AIMS Mathematics, 10(8): 19460–19494.
DOI: 10.3934/math.2025869
Received: 23 December 2024
Revised: 20 July 2025
Accepted: 14 August 2025
Published: 26 August 2025

https://www.aimspress.com/journal/Math

Research article

Online reservation queueing–inventory system with two distinct services and client types

N. Suresh Kumar[1], N. Anbazhagan[1], S. Amutha[2], Gyanendra Prasad Joshi[3,*] and Woong Cho[3,*]

[1] Department of Mathematics, Alagappa University, karaikudi-630003, India
[2] Ramanujan Centre for Higher Mathematics, Alagappa University, karaikudi-630003, India
[3] Department of Electronic and AI System Engineering, Kangwon National University, Samcheok 25913, Republic of Korea

* **Correspondence:** Email: joshi@kangwon.ac.kr; wcho@kangwon.ac.kr.

Abstract: This study examines a queueing–inventory system offering premium and non-premium services for a single commodity, where both services have individual waiting areas with finite capacities. To enhance premium service, the system includes an online reservation facility with a limited level. An online client initially pre-books a spot in the premium waiting area (PWA) and, after a random duration, either joins the PWA or cancels the reservation. An offline client directly visits the system and chooses any one of the services based on their needs. The arrivals of both client types follow independent Markovian arrival processes (MAPs). Further, the waiting times of a client in the premium and non-premium waiting areas are derived using the Laplace–Stieltjes transform. The steady-state probabilities are computed, and the system's essential performance metrics are calculated. Subsequently, the optimal total expected cost is determined through numerical analysis and visually represented in a graph.

Keywords: online client; offline client; online reservation; Markovian arrival process; premium service; non-premium service
Mathematics Subject Classification: 60K25, 90B05, 91B70

1. Introduction

In our daily lives, we often encounter two distinct types of services: premium and non-premium. These labels are not just about cost; they represent different levels of quality, features, and overall user experience. Consider the following examples. **Non-premium services:** Imagine being at a local coffee shop, ordering a regular cup of coffee. This straightforward service is an example of a non-premium experience, offering the essential product without any frills. In this scenario, a caffeine fix is

obtained quickly and at an affordable price. Non-premium services are akin to the economy class of experiences—functional and cost-effective. **Premium services:** Imagine visiting a high-end specialty cafe. Opting for a premium coffee experience means more than just receiving a beverage—it is an investment in an elevated experience. The coffee beans are carefully selected, the brewing process is meticulous, and the ambiance is crafted for comfort and luxury. Premium services often come with added benefits, meticulous attention to detail, and a higher price tag. Across various industries, from streaming platforms to airlines, the choice between premium and non-premium services depends on individual preferences, needs, and budgets. Some individuals prefer the simplicity and affordability of non-premium services, while others seek the added value and quality associated with premium services. We recommend that readers refer to the works of Anbazhagan et al. [1], Krishnamoorthy et al. [2], and Sivakumar et al. [3] for a more in-depth exploration of inventory systems integrated with service facilities.

Many researchers have analyzed queueing–inventory systems (QISs) that incorporate various types of services. Jeganathan et al. [4] analyzed a queueing–inventory system (QIS) where demands originate from a finite homogeneous population, with a single server offering two types of services. Upon arrival, clients can choose either type of service with predefined probabilities. Mathew et al. [5] examined an inventory system with positive service time, featuring two types of service channels: Channel 1 operates as a single-server facility, while Channel 2 functions as a bulk service facility. Jeganathan et al. [6] explored a QIS with a heterogeneous service rate, where the service rate depends on the current queue length. Kocer and Ozkar [7] examined a QIS involving server breakdowns, which could be either minor or major. After recovery, priority is given to higher-class customers for service. Jeganathan et al. [8] considered a stochastic model combining interconnected queueing and queueing–inventory systems, featuring dual service stations for non-commodity and commodity services. Bhuvaneshwari et al. [9] investigated a QIS featuring multiple optional services, where arriving clients can initially request a regular service and have the option to additionally request several services, such as type-i, where $i = 1, 2, ..., N$. Jeganathan et al. [10] analyzed an inventory system with two service channels in a multi-server setup, where channel 1 contains n identical servers and channel 2 contains m identical servers. This system features optional service connections that interconnect channel 1 and channel 2. In our proposed model, we consider two types of services: premium and non-premium, with each service provided by a dedicated server having a heterogeneous service rate, where the premium service rate is slower compared to the non-premium service rate.

Online reservation offers several advantages in a queueing–inventory system. First, it provides convenience for clients to book services anytime and anywhere, eliminating the need to visit a physical location. Second, it allows for easy comparison of available options and prices, enabling clients to make more informed decisions. Additionally, online reservations often include features such as real-time availability updates, confirmation notifications, and the flexibility to modify or cancel reservations—each of which enhances client satisfaction. The reference [11–13] gave a brief idea about online reservation and cancellation. Shajin et al. [14] analyzed a single-server QIS with the ability to reserve services in advance for the next K time frames (days). Baron et al. [15] considered clients who plan to book a service online and are provided with information about their position in the queue at the time of booking. This enables them to decide whether to enter the queue based on their travel time and anticipated waiting time. In our proposed model, an online reservation facility is considered to pre-book the premium waiting area in order to avoid the loss of getting a premium

service.

This study examines an online reservation facility with two types of clients: online and offline. Client arrivals follow independent Markovian arrival processes. In many real-world scenarios, demand does not strictly follow a renewal process. Consequently, the Markovian arrival process (MAP) is particularly well-suited for modeling both renewal and non-renewal cases. In this system, the MAP allows us to capture realistic arrival patterns, accounting for dependencies and correlations between arrivals. Additionally, the MAP is applicable to both discrete and continuous cases, though we focus on the continuous-time scenario. For a more comprehensive understanding of the MAP and its properties, refer to Neuts [16] and Chakravarthy [17].

Wang et al. [18] examined Markov models with uniform service rates for two types of clients, each with different service priorities. The arrivals for both client classes were modeled using independent MAPs. Krishnamoorthy et al. [19] analyzed a QIS with a single server, where arrivals followed a batch MAP and services were delivered in batches according to a batch Markovian service process. Manuel et al. [20] considered two client types: ordinary and negative, both arriving according to an MAP. An ordinary client enters the queue, while a negative client removes an ordinary client from the queue instead of joining the queue. Hanukov [21] explored a QIS involving skeptical and trusting clients. AlMaqbali et al. [22] studied a QIS with multi-class customers and multi-server batch service facilities. Melikov et al. [23] analyzed a QIS involving negative clients and warehouse catastrophes at the service facility.

Ozkar et al. [24] analyzed a QIS with two client types: priority and non-priority clients. Priority clients purchase commodity I, while non-priority clients purchase commodity II. Wang [25] investigated a multi-server QIS with two classes of demand. The impatience of low-priority clients was modeled using Bernoulli reneging probabilities. Jeganathan et al. [26] examined a QIS that provides two priority levels for clients: first priority and second priority. Vinitha et al. [27] analyzed a QIS with two distinct client classes: impulse clients, who enter the system without a predetermined purchase plan, and ordinary clients, who arrive with a predefined plan to make a purchase. Harikrishnan et al. [28] discussed a finite-source, stock-dependent QIS with multiple servers and a retrial facility, where primary customers are served by multi-servers. queueing–inventory models involving various customer types have been investigated by Otten et al. [29], Shajin et al. [30], Dissa et al. [31], and Ozkar et al. [32].

This study reflects the author's experience at a restaurant, particularly during a recent visit. The restaurant offers two kinds of service: premium and non-premium. The premium services aim to provide a top-notch dining experience, going beyond the basics, and include special attention, a fancy atmosphere, and the use of technology, making clients lean toward choosing premium over non-premium. Non-premium services are standard offerings without exclusive features, including regular seating, no special reservations, and standard billing. Additionally, the premium service allows online reservations for clients to book in advance. If a reserved client does not show up, the restaurant supervisor cancels their reservation after a random time. The restaurant allows clients to select the service that best suits their needs. These real-life experiences motivated the author to develop a mathematical model that incorporates two types of service with online reservation facilities in a QIS.

Beyond the restaurant context, similar queueing–inventory structures are found in hospitals (e.g., scheduled vs. walk-in consultations), airport lounges (e.g., business-class vs. economy services), and high-demand service centers (e.g., express vs. standard repair services). In each of these settings, the

ability to pre-book premium services while managing inventory and balancing walk-in clients aligns well with the proposed model's structure and analysis.

This study makes the following key contributions:

- This study investigates two types of clients, two types of services for a single commodity, and an online reservation facility.
- The arrivals of both client types follow independent Markovian arrival processes (MAPs), and the service duration for both services follows independent exponential distributions.
- The steady-state probabilities are derived using the Gaver algorithm, and waiting time distributions are obtained using the Laplace–Stieltjes transform.
- We determine the optimal total expected cost and the optimal expected waiting time based on the variation of certain system parameters.

The paper is organized as follows: We present a descriptive investigation in Section 2. Steady-state evaluation is discussed in Section 3. Section 4 presents the system performance metrics. Section 5 discusses the analysis of waiting time. Section 6 provides numerical illustrations, while Section 7 presents the conclusions about the proposed system.

2. Descriptive investigation

2.1. Notation and abbreviations

$[D]_{gh}$	:	The (g,h)-th entry of matrix D
$\mathbf{e}$	:	A column vector of ones with appropriate dimensions
$G \oplus H$	:	Kronecker sum of matrices G and H
$G \otimes H$	:	Kronecker product of matrices G and H
$\mathbf{I}$	:	Identity matrix
$\mathbf{0}$	:	Zero matrix
$\mathbb{W}$	:	The set that consists of all whole numbers
V_p^q	:	$\{p, p+1, p+2, \cdots, q\}$, where $p, q \in \mathbb{W}$
δ_{nm}	:	$\begin{cases} 1, & \text{if } m = n, \\ 0, & \text{otherwise} \end{cases}$
$\bar{\delta}_{nm}$	:	$1 - \delta_{nm}$
$H(z)$	:	$\begin{cases} 1, & \text{if } z \geq 0, \\ 0, & \text{otherwise} \end{cases}$ is the Heaviside function
QIS	:	queueing–inventory system
MAP	:	Markovian arrival process
SR_1	:	Server 1
SR_2	:	Server 2
PWA	:	Premium waiting area

NPWA	:	Non-premium waiting area
OR	:	Online reservation
ONC	:	Online client
OFC	:	Offline client
FCFS	:	First come, first serve
HEPL	:	Hyper-exponential
ERG	:	Erlang
NCR	:	Negative correlation
PCR	:	Positive correlation

2.2. Model overview

This study explores a queueing–inventory system that offers two types of services for a single commodity, where the storage limit for the item is denoted by S. Two dedicated servers are assigned to provide premium and non-premium services. Both services have individual waiting areas: the premium waiting area (PWA) with a capacity of N_1 is allocated for premium service, and the non-premium waiting area with a capacity of N_2 is allocated for non-premium service.

An offline client (OFC) directly visits the system and, based on their needs, either joins the PWA, the NPWA, or balks (if dissatisfied with the system) with probabilities p, q, or r, respectively, where $p+q+r = 1$. The system features an online reservation (OR) facility for pre-booking the spot in PWA, and L ($< N_1$) is the maximum level of OR allowed in the PWA.

An online client (ONC) pre-books the PWA if there is a vacancy in the PWA and the current level of OR is less than L; otherwise, the ONC client is considered lost. A reserved ONC either joins the PWA or cancels their booking after a random duration, which follows independent exponential distributions with rates $g_1\alpha$ or $g_1\beta$, respectively, where g_1 denotes the current level of OR.

The arrivals of both client types are governed by independent MAPs. In this system, the ONC arrival process is denoted as (E_0, E_1), where E_0 and E_1 are matrices of dimension $m_1 \times m_1$. E_0 governs transitions when no arrival occurs, while E_1 governs transitions when an arrival happens. The Markov chain $R_5(t)$ has a generator E, which is a matrix of dimension $m_1 \times m_1$ and is expressed as $E = E_0 + E_1$. For an ONC, the stationary rate λ_1 is defined as $\lambda_1 = \eta_1 E_1 \mathbf{e}$. Here, the stationary row vector η_1 of size $1 \times m_1$ is determined by $\eta_1 E = \mathbf{0}$ and $\eta_1 \mathbf{e} = 1$. Likewise, the OFC arrival process is denoted as (D_0, D_1). D_0 and D_1 are matrices of dimension $m_2 \times m_2$, where D_0 determines the transitions when no arrival occurs, and D_1 determines the transitions when an arrival happens. The Markov chain $R_6(t)$ has a generator D, which is a matrix of dimension $m_2 \times m_2$, expressed as $D = D_0 + D_1$. For an OFC, the stationary rate λ_2 is defined as $\lambda_2 = \eta_2 D_1 \mathbf{e}$. Here, the stationary row vector η_2 of size $1 \times m_2$ is determined by $\eta_2 D = \mathbf{0}$ and $\eta_2 \mathbf{e} = 1$.

In this system, server-1 (SR_1) and server-2 (SR_2) are assigned to serve clients in the premium and non-premium waiting areas, with service times following independent exponential distributions with rates μ_1 and μ_2 ($> \mu_1$), respectively. Here, the premium service rate is slower than the non-premium service rate, even though both services offer the same commodity (e.g., in a restaurant, the premium service is intentionally slower to provide a luxurious and memorable dining experience, in comparison

with a non-premium service that focuses on speed and efficiency). In this context, clients in both waiting areas are served on a first-come, first-served (FCFS) basis. Upon the completion of each service, a single item is delivered to the client. The system operates under the (s, Q) ordering policy. In accordance with this policy, when the inventory position decreases to its designated reorder point s, an order is initiated for a quantity of $Q\ (= (S - s) \geq s + 1)$ items. The replenishment time adheres to an exponential distribution with the rate η. A schematic overview of the proposed model is illustrated in Figure 1.

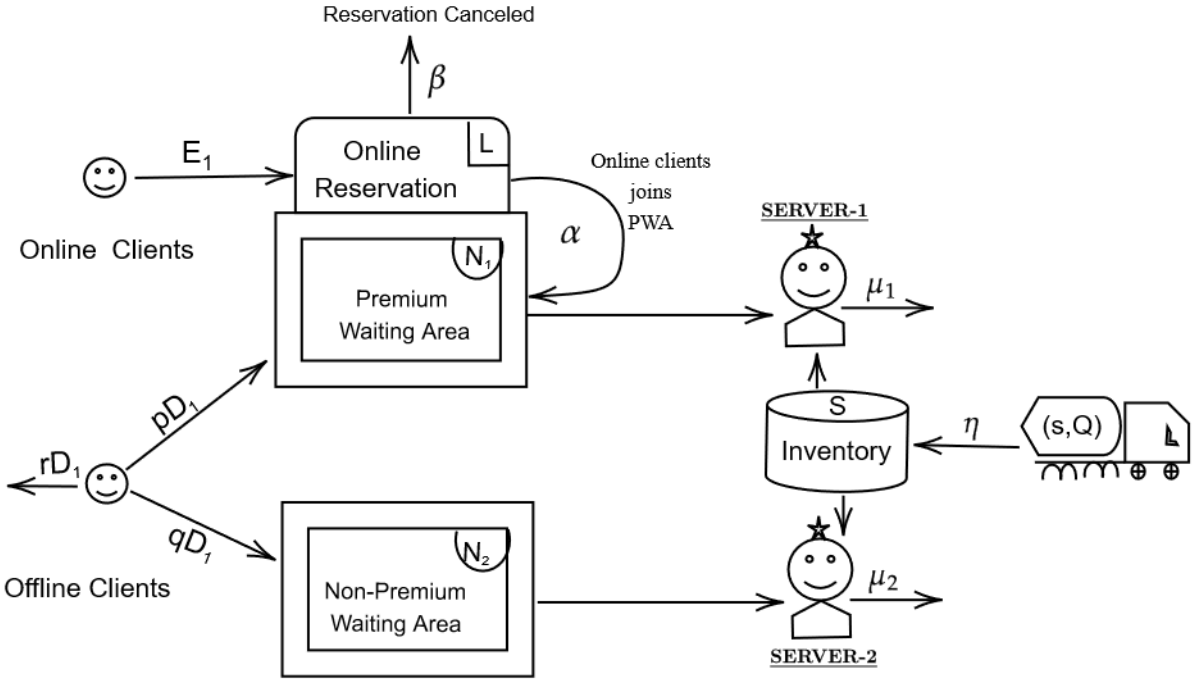

Figure 1. Online Reservation queueing–inventory System with Two Distinct Services and Client Types.

2.3. Matrix formulation

Let $R_1(t)$, $R_2(t)$, $R_3(t)$, $R_4(t)$, $R_5(t)$, and $R_6(t)$ denote, respectively, the current level of online reservations in the PWA, the count of clients in the PWA, the count of clients in the NPWA, the inventory count, the phase of the online client arrival process, and the phase of the offline client arrival process at time t. The assumptions established regarding the birth and death process of a QIS in the descriptive investigation (Subsection 2.2) constitute a stochastic process $R(t) = \{(R_1(t),R_2(t),R_3(t),R_4(t),R_5(t),R_6(t)),\ t \geq 0\}$. This process is also referred to as a continuous-time Markov chain (CTMC) with the following state space F:

$$\mathrm{F} = \{(g_1, g_2, g_3, g_4, g_5, g_6) \mid g_1 \in V_0^L; g_2 \in V_0^{N_1 - g_1}; g_3 \in V_0^{N_2}; g_4 \in V_0^S; g_5 \in V_1^{m_1}; g_6 \in V_1^{m_2}\}.$$

This Markov chain forms a transition matrix in which the coordinates $(g_1, g_2, g_3, g_4, g_5, g_6)$ and $(h_1, h_2, h_3, h_4, h_5, h_6)$ represent the row and column indices, respectively.

Theorem 2.3.1. *Let the infinitesimal generator matrix $\mathbb{P}$ with state space F and the CTMC $\{R(t), t \geq 0\}$, determined by*

$$\mathbb{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ L-1 \\ L \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \dots & L-1 & L \end{array} \\ \begin{pmatrix} A_{00} & A_{01} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ A_{10} & A_{11} & A_{12} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & A_{21} & A_{22} & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & A_{L-1L-1} & A_{L-1L} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & A_{LL-1} & A_{LL} \end{pmatrix} \end{array},$$

where $g_1 \in V_0^{L-1}$,

$$[A_{g_1 g_1+1}]_{((g_2,g_3,g_4),\ (h_2,h_3,h_4))} = \begin{cases} E_1 \otimes I_{m_2}, & h_2 = g_2, \quad h_3 = g_3, \quad h_4 = g_4, \\ & g_2 \in V_0^{N_1-1-g_1} \quad g_3 \in V_0^{N_2}, \quad g_4 \in V_0^{S}, \\ \mathbf{0} & \textit{otherwise.} \end{cases} \tag{2.1}$$

if $g_1 \in V_1^{L}$,

$$[A_{g_1 g_1-1}]_{((g_2,g_3,g_4),\ (h_2,h_3,h_4))} = \begin{cases} g_1 \alpha I_{m_1} \otimes I_{m_2}, & h_2 = g_2 + 1, \quad h_3 = g_3, \quad h_4 = g_4, \\ & g_2 \in V_0^{N_1-g_1}, \quad g_3 \in V_0^{N_2}, \quad g_4 \in V_0^{S}, \\ g_1 \beta I_{m_1} \otimes I_{m_2}, & h_2 = g_2, \quad h_3 = g_3, \quad h_4 = g_4, \\ & g_2 \in V_0^{N_1-g_1}, \quad g_3 \in V_0^{N_2}, \quad g_4 \in V_0^{S}, \\ \mathbf{0} & \textit{otherwise.} \end{cases} \tag{2.2}$$

where $g_1 \in V_0^L$,

$$[A_{g_1g_1}]_{((g_2,g_3,g_4),\,(h_2,h_3,h_4))} = \begin{cases} I_{m_1} \otimes pD_1 & h_2 = g_2 + 1, \quad h_3 = g_3, \quad h_4 = g_4, \\ & g_2 \in V_0^{N_1-1-g_1}, \quad g_3 \in V_0^{N_2}, \quad g_4 \in V_0^S, \\ \mu_1 I_{m_1} \otimes I_{m_2} & h_2 = g_2 - 1, \quad h_3 = g_3, \quad h_4 = g_4 - 1, \\ & g_2 \in V_1^{N_1-g_1}, \quad g_3 \in V_0^{N_2}, \quad g_4 \in V_1^S, \\ I_{m_1} \otimes qD_1 & h_2 = g_2, \quad h_3 = g_3 + 1, \quad h_4 = g_4, \\ & g_2 \in V_0^{N_1-g_1}, \quad g_3 \in V_0^{N_2-1}, \quad g_4 \in V_0^S, \\ \mu_2 I_{m_1} \otimes I_{m_2} & h_2 = g_2, \quad h_3 = g_3 - 1, \quad h_4 = g_4 - 1, \\ & g_2 \in V_0^{N_1-g_1}, \quad g_3 \in V_1^{N_2}, \quad g_4 \in V_1^S, \\ \eta I_{m_1} \otimes I_{m_2} & h_2 = g_2, \quad h_3 = g_3, \quad h_4 = g_4 + Q, \\ & g_2 \in V_0^{N_1-g_1}, \quad g_3 \in V_0^{N_2}, \quad g_4 \in V_0^s, \\ a & h_2 = g_2, \quad h_3 = g_3, \quad h_4 = g_4, \\ & g_2 \in V_0^{N_1-g_1}, \quad g_3 \in V_0^{N_2}, \quad g_4 \in V_0^S, \\ \mathbf{0} & \textit{otherwise.} \end{cases} \tag{2.3}$$

Where,

$$\begin{aligned} a = \; & (\bar{\delta}_{Lg_1}(E_0 + \delta_{(N_1-g_1)g_2}E_1) + \delta_{Lg_1}E) \otimes I_{m_2} + I_{m_1} \otimes (\delta_{(N_1-g_1)g_2}pD_1 + \delta_{N_2g_3}qD_1 + rD_1 + D_0) \\ & -(H(s-g_4)\eta + g_1\alpha + g_1\beta + \bar{\delta}_{0g_2}\bar{\delta}_{0g_4}\mu_1 + \bar{\delta}_{0g_3}\bar{\delta}_{0g_4}\mu_2)I_{m_1} \otimes I_{m_2}. \end{aligned}$$

The generator matrix $\mathbb{P}$ is constructed based on the transitions of the six-dimensional continuous-time Markov chain (CTMC) representing the system state. Each sub-matrix $A_{g_1g_1+1}$, $A_{g_1g_1-1}$, and $A_{g_1g_1}$ corresponds to specific transitions such as arrivals, reservation confirmations or cancellations, service completions, and replenishment.

The block structure of $\mathbb{P}$ reflects the layered dynamics of the system:

- $E_1 \otimes I_{m_2}$ captures the reservation arrival process from online clients.
- $g_1\alpha$ and $g_1\beta$ account for reservation confirmations and cancellations, respectively.
- pD_1, qD_1, and rD_1 represent offline client joining decisions and arrival intensities.
- μ_1 and μ_2 correspond to premium and non-premium service rates, respectively.
- η denotes the replenishment rate under the (s, Q) policy.

Proof. By the assumptions of the proposed model, let $A_{g_1g_1+1}$, $g_1 \in V_0^{L-1}$, be a matrix with dimension $[((N_1+1)-g_1)(N_2+1)(S+1)m_1m_2] \times [((N_1+1)-(g_1+1))(N_2+1)(S+1)m_1m_2]$. The elements of this matrix are determined by transitions resulting from online clients pre-booking the PWA with phase-type parameter E_1 (ONC phase-type arrival rate), as follows:

$$(g_1, g_2, g_3, g_4) \xrightarrow{E_1 \otimes I_{m_2}} (g_1 + 1, g_2, g_3, g_4) \; : \; g_1 \in V_0^{L-1};\; g_2 \in V_0^{(N_1-1)-g_1};\; g_3 \in V_0^{N_2};\; g_4 \in V_0^S. \tag{2.4}$$

Equation (2.4) yields Eq (2.1).

Then, we have $A_{g_1g_1-1}$, $g_1 \in V_1^L$, is a matrix with dimension $[((N_1+1)-g_1)(N_2+1)(S+1)m_1m_2] \times [((N_1+1)-(g_1-1))(N_2+1)(S+1)m_1m_2]$. Its elements represent transitions due to reserved ONCs joining the PWA, governed by the parameter α,

$$(g_1, g_2, g_3, g_4) \xrightarrow{g_1\alpha I_{m_1}\otimes I_{m_2}} (g_1-1, g_2+1, g_3, g_4) \ : \ g_1 \in V_1^L;\ g_2 \in V_0^{N_1-g_1};\ g_3 \in V_0^{N_2};\ g_4 \in V_0^S. \quad (2.5)$$

and OR cancellations, governed by the parameter β,

$$(g_1, g_2, g_3, g_4) \xrightarrow{g_1\beta I_{m_1}\otimes I_{m_2}} (g_1-1, g_2, g_3, g_4) \ : \ g_1 \in V_1^L;\ g_2 \in V_0^{N_1-g_1};\ g_3 \in V_0^{N_2};\ g_4 \in V_0^S. \quad (2.6)$$

From Eqs (2.5) and (2.6), Equation (2.2) is obtained.

The diagonal matrix $A_{g_1g_1}$, where $g_1 \in V_0^L$, is a square matrix of order $[(N_1+1-g_1)(N_2+1)(S+1)m_1m_2]$ and contains elements representing the transition rates, as detailed below.

- If OFCs wish to join the premium waiting area, their transition is governed by the parameter pD_1 (where p is the probability and D_1 is the OFC phase-type arrival rate), as follows:

$$(g_1, g_2, g_3, g_4) \xrightarrow{I_{m_1}\otimes pD_1} (g_1, g_2+1, g_3, g_4) \ : \ g_1 \in V_0^L;\ g_2 \in V_0^{(N_1-1)-g_1};\ g_3 \in V_0^{N_2};\ g_4 \in V_0^S. \quad (2.7)$$

- If OFCs wish to join the non-premium waiting area, their transition is governed by the parameter qD_1 (where q is the probability and D_1 is the OFC phase-type arrival rate), as follows:

$$(g_1, g_2, g_3, g_4) \xrightarrow{I_{m_1}\otimes qD_1} (g_1, g_2, g_3+1, g_4) \ : \ g_1 \in V_0^L;\ g_2 \in V_0^{N_1-g_1};\ g_3 \in V_0^{N_2-1};\ g_4 \in V_0^S. \quad (2.8)$$

- The transition due to a client receiving premium service is governed by the parameter μ_1 (premium service rate), as follow :

$$(g_1, g_2, g_3, g_4) \xrightarrow{\mu_1 I_{m_1}\otimes I_{m_2}} (g_1, g_2-1, g_3, g_4-1) \ : \ g_1 \in V_0^L;\ g_2 \in V_1^{N_1-g_1};\ g_3 \in V_0^{N_2};\ g_4 \in V_1^S. \quad (2.9)$$

- The transition due to a client receiving non-premium service is governed by the parameter μ_2 (non-premium service rate), as follows:

$$(g_1, g_2, g_3, g_4) \xrightarrow{\mu_2 I_{m_1}\otimes I_{m_2}} (g_1, g_2, g_3-1, g_4-1) \ : \ g_1 \in V_0^L;\ g_2 \in V_0^{N_1-g_1};\ g_3 \in V_1^{N_2};\ g_4 \in V_1^S. \quad (2.10)$$

- The transition due to replenishment is governed by the parameter η (replenishment rate), as follows:

$$(g_1, g_2, g_3, g_4) \xrightarrow{\eta I_{m_1}\otimes I_{m_2}} (g_1, g_2, g_3, g_4+Q) \ : \ g_1 \in V_0^L;\ g_2 \in V_0^{N_1-g_1};\ g_3 \in V_0^{N_2};\ g_4 \in V_0^s. \quad (2.11)$$

- The diagonal elements are filled as follow :

$$(g_1, g_2, g_3, g_4) \xrightarrow{a} (g_1, g_2, g_3, g_4) \ : \ g_1 \in V_0^L;\ g_2 \in V_0^{N_1-g_1};\ g_3 \in V_0^{N_2};\ g_4 \in V_0^S. \quad (2.12)$$

where a is the sum of the corresponding rows of entries in $A_{g_1g_1}$, $A_{g_1g_1+1}$, and $A_{g_1g_1-1}$, ensuring that each row of the matrix $\mathbb{P}$ sums to zero. From Equations (2.7)–(2.12), Equation (2.3) is obtained. As a result, each sub-matrices derived from the various transitions together makes up the infinitesimal generator matrix $\mathbb{P}$ of order $[\frac{L+1}{2}(2N_1-L+2)](N_2+1)(S+1)m_1m_2$. □

3. Steady-state evaluation

From the structure of the infinitesimal generator matrix $\mathbb{P}$, the CTMC $\{R(t), t \geq 0\}$ on the finite state space F is irreducible, aperiodic, and non-null-persistent. Consequently, the limiting distribution

$$\psi^{(g_1,g_2,g_3,g_4,g_5,g_6)} = \lim_{t\to\infty} Pr[R_1(t) = g_1, R_2(t) = g_2, R_3(t) = g_3, R_4(t) = g_4, R_5(t) = g_5, R_6(t) = g_6 \\ |R_1(0), R_2(0), R_3(0), R_4(0), R_5(0), R_6(0)]$$

exists and is independent of the initial state. Let the probability vector $\mathbf{\Psi} = \{\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}^{(1)}, ..., \boldsymbol{\psi}^{(L-1)}, \boldsymbol{\psi}^{(L)}\}$ satisfy the following equations:

$$\mathbf{\Psi}\mathbb{P} = \mathbf{0}. \tag{3.1}$$

and

$$\mathbf{\Psi}\boldsymbol{e} = 1. \tag{3.2}$$

Further, the vector $\boldsymbol{\psi}^{(g_1)}$, $g_1 \in V_0^L$ is partitioned as follows,

$$\begin{aligned}
\boldsymbol{\psi}^{(g_1)} &= \{\boldsymbol{\psi}^{(g_1,0)}, \boldsymbol{\psi}^{(g_1,1)}, ..., \boldsymbol{\psi}^{(g_1,N_1-g_1)}\},\ g_1 \in V_0^L, \\
\boldsymbol{\psi}^{(g_1,g_2)} &= \{\boldsymbol{\psi}^{(g_1,g_2,0)}, \boldsymbol{\psi}^{(g_1,g_2,1)}, ..., \boldsymbol{\psi}^{(g_1,g_2,N_2)}\},\ g_1 \in V_0^L,\ g_2 \in V_0^{N_1-g_1}, \\
\boldsymbol{\psi}^{(g_1,g_2,g_3)} &= \{\boldsymbol{\psi}^{(g_1,g_2,g_3,0)}, \boldsymbol{\psi}^{(g_1,g_2,g_3,1)}, ..., \boldsymbol{\psi}^{(g_1,g_2,g_3,S)}\},\ g_1 \in V_0^L,\ g_2 \in V_0^{N_1-g_1},\ g_3 \in V_0^{N_2}, \\
\boldsymbol{\psi}^{(g_1,g_2,g_3,g_4)} &= \{\boldsymbol{\psi}^{(g_1,g_2,g_3,g_4,1)}, \boldsymbol{\psi}^{(g_1,g_2,g_3,g_4,2)}, ..., \boldsymbol{\psi}^{(g_1,g_2,g_3,g_4,m_1)}\}, \\
&\quad g_1 \in V_0^L,\ g_2 \in V_0^{N_1-g_1},\ g_3 \in V_0^{N_2},\ g_4 \in V_0^S, \\
\boldsymbol{\psi}^{(g_1,g_2,g_3,g_4,g_5)} &= \{\psi^{(g_1,g_2,g_3,g_4,g_5,1)}, \psi^{(g_1,g_2,g_3,g_4,g_5,2)}, ..., \psi^{(g_1,g_2,g_3,g_4,g_5,m_2)}\}, \\
&\quad g_1 \in V_0^L,\ g_2 \in V_0^{N_1-g_1},\ g_3 \in V_0^{N_2},\ g_4 \in V_0^S,\ g_5 \in V_1^{m_1}.
\end{aligned}$$

Our infinitesimal generator matrix $\mathbb{P}$ shares the same structure as described in Gaver [33], allowing us to utilize similar arguments to derive the limiting probability vectors. Here, we describe the Gaver algorithm.

1. Initialize Z_0 and determine the matrices Z_n recursively, as follows:
   * $Z_0 = A_{00}$,
   * $Z_n = A_{nn} + A_{nn-1}(-Z_{n-1}^{-1})A_{n-1n}$, $n \in V_1^L$.
2. Obtain the limiting probability vectors,

$$\boldsymbol{\psi}^{(n)} = \boldsymbol{\psi}^{(n+1)}A_{n+1n}(-Z_n^{-1}),\ n \in V_0^{L-1}.$$

3. Determine the system of equations

$$\boldsymbol{\psi}^{(L)}Z_L = \mathbf{0}; \tag{3.3}$$

$$\sum_{n=0}^{L} \boldsymbol{\psi}^{(n)}\mathbf{e} = 1. \tag{3.4}$$

In Eq (3.3), the vector $\boldsymbol{\psi}^{(L)}$ can be found distinctively up to a multiplicative constant. This constant is determined by solving $\boldsymbol{\psi}^{(n)} = \boldsymbol{\psi}^{(n+1)}A_{n+1n}(-Z_n^{-1}),\ n \in V_0^{L-1}$, and $\sum\limits_{n=0}^{L} \boldsymbol{\psi}^{(n)}\mathbf{e} = 1$.

4. System performance metrics

In this section, we derive various important measures to analyze the system's characteristics.

- Expected inventory count: Let E_I denote the expected inventory count of the system in steady state. It is defined using the vector $\boldsymbol{\Psi}$ along with the positive inventory levels as follows:

$$E_I = \sum_{g_1=0}^{L}\sum_{g_2=0}^{N_1-g_1}\sum_{g_3=0}^{N_2}\sum_{g_4=1}^{S} g_4\boldsymbol{\psi}^{(g_1,g_2,g_3,g_4)}\mathbf{e}.$$

- Expected reorder rate: Let E_R denote the expected reorder rate in the steady state. This metric captures how frequently the inventory level reaches the reorder point s and triggers a replenishment process. The reorder is initiated whenever the inventory level drops from $s+1$ to s following a service completion (either premium or non-premium).

$$\begin{aligned} E_R = &\sum_{g_1=0}^{L}\sum_{g_2=1}^{N_1-g_1}\sum_{g_3=0}^{N_2}\boldsymbol{\psi}^{(g_1,g_2,g_3,s+1)}(\mu_1 I_{m_1}\otimes I_{m_2})\mathbf{e} \\ &+\sum_{g_1=0}^{L}\sum_{g_2=0}^{N_1-g_1}\sum_{g_3=1}^{N_2}\boldsymbol{\psi}^{(g_1,g_2,g_3,s+1)}(\mu_2 I_{m_1}\otimes I_{m_2})\mathbf{e}. \end{aligned}$$

The first term accounts for reorder triggers caused by a premium service completion when inventory is at $s+1$, while the second term captures reorder events due to a non-premium service completion. It can be calculated at both premium and non-premium services.

- Expected count of online reservations in the PWA: The maximum capacity of the online reservation is L. Let E_{OR} be the expected count of online reservations in the PWA. Subsequently, we have

$$E_{OR} = \sum_{g_1=1}^{L} g_1\boldsymbol{\psi}^{(g_1)}\mathbf{e}.$$

The formula multiplies each possible reservation count g_1 by the steady-state probability of the system being in a state with that count. The result is the E_{OR} under steady-state conditions.

- Expected count of clients in the PWA: Let E_{PWA} denote the expected count of clients in the PWA under steady-state. This includes clients who are waiting for premium service, excluding those who have only reserved but not yet arrived.

$$E_{PWA} = \sum_{g_1=0}^{L} \sum_{g_2=1}^{N_1-g_1} g_2 \boldsymbol{\psi}^{(g_1,g_2)} \mathbf{e}.$$

- Expected count of clients in the NPWA: Let E_{NPWA} denote the expected count of clients waiting for non-premium service under steady-state conditions. This includes all clients physically present in the NPWA.

$$E_{NPWA} = \sum_{g_1=0}^{L} \sum_{g_2=0}^{N_1-g_1} \sum_{g_3=1}^{N_2} g_3 \boldsymbol{\psi}^{(g_1,g_2,g_3)} \mathbf{e}.$$

- Expected count of online clients lost in the system: Let E_{ONL} denote the expected count of online clients lost in the system under steady-state conditions. These losses occur when either the PWA is full or the reservation limit L is reached.

$$\begin{aligned} E_{ONL} = \frac{1}{\lambda_1} \Bigg[& \sum_{g_1=0}^{L-1} \sum_{g_3=0}^{N_2} \sum_{g_4=0}^{S} \boldsymbol{\psi}^{(g_1,N_1-g_1,g_3,g_4)} (E_1 \otimes I_{m_2}) \mathbf{e} \\ & + \sum_{g_2=0}^{N_1-L} \sum_{g_3=0}^{N_2} \sum_{g_4=0}^{S} \boldsymbol{\psi}^{(L,g_2,g_3,g_4)} (E_1 \otimes I_{m_2}) \mathbf{e} \Bigg]. \end{aligned}$$

  First term: Captures the expected loss of online clients when the number of reservations g_1 is less than L, but the PWA is already full.
  Second term: Accounts for the loss when the reservations limit L has been reached and additional online reservations cannot be accommodated.

- Expected count of online reservations being canceled: Let E_{ORC} denote the expected count of online reservations canceled by clients before they arrive at the PWA. These cancellations occur when clients abandon their reservations at a given cancellation rate β prior to utilizing the service.

$$E_{ORC} = \sum_{g_1=1}^{L} \sum_{g_2=0}^{N_1-g_1} \sum_{g_3=0}^{N_2} \sum_{g_4=0}^{S} \boldsymbol{\psi}^{(g_1,g_2,g_3,g_4)} (g_1 \beta I_{m_1} \otimes I_{m_2}) \mathbf{e}.$$

- Expected count of offline clients lost in the system: Let E_{OFL} represent the expected count of offline clients lost in the system under steady-state conditions.

$$E_{OFL} = \frac{1}{\lambda_2} \Bigg[\sum_{g_1=0}^{L} \sum_{g_2=0}^{N_1-1-g_1} \sum_{g_3=0}^{N_2} \sum_{g_4=0}^{S} \boldsymbol{\psi}^{(g_1,g_2,g_3,g_4)} (I_{m_1} \otimes rD_1) \mathbf{e}$$

$$+\sum_{g_1=0}^{L}\sum_{g_3=0}^{N_2-1}\sum_{g_4=0}^{S}\boldsymbol{\psi}^{(g_1,N_1-g_1,g_3,g_4)}(I_{m_1}\otimes rD_1)\mathbf{e}$$
$$+\sum_{g_1=0}^{L}\sum_{g_4=0}^{S}\boldsymbol{\psi}^{(g_1,N_1-g_1,N_2,g_4)}(I_{m_1}\otimes D_1)\mathbf{e}\Big].$$

Offline clients are considered as lost in the following situations 1. When there is no available space in either PWA or NPWA or both upon their arrival. 2. Even when there are vacancies in both waiting areas, a customer may abandon the system with probability r.

5. Waiting time analysis

In this section, we derive the waiting time for a client in both the PWA and NPWA using the Laplace–Stieltjes Transform (LST). The time interval between a client's arrival and the instant their service is completed is called the waiting time.

5.1. Waiting time for a client in the PWA

Let W_1 be a continuous-time random variable representing the waiting time for a client in the PWA. To derive the distribution of W_1, we consider a Markov chain at an arbitrary time t, and the state space F is redefined as follows: $F_1 = \{(g_1, g_2, g_3, g_4, g_5, g_6) \mid g_1 \in V_0^L; g_2 \in V_1^{N_1-g_1}; g_3 \in V_0^{N_2}; g_4 \in V_0^S; g_5 \in V_1^{m_1}; g_6 \in V_1^{m_2}\}$.

Theorem 5.1.1. *The probability that a client does not wait in the PWA is given as*

$$P\{W_1 = 0\} = 1 - \sum_{g_1=0}^{L}\sum_{g_2=0}^{N_1-1-g_1}\sum_{g_3=0}^{N_2}\sum_{g_4=0}^{S}\sum_{g_5=1}^{m_1}\sum_{g_6=1}^{m_2}\psi^{(g_1,g_2,g_3,g_4,g_5,g_6)}. \tag{5.1}$$

Proof. The sum of the probabilities of zero waiting time and positive waiting time is equal to one. Then,

$$P\{W_1 = 0\} + P\{W_1 > 0\} = 1. \tag{5.2}$$

From Eq (5.2), the probability of a positive waiting time is obtained as follows:

$$P\{W_1 > 0\} = \sum_{g_1=0}^{L}\sum_{g_2=0}^{N_1-1-g_1}\sum_{g_3=0}^{N_2}\sum_{g_4=0}^{S}\sum_{g_5=1}^{m_1}\sum_{g_6=1}^{m_2}\psi^{(g_1,g_2,g_3,g_4,g_5,g_6)}. \tag{5.3}$$

By substituting Eq (5.1) into Eq (5.2), we obtain Eq (5.3).

□

In order to derive the distribution of W_1, some auxiliary variables are defined. Consider a Markov chain at an arbitrary time t, and assume that it is in a state $(g_1, g_2, g_3, g_4, g_5, g_6), g_2 > 0$.

1. $W_1(g_1, g_2, g_3, g_4, g_5, g_6)$ represents the time until the demand of a tagged client is satisfied.

2. $W_1^*(x) = E[e^{xW_1}]$ be a corresponding LST for unconditional waiting time ($UCWT$).
3. $W_1^*(g_1, g_2, g_3, g_4, g_5, g_6)(x) = E[e^{xW_1(g_1,g_2,g_3,g_4,g_5,g_6)}]$ be a corresponding LST for conditional waiting time (CWT).

Theorem 5.1.2. *The LST* $\{W_1^*(g_1, g_2, g_3, g_4, g_5, g_6)(x), (g_1, g_2, g_3, g_4, g_5, g_6) \in F_1^c\}$, *where* $F_1^c = F_1 \cup \{c\}$ *satisfies the following system:*

$$Z_1(x)W_1^*(x) = -\mu_1 \boldsymbol{e}(g_1, g_2, g_3, g_4, g_5, g_6), \tag{5.4}$$

where $g_1 \in V_0^L$, $g_2 \in V_1^{N_1-g_1}$, $g_3 \in V_0^{N_2}$, $g_4 \in V_1^S$, $g_5 \in V_1^{m_1}$, $g_6 \in V_1^{m_2}$.

$Z_1(x) = (A - xI)$, *where the matrix A is determined from* $\mathbb{P}$ *by removing the state* $(g_1, 0, g_3, g_4, g_5, g_6)$, $g_1 \in V_0^L$, $g_3 \in V_0^{N_2}$, $g_4 \in V_1^S$, $g_5 \in V_1^{m_1}$, $g_6 \in V_1^{m_2}$. *Let* $\{c\}$ *be the absorbing state of the system, which occurs if the tagged client demand is satisfied.*

Proof. We apply the first-step argument to determine the CWT as follows:

For $g_1 \in V_0^L$, $g_2 \in V_1^{N_1-g_1}$, $g_3 \in V_0^{N_2}$, $g_4 = 0$, $g_5 \in V_1^{m_1}$, $g_6 \in V_1^{m_2}$, $m = m_1 m_2$. Then,

$$\begin{aligned} &a[W_1^*(g_1, g_2, g_3, 0, g_5, g_6)(x)] - \bar{\delta}_{N_1-g_1g_2} I_{m_1} \otimes pD_1 W_1^*(g_1, g_2 + 1, g_3, 0, g_5, g_6)(x) \\ &-\bar{\delta}_{N_2g_3} I_{m_1} \otimes qD_1 W_1^*(g_1, g_2, g_3 + 1, 0, g_5, g_6)(x) - \bar{\delta}_{Lg_1}\bar{\delta}_{N_1-g_1g_2} E_1 \otimes I_{m_2} W_1^*(g_1 + 1, g_2, g_3, 0, g_5, g_6)(x) \\ &-g_1\alpha I_m W_1^*(g_1 - 1, g_2 + 1, g_3, 0, g_5, g_6)(x) - g_1\beta I_m W_1^*(g_1 - 1, g_2, g_3, 0, g_5, g_6)(x) \\ &-\eta I_m W_1^*(g_1, g_2, g_3, Q, g_5, g_6)(x) = 0, \end{aligned} \tag{5.5}$$

$$a = (xI_m + \bar{\delta}_{N_1-g_1g_2} I_{m_1} \otimes pD_1 + \bar{\delta}_{N_2g_3} I_{m_1} \otimes qD_1 + \bar{\delta}_{Lg_1}\bar{\delta}_{N_1-g_1g_2} E_1 \otimes I_{m_2} + g_1\alpha I_m + g_1\beta I_m + \eta I_m).$$

For $g_1 \in V_0^L$, $g_2 \in V_1^{N_1-g_1}$, $g_3 \in V_0^{N_2}$, $g_4 \in V_1^S$, $g_5 \in V_1^{m_1}$, $g_6 \in V_1^{m_2}$, $m = m_1 m_2$,

$$\begin{aligned} &b[W_1^*(g_1, g_2, g_3, g_4, g_5, g_6)(x)] - \bar{\delta}_{N_1-g_1g_2} I_{m_1} \otimes pD_1 W_1^*(g_1, g_2 + 1, g_3, g_4, g_5, g_6)(x) \\ &-\bar{\delta}_{N_2g_3} I_{m_1} \otimes qD_1 W_1^*(g_1, g_2, g_3 + 1, g_4, g_5, g_6)(x) - \bar{\delta}_{Lg_1}\bar{\delta}_{N_1-g_1g_2} E_1 \otimes I_{m_2} W_1^*(g_1 + 1, g_2, g_3, g_4, g_5, g_6)(x) \\ &-g_1\alpha I_m W_1^*(g_1 - 1, g_2 + 1, g_3, g_4, g_5, g_6)(x) - g_1\beta I_m W_1^*(g_1 - 1, g_2, g_3, g_4, g_5, g_6)(x) \\ &-H(s - g_4)\eta I_m W_1^*(g_1, g_2, g_3, g_4 + Q, g_5, g_6)(x) - \bar{\delta}_{1g_2}\mu_1 I_m W_1^*(g_1, g_2 - 1, g_3, g_4 - 1, g_5, g_6)(x) \\ &-\bar{\delta}_{0g_3}\mu_2 I_m W_1^*(g_1, g_2, g_3 - 1, g_4 - 1, g_5, g_6)(x) = \mu_1 I_m, \end{aligned} \tag{5.6}$$

$$\begin{aligned} b = (&xI_m + \bar{\delta}_{N_1-g_1g_2} I_{m_1} \otimes pD_1 + \bar{\delta}_{N_2g_3} I_{m_1} \otimes qD_1 + \bar{\delta}_{Lg_1}\bar{\delta}_{N_1-g_1g_2} E_1 \otimes I_{m_2} + g_1\alpha I_m \\ &+g_1\beta I_m + H(s - g_4)\eta I_m + \mu_1 I_m + \bar{\delta}_{1g_2}\mu_1 I_m + \bar{\delta}_{0g_3}\mu_2 I_m). \end{aligned}$$

From Eqs (5.5) and (5.6), we obtain a coefficient matrix for the unknowns, which is block tri-diagonal, yielding the stated result. □

Theorem 5.1.3. *The* n^{th} *moments of CWT are expressed as*

$$Z_1(x)\frac{d^{n+1}}{dx^{n+1}}W_1^*(x) - (n+1)\frac{d^n}{dx^n}W_1^*(x) = 0 \tag{5.7}$$

and

$$\frac{d^{n+1}}{dx^{n+1}}W_1^*(g_1,g_2,g_3,g_4,g_5,g_6)(x)|_{x=0} = E[W_1^{n+1}(g_1,g_2,g_3,g_4,g_5,g_6)], \quad (g_1,g_2,g_3,g_4,g_5,g_6) \in F_1^c. \tag{5.8}$$

Proof. The Eqs (5.5) and (5.6) can be exploited to get a recursive algorithm to compute moments for both CWT and UCWT.

By differentiating Eqs (5.5) and (5.6) $(n+1)$ times and then evaluating the results at $x = 0$, we obtain the following expressions:

For $g_1 \in V_0^L$, $g_2 \in V_1^{N_1-g_1}$, $g_3 \in V_0^{N_2}$, $g_4 = 0$, $g_5 \in V_1^{m_1}$, $g_6 \in V_1^{m_2}$, $m = m_1m_2$,

$$\begin{aligned}
&a\Big[E[W_1^{n+1}(g_1,g_2,g_3,0,g_5,g_6)]\Big] - \bar{\delta}_{N_1-g_1g_2}I_{m_1} \otimes pD_1E[W_1^{n+1}(g_1,g_2+1,g_3,0,g_5,g_6)] \\
&-\bar{\delta}_{N_2g_3}I_{m_1} \otimes qD_1E[W_1^{n+1}(g_1,g_2,g_3+1,0,g_5,g_6)] - \bar{\delta}_{Lg_1}\bar{\delta}_{N_1-g_1g_2}E_1 \otimes I_{m_2}E[W_1^{n+1}(g_1+1,g_2,g_3,0,g_5,g_6)] \\
&-g_1\alpha I_mE[W_1^{n+1}(g_1-1,g_2+1,g_3,0,g_5,g_6)] - g_1\beta I_mE[W_1^{n+1}(g_1-1,g_2,g_3,0,g_5,g_6)] \\
&-\eta I_mE[W_1^{n+1}(g_1,g_2,g_3,Q,g_5,g_6)] \\
&= (n+1)E[W_1^n(g_1,g_2,g_3,0,g_5,g_6)]
\end{aligned} \tag{5.9}$$

$$a = (\bar{\delta}_{N_1-g_1g_2}I_{m_1} \otimes pD_1 + \bar{\delta}_{N_2g_3}I_{m_1} \otimes qD_1 + \bar{\delta}_{Lg_1}\bar{\delta}_{N_1-g_1g_2}E_1 \otimes I_{m_2} + g_1\alpha I_m + g_1\beta I_m + \eta I_m).$$

For $g_1 \in V_0^L$, $g_2 \in V_1^{N_1-g_1}$, $g_3 \in V_0^{N_2}$, $g_4 \in V_1^S$, $g_5 \in V_1^{m_1}$, $g_6 \in V_1^{m_2}$, $m = m_1m_2$,

$$\begin{aligned}
&b\Big[E[W_1^{n+1}(g_1,g_2,g_3,g_4,g_5,g_6)]\Big] - \bar{\delta}_{N_1-g_1g_2}I_{m_1} \otimes pD_1E[W_1^{n+1}(g_1,g_2+1,g_3,g_4,g_5,g_6)] \\
&-\bar{\delta}_{N_2g_3}I_{m_1} \otimes qD_1E[W_1^{n+1}(g_1,g_2,g_3+1,g_4,g_5,g_6)] \\
&-\bar{\delta}_{Lg_1}\bar{\delta}_{N_1-g_1g_2}E_1 \otimes I_{m_2}E[W_1^{n+1}(g_1+1,g_2,g_3,g_4,g_5,g_6)] \\
&-g_1\alpha I_mE[W_1^{n+1}(g_1-1,g_2+1,g_3,g_4,g_5,g_6)] - g_1\beta I_mE[W_1^{n+1}(g_1-1,g_2,g_3,g_4,g_5,g_6)] \\
&-H(s-g_4)\eta I_mE[W_1^{n+1}(g_1,g_2,g_3,g_4+Q,g_5,g_6)] \\
&-\bar{\delta}_{1g_2}\mu_1I_mE[W_1^{n+1}(g_1,g_2-1,g_3,g_4-1,g_5,g_6)] \\
&-\bar{\delta}_{0g_3}\mu_2I_mE[W_1^{n+1}(g_1,g_2,g_3-1,g_4-1,g_5,g_6)] \\
&= (n+1)E[W_1^n(g_1,g_2,g_3,g_4,g_5,g_6)],
\end{aligned} \tag{5.10}$$

$$\begin{aligned}
b = (&\bar{\delta}_{N_1-g_1g_2}I_{m_1} \otimes pD_1 + \bar{\delta}_{N_2g_3}I_{m_1} \otimes qD_1 + \bar{\delta}_{Lg_1}\bar{\delta}_{N_1-g_1g_2}E_1 \otimes I_{m_2} + g_1\alpha I_m + g_1\beta I_m \\
&+H(s-g_4)\eta I_m + \bar{\delta}_{1g_2}\mu_1I_m + \mu_1I_m + \bar{\delta}_{0g_3}\mu_2I_m).
\end{aligned}$$

The Eqs (5.9) and (5.10) enable us to ascertain the unknowns $E[W_1^{n+1}(g_1,g_2,g_3,g_4,g_5,g_6)]$, $(g_1,g_2,g_3,g_4,g_5,g_6) \in F_1^c$ based on the moments of one order less. Noticing that $E[W_1^n(g_1,g_2,g_3,g_4,g_5,g_6)] = 1$ for $n = 0$, we can obtain the moments up to the desired order in a recursive way. □

Theorem 5.1.4. *The LST of UCWT for a client in the PWA is expressed as*

$$W_1^*(x) = 1 - \sum_{g_1=0}^{L} \sum_{g_2=0}^{(N_1-1)-g_1} \sum_{g_3=0}^{N_2} \sum_{g_4=0}^{S} \sum_{g_5=1}^{m_1} \sum_{g_6=1}^{m_2} \psi^{(g_1,g_2,g_3,g_4,g_5,g_6)} + \sum_{g_1=0}^{L} \sum_{g_2=0}^{(N_1-1)-g_1} \sum_{g_3=0}^{N_2} \sum_{g_4=0}^{S} \sum_{g_5=1}^{m_1} \sum_{g_6=1}^{m_2} \psi^{(g_1,g_2,g_3,g_4,g_5,g_6)} W_1^*(g_1, g_2 + 1, g_3, g_4, g_5, g_6)(x). \tag{5.11}$$

Proof. Employing the Poisson arrival time average property, the LST of W_1 is determined as follows:

$$W_1^*(x) = \boldsymbol{\psi}^{(g_1)} W_1^*(g_1, g_2 + 1, g_3, g_4, g_5, g_6)(x), \tag{5.12}$$

where $g_1 \in V_0^L$, $g_2 \in V_0^{N_1-1-g_1}$, $g_3 \in V_0^{N_2}$, $g_4 \in V_0^S$, $g_5 \in V_1^{m_1}$, $g_6 \in V_1^{m_2}$.

By Eq (5.12) $W_1^*(x)$ is obtained for given x. This enables the use of the Euler and Post-Widder algorithms as detailed by Abate and Whitt [34], for the numerical inversion of $W_1^*(x)$.

□

Corollary 5.1.1. *The n^{th} moments of UCWT can be expressed as*

$$E[W_1^n] = \Bigg(\delta_{0n} + (1 - \delta_{0n}) \sum_{g_1=0}^{L} \sum_{g_2=0}^{(N_1-1)-g_1} \sum_{g_3=0}^{N_2} \sum_{g_4=0}^{S} \sum_{g_5=1}^{m_1} \sum_{g_6=1}^{m_2} \psi^{(g_1,g_2,g_3,g_4,g_5,g_6)} E[W_1^n(g_1, g_2 + 1, g_3, g_4, g_5, g_6)]\Bigg). \tag{5.13}$$

Proof. The moments of W_1 can be computed by differentiating Theorem 5.1.4 n times and evaluating the result at $x = 0$. This approach gives the n^{th} moments of $UCWT$ in terms of the CWT of the same order. □

Corollary 5.1.2. *The expected waiting time of a client in the PWA is expressed as,*

$$E[W_1] = \sum_{g_1=0}^{L} \sum_{g_2=0}^{(N_1-1)-g_1} \sum_{g_3=0}^{N_2} \sum_{g_4=0}^{S} \sum_{g_5=1}^{m_1} \sum_{g_6=1}^{m_2} \psi^{(g_1,g_2,g_3,g_4,g_5,g_6)} E[W_1(g_1, g_2 + 1, g_3, g_4, g_5, g_6)]. \tag{5.14}$$

Proof. Substituting $n = 1$ into equation (5.13) in Corollary 5.1.1, we get the result as equation (5.14). □

5.2. Waiting time for a client in the NPWA

Let W_2 be a continuous-time random variable representing the waiting time for a client in the NPWA. To derive the distribution of W_2, we consider a Markov chain at an arbitrary time t, and the state space F is redefined as follows: $F_2 = \{(g_1, g_2, g_3, g_4, g_5, g_6) \mid g_1 \in V_0^L; g_2 \in V_0^{N_1-g_1}; g_3 \in V_1^{N_2}; g_4 \in V_0^S; g_5 \in V_1^{m_1}; g_6 \in V_1^{m_2}\}$.

Theorem 5.2.1. *The probability that a client does not wait in the NPWA is given as*

$$P\{W_2 = 0\} = 1 - \sum_{g_1=0}^{L} \sum_{g_2=0}^{N_1-g_1} \sum_{g_3=0}^{N_2-1} \sum_{g_4=0}^{S} \sum_{g_5=1}^{m_1} \sum_{g_6=1}^{m_2} \psi^{(g_1,g_2,g_3,g_4,g_5,g_6)}. \tag{5.15}$$

Proof. The approach used to prove this theorem is similar to that of Theorem 5.1.1. □

Theorem 5.2.2. *The LST* $\{W_2^*(g_1, g_2, g_3, g_4, g_5, g_6)(x), (g_1, g_2, g_3, g_4, g_5, g_6) \in F_2^c\}$*, where* $F_2^c = F_2 \cup \{c\}$*, satisfies the following system:*

$$Z_2(x)W_2^*(x) = -\mu_2 \boldsymbol{e}(g_1, g_2, g_3, g_4, g_5, g_6), \tag{5.16}$$

where $g_1 \in V_0^L$, $g_2 \in V_0^{N_1-g_1}$, $g_3 \in V_1^{N_2}$, $g_4 \in V_1^S$, $g_5 \in V_1^{m_1}$, $g_6 \in V_1^{m_2}$.

$Z_2(x) = (B - xI)$*, where the matrix* B *is determined from* $\mathbb{P}$ *by removing the state* $(g_1, g_2, 0, g_4, g_5, g_6)$*,* $g_1 \in V_0^L$, $g_2 \in V_0^{N_1-g_1}$, $g_4 \in V_1^S$, $g_5 \in V_1^{m_1}$, $g_6 \in V_1^{m_2}$*. Let* $\{c\}$ *be the absorbing state of the system, which occurs if the tagged client demand is expired.*

Proof. The approach used to prove this theorem is similar to that of Theorem 5.1.2.

□

Theorem 5.2.3. *The* n^{th} *moment of CWT is expressed as*

$$Z_2(x)\frac{d^{n+1}}{dx^{n+1}}W_2^*(x) - (n+1)\frac{d^n}{dx^n}W_2^*(x) = 0 \tag{5.17}$$

and

$$\frac{d^{n+1}}{dx^{n+1}}W_2^*(g_1, g_2, g_3, g_4, g_5, g_6)(x)|_{x=0} = E[W_2^{n+1}(g_1, g_2, g_3, g_4, g_5, g_6)], (g_1, g_2, g_3, g_4, g_5, g_6) \in F_2^c. \tag{5.18}$$

Proof. The approach used to prove this theorem is similar to that of Theorem 5.1.3.

□

Theorem 5.2.4. *The LST of UCWT for a client in the NPWA is given by*

$$\begin{aligned} W_2^*(x) \;=\; & 1 - \sum_{g_1=0}^{L} \sum_{g_2=0}^{N_1-g_1} \sum_{g_3=0}^{N_2-1} \sum_{g_4=0}^{S} \sum_{g_5=1}^{m_1} \sum_{g_6=1}^{m_2} \psi^{(g_1,g_2,g_3,g_4,g_5,g_6)} \\ & + \sum_{g_1=0}^{L} \sum_{g_2=0}^{N_1-g_1} \sum_{g_3=0}^{N_2-1} \sum_{g_4=0}^{S} \sum_{g_5=1}^{m_1} \sum_{g_6=1}^{m_2} \psi^{(g_1,g_2,g_3,g_4,g_5,g_6)} W_2^*(g_1, g_2, g_3 + 1, g_4, g_5, g_6)(x). \end{aligned} \tag{5.19}$$

Proof. The approach used to prove this theorem is similar to that of Theorem 5.1.4.

□

Corollary 5.2.1. *The n^{th} moment of UCWT is given by*

$$E[W_2^n] = \Big(\delta_{0n} + (1-\delta_{0n}) \sum_{g_1=0}^{L} \sum_{g_2=0}^{N_1-g_1} \sum_{g_3=0}^{N_2-1} \sum_{g_4=0}^{S} \sum_{g_5=1}^{m_1} \sum_{g_6=1}^{m_2} \psi^{(g_1,g_2,g_3,g_4,g_5,g_6)} E[W_2^n(g_1, g_2, g_3+1, g_4, g_5, g_6)]\Big). \tag{5.20}$$

Proof. The proof of this corollary follows a technique similar to that in Corollary 5.1.1. □

Corollary 5.2.2. *The expected waiting time of a client in the NPWA is given by*

$$E[W_2] = \sum_{g_1=0}^{L} \sum_{g_2=0}^{N_1-g_1} \sum_{g_3=0}^{N_2-1} \sum_{g_4=0}^{S} \sum_{g_5=1}^{m_1} \sum_{g_6=1}^{m_2} \psi^{(g_1,g_2,g_3,g_4,g_5,g_6)} E[W_2(g_1, g_2, g_3+1, g_4, g_5, g_6)]. \tag{5.21}$$

Proof. The proof of this corollary follows a technique similar to that in Corollary 5.1.2.

□

6. Numerical illustration

In this section, our primary objective is to examine the impact of the variables S and s on the system's total expected cost rate (TEC). The MAPs for the appearance of online and offline clients are

1. Hyper-exponential (HEPL):

$$D_0 = \begin{bmatrix} -10 & 0 \\ 0 & -1 \end{bmatrix}, \quad D_1 = \begin{bmatrix} 9 & 1 \\ 0.9 & 0.1 \end{bmatrix};$$

$$E_0 = \begin{bmatrix} -1.90 & 0 \\ 0 & -0.19 \end{bmatrix}, \quad E_1 = \begin{bmatrix} 1.710 & 0.190 \\ 0.171 & 0.019 \end{bmatrix}.$$

2. Erlang (ERG):

$$D_0 = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}, \quad D_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix};$$

$$E_0 = \begin{bmatrix} -4.5000 & 4.5000 & 0 \\ 0 & -4.5000 & 4.5000 \\ 0 & 0 & -4.5000 \end{bmatrix}, \quad E_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 4.5000 & 0 & 0 \end{bmatrix}.$$

3. Negative correlation (NCR):

$$D_0 = \begin{bmatrix} -2 & 2 & 0 \\ 0 & -81 & 0 \\ 0 & 0 & -81 \end{bmatrix}, \quad D_1 = \begin{bmatrix} 0 & 0 & 0 \\ 25.25 & 0 & 55.75 \\ 55.75 & 0 & 25.25 \end{bmatrix};$$

$$E_0 = \begin{bmatrix} -1.00222 & 1.00222 & 0 \\ 0 & -1.00222 & 0 \\ 0 & 0 & -225.75 \end{bmatrix}, \quad E_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0.01002 & 0 & 0.9922 \\ 223.4925 & 0 & 2.2575 \end{bmatrix}.$$

4. Positive correlation (PCR):

$$D_0 = \begin{bmatrix} -2 & 2 & 0 \\ 0 & -81 & 0 \\ 0 & 0 & -81 \end{bmatrix}, \quad D_1 = \begin{bmatrix} 0 & 0 & 0 \\ 55.25 & 0 & 25.75 \\ 25.75 & 0 & 55.25 \end{bmatrix};$$

$$E_0 = \begin{bmatrix} -1.00222 & 1.00222 & 0 \\ 0 & -1.00222 & 0 \\ 0 & 0 & -225.75 \end{bmatrix}, \quad E_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0.9922 & 0 & 0.01002 \\ 2.2575 & 0 & 223.4925 \end{bmatrix}.$$

The online client process exhibits a positive (negative) correlated arrival pattern, with a coefficient of variance $c_{var} = 2\lambda_1\eta_1(-E_0)^{-1}\mathbf{e} - 1 = 1.9868$ (1.9868) and a coefficient of correlation $c_{cor} = (\lambda_1\eta_1(-E_0)^{-1}E_1(-E_0)^{-1}\mathbf{e} - 1)/c_{var} = 0.4889$ (−0.4889), where the arrival rate $\lambda_1 = 1.00$.

The offline client process exhibits a positive (negative) correlated arrival pattern, with $c_{var} = 2\lambda_2\eta_2(-D_0)^{-1}\mathbf{e} - 1 = 2.7265$ (2.7265) and $c_{cor} = (\lambda_2\eta_2(-D_0)^{-1}D_1(-D_0)^{-1}\mathbf{e} - 1)/c_{var} = 0.1213$ (−0.1213), where the arrival rate $\lambda_2 = 3.81$.

6.1. Total expected cost

The total expected cost rate (TEC) is calculated by considering the following cost:

$$\text{TEC} = c_hE_I + c_sE_R + c_{PWA}E_{PWA} + c_{NPWA}E_{NPWA} + c_{OR}E_{OR} + c_{cl}(E_{ONL} + E_{OFL}) + c_{ORC}E_{ORC},$$

where

c_h : The inventory holding cost per unit item per unit time t.
c_s : The setup cost for every order.
c_{PWA} : The waiting cost for a client in the PWA per unit time t.
c_{NPWA} : The waiting cost for a client in the NPWA per unit time t.
c_{OR} : The online reservation holding cost per unit time t.
c_{cl} : The lost cost per client per unit time t.
c_{ORC} : The online reservation cancellation cost per unit time t.

The primary goal of any business is to increase its profits by minimizing the total expected cost (TEC). Analyzing the total expected cost function involves adjusting the model's parameters to optimize operational efficiency and support informed decision-making, ultimately ensuring sustained profitability. For the numerical analysis, we begin by setting the parameters as follows: $S = 28$, $s = 8$, $\lambda_1 = 1$, $\lambda_2 = 3.7$, $\alpha = 2.5$, $\beta = 1.4$, $\mu_1 = 6.5$, $\mu_2 = 7.5$, $\eta = 8.4$, $N_1 = 10$, $N_2 = 8$, $L = 5$, $p = 0.5$, $q = 0.3$, $r = 1 - p - q$, and with the cost values $c_h = 1.06$, $c_s = 10$, $c_{PWA} = 8$, $c_{NPWA} = 5$, $c_{OR} = 2.6$, $c_{cl} = 0.5$, and $c_{ORC} = 0.5$.

- The local convexity points of the total expected cost (TEC) are explored by simultaneously varying both S and s, as shown in Table 1 and Figure 2. It is important to note the existence of a minimum TEC in each row and column. Specifically, the minimum TEC values are denoted by underlined numbers in the rows and bold numbers in the columns. The intersection of an underlined number and a bold number represents the optimal cost value $TEC^*(S^*, s^*)$. In this context, $TEC^*(S^*, s^*) = 8.243018$ is achieved at $S^* = 28$ and $s^* = 8$. This result provides a practical decision basis for inventory managements; by setting the maximum inventory level to $S^* = 28$ and the reorder point to $s^* = 8$, the system achieves its lowest expected cost. This helps management balance inventory holding and ordering costs effectively.
- Figure 3 depicts the values of $S \in [26, 31]$, $s \in [6, 12]$, $N_1 = 9, 10$, and $L = 5$ with $\eta = 8.6$ while keeping all other parameters constant. Increasing the maximum capacity of PWA (N_1) leads to a rise in the optimal cost value $TEC^*(S^*, s^*)$. This means that, if the size of the PWA increases, the client count in the PWA rises, which in turn causes the TEC to increase. Increasing PWA capacity attracts more clients but may raise operational costs, so expansion should be carefully evaluated against demand.
- Compared to Figure 3, the optimal cost value $TEC^*(S^*, s^*)$ in Figure 4 increases as the maximum level of the OR (L) has increased. This implies that overbooking may lead to higher cancellations and holding costs.
- Compared to Figures 3 and 4, the optimal cost value $TEC^*(S^*, s^*)$ in Figure 5 increases due to the rise in the maximum level (L) of the OR. Reservation levels have a significant impact on cost; overbooking should be avoided unless sufficient inventory and staffing are available to support it.
- As the cost values of c_h, c_{PWA}, c_{NPWA}, c_s, and c_{cl} increase, the total expected cost also rises, as indicated in Tables 2 and 3. This highlights the importance of cost control—minimizing holding, waiting, setup, and loss-related costs can significantly reduce the overall TEC. Inventory management should regularly assess and optimize these cost components to maintain system efficiency.

Table 1. Impact of S and s on $TEC(S, s)$.

S/s	6	7	8	9	10	11	12
26	10.417744	9.660536	9.075629	9.030448	9.329161	9.818025	10.437805
27	9.776801	8.998868	8.462639	8.482110	8.765466	9.196012	9.732667
28	9.409467	**8.734543**	**8.243018**	**8.258543**	**8.511385**	**8.895220**	9.369460
29	**9.399509**	8.762253	8.294998	8.307935	8.544775	8.900918	**9.357820**
30	9.604686	9.000279	8.559727	8.580022	8.829237	9.200824	9.654875
31	9.982851	9.413633	9.021309	9.080195	9.348611	9.729000	10.184926

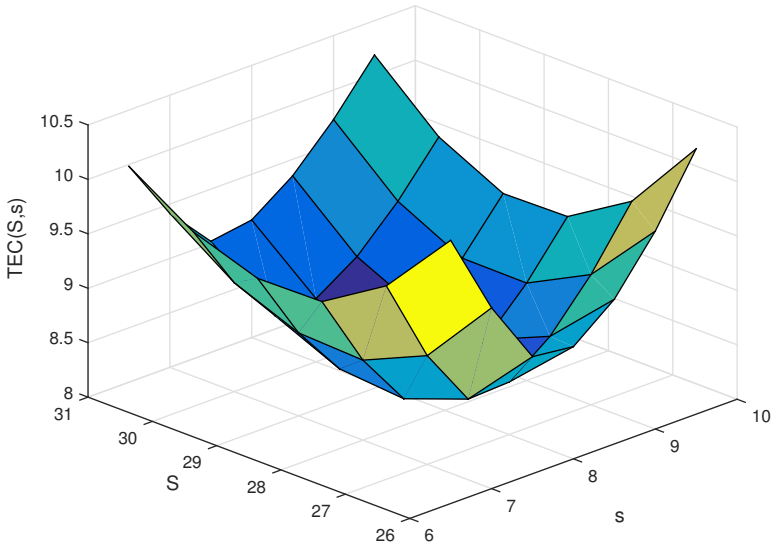

Figure 2. $TEC(S, s)$.

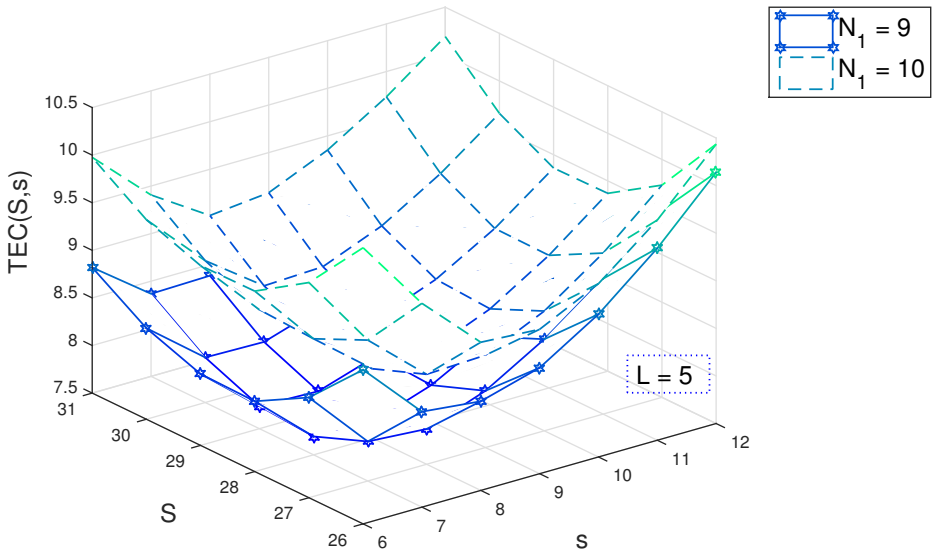

Figure 3. Three-dimensional plot illustrating the convexity of $TEC(S, s)$ ($\eta = 8.6$ and $L = 5$).

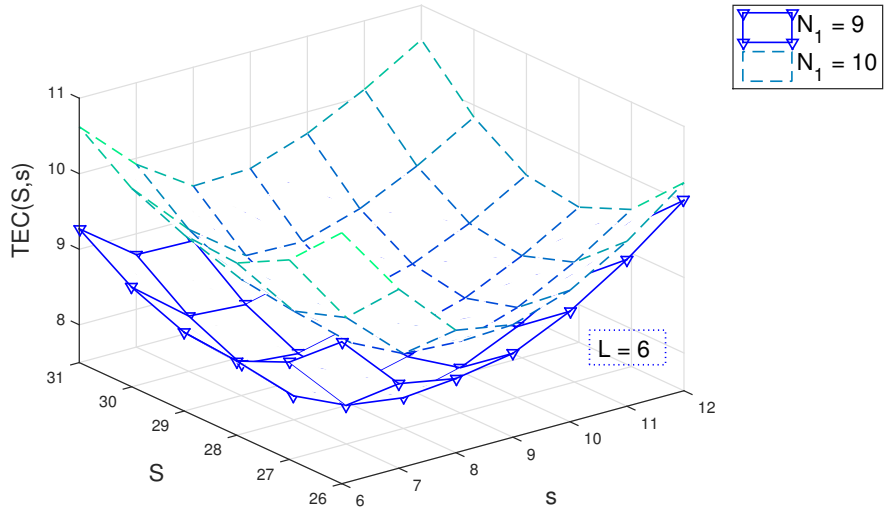

Figure 4. Three-dimensional plot illustrating the convexity of $TEC(S, s)$ ($\eta = 8.6$ and $L = 6$).

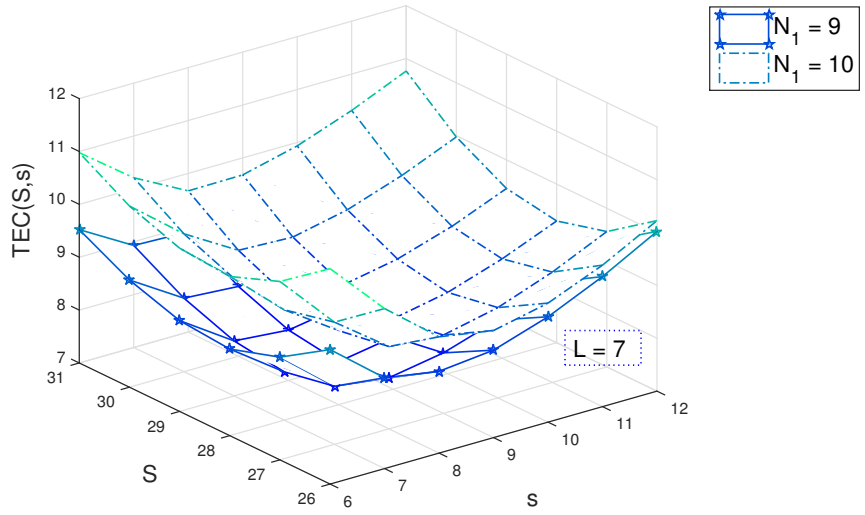

Figure 5. Three-dimensional plot illustrating the convexity of $TEC(S, s)$ ($\eta = 8.6$ and $L = 7$).

Table 2. Impact of c_h, c_{PWA}, and c_{NPWA} on TEC.

	c_h=1.02				c_h=1.03			
c_{PWA}/c_{NPWA}	5.25	5.5	5.75	6	5.25	5.5	5.75	6
7	7.7222	7.8956	8.0691	8.2425	8.0008	8.1742	8.3477	8.5211
7.25	7.7801	7.9536	8.1270	8.3005	8.0587	8.2322	8.4056	8.5791
7.5	7.8381	8.0115	8.1850	8.3585	8.1167	8.2902	8.4636	8.6371
7.75	7.8961	8.0695	8.2430	8.4164	8.1747	8.3481	8.5216	8.6950
	c_h=1.04				c_h=1.05			
c_{PWA}/c_{NPWA}	5.25	5.5	5.75	6	5.25	5.5	5.75	6
7	8.2794	8.4528	8.6263	8.7997	8.5580	8.7315	8.9049	9.0784
7.25	8.3374	8.5108	8.6843	8.8577	8.6160	8.7894	8.9629	9.1363
7.5	8.3953	8.5688	8.7422	8.9157	8.6739	8.8474	9.0208	9.1943
7.75	8.4533	8.6268	8.8002	8.9737	8.7319	8.9054	9.0788	9.2523

Table 3. Impact of c_h, c_s, and c_{cl} on TEC.

	c_h=1.02				c_h=1.03			
c_s/c_{cl}	0.35	0.4	0.45	0.5	0.35	0.4	0.45	0.5
7.05	5.2749	5.3354	5.3956	5.4564	5.5535	5.6140	5.6745	5.7350
7.25	5.8927	5.9532	6.0137	6.0742	6.1713	6.2318	6.2923	6.3528
7.45	6.5105	6.5710	6.6315	6.6920	6.7891	6.8496	6.9102	6.9707
7.65	7.1283	7.1888	7.2494	7.3099	7.4069	7.4675	7.5280	7.5885
	c_h=1.04				c_h=1.05			
c_s/c_{cl}	0.35	0.4	0.45	0.5	0.35	0.4	0.45	0.5
7.05	6.1107	6.1712	6.2318	6.2923	6.3893	6.4499	6.5104	6.5709
7.25	6.7285	6.7891	6.8496	6.9101	7.0072	7.0677	7.1282	7.1887
7.45	7.3464	7.4069	7.4674	7.5279	7.6250	7.6855	7.7460	7.8065
7.65	7.9642	8.0247	8.0852	8.1457	8.2428	8.3033	8.3638	8.4243

6.2. Effects of varying parameters on certain performance measures

- Table 4 demonstrates that an increase in the non-premium service rate (μ_2) results in a decrease in E_I. Similarly, as the replenishment rate (η) increases, the E_I also increases. This means that an increase in the service rate leads to a faster checkout and reduces the inventory level. Conversely, if the inventory restocking duration decreases, then the inventory level increases. Inventory management should strategically adjust service and restocking rates to maintain optimal inventory levels—accelerating service to reduce overstock and increasing replenishment speed to prevent stock-outs.
- Table 5 examines the impacts of the probability p on E_{PWA}, E_{NPWA}, and E_{OR}. If the probability (p) of offline arrivals choosing the PWA increases, the expected client count in the PWA (E_{PWA}) and the total expected cost (TEC) also increase. Conversely, the expected client count in the NPWA (E_{NPWA}) and the expected count of online reservations in the PWA (E_{OR}) are slightly decreased.

Management should monitor and control the flow of offline clients toward premium services, as a higher preference for PWA may lead to congestion and increased operational costs. Balancing service allocation can improve efficiency and cost-effectiveness.

- Figure 6 illustrates that an increase in the rate (α) of a reserved online client joining the PWA leads to a reduction in E_{OR}. Similarly, an increase in the rate (β) of online reservations being canceled leads to a reduction in E_{OR}. To manage reservation load effectively, operators can influence client behavior—encouraging timely arrivals (higher α) or enabling flexible cancellations (higher β) to avoid reservation backlog and improve service utilization.
- Figure 7 explains the impact on E_{PWA} due to the rates μ_1 and α. When the premium service rate (μ_1) increases, the expected client count in the PWA (E_{PWA}) decreases. This suggests that, as the service time decreases with an increase in the rate, customers experience faster service. Similarly, when the rate of reserved clients joining the PWA (α) rises, the expected client count in the PWA (E_{PWA}) also increases. This indicates that reserved clients join the PWA within shorter intervals of time when the rate increases. Enhancing premium service speed can help reduce waiting area congestion, while a higher rate of reservation fulfillment should be managed to avoid overcrowding and maintain service quality.

Table 4. Impact of μ_2 and η on E_I.

μ_2/η	8.4	8.5	8.6	8.7	8.8	8.9
	E_I					
7.1	0.740472	0.764460	0.787888	0.810769	0.833119	0.854955
7.2	0.655943	0.679300	0.702109	0.724385	0.746142	0.767397
7.3	0.572785	0.595531	0.617742	0.639432	0.660616	0.681309
7.4	0.490969	0.513124	0.534756	0.555880	0.576509	0.596658
7.5	0.410466	0.432049	0.453121	0.473697	0.493789	0.513413
7.6	0.331249	0.352279	0.372809	0.392853	0.412426	0.431541

Table 5. Impact of the probability p on various performance measures.

p	E_{PWA}	E_{NPWA}	E_{OR}	TEC(S,s)
0.4	0.200229	0.208279	0.000057	9.561180
0.45	0.216346	0.196493	0.000055	9.813898
0.5	0.230531	0.183859	0.000051	10.04449
0.55	0.242290	0.170505	0.000050	10.25123
0.6	0.251222	0.156595	0.000047	10.43192
0.65	0.257029	0.142336	0.000044	10.58404

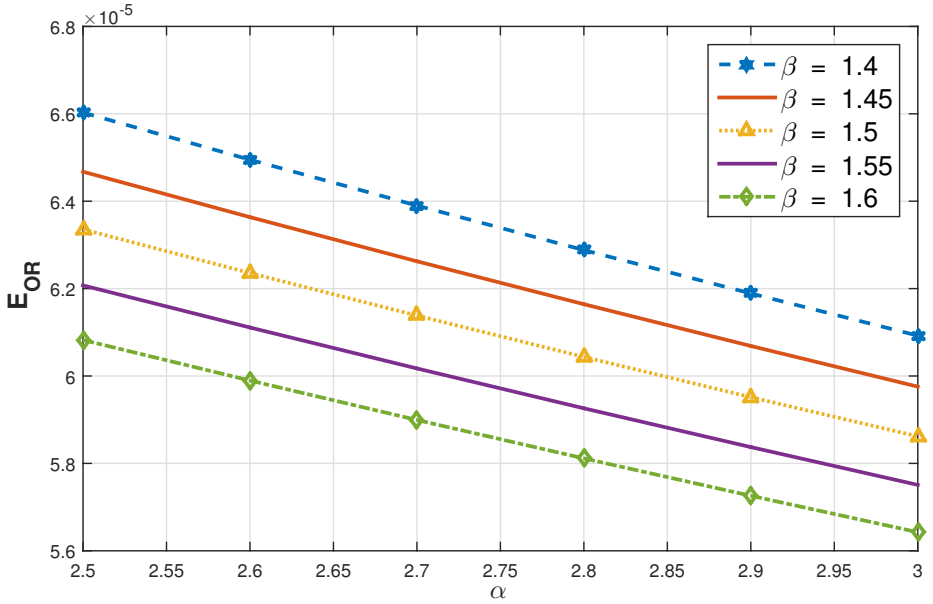

Figure 6. Impact of α and β on E_{OR}.

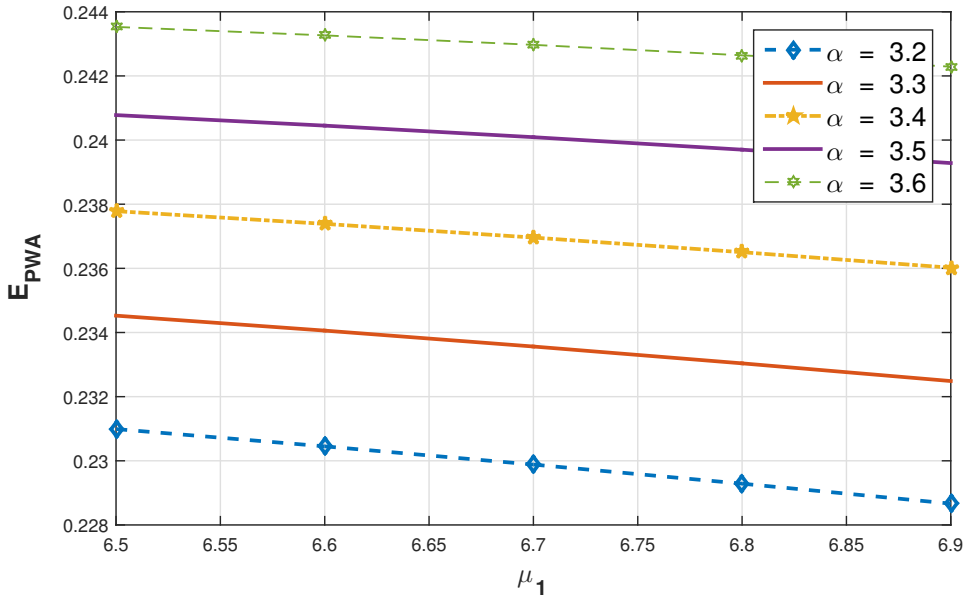

Figure 7. Impact of μ_1 and α on E_{PWA}.

6.3. Analyzing the impact of parameter variations under different MAP

- Table 6 indicates that as the rate α increases, the expected count of online reservations in the PWA (E_{OR}) decreases. Conversely, the expected client count in the PWA (E_{PWA}), the expected number of online clients lost (E_{ONL}), the expected client count in the NPWA (E_{NPWA}), and the expected number of offline clients lost (E_{OFL}) increase. These effects are observed across various distributions. While encouraging prompt arrival of online clients reduces reservation backlog, it may lead to system congestion and higher client loss. Management should optimize α to balance service availability with system capacity.
- Table 7 demonstrates that an escalation in the lead time rate (η) corresponds to an increase in the expected inventory count (E_I) and expected reorder rate (E_R). Conversely, there is a decrease in the expected client count in the PWA (E_{PWA}) and expected client count in the NPWA (E_{NPWA}) due to stock availability. These effects are observed across various distributions. Faster replenishment boosts inventory levels and turnover but also frees up waiting-area capacity, reducing customer congestion. Management should calibrate ordering speed to maintain service levels without inflating holding costs.
- Table 8 illustrates that increasing the premium service rate (μ_1) leads to a decrease in both the expected inventory count (E_I) and the expected client count in the PWA (E_{PWA}). Conversely, there is an increase in E_{OR} due to the faster depletion of inventory, and these effects are observed across various distributions. While speeding up premium service reduces waiting and inventory levels, it may require better coordination with reservation capacity to avoid excess demand or stockouts. Management should align service speed with inventory replenishment and booking controls.
- Table 9 shows the presence of online and offline clients under the Erlang distribution, along with negative and positive correlations on the optimal cost values. Understanding how correlated arrival patterns impact cost under realistic distributions like Erlang helps management fine-tune system parameters and client prioritization strategies to achieve cost efficiency.
- Table 10 presents the occurrence of offline and online clients across various distributions (Erlang, negative correlation, and positive correlation) on some performance measure. Recognizing how different client arrival patterns impact performance metrics allows management to adapt service strategies and resource allocation based on the underlying demand variability.

Table 6. Effects of the MAP with ERG, NCR, and PCR on performance measures.

Client arrival	α	E_{OR}	E_{PWA}	E_{NPWA}	E_{ONL}	E_{OFL}
MAP with ERG	3.2	0.002997	0.004555	0.003460	0.000193	0.000122
	3.3	0.002300	0.005545	0.006770	0.000495	0.000125
	3.4	0.001631	0.008336	0.014405	0.001471	0.000233
	3.5	0.001075	0.014329	0.023855	0.002674	0.000380
MAP with NCR	3.2	0.017270	0.034124	0.022543	0.004862	0.001412
	3.3	0.013856	0.056668	0.023345	0.005747	0.001416
	3.4	0.012140	0.075949	0.023431	0.007033	0.001626
	3.5	0.011589	0.093509	0.025585	0.011574	0.002136
MAP with PCR	3.2	0.000803	0.010454	0.034190	0.005479	0.000877
	3.3	0.000693	0.016218	0.047212	0.01190	0.002922
	3.4	0.000616	0.040793	0.056397	0.018915	0.015442
	3.5	0.000559	0.055609	0.057733	0.035317	0.025118

Table 7. Effects of the MAP with HEPL, ERG, NCR, and PCR.

Client arrival	η	E_I	E_R	E_{PWA}	E_{NPWA}
MAP with HEPL	8.8	0.174310	0.168781	0.002696	0.062596
	8.9	0.178880	0.173710	0.002589	0.061230
	9	0.182961	0.183509	0.002380	0.058440
	9.1	0.186570	0.188353	0.002278	0.057024
MAP with ERG	8.8	2.198943	0.392753	0.103729	0.310223
	8.9	2.199663	0.392899	0.103709	0.309933
	9	2.201199	0.393204	0.103671	0.309373
	9.1	2.202008	0.393363	0.103653	0.309104
MAP with NCR	8.8	3.311371	0.248782	0.011272	0.052170
	8.9	3.353101	0.250233	0.010228	0.051344
	9	3.393691	0.253521	0.008252	0.049870
	9.1	3.434086	0.255516	0.007305	0.049218
MAP with PCR	8.8	1.411492	0.036408	0.040941	0.016812
	8.9	1.414364	0.038132	0.040841	0.013570
	9	1.418938	0.039839	0.040729	0.007100
	9.1	1.420669	0.041524	0.040607	0.003868

Table 8. Effects of the MAP with HEPL, ERG, NCR, and PCR.

Client arrival	μ_1	E_I	E_{OR}	E_{PWA}
MAP with HEPL	6.5	0.347937	0.000153	0.004459
	6.6	0.328664	0.000163	0.004012
	6.7	0.288655	0.000284	0.003304
	6.8	0.233663	0.000386	0.002696
MAP with ERG	6.5	2.202008	0.000034	0.103653
	6.6	2.183082	0.000035	0.103595
	6.7	2.163971	0.000036	0.103535
	6.8	2.144756	0.000037	0.103473
MAP with NCR	6.5	3.311371	0.160274	0.021293
	6.6	3.213376	0.172271	0.016191
	6.7	2.781790	0.181240	0.011272
	6.8	1.877155	0.203387	0.010230
MAP with PCR	6.5	2.387458	0.000459	0.056419
	6.6	2.146230	0.000498	0.052986
	6.7	1.903801	0.000536	0.049265
	6.8	1.661274	0.000573	0.045270

Table 9. Impact of online and offline arrivals on the optimal cost value.

Online/Offline Arrival	MAP with ERG S* s* TEC(S*,s*)	MAP with NCR	MAP with PCR
MAP with ERG	27 8 12.360084	27 7 10.009229	27 8 11.444774
MAP with NCR	27 8 10.933121	27 6 10.088862	28 8 10.449729
MAP with PCR	28 7 14.688656	28 8 12.729172	27 7 13.705695

Table 10. Impact of online and offline client arrivals on certain performance measures.

Online Client Arrival	E_I	E_R	E_{PWA}	E_{NPWA}	E_{OR}
Offline client arrival with ERG					
MAP with ERG	4.621442	0.423645	0.002354	0.118958	0.000013
MAP with NCR	1.258070	0.616003	0.183491	1.783972	0.618471
MAP with PCR	5.319139	0.903415	0.169887	1.024294	0.000231
Offline client arrival with NCR					
MAP with ERG	4.022059	0.383683	0.112221	0.439154	0.000015
MAP with NCR	1.274188	0.479374	0.053132	0.360379	0.338045
MAP with PCR	2.386798	0.725307	0.142748	0.874890	0.000259
Offline client arrival with PCR					
MAP with ERG	3.865297	0.392535	0.110237	0.432563	0.000014
MAP with NCR	1.200502	0.115720	0.041943	0.220459	0.056056
MAP with PCR	2.236128	0.717345	0.138985	0.881299	0.000246

6.4. Analyzing the expected waiting time of a client in the PWA and NPWA

- Figure 8 illustrates that when the maximum capacity of the PWA (N_1) and the OR level (L) increase, the expected waiting time for a client in the PWA $E[W_1]$ also increases. This means that a higher number of clients in the waiting area leads to longer waiting times to receive the service. Management should carefully regulate PWA capacity and reservation limits to avoid excessive congestion, ensuring a balance between premium service availability and acceptable waiting times.
- Compared to Figure 8, the expected waiting time for a client in the PWA ($E[W_1]$) decreases as the service rate μ_1 increases, as shown in Figure 9. Enhancing service speed is an effective strategy to manage congestion and improve client experience in premium service zones.
- Figure 10 illustrates the impact on the expected waiting time for a client in the NPWA ($E[W_2]$) when varying N_2 and λ_2. As the maximum capacity of the NPWA (N_2) and the offline arrival rate (λ_2) increase, $E[W_2]$ also increases. Management should monitor client inflow and adjust service or capacity levels in the NPWA to prevent excessive waiting and ensure smooth service flow for offline clients.
- Compared to Figure 10, the expected waiting time for a client in the NPWA ($E[W_2]$) decreases as the service rate μ_2 increases, as shown in Figure 11. Increasing the non-premium service rate is an effective strategy to improve service efficiency and reduce waiting time for offline clients.
- By setting $s = 3$ and $L = 3$ while keeping all other parameters constant, the optimal expected waiting time of a client in the PWA is analyzed by varying the inventory maximum capacity (S) and the PWA maximum capacity (N_1). In this context, the optimal expected waiting time is attained at $S = 10$ and $N_1 = 8$, as indicated by the bold number in Table 11.
- By setting $s = 3$ and keeping all other parameters constant, the optimal expected waiting time of a client in the NPWA is analyzed by varying the inventory maximum capacity (S) and the NPWA maximum capacity (N_2). In this context, the optimal expected waiting time is achieved at $S = 11$ and $N_2 = 7$, as indicated by the bold number in Table 12.

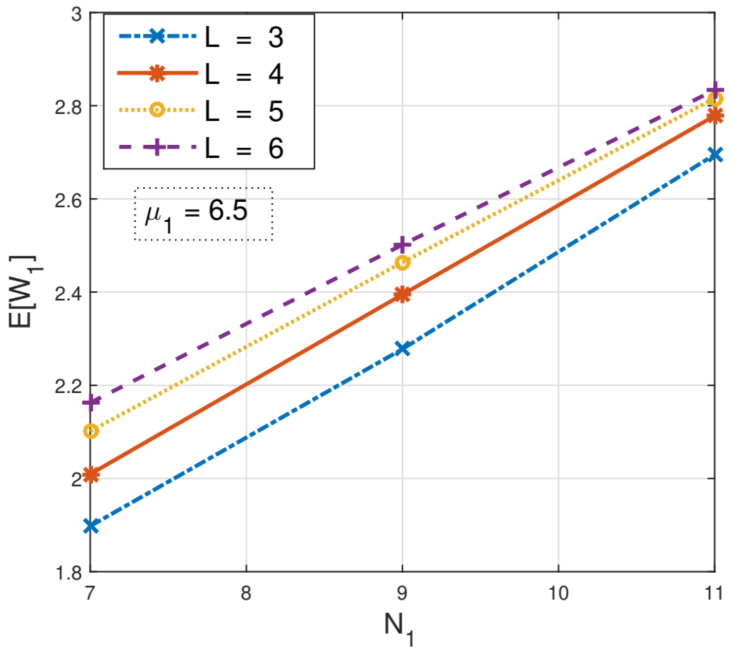

Figure 8. Impact of N_1 vs L on $E[W_1]$.

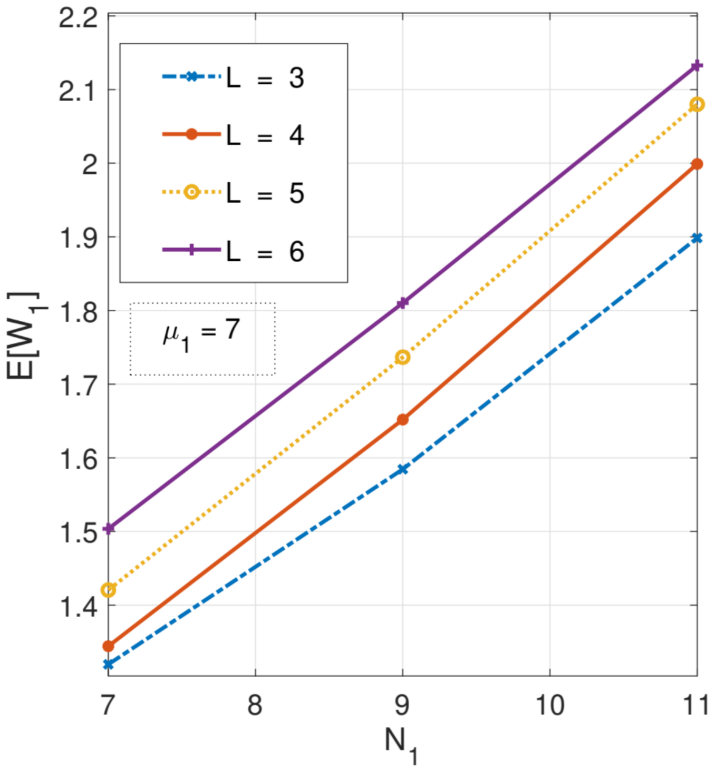

Figure 9. Impact of N_1 vs L on $E[W_1]$.

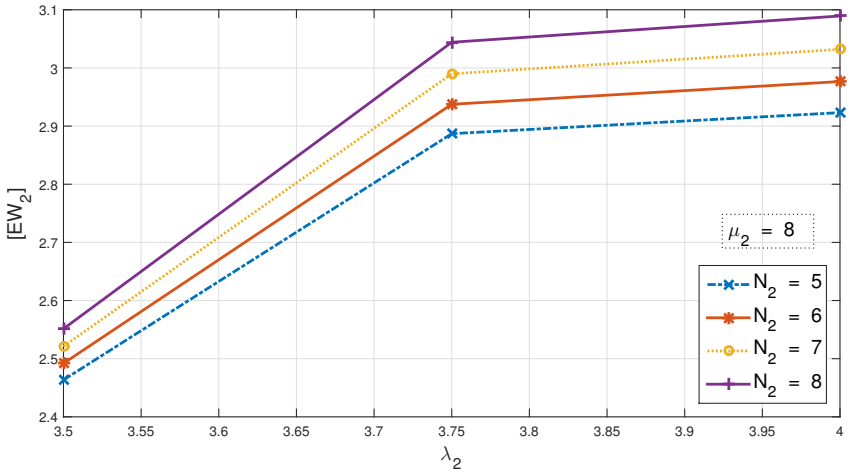

Figure 10. Impact of λ_2 vs N_2 on $E[W_2]$.

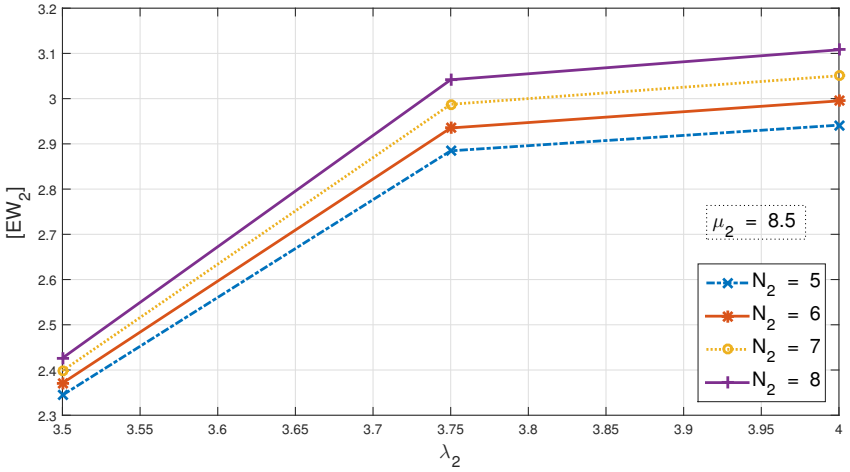

Figure 11. Impact of λ_2 vs N_2 on $E[W_2]$.

Table 11. Optimal expected waiting time of a client in PWA.

S/N_1	6	7	**8**	9	10
8	2.5717	2.5716	2.4009	2.4889	2.5397
9	2.3623	2.3566	2.2873	2.3362	2.3672
10	1.6828	1.6799	**1.6741**	1.6789	1.6795
11	1.9232	1.9229	1.9341	1.9346	1.9359
12	2.1455	2.1399	2.1262	2.1565	2.1750

Table 12. Optimal expected waiting time of a client in NPWA.

S/N_2	5	6	**7**	8	9
8	1.6840	1.6788	1.6774	1.6798	1.6827
9	1.6839	1.6787	1.6772	1.6797	1.6825
10	1.6829	1.6775	1.6760	1.6783	1.6813
11	1.6821	1.6766	**1.6750**	1.6774	1.6808
12	1.6834	1.6781	1.6767	1.6790	1.6818
13	1.6837	1.6784	1.6770	1.6794	1.6823

7. Conclusions

In this study, a stochastic inventory system was examined that addresses the two distinct types of services for a single commodity, two types of clients, and two waiting areas. An online reservation service facility was implemented, enabling clients to pre-book their premium service—a feature encountered during the author's recent visit to a restaurant. The description of the model assumed an infinitesimal generator transition matrix, which was finite; based on this, a steady-state probability vector was calculated. The waiting time of a client in both the PWA and NPWA was analyzed using the Laplace–Stieltjes transform. The impact of parameters on performance measures was analyzed with various distributions for online and offline client arrival patterns. Importantly, the research delved into optimizing the TEC of the system, offering valuable insights for improvement of the efficiency and performance of stochastic inventory management under diverse client types and waiting areas. In future research, we intend to extend this system by incorporating a batch-marked Markovian arrival process (BMMAP) and a phase-type distribution for the service time. This concept can be further expanded by offering an online reservation facility for non-premium services.

Author contributions

Suresh Kumar: Formal Analysis, Investigation, Visualization, and Writing—Review and Editing. Anbazhagan: Conceptualization, Methodology, Writing—Original Draft Preparation, and Supervision. Amutha: Software, Validation and Visualization. G. P. Joshi: Data Curation, Resources and Writing—Review and Editing. Woong Cho: Funding acquisition, Supervision and Project Administration.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

All authors declare no conflicts of interest in this paper.

Acknowledgments

Suresh Kumar, Anbazhagan, and Amutha would like to thank RUSA Phase 2.0 (F 24- 51/2014-U), DST-FIST (SR/FIST/MS-I/2018/17), and the DST-PURSE Second-Phase Program (SR/PURSE Phase 2/38) of the Govt. of India.

References

1. N. Anbazhagan, B. Vigneshwaran, K. Jeganathan, Stochastic inventory system with two types of services, *Int. J. Adv. Appl. Math. Mech.*, **2** (2014), 120–127.
2. A. Krishnamoorthy, N. Anbazhagan, Perishable inventory system at service facilities with *n* policy, *Stoch. Anal. Appl.*, **26** (2007), 120–135. https://doi.org/10.1080/07362990701673096
3. B. Sivakumar, C. Elango, G. Arivarignan, A perishable inventory system with service facilities and batch Markovian demands, *Int. J. Pure Appl. Math.*, **32** (2006), 33–38.
4. K. Jeganathan, A stochastic inventory system with two types of services and a finite population, *Int. J. Math. Appl.*, **3** (2015), 73–81.
5. N. Mathew, V. C. Joshua, A. Krishnamoorthy, A queueing inventory system with two channels of service, *Distributed Computer and Communication Networks: 23rd International Conference, DCCN 2020, Moscow, Russia, September 14–18, 2020, Revised Selected Papers*, **23**, Springer, Cham, (2020), 604–616. https://doi.org/10.1007/978-3-030-66471-8_46
6. K. Jeganathan, S. Selvakumar, N. Anbazhagan, S. Amutha, P. Hammachukiattikul, Stochastic modeling on M/M/1/*n* inventory system with queue-dependent service rate and retrial facility, *AIMS Math.*, **6** (2021), 7386–7420. https://doi.org/10.3934/math.2021433
7. U. U. Kocer, S. Ozkar, A production queueing–inventory system with two-customer and a server subject to breakdown, *Ann. Oper. Res.*, **331** (2023), 1089–1117. https://doi.org/10.1007/s10479-023-05275-9
8. K. Jeganathan, S. Selvakumar, K. Srinivasan, N. Anbazhagan, G. P. Joshi, W. Cho, Two types of service facilities in interconnected stochastic queueing and queueing–inventory system with marked Markovian arrival process, *Ain Shams Eng. J.*, **15** (2024), 102963. https://doi.org/10.1016/j.asej.2024.102963
9. M. Bhuvaneshwari, Stochastic inventory system with multi-optional service and finite source, *Int. J. Math. Appl.*, **6** (2018), 37–44.
10. K. Jeganathan, T. Harikrishnan, S. Selvakumar, N. Anbazhagan, S. Amutha, S. Acharya, R. Dhakal, G. P. Joshi, Analysis of interconnected arrivals on queueing–inventory system with two multi-server service channels and one retrial facility, *Electronics*, **10** (2021), 576. https://doi.org/10.3390/electronics10050576
11. A. Krishnamoorthy, D. Shajin, B. Lakshmy, GI/M/1 type queueing–inventory systems with postponed work, reservation, cancellation and common life time, *Indian J. Pure Appl. Math.*, **47** (2016), 357–388. https://doi.org/10.1007/s13226-016-0192-5

12. D. Shajin, A. Krishnamoorthy, On a queueing–inventory system with impatient customers, advanced reservation, cancellation, overbooking and common life time, *Oper. Res.*, **21** (2021), 1229–1253. https://doi.org/10.1007/s12351-019-00475-3

13. A. Krishnamoorthy, D. Shajin, B. Lakshmy, On a queueing–inventory with reservation, cancellation, common life time and retrial, *Ann. Oper. Res.*, **247** (2016), 365–389. https://doi.org/10.1007/s10479-015-1849-x

14. D. Shajin, A. Krishnamoorthy, A. N. Dudin, V. C. Joshua, V. Jacob, On a queueing–inventory system with advanced reservation and cancellation for the next k time frames ahead: the case of overbooking, *Queueing Syst.*, **94** (2020), 3–37. https://doi.org/10.1007/s11134-019-09631-0

15. O. Baron, O. Berman, L. Wang, Synchronizing travelling and waiting processes: Customer strategy with an online reservation system, *SSRN Preprint*, (2020). https://dx.doi.org/10.2139/ssrn.3536517

16. M. F. Neuts, *Matrix-geometric solutions in stochastic models: An algorithmic approach*, Courier Corporation, (1994).

17. S. R. Chakravarthy, Markovian arrival processes, in: *Wiley Encyclopedia of Operations Research and Management Science*, Wiley, (2010).

18. F. F. Wang, A. Bhagat, T. M. Chang, Analysis of priority multi-server retrial queueing inventory systems with MAP arrivals and exponential services, *Opsearch*, **54** (2017), 44–66. https://doi.org/10.1007/s12597-016-0270-9

19. A. Krishnamoorthy, A. N. Joshua, D. Kozyrev, Analysis of a batch arrival, batch service queueing–inventory system with processing of inventory while on vacation, *Mathematics*, **9** (2021), 419. https://doi.org/10.3390/math9040419

20. P. Manuel, B. Sivakumar, G. Arivarignan, A perishable inventory system with service facilities, MAP arrivals and PH-service times, *J. Syst. Sci. Syst. Eng.*, **16** (2007), 62–73. https://doi.org/10.1007/s11518-006-5025-3

21. G. Hanukov, A queueing–inventory model with skeptical and trusting customers, *Ann. Oper. Res.*, **331** (2023), 763–786. https://doi.org/10.1007/s10479-022-04936-5

22. K. A. K. AlMaqbali, V. C. Joshua, A. Krishnamoorthy, Multi-class, multi-server queueing inventory system with batch service, *Mathematics*, **11** (2023), 830. https://doi.org/10.3390/math11040830

23. A. Melikov, L. Poladova, S. Edayapurath, J. Sztrik, Single-server queueing–inventory systems with negative customers and catastrophes in the warehouse, *Mathematics*, **11** (2023), 2380. https://doi.org/10.3390/math11102380

24. S. Ozkar, U. U. Kocer, Two-commodity queueing–inventory system with two classes of customers, *Opsearch*, **58** (2021), 234–256. https://doi.org/10.1007/s12597-020-00479-0

25. F. F. Wang, Approximation and optimization of a multi-server impatient retrial inventory-queueing system with two demand classes, *Qual. Technol. Quant. Manag.*, **12** (2015), 269–292. https://doi.org/10.1080/16843703.2015.11673381

26. K. Jeganathan, S. Vidhya, R. Hemavathy, N. Anbazhagan, G. P. Joshi, C. Kang, C. Seo, Analysis of M/M/1/n stochastic queueing–inventory system with discretionary priority service and retrial facility, *Sustainability*, **14** (2022), 6370. https://doi.org/10.3390/su14106370

27. V. Vinitha, N. Anbazhagan, S. Amutha, K. Jeganathan, G. P. Joshi, W. Cho, S. Seo, Steady state analysis of impulse customers and cancellation policy in queueing–inventory system, *Processes*, **9** (2021), 2146. https://doi.org/10.3390/pr9122146

28. T. Harikrishnan, K. Jeganathan, S. Redkar, G. Umamaheswari, B. Pattanaik, K. Loganathan, A finite source retrial queueing inventory system with stock-dependent arrival and heterogeneous servers, *Sci. Rep.*, **14** (2024), 30588. https://doi.org/10.1038/s41598-024-81593-7

29. S. Otten, H. Daduna, Stability of queueing–inventory systems with customers of different priorities, *Ann. Oper. Res.*, **331** (2023), 963–983. https://doi.org/10.1007/s10479-022-05140-1

30. D. Shajin, A. Melikov, queueing–inventory system with return of purchased items and customer feedback, *RAIRO-Oper. Res.*, **59** (2025), 1443–1473. https://doi.org/10.1051/ro/2025042

31. S. Dissa, P. V. Ushakumari, Two-commodity queueing–inventory system with random common lifetime, two demand classes and pool of customers, *Heliyon*, **9** (2023), 11. https://doi.org/10.1016/j.heliyon.2023.e21478

32. S. Ozkar, Two-commodity queueing–inventory system with phase-type distribution of service times, *Ann. Oper. Res.*, **331** (2023), 711–737. https://doi.org/10.1007/s10479-022-04865-3

33. D. P. Gaver, P. A. Jacobs, G. Latouche, Finite birth-and-death models in randomly changing environments, *Adv. Appl. Probab.*, **16** (1984), 715–731.

34. J. Abate, W. Whitt, Numerical inversion of Laplace transforms of probability distributions, *ORSA J. Comput.*, **7** (1995), 36–43.