t), \Phi(t), \kappa_0(t), \kappa_1(t), \kappa_2(t), \kappa_3(t), \kappa_4(t)\}.$$

The variable $\Upsilon(t)$ indicates the state of the server at time t, defined as follows:

$$\Upsilon(t) = \begin{cases} 0 & \mapsto \text{Idle}, \\ 1 & \mapsto \text{Busy by a high-patience customer}, \\ 2 & \mapsto \text{Busy by a low-patience customer}, \\ 3 & \mapsto \text{Under repair}, \\ 4 & \mapsto \text{Reserved by a low-patience customer}. \end{cases}$$

The variable $\Phi(t)$ indicates the number of high-patience customers waiting for service in the retrial group (orbit). The variable $\Psi(t)$ represents the state of the customer in service after an active breakdown and takes the following values:

$$\Psi(t) = \begin{cases} 0 & \mapsto \text{ the high-patience customer remains in service position}, \\ 1 & \mapsto \text{ the low-patience customer remains in service position}, \\ 2 & \mapsto \text{ the low-patience customer enters a service orbit}. \end{cases}$$

The variable $\kappa_i(t)$, $i = 0, 1, 2, 3, 4$, represents the supplementary variable at time t in the retrial queue, where:
$\kappa_0(t)$: the elapsed retrial time of a high-patience customer,
$\kappa_1(t)$: the elapsed service time of a high-patience customer,
$\kappa_2(t)$: the elapsed service time of a low-patience customer,
$\kappa_3(t)$: the time taken to repair the server after a failure occurs,
$\kappa_4(t)$: the elapsed reserved time of a low-patience customer.

The transient state probabilities are denoted as follows: $P_{0,0}(t) = P(\Upsilon(t) = 0, \Phi(t) = 0)$ is the probability that the system is empty. $P_{0,n}(t, w)\partial w = P(\Upsilon(t) = 0, \Phi(t) = n, w \leq \kappa_0(t) < w + \partial w), n \geq 1$ is the probability that the server remains idle throughout the retrial period, since there are n high-patience customers in the orbit. $P_{1,n}(t, x)\partial x = P(\Upsilon(t) = 1, \Phi(t) = n, x \leq \kappa_1(t) < x + \partial x)$ (resp. $P_{2,n}(t, x)\partial x = P(\Upsilon(t) = 2, \Phi(t) = n, x \leq \kappa_2(t) < x + \partial x)$) is the probability that the server is occupied by a high-patience (resp. low-patience) customer throughout the retrial period, since there are n high-patience customers in the orbit. $P_{3,0,n}(t, x, y)\partial x\partial y = P(\Upsilon(t) = 3, \Psi(t) = 0, \Phi(t) = n, x \leq \kappa_1(t) < x + \partial x, y \leq \kappa_3(t) < y + \partial y)$ (resp. $P_{3,1,n}(t, x, y)\partial x\partial y = P(\Upsilon(t) = 3, \Psi(t) = 1, \Phi(t) = n, x \leq \kappa_2(t) < x + \partial x, y \leq \kappa_3(t) < y + \partial y)$) is the probability that the server is in repair and the high-patience (resp. low-patience) customer in service remains in the service zone, as there are n high-patience customers in the orbit. $P_{3,2,n}(t, x, y)\partial x\partial y = P(\Upsilon(t) = 3, \Psi(t) = 2, \Phi(t) = n, x \leq \kappa_2(t) < x + \partial x, y \leq \kappa_3(t) < y + \partial y)$ is the probability that the server is in repair and that the low-patience customer occupying the server enters the service orbit, given that there are n high-patience customers in the orbit. $P_{4,n}(t,x,v)\partial x\partial v = P(\Upsilon(t) = 4, \Phi(t) = n, x \le \kappa_2(t) < x+\partial x, v \le \kappa_4(t) < v+\partial v)$ is the probability that the server is reserved by a low-patience customer during the retrial period, since there are n high-patience customers in the orbit.

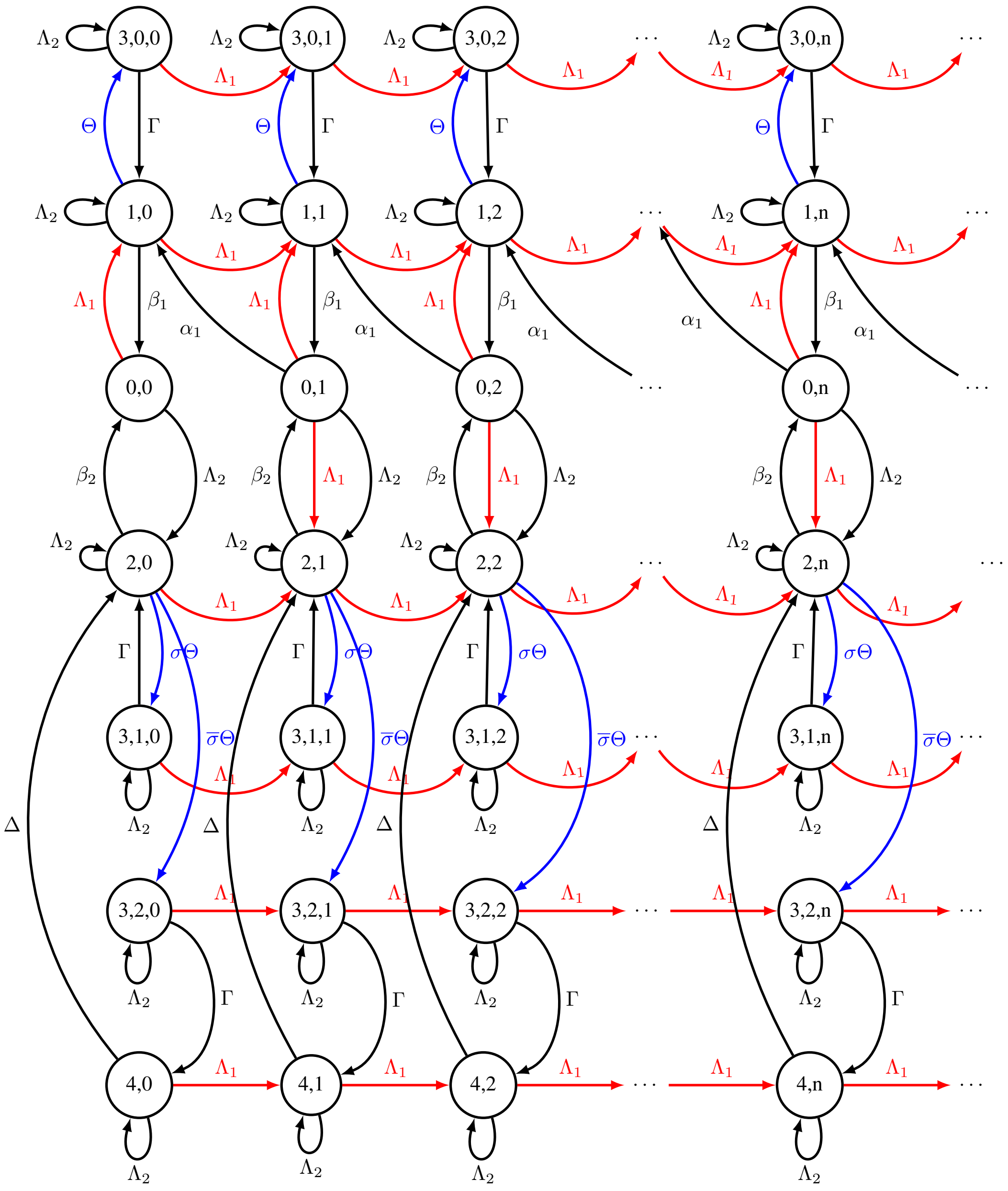


Figure 2. State transition rate diagram.

3.1. The steady-state solution

Theorem 3.1. *Inequality $\rho < 1$ is a sufficient condition for ergodicity.*

Proof. See Remark 3.3. □

Now, we introduce the following probability-generating functions:

$$
\begin{aligned}
P_0(z,w) &= \sum_{n=1}^{\infty} P_{0,n}(w)z^n, \\
P_i(z,x) &= \sum_{n=0}^{\infty} P_{i,n}(x)z^n,\ i=1,2, \\
P_{3,j}(z,x,y) &= \sum_{n=0}^{\infty} P_{3,j,n}(x,y)z^n,\ j=0,1,2, \\
P_4(z,x,v) &= \sum_{n=0}^{\infty} P_{4,n}(x,v)z^n,
\end{aligned}
$$

which are convergent for each $w \geq 0$, $x \geq 0$, $y \geq 0$, $v \geq 0$ and for all $|z| \leq 1$.

Corollary 3.2. *If $\rho < 1$, the generating functions of the server states are expressed as follows:*

(i) If the server is idle, the generating function is given by

$$
P_0(z,w) = \frac{\left\{ zP_{0,0}\Big[\Lambda_1 + \Lambda_2\Big(1 - L_{S_2}[F(z)]\Big)\Big] e^{-(\Lambda_1+\Lambda_2)w}\overline{R}_1(w) \right\}}{\left\{ \begin{array}{l} L_{S_1}[J(z)]\bigg(\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1 z + L_{R_1}[\Lambda_1+\Lambda_2]\bigg) \\ + L_{S_2}[F(z)]\bigg(1 - L_{R_1}[\Lambda_1+\Lambda_2]\bigg)z - z \end{array} \right\}}. \tag{3.1}
$$

(ii) When the server is occupied by a high-patience customer, the generating function equals to

$$
P_1(z,x) = \frac{\left\{ \begin{array}{l} P_{0,0}\Big[\Lambda_1 + \Lambda_2\Big(1 - L_{S_2}[F(z)]\Big)\Big]\Big[\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1 z + L_{R_1}[\Lambda_1+\Lambda_2]\Big] \\ \times\, e^{-J(z)x}\overline{S}_1(x) \end{array} \right\}}{\left\{ \begin{array}{l} L_{S_1}[J(z)]\bigg(\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1 z + L_{R_1}[\Lambda_1+\Lambda_2]\bigg) \\ + L_{S_2}[F(z)]\bigg(1 - L_{R_1}[\Lambda_1+\Lambda_2]\bigg)z - z \end{array} \right\}}. \tag{3.2}
$$

(iii) When the server is occupied by a low-patience customer, the generating function is defined as

$$
P_2(z,x) = \frac{\left\{ \begin{array}{l} P_{0,0}\Big[z(\Lambda_1+\Lambda_2)\Big(1 - L_{R_1}[\Lambda_1+\Lambda_2]\Big) \\ \quad + \Lambda_2 L_{S_1}[J(z)]\bigg(\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1 z + L_{R_1}[\Lambda_1+\Lambda_2]\bigg) \\ \quad - z\Big] e^{-F(z)x}\overline{S}_2(x) \end{array} \right\}}{\left\{ \begin{array}{l} L_{S_1}[J(z)]\bigg(\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1 z + L_{R_1}[\Lambda_1+\Lambda_2]\bigg) \\ + L_{S_2}[F(z)]\bigg(1 - L_{R_1}[\Lambda_1+\Lambda_2]\bigg)z - z \end{array} \right\}}. \tag{3.3}
$$

(iv) If the server is under repair and the high-patience customer in service chooses to remain in the service area, we obtain the generating function as

$$P_{3,0}(z,x,y)=\frac{\left\{\begin{array}{l}\Theta P_{0,0}\Big[\Lambda_1+\Lambda_2\Big(1-L_{S_2}[F(z)]\Big)\Big]\Big[\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1 z\\ +L_{R_1}[\Lambda_1+\Lambda_2]\Big]e^{-J(z)x}\overline{S}_1(x)e^{-\Lambda_1\bar{z}y}\overline{M}(y)\end{array}\right\}}{\left\{\begin{array}{l}L_{S_1}[J(z)]\Big(\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1 z+L_{R_1}[\Lambda_1+\Lambda_2]\Big)\\ +L_{S_2}[F(z)]\Big(1-L_{R_1}[\Lambda_1+\Lambda_2]\Big)z-z\end{array}\right\}}. \tag{3.4}$$

(v) When the server is under repair and the low-patience customer in service chooses to remain in the service area, we have

$$P_{3,1}(z,x,y)=\frac{\left\{\begin{array}{l}P_{0,0}\Big[(\Lambda_1+\Lambda_2)z\Big(1-L_{R_1}[\Lambda_1+\Lambda_2]\Big)\\ +\Lambda_2 L_{S_1}[J(z)]\Big(\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1 z+L_{R_1}[\Lambda_1+\Lambda_2]\Big)\\ -z\Big]\sigma\Theta e^{-F(z)x}\overline{S}_2(x)e^{-\Lambda_1\bar{z}y}\overline{M}(y)\end{array}\right\}}{\left\{\begin{array}{l}L_{S_1}[J(z)]\Big(\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1 z+L_{R_1}[\Lambda_1+\Lambda_2]\Big)\\ +L_{S_2}[F(z)]\Big(1-L_{R_1}[\Lambda_1+\Lambda_2]\Big)z-z\end{array}\right\}}. \tag{3.5}$$

(vi) If the server is under repair and the low-patience customer currently in service enters the service orbit, we define the generating function as

$$P_{3,2}(z,x,y)=\frac{\left\{\begin{array}{l}P_{0,0}\Big[(\Lambda_1+\Lambda_2)z\Big(1-L_{R_1}[\Lambda_1+\Lambda_2]\Big)\\ +\Lambda_2 L_{S_1}[J(z)]\Big(\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1 z+L_{R_1}[\Lambda_1+\Lambda_2]\Big)\\ -z\Big]\overline{\sigma}\Theta e^{-F(z)x}\overline{S}_2(x)e^{-\Lambda_1\bar{z}y}\overline{M}(y)\end{array}\right\}}{\left\{\begin{array}{l}L_{S_1}[J(z)]\Big(\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1 z+L_{R_1}[\Lambda_1+\Lambda_2]\Big)\\ +L_{S_2}[F(z)]\Big(1-L_{R_1}[\Lambda_1+\Lambda_2]\Big)z-z\end{array}\right\}}. \tag{3.6}$$

(vii) When the server is reserved by low-patience customer, the generating function can be defined as

$$P_4(z,x,v)=\frac{\left\{\begin{array}{l}P_{0,0}\Big[(\Lambda_1+\Lambda_2)z\Big(1-L_{R_1}[\Lambda_1+\Lambda_2]\Big)\\ +\Lambda_2L_{S_1}[J(z)]\Big(\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1z+L_{R_1}[\Lambda_1+\Lambda_2]\Big)\\ -z\Big]\overline{\sigma}\Theta e^{-F(z)x}\overline{S}_2(x)L_M[\Lambda_1\overline{z}]e^{-\Lambda_1\overline{z}v}\overline{V}(v)\end{array}\right\}}{\left\{\begin{array}{l}L_{S_1}[J(z)]\Big(\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1z+L_{R_1}[\Lambda_1+\Lambda_2]\Big)\\ +L_{S_2}[F(z)]\Big(1-L_{R_1}[\Lambda_1+\Lambda_2]\Big)z-z\end{array}\right\}}. \tag{3.7}$$

Here, we have

$$\begin{aligned}J(z)&=\Lambda_1\overline{z}+\Theta-\Theta L_M[\Lambda_1\overline{z}],\\ F(z)&=\Lambda_1\overline{z}+\Theta-\Theta L_M[\Lambda_1\overline{z}]\Big(\sigma+\overline{\sigma}L_V[\Lambda_1\overline{z}]\Big).\end{aligned}$$

Proof. The supplementary variable technique (SVT) and generating functions are employed to derive the following server state equations:

(i) Idle state of the server:

$$[2\Lambda_1+\Lambda_2+\alpha_1(w)+\frac{\partial}{\partial w}]P_0(z,w) = 0. \tag{3.8}$$

(ii) Busy state of the server:

$$[\Theta+\Lambda_1+\beta_1(x)+\frac{\partial}{\partial x}]P_1(z,x) = \int_0^\infty\Gamma(y)P_{3,0}(z,x,y)\,dy+\Lambda_1zP_1(z,x), \tag{3.9}$$

$$\begin{aligned}\left[\Theta+\Lambda_1+\beta_2(x)+\frac{\partial}{\partial x}\right]P_2(z,x) &= \int_0^\infty\Gamma(y)P_{3,1}(z,x,y)\,dy\\ &+ \int_0^\infty\Delta(v)P_4(z,x,v)\,dv+\Lambda_1zP_2(z,x).\end{aligned} \tag{3.10}$$

(iii) Repair state of the server:

$$\left[\Lambda_1+\Gamma(y)+\frac{\partial}{\partial x}\right]P_{3,0}(z,x,y) = \Lambda_1zP_{3,0}(z,x,y), \tag{3.11}$$

$$\left[\Lambda_1+\Gamma(y)+\frac{\partial}{\partial x}\right]P_{3,1}(z,x,y) = \Lambda_1zP_{3,1}(z,x,y), \tag{3.12}$$

$$\left[\Lambda_1+\Gamma(y)+\frac{\partial}{\partial x}\right]P_{3,2}(z,x,y) = \Lambda_1zP_{3,2}(z,x,y). \tag{3.13}$$

(iv) Reservation state of the server:

$$\left[\Lambda_1+\Delta(v)+\frac{\partial}{\partial x}\right]P_4(z,x,v) = \Lambda_1zP_4(z,x,v). \tag{3.14}$$

The following expressions are derived by solving Equations (3.8)–(3.14) with respect to the provided boundary conditions as

$$P_0(z,0) = \int_0^\infty\beta_1(x)P_1(z,x)\,dx+\int_0^\infty\beta_2(x)P_2(z,x)\,dx-(\Lambda_1+\Lambda_2)P_{0,0}, \tag{3.15}$$

$$P_1(z,0) = \Lambda_1 \int_0^\infty P_0(z,w)\,dw + \frac{1}{z}\int_0^\infty \alpha_1(w)P_0(z,w)\,dw, \tag{3.16}$$

$$P_2(z,0) = (\Lambda_1+\Lambda_2)\int_0^\infty P_0(z,w)\,dw + \Lambda_2 P_{0,0}, \tag{3.17}$$

$$P_{3,0}(z,x,0) = \Theta P_1(z,x), \tag{3.18}$$

$$P_{3,1}(z,x,0) = \sigma\Theta P_2(z,x), \tag{3.19}$$

$$P_{3,2}(z,x,0) = \overline{\sigma}\Theta P_2(z,x), \tag{3.20}$$

$$P_4(z,x,0) = \int_0^\infty \Gamma(y)P_{3,2}(z,x,y)\,dy. \tag{3.21}$$

The normalization equation is given by

$$\begin{aligned} &P_{0,0} + \int_0^\infty P_0(1,w)\,dw + \int_0^\infty P_1(1,x)\,dx + \int_0^\infty P_2(1,x)\,dx \\ &+ \int_0^\infty\int_0^\infty P_{3,0}(1,x,y)\,dxdy + \int_0^\infty\int_0^\infty P_{3,1}(1,x,y)\,dxdy \\ &+ \int_0^\infty\int_0^\infty P_{3,2}(1,x,y)\,dxdy + \int_0^\infty\int_0^\infty P_4(1,x,v)\,dxdv = 1. \end{aligned} \tag{3.22}$$

We obtain the following equation by substituting Equation (3.18) into (3.11) as

$$P_{3,0}(z,x,y) = \Theta P_1(z,x)e^{-\Lambda_1\overline{z}y}(1-M(y)). \tag{3.23}$$

Now, replacing Equation (3.19) in Equation (3.12), we get

$$P_{3,1}(z,x,y) = \sigma\Theta P_2(z,x)e^{-\overline{z}\Lambda_1 y}(1-M(y)). \tag{3.24}$$

Then, substituting Equation (3.20) into Equation (3.13), we have

$$P_{3,2}(z,x,y) = \overline{\sigma}\Theta P_2(z,x)e^{-\overline{z}\Lambda_1 y}(1-M(y)). \tag{3.25}$$

We derive $P_4(z,x,v)$ substituting Equations (3.21) and (3.25) into Equation (3.14) as

$$P_4(z,x,v) = \overline{\sigma}\Theta L_M[\overline{z}\Lambda_1]P_2(z,x)e^{-\overline{z}\Lambda_1 v}(1-V(v)). \tag{3.26}$$

$P_1(z,x)$ is obtained by substituting Equations (3.23) and (3.16) into Equation (3.9) as

$$P_1(z,x) = \left[\Lambda_1\int_0^\infty P_0(z,w)\,dw + \frac{1}{z}\int_0^\infty \alpha_1(w)P_0(z,w)\,dw\right]e^{-J(z)x}\overline{S}_1(x). \tag{3.27}$$

Substituting Equations (3.24), (3.26) and (3.17) into Equation (3.10), we get

$$P_2(z,x) = \left[(\Lambda_1+\Lambda_2)\int_0^\infty P_0(z,w)\,dw + \Lambda_2 P_{0,0}\right]e^{-F(z)x}\overline{S}_2(x). \tag{3.28}$$

From Equations (3.23) and (3.27), we get

$$\begin{aligned} P_{3,0}(z,x,y) &= \Theta\left[\Lambda_1\int_0^\infty P_0(z,w)\,dw + \frac{1}{z}\int_0^\infty \alpha_1(w)P_0(z,w)\,dw\right]e^{-J(z)x}\overline{S}_1(x) \\ &\times\ e^{-\Lambda_1\overline{z}y}\overline{M}(y). \end{aligned} \tag{3.29}$$

Using Equations (3.24) and (3.28), we find

$$\begin{aligned} P_{3,1}(z,x,y) &= \sigma\Theta\left[(\Lambda_1+\Lambda_2)\int_0^\infty P_0(z,w)\,dw + \Lambda_2 P_{0,0}\right]e^{-F(z)x}\overline{S}_2(x) \\ &\times\ e^{-\Lambda_1\overline{z}y}\overline{M}(y). \end{aligned} \tag{3.30}$$

According to Equations (3.25) and (3.28), we obtain

$$\begin{aligned} P_{3,2}(z,x,y) &= \overline{\sigma}\Theta\Big[(\Lambda_1+\Lambda_2)\int_0^\infty P_0(z,w)\,dw + \Lambda_2 P_{0,0}\Big] e^{-F(z)x}\overline{S}_2(x) \\ &\times \; e^{-\Lambda_1\overline{z}y}\overline{M}(y). \end{aligned} \tag{3.31}$$

From Equations (3.26) and (3.28), we have

$$\begin{aligned} P_4(z,x,v) &= \overline{\sigma}\Theta L_M[\Lambda_1\overline{z}]\Big[(\Lambda_1+\Lambda_2)\int_0^\infty P_0(z,w)\,dw + \Lambda_2 P_{0,0}\Big] e^{-F(z)x}\overline{S}_2(x) \\ &\times \; e^{-\Lambda_1\overline{z}v}\overline{V}(v). \end{aligned} \tag{3.32}$$

From Equations (3.15), (3.27) and (3.28), we obtain

$$P_0(z,0) = \frac{\Big\{zP_{0,0}\Big[\Lambda_1+\Lambda_2\Big(1-L_{S_2}[F(z)]\Big)\Big]\Big\}}{\left\{\begin{array}{l} L_{S_1}[J(z)]\Big(\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1 z + L_{R_1}[\Lambda_1+\Lambda_2]\Big) \\ + L_{S_2}[F(z)]\Big(1-L_{R_1}[\Lambda_1+\Lambda_2]\Big)z - z \end{array}\right\}}. \tag{3.33}$$

Next, by substituting Equation (3.33) into Equation (3.8), we find

$$P_0(z,w) = \frac{\Big\{zP_{0,0}\Big[\Lambda_1+\Lambda_2\Big(1-L_{S_2}[F(z)]\Big)\Big]e^{-(\Lambda_1+\Lambda_2)w}\overline{R}_1(w)\Big\}}{\left\{\begin{array}{l} L_{S_1}[J(z)]\Big(\overline{L}_{R_1}[\Lambda_1+\Lambda_2]\Lambda_1 z + L_{R_1}[\Lambda_1+\Lambda_2]\Big) \\ + L_{S_2}[F(z)]\Big(1-L_{R_1}[\Lambda_1+\Lambda_2]\Big)z - z \end{array}\right\}}. \tag{3.34}$$

□

Remark 3.3. The quantity $P_{0,0}$ can be found using Equation (3.22), which is given by

$$P_{0,0} \;=\; 1-\lim_{z\to 1}\Big(P_0(z)+P_1(z)+P_2(z)+P_{30}(z)+P_{31}(z)+P_{32}(z)+P_4(z)\Big).$$

Thus, $P_{0,0}$ is obtained by

$$P_{0,0} = \frac{\Big\{\Lambda_1\Big(1-L_{R_1}[\Lambda_1+\Lambda_2]\Big)\Big\}}{\left\{\begin{array}{l} \Big(2\Lambda_1\Big(1-L_{R_1}[\Lambda_1+\Lambda_2]\Big)+\Lambda_1\beta_{11}\Big[1+\Theta\Gamma_1\Big]\Big[\Lambda_1+\Lambda_2 L_{R_1}[\Lambda_1+\Lambda_2]\Big] \\ +\beta_{21}\Big[1+\Theta\Gamma_1+\overline{\sigma}\Theta\Delta_1\Big]\Big[\Big((\Lambda_1+\Lambda_2)^2+\Lambda_1\Lambda_2\Big)\Big(1-L_{R_1}[\Lambda_1+\Lambda_2]\Big) \\ +(\Lambda_1+\Lambda_2)\Big(\Lambda_2 L_{R_1}[\Lambda_1+\Lambda_2]-1\Big)\Big]\Big) \end{array}\right\}}. \tag{3.35}$$

To ensure the stability of the system under study, condition $\rho < 1$ must be satisfied. Here, $\rho = (1-P_{0,0})$ represents the traffic intensity of the system. Maintaining a traffic intensity of $\rho < 1$ prevents the system from being overwhelmed by arrivals and allows it to handle incoming workload efficiently. This condition ensures that the system operates within its capacity and avoids excessive congestion or delays. When condition $\rho < 1$ is met, the system has a positive probability $P_{0,0}$ for the initial state, indicating that the system can start in an empty state and transition between different states. This allows for a balanced distribution of probabilities among the states, promoting stability and preventing the system from remaining in a single state indefinitely.

From Equation (3.35), we derive the expression for ρ as

$$\rho=\frac{\left\{\begin{array}{l}\Big(\Lambda_1\Big(1-L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)+\Lambda_1\beta_{11}\Big[1+\Theta\Gamma_1\Big]\Big[\Lambda_1+\Lambda_2 L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big]\\ +\beta_{21}\Big[1+\Theta\Gamma_1+\overline{\sigma}\Theta\Delta_1\Big]\Big[\Big((\Lambda_1+\Lambda_2)^2+\Lambda_1\Lambda_2\Big)\Big(1-L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\\ +(\Lambda_1+\Lambda_2)\Big(\Lambda_2 L_{R_1}[\Lambda_1+\Lambda_2]-1\Big)\Big]\Big)\end{array}\right\}}{\left\{\begin{array}{l}\Big(2\Lambda_1\Big(1-L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)+\Lambda_1\beta_{11}\Big[1+\Theta\Gamma_1\Big]\Big[\Lambda_1+\Lambda_2 L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big]\\ +\beta_{21}\Big[1+\Theta\Gamma_1+\overline{\sigma}\Theta\Delta_1\Big]\Big[\Big((\Lambda_1+\Lambda_2)^2+\Lambda_1\Lambda_2\Big)\Big(1-L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\\ +(\Lambda_1+\Lambda_2)\Big(\Lambda_2 L_{R_1}[\Lambda_1+\Lambda_2]-1\Big)\Big]\Big)\end{array}\right\}}. \tag{3.36}$$

4. System performance measures

The main objective of this section is to derive explicit formulas for the probabilities of state of the server, as well as some performance measures.

Corollary 4.1. *The server state probabilities are expressed as follows:*

(i) If the server is idle, we have

$$P_0=\frac{\Big\{\Lambda_1\Big(1-L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\Big\}}{\left\{\begin{array}{l}\Big(2\Lambda_1\Big(1-L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)+\Lambda_1\beta_{11}\Big[1+\Theta\Gamma_1\Big]\Big[\Lambda_1+\Lambda_2 L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big]\\ +\beta_{21}\Big[1+\Theta\Gamma_1+\overline{\sigma}\Theta\Delta_1\Big]\Big[\Big((\Lambda_1+\Lambda_2)^2+\Lambda_1\Lambda_2\Big)\Big(1-L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\\ +(\Lambda_1+\Lambda_2)\Big(\Lambda_2 L_{R_1}[\Lambda_1+\Lambda_2]-1\Big)\Big]\Big)\end{array}\right\}}. \tag{4.1}$$

(ii) When the server is busy by a high-patience customer, we define

$$P_1=\frac{\Big\{P_{0,0}\Lambda_1\beta_{11}\Big(\Lambda_1+\Lambda_2 L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\Big\}}{\Big\{\Lambda_1\Big(1-L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\Big\}}. \tag{4.2}$$

(iii) If the server is busy by a low-patience customer, the probability is given by

$$P_2=\frac{\left\{\begin{array}{l}P_{0,0}\beta_{21}\Big[\Big((\Lambda_1+\Lambda_2)^2+\Lambda_1\Lambda_2\Big)\Big(1-L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\\ +\Big(\Lambda_1+\Lambda_2\Big)\Big(\Lambda_2 L_{R_1}\left[\Lambda_1+\Lambda_2\right]-1\Big)\Big]\end{array}\right\}}{\Big\{\Lambda_1\Big(1-L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\Big\}}. \tag{4.3}$$

(iv) When the server is under repair and the high-patience customer in service remains in the service area, we have

$$P_{3,0} = \frac{\Big\{P_{0,0}\Theta\Gamma_1\Lambda_1\beta_{11}\Big(\Lambda_1 + \Lambda_2 L_{R_1}\left[\Lambda_1 + \Lambda_2\right]\Big)\Big\}}{\Big\{\Lambda_1\Big(1 - L_{R_1}\left[\Lambda_1 + \Lambda_2\right]\Big)\Big\}}. \tag{4.4}$$

(v) If the server is under repair and the low-patience customer in service remains in the service area, the probability is given by

$$P_{3,1} = \frac{\left\{\begin{array}{l}\sigma\Theta P_{0,0}\beta_{21}\Gamma_1\Big[\Big((\Lambda_1+\Lambda_2)^2 + \Lambda_1\Lambda_2\Big)\Big(1 - L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\\ +\Big(\Lambda_1+\Lambda_2\Big)\Big(\Lambda_2 L_{R_1}\left[\Lambda_1+\Lambda_2\right] - 1\Big)\Big]\end{array}\right\}}{\Big\{\Lambda_1\Big(1 - L_{R_1}\left[\Lambda_1 + \Lambda_2\right]\Big)\Big\}}. \tag{4.5}$$

(vi) If the server is under repair and the low-patience customer who occupies the server enters into the service orbit, we obtain

$$P_{3,2} = \frac{\left\{\begin{array}{l}\overline{\sigma}\Theta P_{0,0}\beta_{21}\Gamma_1\Big[\Big((\Lambda_1+\Lambda_2)^2 + \Lambda_1\Lambda_2\Big)\Big(1 - L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\\ +\Big(\Lambda_1+\Lambda_2\Big)\Big(\Lambda_2 L_{R_1}\left[\Lambda_1+\Lambda_2\right] - 1\Big)\Big]\end{array}\right\}}{\Big\{\Lambda_1\Big(1 - L_{R_1}\left[\Lambda_1 + \Lambda_2\right]\Big)\Big\}}. \tag{4.6}$$

(vii) When the server is reserved by a low-patience customer, the probability is given by

$$P_4 = \frac{\left\{\begin{array}{l}\overline{\sigma}\Theta P_{0,0}\beta_{21}\Delta_1\Big[\Big((\Lambda_1+\Lambda_2)^2 + \Lambda_1\Lambda_2\Big)\Big(1 - L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\\ +\Big(\Lambda_1+\Lambda_2\Big)\Big(\Lambda_2 L_{R_1}\left[\Lambda_1+\Lambda_2\right] - 1\Big)\Big]\end{array}\right\}}{\Big\{\Lambda_1\Big(1 - L_{R_1}\left[\Lambda_1 + \Lambda_2\right]\Big)\Big\}}. \tag{4.7}$$

(viii) If the server is busy, the probability is given by

$$\Pi_{Busy} = \frac{\left\{\begin{array}{l}P_{0,0}\Bigg(\Lambda_1\beta_{11}\Big(\Lambda_1 + \Lambda_2 L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\\ +\beta_{21}\Big[\Big((\Lambda_1+\Lambda_2)^2 + \Lambda_1\Lambda_2\Big)\Big(1 - L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\\ +\Big(\Lambda_1+\Lambda_2\Big)\Big(\Lambda_2 L_{R_1}\left[\Lambda_1+\Lambda_2\right] - 1\Big)\Big]\Bigg)\end{array}\right\}}{\Big\{\Lambda_1\Big(1 - L_{R_1}\left[\Lambda_1 + \Lambda_2\right]\Big)\Big\}}. \tag{4.8}$$

(ix) When the server is blocked by a high-patience customer, we have

$$\Pi_{Blocked_{HP}} = \frac{\Big\{P_{0,0}\Lambda_1\beta_{11}\Big(\Lambda_1 + \Lambda_2 L_{R_1}[\Lambda_1 + \Lambda_2]\Big)\Big(1 + \Theta\Gamma_1\Big)\Big\}}{\Big\{\Lambda_1\Big(1 - L_{R_1}\left[\Lambda_1 + \Lambda_2\right]\Big)\Big\}}. \tag{4.9}$$

(x) We obtain the probability when the server is blocked by a low-patience customer as follows:

$$\Pi_{Blocked_{LP}} = \frac{\left\{\begin{array}{l} P_{0,0}\beta_{21}\Big[\Big((\Lambda_1+\Lambda_2)^2+\Lambda_1\Lambda_2\Big)\Big(1-L_{R_1}[\Lambda_1+\Lambda_2]\Big) \\ +\Big(\Lambda_1+\Lambda_2\Big)\Big(\Lambda_2 L_{R_1}[\Lambda_1+\Lambda_2]-1\Big)\Big]\Big[1+\Theta\Gamma_1+\overline{\sigma}\Theta\Delta_1\Big] \end{array}\right\}}{\Big\{\Lambda_1\Big(1-L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\Big\}}. \tag{4.10}$$

(xi) If the server is under repair, we obtain

$$\Pi_{Repair} = \frac{\left\{\begin{array}{l} \Theta\Gamma_1 P_{0,0}\Big[\Lambda_1\beta_{11}\Big[\Lambda_1+\Lambda_2 L_{A_1}\left[\Lambda_1+\Lambda_2\right]\Big] \\ +\beta_{21}\Big[\Big((\Lambda_1+\Lambda_2)^2+\Lambda_1\Lambda_2\Big)\Big(1-L_{A_1}[\Lambda_1+\Lambda_2]\Big) \\ +(\Lambda_1+\Lambda_2)\Big(\Lambda_2 L_{A_1}[\Lambda_1+\Lambda_2]-1\Big)\Big]\Big] \end{array}\right\}}{\Big\{\Lambda_1\Big(1-L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\Big\}}. \tag{4.11}$$

(xii) When the server is blocked, the probability is given by

$$\Pi_{Blocked} \ = \ \Pi_{Blocked_{HP}} + \Pi_{Blocked_{LP}}. \tag{4.12}$$

Proof. After a few algebraic manipulations and the application of l'Hôpital's rule, we obtain the desired result. In fact, we have

$$P_l = P_l(1), \quad l=0,1,2,4; \quad P_{3,m} = P_{3,m}(1), \quad m=0,1,2; \quad \Pi_{\text{Blocked}_{HP}} = P_1 + P_{3,0},$$

$$\Pi_{\text{Busy}} = P_1 + P_2, \quad \Pi_{\text{Blocked}_{LP}} = P_2 + P_{3,1} + P_{3,2} + P_4, \quad \Pi_{\text{Repair}} = P_{3,0} + P_{3,1} + P_{3,2},$$

$$\Pi_{\text{Blocked}} = P_1 + P_2 + P_{3,0} + P_{3,1} + P_{3,2} + P_4.$$

□

Corollary 4.2. *The generating functions of the number of customers in the orbit* $(\Pi_o(z))$ *and in the system* $(\Pi_s(z))$*, are given by*

$$\begin{aligned}\Pi_o(z) = P_{0,0} + P_0(z) + P_1(z)\Big(1+\Theta\overline{L}_M[\Lambda_1(1-z)]\Big) + P_2(z)\Big(\overline{L}_M[\Lambda_1(1-z)] \\ +\overline{\sigma}\Theta L_M[\Lambda_1(1-z)]\overline{L}_V[\Lambda_1(1-z)]+1\Big),\end{aligned} \tag{4.13}$$

$$\begin{aligned}\Pi_s(z) = P_{0,0} + P_0(z) + zP_1(z)\Big(1+\Theta\overline{L}_M[\Lambda_1(1-z)]\Big) + zP_2(z)\Big(\overline{L}_M[\Lambda_1(1-z)] \\ +\overline{\sigma}\Theta L_M[\Lambda_1(1-z)]\overline{L}_V[\Lambda_1(1-z)]+1\Big).\end{aligned} \tag{4.14}$$

Proof. The previous results are based on the following relationships

$$\Pi_o(z) = P_{0,0} + P_0(z) + P_1(z) + P_2(z) + P_{3,0}(z) + P_{3,1}(z) + P_{3,2}(z) + P_4(z),$$

$$\Pi_s(z) = P_{0,0} + P_0(z) + z\Bigg(P_1(z) + P_2(z) + P_{3,0}(z) + P_{3,1}(z) + P_{3,2}(z) + P_4(z)\Bigg).$$

□

Corollary 4.3. *The mean performance measures can be expressed as follows:*
(i) The mean number of high-patience customers in the orbit (L_o) is given by

$$
\begin{aligned}
L_o = P_0^{'}(1) + P_1^{'}(1)\Big(1+\Theta\Gamma_1\Big) + \Theta P_1\eta_1(1) + P_2^{'}(1)\Big(1+\Gamma_1+(1-\sigma)\Theta\Delta_1\Big)& \\
+ P_2\Big(\eta_1(1)+(1-\sigma)\Theta\eta_2(1)\Big).&
\end{aligned} \tag{4.15}
$$

(ii) The mean waiting time of high-patience customers in the orbit (W_o) is defined as

$$W_o = \frac{L_o}{\lambda_1}. \tag{4.16}$$

(iii) The mean number of customers in the system (L_s) is obtained as

$$
\begin{aligned}
L_s = P_0^{'}(1) + \Big(P_1 + P_1^{'}(1)\Big)\Big(1+\Theta\Gamma_1\Big) + \Theta P_1\eta_1(1) + \Big(P_2 + P_2^{'}(1)\Big)& \\
\times\Big(1+\Gamma_1+(1-\sigma)\Theta\Delta_1\Big) + P_2\Big(\eta_1(1)+(1-\sigma)\Theta\eta_2(1)\Big).&
\end{aligned} \tag{4.17}
$$

(iii) The mean waiting time of high-patience customers in the system (W_s) is given by

$$W_s = \frac{L_s}{\lambda_1}. \tag{4.18}$$

where, we define

$$P_1^{'}(1) = K_1^{'}(1)\beta_{11} + K_1(1)\bigg(\frac{(\Theta\Gamma_1+1)\Lambda_1\beta_{12}}{2}\bigg),$$

$$P_2^{'}(1) = K_2^{'}(1)\beta_{21} + K_2(1)\bigg(\frac{(\Theta\Gamma_1+1)\Lambda_1\beta_{22}}{2}\bigg),$$

$$
P_0^{'}(1) = \frac{\left\{\begin{aligned}
&P_{0,0}\Big(\beta_{21}\Gamma_1 L_{A_1}[\Lambda_1+\Lambda_2]\Theta\Lambda_1^2 + 2\beta_{21}\Gamma_1 L_{A_1}[\Lambda_1+\Lambda_2]\Theta\Lambda_1\Lambda_2 \\
&-\beta_{11}\Gamma_1 L_{A_1}[\Lambda_1+\Lambda_2]\Theta\Lambda_1\Lambda_2 - \beta_{21}\Gamma_1\Theta\Lambda_1^2 - 2\beta_{21}\Gamma_1\Theta\Lambda_1\Lambda_2 \\
&-\beta_{11}\Gamma_1\Theta\Lambda_1^2 + \beta_{21}L_{A_1}[\Lambda_1+\Lambda_2]\Lambda^2 + 2\beta_{21}L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_1\Lambda_2 \\
&-\beta_{11}L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_1\Lambda_2 - \beta_{21}\Lambda_1^2 - 2\beta_{21}\Lambda_1\Lambda_2 - \beta_{11}\Lambda_1^2 + L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_1 \\
&+L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_2\Big)
\end{aligned}\right\}}{\Big\{\Lambda_1\Big(1-L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\Big\}},
$$

$$\eta_1(1) = \frac{\Lambda_1\Gamma_2}{2}, \quad \eta_2(1) = \frac{\Lambda_1\Big(2\Delta_1\Gamma_1+\Delta_2\Big)}{2},$$

$$K_1(1) = P_{0,0}\left(\frac{L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_2+\Lambda_1}{\Big(1-L_{A_1}[\Lambda_1+\Lambda_2]\Big)}\right),$$

$$
K_2(1) = \frac{\left\{\begin{aligned}
&P_{0,0}\Big((\Lambda_1+\Lambda_2)\Big(1-L_{A_1}[\Lambda_1+\Lambda_2]\Big) \\
&+\Lambda_2\Big(\Lambda_1\overline{L}_{A_1}[\Lambda_1+\Lambda_2] + L_{A_1}[\Lambda_1+\Lambda_2]\Big) - 1\Big)
\end{aligned}\right\}}{\Big\{\Lambda_1\Big(1-L_{A_1}\left[\Lambda_1+\Lambda_2\right]\Big)\Big\}},
$$

$$K_1^{'}(1) = \frac{\left\{\begin{array}{l} P_{0,0}\Bigg((L_{A_1}[\Lambda_1+\Lambda_2])^2\beta_{21}\Gamma_1\Theta\Lambda_1^2\Lambda_2 + 2(L_{A_1}[\Lambda_1+\Lambda_2])^2\beta_{21}\Gamma_1\Theta\Lambda_1\Lambda_2^2 \\ -(L_{A_1}[\Lambda_1+\Lambda_2])^2\Gamma_1\beta_{11}\ \Theta\Lambda_1\Lambda_2 + L_{A_1}[\Lambda_1+\Lambda_2]\beta_{21}\Gamma_1\Theta\Lambda_1^3 \\ +L_{A_1}[\Lambda_1+\Lambda_2]\beta_{21}\Gamma_1\Theta\Lambda_1^2\Lambda_2 - 2L_{A_1}[\Lambda_1+\Lambda_2]\beta_{21}\Gamma_1\Theta\Lambda_1\Lambda_2^2 \\ -2L_{A_1}[\Lambda_1+\Lambda_2]\Gamma_1\beta_{11}\Theta\Lambda_1^2\Lambda_2 + (L_{A_1}[\Lambda_1+\Lambda_2])^2\beta_{21}\Lambda_1\Lambda_2^2 \\ +2(L_{A_1}[\Lambda_1+\Lambda_2])^2\beta_{21}\Lambda_1\Lambda_2^2 - (L_{A_1}[\Lambda_1+\Lambda_2])^2\beta_{11}\Lambda_1\Lambda_2^2 \\ -\beta_{21}\Gamma_1\Theta\Lambda_1^3 - 2\beta_{21}\Gamma_1\Theta\Lambda_1^2\Lambda_2 - \Gamma_1\beta_{11}\Theta\Lambda_1^3 + L_{A_1}[\Lambda_1+\Lambda_2]\beta_{21}\Lambda_1^3 \\ +L_{A_1}[\Lambda_1+\Lambda_2]\beta_{21}\Lambda_1^2\Lambda_2 - 2L_{A_1}[\Lambda_1+\Lambda_2]\beta_{21}\Lambda_1\Lambda_2^2 \\ -2L_{A_1}[\Lambda_1+\Lambda_2]\beta_{11}\Lambda_1^2\Lambda_2 + (L_{A_1}[\Lambda_1+\Lambda_2])^2\Lambda_1^2 \\ +2(L_{A_1}[\Lambda_1+\Lambda_2])^2\Lambda_1\Lambda_2 + (L_{A_1}[\Lambda_1+\Lambda_2])^2\Lambda_2^2 - \beta_{21}\Lambda_1^3 \\ -2\beta_{21}\Lambda_1^2\Lambda_2 - \beta_{11}\Lambda_1^3\Bigg) \end{array}\right\}}{\left\{\Lambda_1\Big(1 - L_{R_1}\left[\Lambda_1+\Lambda_2\right]\Big)\right\}},$$

$$K_2^{'}(1) = \frac{\left\{\begin{array}{l} P_{0,0}(\Lambda_1+\Lambda_2)\Bigg(\beta_{21}\Gamma_1\Theta\Lambda_1^3 + \Gamma_1\beta_{11}\Theta\Lambda_1^3 - 5L_{A_1}[\Lambda_1+\Lambda_2]\beta_{21}\Lambda_1^2\Lambda_2 \\ -L_{A_1}[\Lambda_1+\Lambda_2]\beta_{21}\Lambda_1\Lambda_2^2 + L_{A_1}[\Lambda_1+\Lambda_2]\beta_{11}\Lambda_1^2\Lambda_2 + L_{A_1}[\Lambda_1+\Lambda_2]\beta_{11}\Lambda_1\Lambda_2^2 \\ +\beta_{21}\Gamma_1L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_1\Lambda_2\Theta - \Gamma_1\beta_{11}L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_1\Lambda_2\Theta \\ +2\beta_{21}\Gamma_1(L_{A_1}[\Lambda_1+\Lambda_2])^2\Lambda_1^2\Lambda_2\Theta - \Gamma_1\beta_{11}(L_{A_1}[\Lambda_1+\Lambda_2])^2\Lambda_1^2\Lambda_2\Theta \\ -5L_{A_1}[\Lambda_1+\Lambda_2]\beta_{21}\Gamma_1\Theta\Lambda_1^2\Lambda_2 - L_{A_1}[\Lambda_1+\Lambda_2]\beta_{21}\Gamma_1\Theta\Lambda_1\Lambda_2^2 \\ +L_{A_1}[\Lambda_1+\Lambda_2]\Gamma_1\beta_{11}\Theta\Lambda_1^2\Lambda_2 + L_{A_1}[\Lambda_1+\Lambda_2]\Gamma_1\beta_{11}\Theta\Lambda_1\Lambda_2^2 \\ -\beta_{21}\Gamma_1\Lambda_1^2\Theta - \Gamma_1\beta_{11}\Lambda_1^2\Theta + 2(L_{A_1}[\Lambda_1+\Lambda_2])^2\Lambda_1^2\Lambda_2\beta_{21} \\ -\beta_{11}(L_{A_1}[\Lambda_1+\Lambda_2])^2\Lambda_1^2\Lambda_2 + \beta_{21}L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_1\Lambda_2 \\ -\beta_{11}L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_1\Lambda_2 - \beta_{21}\Lambda_1\Lambda_2 + (L_{A_1}[\Lambda_1+\Lambda_2])^2\Lambda_1\Lambda_2 \\ +\beta_{21}(L_{A_1}[\Lambda_1+\Lambda_2])^2\Lambda_1^3 - \beta_{11}\Lambda_1^3L_{A_1}[\Lambda_1+\Lambda_2]\beta_{21}L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_1^2 \\ -2L_{A_1}[\Lambda_1+\Lambda_2]\beta_{21}\Lambda_1^3 + 3\beta_{21}\Lambda_1^2\Lambda_2 + \beta_{21}\Lambda_1\Lambda_2^2 + \beta_{11}\Lambda_1^2\Lambda_2 \\ -2\Lambda_1\Lambda_2L_{A_1}[\Lambda_1+\Lambda_2] - \beta_{21}\Gamma_1\Lambda_1\Lambda_2\Theta - \Gamma_1\beta_{11}L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_1^3\Theta \\ +\beta_{21}\Gamma_1L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_1^2\Theta + \beta_{21}\Gamma_1(L_{A_1}[\Lambda_1+\Lambda_2])^2\Lambda_1^3\Theta \\ -2\Gamma_1\beta_{21}L_{A_1}[\Lambda_1+\Lambda_2]\Theta\Lambda_1^3 + 3\beta_{21}\Gamma_1\Theta\Lambda_1^2\Lambda_2 + \beta_{21}\Gamma_1\Theta\Lambda_1\Lambda_2^2 \\ +\Gamma_1\beta_{11}\Theta\Lambda_1^2\Lambda_2 - \beta_{21}\Lambda_1^2 - \beta_{11}\Lambda_1^2 + (L_{A_1}[\Lambda_1+\Lambda_2])^2\Lambda_1^2 \\ +\Lambda_1L_{A_1}[\Lambda_1+\Lambda_2] + L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_2 + \beta_{21}\Lambda_1^3 + \beta_{11}\Lambda_1^3 \\ -L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_1^2 - L_{A_1}[\Lambda_1+\Lambda_2]\Lambda_2^2\Bigg) \end{array}\right\}}{\left\{-\Lambda_1^2(\Big(1 - L_{A_1}[\Lambda_1+\Lambda_2])^2\right\}}.$$

Proof. By definition and through the application of l'Hôpital's rule, we obtain the following results

$$L_o = \Pi_o'(1), \quad \text{and } L_s = \Pi_s'(1).$$

The quantities W_s and W_o are calculated using Little's formula. □

5. Numerical results

Numerical analysis is employed as a means to validate the precision and reliability of retrial queueing models. This analysis involves evaluating crucial performance metrics such as the mean number of customers in the system (L_s) and the probability that the system is empty ($P_{0,0}$). The objective is to provide valuable information that informs decision making within the systems under examination. Furthermore, MATLAB can be used to encode essential elements, streamlining the validation of these models. We assume that all introduced distribution functions follow exponential distributions and the values of the system parameters have been chosen arbitrarily to ensure the stability of the system.

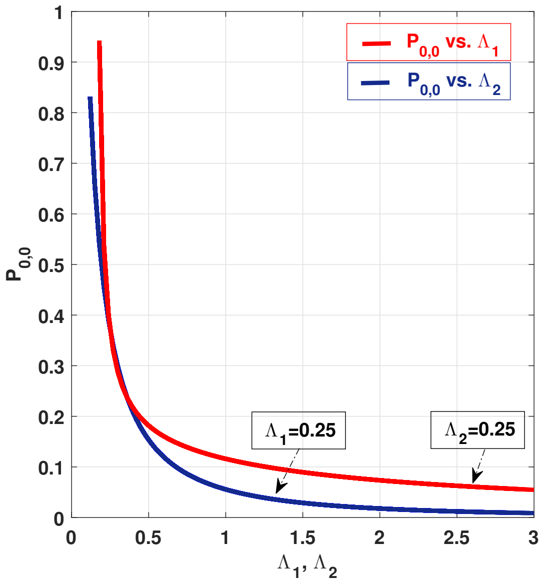


Figure 3. $P_{0,0}$ vs. Λ_1 and Λ_2, when $\sigma = 0.45$ and $\alpha_1 = \beta_1 = \beta_2 = \Theta = \Gamma = \Delta = 1$

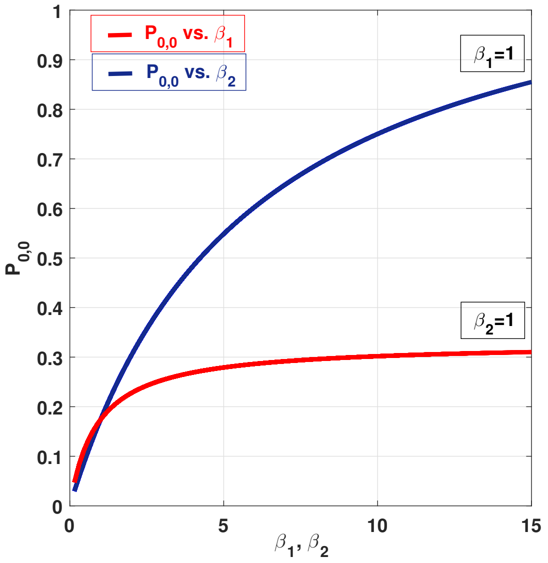


Figure 4. $P_{0,0}$ vs. β_1 and β_2, when $\sigma = 0.2$, $\Lambda_1 = \Lambda_2 = 0.35$ and $\alpha_1 = \Theta = \Gamma = \Delta = 1$

Figures 3–5 illustrate how variations in system parameters, such as different arrival rates, service rates, failure and repair rates, and reserved rates, impact the probability of the system being empty ($P_{0,0}$). Upon analyzing the figures, we observe that as the arrival rates (Λ_1 and Λ_2) and the failure rate (Θ) increase, the probability $P_{0,0}$ decreases. On the other hand, when the service rates (β_1 and β_2), the repair rate (Γ) and the reserved rate (Δ) increase, the probability $P_{0,0}$ increases. These observations suggest that higher arrival rates and failure rates decrease the probability $P_{0,0}$, while higher service rates, repair rates, and reserved rates increase it.

Figures 6 and 7 provide valuable information on the relationship between the performance metric L_s and the parameters Γ, Θ, σ and Δ. According to Figure 6, an increase in the reservation rate (Δ) and the parameter (σ) results in a noticeable upward trend in the characteristic L_s. This suggests that higher values of Δ and σ lead to more time allocated to customers with low patience, both in the waiting state and in the service orbit. In addition, customers with greater patience in the orbit also benefit from increased time allocation. Consequently, the total number of customers in the system increases. Figure 7 shows that a higher repair rate (Γ) corresponds to a decrease in L_s. This indicates that when the repair rate is higher, the system can quickly address failures, resulting in fewer customers being present within the system. Furthermore, the failure rate (Θ) significantly impacts L_s. A higher failure rate leads to an increase in L_s, indicating that more customers are present in the system due to the increased frequency of failures.

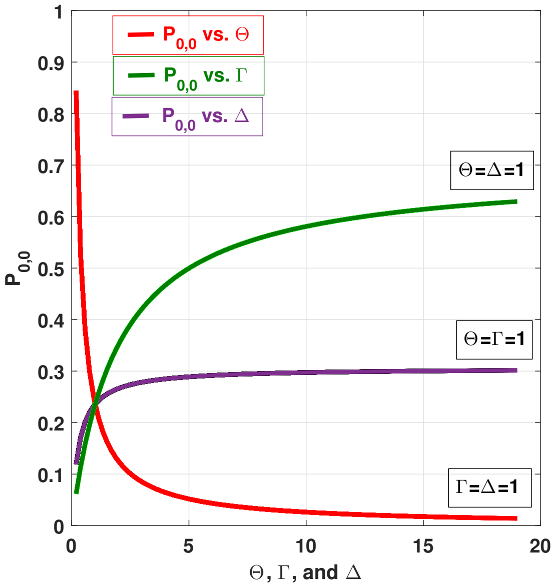
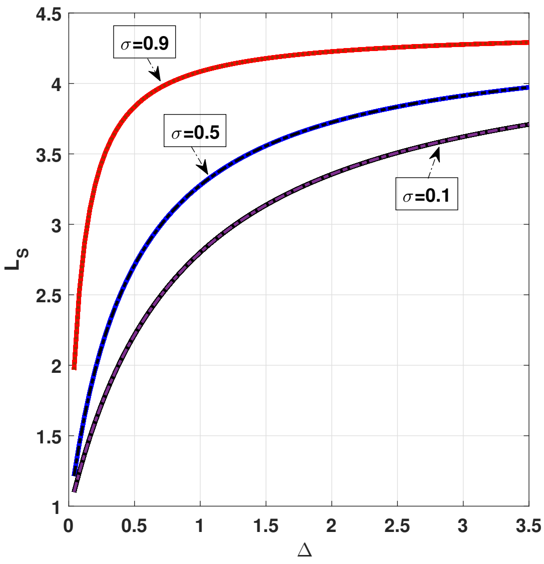


Figure 5. $P_{0,0}$ vs. Θ, Γ and Δ, when $\sigma = 0.35$, $\Lambda_1 = \Lambda_2 = 0.3$ and $\alpha_1 = \beta_1 = \beta_2 = 1$.



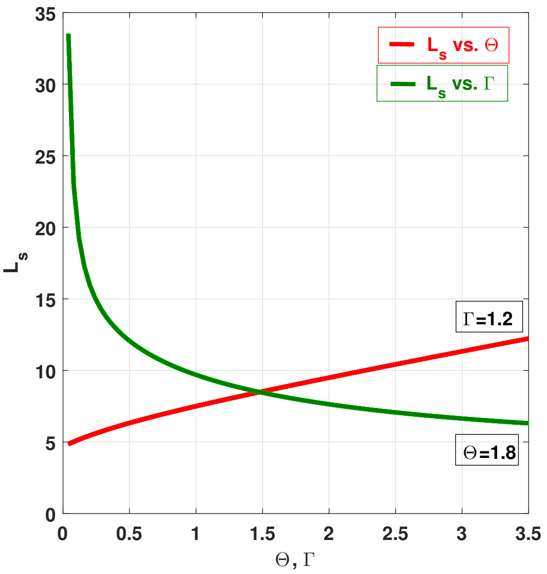


Figure 6. L_s vs. Δ and σ, when $\Lambda_1 = 0.01$ and $\alpha_1 = \Lambda_2 = \beta_1 = \beta_2 = \Theta = \Gamma = 1$

Figure 7. L_s vs. Θ and Γ, when $\Lambda_1 = \Lambda_2 = 0.3$, $\alpha_1 = \Delta = 1$, $\beta_1 = \beta_2 = 0.2$ and $\sigma = 0.3$

6. Economic analysis

This section presents a cost model for our retrial queueing system. The total cost of the system, denoted as C_T, is determined using the following formula

$$C_T = C_I(P_{0,0} + P_0) + C_W L_o + C_B \Pi_{Blocked}, \tag{6.1}$$

where C_I is the cost of server preparation work when the system is empty; C_B is the cost per unit time when the server is blocked; C_W is the cost per unit of time that a customer spends waiting in the orbit. These notations enable us to measure different cost elements in the system, including preparation work costs, waiting costs, service costs, repair costs, and reservation costs. By incorporating these elements, we can assess the overall economic impact of the queueing system and make informed decisions about resource allocation, service level agreements, and system optimization. The cost coefficients were chosen to maintain the stability of the system while reflecting the economic impact of different states of the system. Specifically, we set $C_I = 150$, $C_B = 500$, and $C_W = 20$ to balance operational costs and service efficiency. The value of C_B is relatively high because server blockages generate significant inefficiencies and disrupt service continuity. In contrast, the waiting cost C_W is lower since waiting in orbit does not directly affect system performance but still contributes to customer dissatisfaction. The setup cost C_I was set to a moderate value to balance operational readiness without excessive overhead. The remaining parameters Λ_1, Λ_2, α_1, β_1, β_2, Θ, Γ, Δ and σ were chosen in a way that ensures the stability of the model. These parameter values are selected on the basis of the specific characteristics and requirements of our retrial queueing system. Furthermore, Figures 8 and 9 provide insight into the impact of key system parameters on total cost. These figures illustrate how variations in the breakdown rate (Θ), the repair rate (Γ) and the reservation parameter (σ) influence C_T. They indirectly reflect how cost-related factors affect overall expenses.

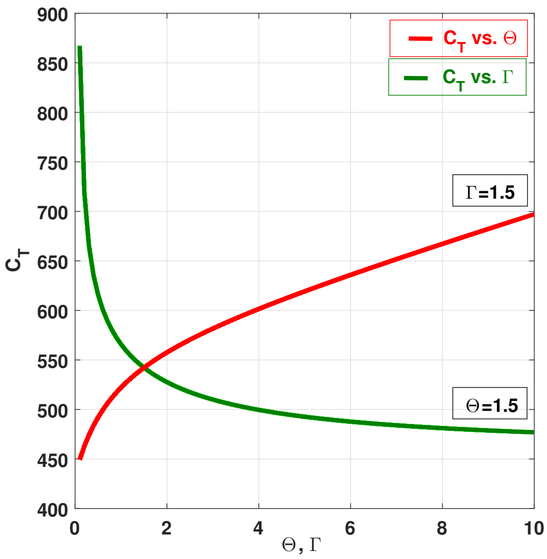


Figure 8. C_T vs. Θ and Γ, when $\sigma = 0.5$, $\Lambda_1 = \Lambda_2 = \beta_1 = \beta_2 = 0.3$ and $\alpha_1 = \Delta = 1$

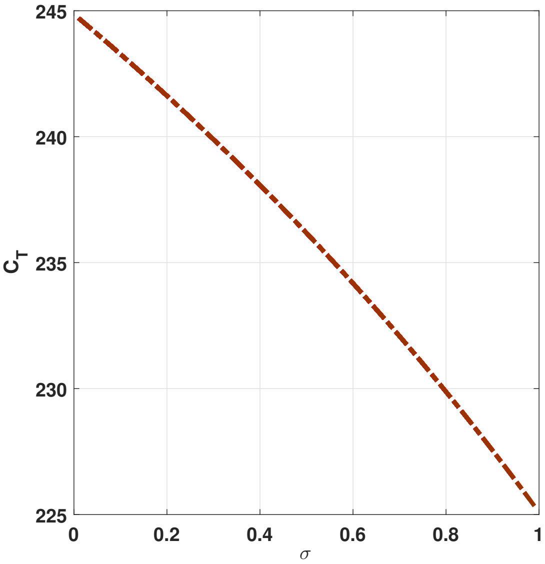


Figure 9. C_T vs. σ, when $\Lambda_1 = \Lambda_2 = 0.8$, $\beta_1 = \beta_2 = 0.4$, $\Theta = 2$, $\Gamma = 0.5$ and $\alpha_1 = \Delta = 1$

Figure 8 shows the relationship between the total cost (C_T) and the repair rate (Γ). As the parameter Γ increases, the total cost decreases. This inverse relationship can be explained by the fact that a higher repair rate results in a reduced number of customers in the system. Consequently, overall cost decreases, indicating improved system efficiency and lower maintenance expenses. Furthermore, the figure examines the impact of the breakdown rate (Θ) on C_T. It demonstrates that an increase in the parameter (Θ) results in a higher total cost. This suggests that a higher breakdown rate leads to increased costs as a result of the resulting increase in the number of customers in the system. Figure 9 shows an inverse relationship between the total cost (C_T) and the reservation parameter

(σ). As σ increases, C_T decreases. This effect is due to σ influencing the reserved time: When low-patience customers choose to remain in the service area after a server failure, reserved time is reduced, leading to lower costs.

The retrial queueing model also presents following managerial insights:

(i) Understanding cost dynamics: The model allows managers to analyze various cost elements such as preparation work costs, waiting costs, service costs, repair costs, and reservation costs. By quantifying these costs, managers can gain a comprehensive understanding of the economic impact of the queueing system.

(ii) Resource allocation: By assessing different cost components, managers can make informed decisions regarding resource allocation. For example, they can allocate resources to minimize waiting costs or optimize repair processes to reduce repair costs.

(iii) Service level agreements: Managers can use the insights from the model to establish effective service-level agreements. Understanding the relationship between system parameters and costs can help set realistic service-level targets while balancing cost considerations.

(iv) System optimization: The model provides insights into how the parameters of the system affect overall costs. Managers can use this information to optimize system performance by adjusting parameters such as repair rates or reservation policies to minimize total costs while maintaining service quality.

(v) Predictive maintenance: By analyzing the impact of breakdown rates on total costs, managers can develop predictive maintenance strategies. Identifying patterns in breakdown rates can help anticipate failures and schedule preventive maintenance activities, thereby reducing downtime and repair costs.

(vi) Customer service strategy: Understanding the behavior of different types of customers, such as highly patient and low-patience customers, enables managers to tailor customer service strategies. For example, they can implement strategies to reduce wait times for high-patience customers or minimize customer attrition for low-patience customers.

In general, the queueing model provides valuable information that can inform decision-making processes, leading to improved system efficiency, cost savings, and improved customer satisfaction.

7. Conclusion and future scope

In this paper, we presented a single-server retrial queue with customers exhibiting different levels of patience (high-patience and low-patience). This model is relevant for various applications, including healthcare systems, web traffic management, and telecommunication networks. It integrates several realistic features such as active server breakdowns, corrective repairs, retrial times, reserved times, customer balking, and distinct service rates for both customer types. Using the supplementary variable technique, we derived probability-generating functions for different system states, including idle, busy (with high-patience and low-patience customers), under repair and in reservation. We also established the stability condition and obtained closed-form expressions for steady-state probabilities and key performance measures. Furthermore, we conducted an economic analysis to assess the impact of system parameters on performance metrics and total expected cost.

7.1. Comparative contribution and study limitations

Our study advances the understanding of retrial queueing systems by explicitly incorporating heterogeneous customer patience, an aspect often neglected in classical models. Unlike previous research that assumes uniform customer behavior, our approach provides a more refined analysis of impatience and its effects on system performance. Furthermore, while most studies focus on queueing dynamics under stable conditions, our model explicitly accounts for server breakdowns and corrective repairs, enhancing its applicability to real-world service systems.

A key contribution of this study is its economic optimization framework, which evaluates trade-offs between performance measures and operational costs. This aspect is rarely addressed in retrial queueing models, yet it provides valuable insights into cost-effective system management. Most existing research primarily examines performance indicators such as queue length, waiting time and system stability, often overlooking the financial consequences of service interruptions and retrial dynamics. By integrating economic considerations, our study bridges this gap and emphasizes the importance of balancing efficiency with cost constraints. This perspective is particularly relevant for service providers who must optimize system performance while ensuring financial sustainability. However, our model has certain limitations that suggest directions for future research. One key limitation is the assumption of single-customer arrivals, which simplifies the analysis but does not fully capture scenarios where customers arrive in groups. Extending the model to batch arrivals, where both high-patience and low-patience customers may arrive simultaneously, would increase its applicability to large-scale telecommunication and manufacturing systems.

Our model does not incorporate priority-based queuing disciplines, which are critical in systems where certain customers require immediate service, such as healthcare and emergency call centers. Future research could investigate how priority mechanisms influence system stability, performance, and cost optimization. Another important extension would be to consider a general arrival process beyond the traditional Poisson assumption, as non-exponential inter-arrival times may better reflect dynamic environments such as web traffic and smart grid networks. Additionally, integrating service control policies, such as the N-policy (where the server activates or deactivates based on system occupancy levels), could significantly improve resource utilization and provide deeper insight into balancing performance and cost efficiency in unreliable queueing systems. Finally, our model does not explicitly account for complex customer behaviors, such as strategic decision-making, patience thresholds, or adaptive retrial strategies. Investigating how customers adjust their retrial attempts based on queue length, service delays, or past experiences would enhance the model's realism and practical applicability.

7.2. Future research directions

For future research, an interesting extension would be to explore the same model with batch arrivals, where some groups consist of high-patience customers, while others include low-patience customers. This modification would broaden the model's applicability to large-scale service systems, such as manufacturing and telecommunications, where arrivals typically occur in clusters rather than individually. Additionally, incorporating priority-based queuing, more sophisticated customer behavior models, service control policies, and generalized arrival processes would further strengthen the model's robustness and practical relevance. These enhancements would contribute to the theoretical advancement of retrial queueing systems while offering practical solutions to optimize real-world service operations.

Acknowledgements

We sincerely appreciate the insightful and constructive feedback from the editor and reviewers. Their valuable comments have significantly enhanced the clarity, coherence, and overall quality of our paper.

Author contributions. All co-authors have contributed equally to every aspect of the preparation of this submission.

Conflict of interest statement. There is no conflict of interest.

Funding. No funding.

Data availability. Not applicable.

References

[1] A. Aissani, F. Lounis, D. Hamadouche and S. Taleb, *Analysis of customers' impatience in a repairable retrial queue under postponed preventive actions*, American Journal of Mathematical and Management Sciences, **38** (2), 125-150, 2019, https://doi.org/10.1080/01966324.2018.1486763.

[2] L.M. Alem, M. Boualem and D. Aïssani, *Bounds of the stationary distribution in $M/G/1$ retrial queue with two-way communication and n types of outgoing calls*, Yugoslav Journal of Operations Research, **29** (3), 375-39, 2019, https://doi.org/10.2298/YJOR180715012A.

[3] L.M. Alem, M. Boualem and D. Aïssani, *Stochastic comparison bounds for an $M_1, M_2/G_1, G_2/1$ retrial queue with two way communication*, Hacettepe Journal of Mathematics and Statistics, **48** (4), 1185-1200, 2019, https://dergipark.org.tr/en/pub/hujms/issue/47862/604504.

[4] J. Artalejo and A. Gomez-Corral, *Retrial queueing systems: A Computational Approach*, Springer-Verlag, Berlin, 2008, https://api.semanticscholar.org/CorpusID:60225921.

[5] G. Ayyappan and P. Thamizhselvi, *Transient analysis of $M^{[X_1]}, M^{[X_2]}/G_1, G_2/1$ retrial queueing system with priority services, working vacations and vacation interruption, emergency vacation, negative arrival and delayed repair*, International Journal of Applied and Computational Mathematics, **4** (2), 2018, https://doi.org/10.1007/s40819-018-0509-7.

[6] G. Ayyappan and J. Udayageetha, *Transient analysis of $M^{[X_1]}, M^{[X_2]}/G_1, G_2/1$ retrial queueing system with priority services, working breakdown, start up/close down time, Bernoulli vacation, reneging and balking*, Pakistan Journal of Statistics and Operation Research, **16** (1), 203-216, 2020, https://doi.org/10.18187/pjsor.v16i1.2181.

[7] M. Boualem, A. Bareche and M. Cherfaoui, *Approximate controllability of stochastic bounds of stationary distribution of an $M/G/1$ queue with repeated attempts and two phase service*, International Journal of Management Science and Engineering Management, **14** (2), 79-85, 2018, https://api.semanticscholar.org/CorpusID:125814082.

[8] A.A. Bouchentouf, M. Boualem, L. Yahiaoui and H. Ahmad, *A multi-station unreliable machine model with working vacation policy and customer impatience*, Quality Technology & Quantitative Management, **19** (6), 766-796, 2022, https://doi.org/10.1080/16843703.2022.2054088.

[9] A.A. Bouchentouf, M. Cherfaoui and M. Boualem, *Performance and economic analysis of a single server feedback queueing model with vacation and impatient customers*, Opsearch, **56** (1), 300-323, 2019, https://doi.org/10.1007/s12597-019-00357-4.

[10] A.A. Bouchentouf, M. Cherfaoui and M. Boualem, *Analysis and performance evaluation of Markovian feedback multi-server queueing model with vacation and impatience*, American Journal of Mathematical and Management Sciences, **40**, 261-282, 2021, https://doi.org/10.1080/01966324.2020.1842271.

[11] M. Cherfaoui, A.A. Bouchentouf and M. Boualem, *Modeling and simulation of Bernoulli feedback queue with general customers impatience under variant vacation policy*, International Journal of Operational Research, **46**, 451-480, 2023, https://doi.org/10.1504/ijor.2023.129959.

[12] G. Choudhury and M. Deka, *A batch arrival unreliable server delaying repair queue with two phases of service and Bernoulli vacation under multiple vacation policy*, Quality Technology & Quantitative Management, **15** (2), 157-186, 2018, https://doi.org/10.1080/16843703.2016.1208934.

[13] A. Dehimi, M. Boualem, A.A. Bouchentouf, S. Ziani and L. Berdjoudj, *Analytical and computational aspects of a multi-server queue with impatience under differentiated working Vacations policy*, Reliability: Theory & Applications **19**, 3 (79), 393407, 2024, https://doi.org/10.24412/1932-2321-2024-379-393-407.

[14] S. Dhar, L.B. Mahanta and K.K. Das, *Estimation of the waiting time of patients in a hospital with simple Markovian model using order statistics*, Hacettepe Journal of Mathematics and Statistics, **48** (1), 274-289, 2019, https://doi.org/10.15672/HJMS.2018.607.

[15] A. Dudin, O. Dudina, S. Dudin and K. Samouylov, *Analysis of single-server multi-class queue with unreliable service, batch correlated arrivals, customers impatience, and dynamical change of priorities*, Mathematics, **9** (11), 1257, 2021, https://doi.org/10.3390/math9111257.

[16] D. Fiems, *Retrial queues with constant retrial times*, Queueing Systems, **103** (3/4), 347-365, 2023, https://doi.org/10.1007/s11134-022-09866-4.

[17] S. Gao, *A preemptive priority retrial queue with two classes of customers and general retrial times*, Operational Research, **15** (2), 233-251, 2015, https://doi.org/10.1007/s12351-015-0175-z.

[18] H. Gao, J. Zhang and X. Wang, *Analysis of a retrial queue with two-type breakdowns and delayed repairs*, IEEE Access, **8**, 172428-172442, 2020, https://doi.org/10.1109/ACCESS.2020.3023191.

[19] H. Hablal, N. Touche, L. Alem, A.A. Bouchentouf and M. Boualem, *Lower and upper stochastic bounds for the joint stationary distribution of a non-preemptive priority retrial queueing system*, Hacettepe Journal of Mathematics and Statistics, **52** (5), 1438-1460, 2023, https://doi.org/10.15672/hujms.1183966.

[20] D. Hamadouche, A. Aissani, F. Lounis. *On the asymptotic behaviour of an unreliable $M/G/1$ retrial queue with impatience*, Authorea, 2024, https://doi.org/10.22541/au.170668021.12989057/v1.

[21] K. C. Hariom, Sharma, K. Singh and D. Singh, *Analysis of an inventory model for time-dependent linear demand rate three levels of production with shortage*, International Journal of Professional Business Review, **9** (4), 2024, https://doi.org/10.26668/businessreview/2024.v9i4.4579.

[22] B. Jagannathan and N. Sivasubramaniam, *Bulk arrival queue with unreliable server, balking and modified Bernoulli vacation*, Hacettepe Journal of Mathematics and Statistics, **53** (1), 289-304, 2024, https://doi.org/10.15672/hujms.1181711.

[23] M. Jain and A. Bhagat, *$M^X/G/1$ retrial vacation queue for multi-optional services, phase repair and reneging*, Quality Technology & Quantitative Management, **13**, 263-288, 2016, https://doi.org/10.1080/16843703.2016.1189025.

[24] B. Kim and J. Kim, *Waiting time distributions in an $M/G/1$ retrial queue with two classes of customers*, Annals of Operations Research, **252** (1), 121-134, 2017, https://doi.org/10.1007/s10479-015-1979-1.

[25] V. Klimenok, A. Dudin, O. Dudina and I. Kochetkova, *Queueing system with two types of customers and dynamic change of a priority*, Mathematics, **8** (5), 824, 2020, https://doi.org/10.3390/math8050824.

[26] B. Krishna Kumar, R. Rukmani, A. Thanikachalam and V. Kanakasabapathi, *Performance analysis of retrial queue with server subject to two types of breakdowns and repairs*, Operational Research, **18**, 521-559, 2018, https://doi.org/10.1007/s12351-016-0275-4.

[27] A. Kumar, M. Boualem and A.A. Bouchentouf, *Optimal analysis of machine interference problem with standby, random switching failure, vacation interruption, and synchronized reneging*, In Applications of Advanced Optimization Techniques in Industrial Engineering, 155-168, 2022, https://doi.org/10.1201/9781003089636-10.

[28] S.K. Lee, S. Dudin, O. Dudina, C.S. Kim and A. Klimenok, *A priority queue with many customer types, correlated arrivals, and changing priorities*, Mathematics, **8**, 1292, 2020, https://doi.org/10.3390/math8081292.

[29] T. Li and L. Zhang, *An $M/G/1$ retrial G-queue with general retrial times and working breakdowns*, Mathematical and Computational Applications, **22**, 15, 2017, https://doi.org/10.3390/mca22010015.

[30] S.P. Madheswari, B.K. Kumar and P. Suganthi, *Analysis of $M/G/1$ retrial queues with second optional service and customer balking under two types of Bernoulli vacation schedule*, RAIRO-Operations Research, **53** (2), 415-443, 2019, https://doi.org/10.1051/ro/2017029.

[31] S. Mahanta, N. Kumar and G. Choudhury, *Study of a two types of general heterogeneous service queueing system in a single server with optional repeated service and feedback queue*, Hacettepe Journal of Mathematics and Statistics, **53**, 3, 851-878, 2024, https://doi.org/10.15672/hujms.1312795.

[32] A. Melikov, S. Aliyeva, J. Sztrik, *Retrial queues with unreliable servers and delayed feedback*, Mathematics, **9** (19), 2415, 2021, https://doi.org/10.3390/math9192415.

[33] S. Muthusamy, N. Devadoss and S.I. Ammar, *Reliability and optimization measures of retrial queue with different classes of customers under a working vacation schedule*, Discrete Dynamics in Nature and Society, 2022, https://doi.org/10.1155/2022/6806104.

[34] D. Singh, *Production inventory model of deteriorating items with holding cost, stock, and selling price with backlog*, International Journal of Mathematics in Operational Research, **14** (2), 290-305, 2019, https://doi.org/10.1504/IJMOR.2019.097760.

[35] D. Singh, M.G. Alharbi, A. Jayswal and A. A. Shaikh, *Analysis of inventory model for quadratic demand with three levels of production*, Intelligent Automation & Soft Computing, **32** (1), 167-182, 2022, https://doi.org/10.32604/iasc.2022.021815.

[36] D. Singh, A. Jayswal, M. G. Alharbi and A. A. Shaikh, *An investigation of a supply chain model for coordination of finished products and raw materials in a production system under different situations*, Sustainability, **13** (22), 12601, 2021, https://doi.org/10.3390/su132212601.

[37] J. Sztrik, A. Tóth, E. Danilyuk, S. Moiseeva, *Analysis of retrial queueing system $M/G/1$ with impatient customers, collisions and unreliable server using simulation*, **1391**, Communications in Computer and Information Science, 291-303, 2021, https://doi.org/10.1007/978-3-030-72247-0_22.

[38] S. Taleb and A. Aissani, *Preventive maintenance in an unreliable $M/G/1$ retrial queue with persistent and impatient customers*, Annals of Operations Research, **247** (1), 291-317, 2016, https://doi.org/10.1007/s10479-016-2217-1.

[39] R. Tian and Y. Zhang, *Analysis of $M/M/1$ queueing systems with negative customers and unreliable repairers*, Communications in Statistics-Theory and Methods, **53** (21), 74917504, 2023, https://doi.org/10.1080/03610926.2023.2265000.

[40] A. Toth and J. Sztrik, *Simulation of two-way communication retrial queueing systems with unreliable server and impatient customers in the orbit*, Stochastic Modeling

and Applied Research of Technology, **3**, 45-50, 2023, https://doi.org/10.57753/SMARTY.2023.39.42.006.

[41] X. Wu, P. Brill, M. Hlynka and J. Wang, *An $M/G/1$ retrial queue with balking and retrials during service*, International Journal of Operational Research, **1** (1/2), 30-51, 2005, https://doi.org/10.1504/IJOR.2005.007432.

[42] M. Yin, M. Yan, Y. Guo and M. Liu, *Analysis of a pre-emptive two-priority queueing system with impatient customers and heterogeneous servers*, Mathematics, **11**, 3878, 2023, https://doi.org/10.3390/math11183878.

[43] Y. Zhang and J. Wang, *Managing retrial queueing systems with boundedly rational customers*, Journal of the Operational Research Society, **74** (3), 748-761, 2022, https://doi.org/10.1080/01605682.2022.2053305.

[44] D. Zirem, M. Boualem, K. Adel-Aissanou and D. Aïssani, *Analysis of a single server batch arrival unreliable queue with balking and general retrial time*, Quality Technology & Quantitative Management, **16**, 672-695, 2019, https://doi.org/10.1080/16843703.2018.1510359.