*Journal of Contemporary Applied Mathematics*
*V. 16, No 1, 2026, January-June*
*ISSN 2222-5498, E-ISSN 3006-3183*
*https://doi.org/10.62476/jcam.161.4*

# A Queueing Inventory System with Two Classes of Customers Demanding Processed or Unprocessed Items

Shajeeb P U , Jaison Jacob *, Achyutha Krishnamoorthy

**Abstract.** We study a single-server queueing inventory system with two types of customers, governed by a positive lead time $(s, S)$ inventory policy. Type 1 customers require processed items and join a queue of infinite capacity for service, while Type 2 customers demand unprocessed items and are served instantly. If the excess inventory level is zero, Type 1 customers are blocked from joining the queue, whereas Type 2 customer demands are backlogged up to a certain limit $b$. Arrivals follow independent Poisson processes, with rates $\lambda_1$ for Type 1 and $\lambda_2$ for Type 2. Service time for a Type 1 customer is exponentially distributed with rate $\mu$ and replenishment occurs after an exponentially distributed lead time with rate $\beta$. The system is modeled as a quasi birth and death process (QBD process) and the matrix geometric method is applied to obtain steady-state distributions. We derive the stability condition, the waiting time distribution for Type 1 customers, and the backlog distribution for Type 2 customers. Also, we develop a cost function based on the performance measures. We provide numerical illustrations to analyze the effect of various parameters on performance measures and the cost function.

**Key Words and Phrases**: Queueing Inventory, Matrix-geometric Method, Excess inventory, Processed item, unprocessed item, Type 1 customer, Type 2 customer

**2010 Mathematics Subject Classifications**: 60J28, 90B22, 90B05, 60K25

## 1. Introduction

A queueing inventory system combines the flow of customer service with the availability and restocking of inventory. For service to continue, there must be both a free server and available stock. In this case, service may be blocked by stockouts, which is different from a classical queueing system where delays are only caused by server congestion. Also, unlike a traditional inventory system

*Corresponding author.

 

where demand immediately depletes supply and unmet demand is either lost or backordered, inventory here is consumed through the service process. When the server is busy or stock is unavailable, customers may build queues. So, the performance of a queueing inventory system depends on both the service capacity and the rules for managing inventory.

Simchi-Levi and Sigman [1] were among the first to introduce the notion of inventory systems with positive service times. Around the same period, Melikov and Molchanov [2] independently developed a similar model. Neuts [3, 4] laid the foundation for structured solution techniques in Markovian environments. Bolch et al. [5] , and Chakravarthy et al.[6] made significant advances, while Asmussen [7] and Gross et al. [8] extended general queueing frameworks to address inventory related problems.

Recently, researchers have investigated models incorporating features such as bulk arrivals, backordering and differentiated services. Schwarz et al. [9] analyzed systems with stochastic lead times and partial backordering. Berman and Kim [10] explored inventory management in service facilities using probabilistic modeling, while Berman and Sapna [11] studied systems with perishable inventory and limited backlog capacity.

Queueing inventory systems with different classes of customers are studied by several researchers. For example, Kocer et al. [12] studied systems where customers are served according to priority rules, while Almaqbali et al. [13] studied models involving multiple customer classes and batch service rules across multiple servers. Melikov et al. [14] investigated a mathematical model for inventory management in counter-stream serving systems with stochastic supply and demand, deriving both exact and approximate methods to determine optimal situational inventory policies. A comprehensive survey on queueing inventory models is provided in [15].

This paper introduces a single server queueing inventory model with two types of customers: Type 1 and Type 2, where:

- Type 1 customers demand processed items and join a queue for service. Their service times are exponentially distributed.
- Type 2 customers demand unprocessed items. If excess inventory is positive, they are served instantly, otherwise their demands are backlogged with a maximum limit of $b$ units.

Inventory is managed with an $(s, S)$ inventory policy with exponentially distributed lead times. When inventory drops to the level $s$ or below, an order is placed to replenish to level $S$ and fulfill all outstanding Type 2 demands.

Zhao et al. [16] have examined a similar queueing inventory system with two classes of customers, where the server needs to prioritize one class of customer

over the other. But here we address a queueing inventory system with two types of customers where one requires a positive service time and the other requires negligible service time. In addition, a backlog of customers is also introduced in the absence of excess inventory in the system.

The model discussed in this paper is motivated by real life scenarios such as:

- Healthcare logistics, where critical treatments (Type 1) need time to prepare, while routine care (Type 2) uses ready-to-use supplies.
- Retail systems, where online orders (Type 1) require processing and packaging, while in-store customers (Type 2) take items directly if available, or wait if not.

We model the system as a continuous-time Markov chain with two state variables and analyze its long-term behavior using matrix-geometric method [3]. This allows us to calculate important performance measures such as the average number of customers, inventory levels, waiting times, and the chances of lost demand. We also develop a cost function that combines contributions from these performance measures and analyze its sensitivity over various system parameters.

The rest of the paper is organized as follows: Section 2 presents the mathematical formulation of the model. Section 3 focuses on the steady-state analysis using matrix-geometric methods. Section 4 discusses key performance measures. Sections 5 and 6 examine the waiting time distribution for Type 1 customers and the backlog behavior of Type 2 customers, respectively. Section 7 introduces a cost function based on the performance measures. Section 8 provides numerical illustrations that show how various parameters affect system performance measures. Section 9 discusses the sensitivity of the cost function over the system parameters.

## 2. Mathematical Formulation of the Model

We consider a single-server queueing inventory system where two distinct types of customers arrive at a service station. Each arriving customer is independently classified as either Type 1, who demand a processed item, or Type 2, who require an unprocessed item. Type 1 customers arrive according to a Poisson process with rate $\lambda_1$, while Type 2 customer arrivals follow a Poisson process with rate $\lambda_2$. Both types of customers require exactly one unit from a common inventory pool.

Upon arrival, Type 1 customers join an infinite-capacity queue. When taken for service, each Type 1 customer is provided with one unit of the processed item. The items remaining in inventory are referred to as **excess inventory**. The

service time for Type 1 customers follows an exponential distribution with rate $\mu$. It is assumed that Type 1 arrivals are blocked whenever the excess inventory level is zero.

A Type 2 customer demands one unit of the unprocessed item and requires negligible service time when the excess inventory level is positive. If the excess inventory level drops to zero, their demands are backlogged, with the backlog limited to a maximum of $b$ units. The system employs an $(s, S)$ inventory replenishment policy, where restocking occurs after an exponentially distributed lead time with rate $\beta$. Once replenishment arrives, all accumulated Type 2 backlogs are cleared instantly, and the excess inventory is replenished up to the level $S$.

We model the system as a two dimensional Continuous Time Markov Chain (CTMC) $\mathcal{X} = \{(X(t), Y(t)) ;\ t \geq 0\}$, where $X(t)$ denotes the number of Type 1 customers in the system at time $t$, and $Y(t)$ is defined as follows:

$$Y(t) = \begin{cases} i & \text{if the excess inventory level is } i > 0 \text{ and the server is busy,} \\ i^* & \text{if the excess inventory level is } i > 0 \text{ and the server is idle,} \\ 0 & \text{if the excess inventory is zero and the server is busy,} \\ 0^* & \text{if the excess inventory is zero and the server is idle,} \\ 0_j & \text{if the excess inventory is zero with } j \text{ backlogged Type 2 demands,} \\ & \text{and the server is busy,} \\ 0_j^* & \text{if the excess inventory is zero with } j \text{ backlogged Type 2 demands,} \\ & \text{and the server is idle.} \end{cases}$$

We organize the state space using the first coordinate $n$ as the level index, and define it as follows:

$$\Omega = \bigcup_{n=0}^{\infty} \mathcal{L}(n)$$

where $\mathcal{L}(n)$ denotes the $n^{th}$ level of the process $\mathcal{X}$ defined for $n = 1, 2, 3, \ldots$ as :

$$\mathcal{L}(n) = \{(n, i) \mid i = 0_b^*, 0_b, 0_{b-1}^*, 0_{b-1}, \ldots, 0_1^*, 0_1, 0^*, 0, 1, 2 \ldots, S\}.$$

The level $\mathcal{L}(0)$ corresponding to zero Type 1 customers is given by:

$$\mathcal{L}(0) = \{(0, i) \mid i = 0_b^*, 0_{b-1}^*, \ldots, 0_1^*, 0_1^*, 0^*, 1^*, 2^* \ldots, S^*\}$$

The process undergoes the following transitions:

1. Transition due to the arrival of Type 1 customers:

$$(n, i) \xrightarrow{\lambda_1} (n+1, i) \quad \text{for } n = 0, 1, 2, \ldots; \quad i \in \{1^*, 1, 2^*, 2, \ldots, S^*, S\}$$

2. Transition due to the arrival of Type 2 customers:

   (a) $(n,i) \xrightarrow{\lambda_2} (n,i-1)$ for $n = 0,1,2,\ldots;\quad i = 1,2,\ldots,S$

   (b) $(n,i^*) \xrightarrow{\lambda_2} (n,(i-1)^*)$ for $n = 0,1,2,\ldots;\quad i = 1,2,\ldots,S$

   (c) $(n,0^*) \xrightarrow{\lambda_2} (n,0_1^*)$ for $n = 0,1,2,\ldots$

   (d) $(n,0) \xrightarrow{\lambda_2} (n,0_1)$ for $n = 0,1,2,\ldots$

   (e) $(n,0_j^*) \xrightarrow{\lambda_2} (n,0_{j+1}^*)$ for $n = 0,1,2,\ldots;\quad j = 1,2,\ldots,b-1$

   (f) $(n,0_j) \xrightarrow{\lambda_2} (n,0_{j+1})$ for $n = 0,1,2,\ldots;\quad j = 1,2,\ldots,b-1$

3. **Transitions due to service of Type 1 customers**

   (a) $(n,i) \xrightarrow{\mu} (n-1,i-1)$ for $n = 1,2,3,\ldots;\quad i = 1,2,\ldots,S$

   (b) $(n,0) \xrightarrow{\mu} (n-1,0^*)$ for $n = 1,2,3,\ldots$

   (c) $(n,0_j) \xrightarrow{\mu} (n-1,0_j^*)$ for $n = 1,2,3,\ldots;\quad j = 1,2,3,\ldots,b$

4. **Transitions due to replenishment**

   (a) $(n,i) \xrightarrow{\beta} (n,S)$ for $n = 1,2,3,\ldots;\quad i \in \{0_b^*,0_b,0_{b-1}^*,0_{b-1},\ldots,0_1^*,0_1,0^*,0,1,2,\ldots\}$

   (b) $(0,i) \xrightarrow{\beta} (0,S^*)$ for $i \in \{0_b^*,0_{b-1}^*,\ldots,0_1^*,0^*,1^*,2^*,\ldots\}$

Let $\mathcal{Q}$ be the infinitesimal generator matrix of the CTMC $\mathcal{X}$. Using lexicographical sequence for the states, the generator matrix $\mathcal{Q}$ can be expressed in terms of sub matrices as follows:

$$\mathcal{Q} = \begin{pmatrix} A_{00} & A_{01} & & & & \\ A_{10} & A_1 & A_0 & & & \\ & A_2 & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}.$$

where $A_{00}$ is a square matrix of order $S+b+1$ , $A_{01}$ is of size $(S+b+1)\times(S+2b+2)$, and $A_{10}$ is of size $(S+2b+2)\times(S+b+1)$. The matrices $A_0, A_1$ and $A_2$ are all square matrices of order $S+2b+2$. The $(i,j)^{\text{th}}$ entry of the sub

matrices are as follows:

$$(A_{00})_{ij} = \begin{cases} \lambda_2 & \text{for } j = i-1;\ i = 2,3,\ldots,S+b+1 \\ \beta & \text{for } i = 1,2,\ldots,s+b+1;\ j = S+b+1 \\ -\beta & \text{for } i = j = 1 \\ -(\lambda_2+\beta) & \text{for } i = j = 2,3,\ldots,b+1 \\ -(\lambda_1+\lambda_2+\beta) & \text{for } i = j = b+2, b+3,\ldots,s+b+1 \\ -(\lambda_1+\lambda_2) & \text{for } i = j = s+b+2, s+b+3,\ldots,S+b+1 \\ 0 & \text{otherwise} \end{cases}$$

$$(A_{01})_{ij} = \begin{cases} \lambda_1 & \text{for } j = i+b+1;\ i = b+2, b+3,\ldots,S+2b+2 \\ 0 & \text{otherwise} \end{cases}$$

$$(A_{10})_{ij} = \begin{cases} \mu & \text{for } j = \frac{i}{2};\ i = 2,4,6,\ldots,2b+2 \\ \mu & \text{for } j = i-b-2;\ i = 2b+3, 2b+4,\ldots,S+2b+2 \\ 0 & \text{otherwise} \end{cases}$$

$$(A_1)_{ij} = \begin{cases} \lambda_2 & \text{for } j = i-2;\ i = 3,4,\ldots,2b+2 \\ \lambda_2 & \text{for } j = i-1;\ i = 2b+3, 2b+4,\ldots,S+2b+2 \\ \beta & \text{for } i = 1,2,\ldots,s+2b+2;\ j = S+2b+2 \\ -\beta & \text{for } i = j = 1 \\ -(\mu+\beta) & \text{for } i = j = 2 \\ -(\lambda_2+\beta) & \text{for } i = j = 3,5,7,\ldots,2b+1 \\ -(\mu+\lambda_2+\beta) & \text{for } i = j = 4,6,8,\ldots,2b+2 \\ -(\lambda_1+\lambda_2+\mu+\beta) & \text{for } i = j = 2b+3, 2b+4,\ldots,s+2b+2 \\ -(\lambda_1+\lambda_2+\mu) & \text{for } i = j = s+2b+3, s+2b+4,\ldots,S+2b+2 \\ 0 & \text{otherwise} \end{cases}$$

$$(A_0)_{ij} = \begin{cases} \lambda_1 & \text{for } i = j;\ i = 2b+3, 2b+4,\ldots,S+2b+2 \\ 0 & \text{otherwise} \end{cases}$$

$$(A_2)_{ij} = \begin{cases} \mu & \text{for } j = i-1;\ i = 2,4,6,\ldots,2b+2 \\ \mu & \text{for } j = i-1;\ i = 2b+3, 2b+4,\ldots,S+2b+2 \\ 0 & \text{otherwise} \end{cases}$$

## 3. Steady state analysis

Let $\boldsymbol{\pi} = (\pi_{0_b^*}, \pi_{0_b}, \pi_{0_{b-1}^*}, \pi_{0_{b-1}}, \ldots, \pi_{0_1^*}, \pi_{0_1}, \pi_{0^*}, \pi_0, \pi_1, \pi_2, \ldots, \pi_s, \pi_{s+1}, \ldots, \pi_S)$ be the steady state probability vector of the matrix $A = A_0 + A_1 + A_2$. Then, We have $\boldsymbol{\pi} A = 0$ and $\boldsymbol{\pi e} = 1$. The components of $\boldsymbol{\pi}$ are obtained as follows.

Define the constants:

$$q = \frac{\lambda_2 + \mu}{\lambda_2 + \beta + \mu}, \quad r = \frac{\lambda_2 + \beta + \mu}{\lambda_2 + \mu}, \quad \alpha = \frac{\lambda_2}{\lambda_2 + \mu}, \quad \gamma = \frac{\lambda_2}{\beta}, \quad \delta = \frac{\mu}{\beta}$$

$$\pi_{0^*} = \frac{\mu}{\beta + \lambda_2}\pi_0$$

$$\pi_{0_1} = \alpha\pi_0, \quad \pi_{0_1^*} = \frac{\mu}{\beta + \lambda_2}\pi_{0_1} = \frac{\mu\lambda_2}{(\beta + \lambda_2)(\lambda_2 + \mu)}\pi_0$$

$$\pi_{0_j} = \alpha^j\pi_0, \quad \pi_{0_j^*} = \frac{\mu\lambda_2^j}{(\beta + \lambda_2)^{j+1}(\lambda_2 + \mu)^j}\pi_0, \quad \text{for } j = 2, \ldots, b-1$$

$$\pi_{0_b} = \frac{\lambda_2}{\beta + \mu}\alpha^{b-1}\pi_0, \quad \pi_{0_b^*} = \delta \cdot \pi_{0_b} = \frac{\mu\lambda_2^b}{\beta(\beta + \mu)(\lambda_2 + \mu)^{b-1}}\pi_0$$

$$\pi_j = \left(\frac{\beta}{\beta(S-s) + \lambda_2 + \mu}\right) r^{-j}\pi_0, \quad \text{for } j = 1, \ldots, s$$

$$\pi_j = \left(\frac{\beta}{\beta(S-s) + \lambda_2 + \mu}\right) r^{-s}\pi_0, \quad \text{for } j = s+1, \ldots, S$$

**Normalization condition:**

$$\sum_{j=1}^{b-1}\pi_{0_j} + \sum_{j=1}^{b-1}\pi_{0_j^*} + \pi_{0_b} + \pi_{0_b^*} + \pi_{0^*} + \pi_{0_1} + \pi_{0_1^*} + \pi_0 + \sum_{j=1}^{S}\pi_j = 1$$

Substituting each term in terms of $\pi_0$, we get:

$$\pi_0 \left[1 + \sum_{j=1}^{b-1}\alpha^j + \sum_{j=1}^{b-1}\frac{\mu\lambda_2^j}{(\beta + \lambda_2)^{j+1}(\lambda_2 + \mu)^j} + \frac{\lambda_2}{\beta + \mu}\alpha^{b-1} + \frac{\mu\lambda_2^b}{\beta(\beta + \mu)(\lambda_2 + \mu)^{b-1}} + \frac{\mu}{\beta + \lambda_2}\right.$$

$$\left. + \alpha + \frac{\mu\lambda_2}{(\beta + \lambda_2)(\lambda_2 + \mu)} + \sum_{j=1}^{s}\left(\frac{\beta}{\beta(S-s) + \lambda_2 + \mu}\right) r^{-j} + (S-s)\left(\frac{\beta}{\beta(S-s) + \lambda_2 + \mu}\right) r^{-s}\right] = 1$$

Hence,

$$\pi_0 = \Bigg[1+\sum_{j=1}^{b-1}\alpha^j+\sum_{j=1}^{b-1}\frac{\mu\lambda_2^j}{(\beta+\lambda_2)^{j+1}(\lambda_2+\mu)^j}+\frac{\lambda_2}{\beta+\mu}\alpha^{b-1}+\frac{\mu\lambda_2^b}{\beta(\beta+\mu)(\lambda_2+\mu)^{b-1}}+\frac{\mu}{\beta+\lambda_2}$$

$$+\alpha+\frac{\mu\lambda_2}{(\beta+\lambda_2)(\lambda_2+\mu)}+\sum_{j=1}^{s}\left(\frac{\beta}{\beta(S-s)+\lambda_2+\mu}\right)r^{-j}+(S-s)\left(\frac{\beta}{\beta(S-s)+\lambda_2+\mu}\right)r^{-s}\Bigg]^{-1}$$

### 3.1. Stability Condition

**Theorem 1.** *The process* $\mathcal{X}=\{(X(t),Y(t))\,;t\geq 0\}$ *is stable if and only if*

$$\lambda_1<\left(\frac{1-\sum_{j=1}^{b}\pi_{0_j^*}-\pi_{0^*}}{1-\sum_{j=1}^{b}\pi_{0_j^*}-\sum_{j=1}^{b}\pi_j-\pi_{0^*}}\right)\mu$$

.

*Proof.* The process under consideration is a Level Independent Quasi Birth Death (LIQBD) process. According to Neuts ( see [3]), stability is achieved if and only if $\pi A_0\mathbf{e}<\pi A_2\mathbf{e}$, where $\pi$ is the steady-state distribution of the generator matrix $A$ and $\mathbf{e}$ is a column vector of the ones with dimension $S+1$. Through computation we obtain $\pi A_0\mathbf{e}=\lambda_1\left(1-\sum_{j=1}^{b}\pi_{0_j^*}-\sum_{j=1}^{b}\pi_j-\pi_{0^*}\right)$ and $\pi A_2\mathbf{e}=\mu\left(1-\sum_{j=1}^{b}\pi_{0_j^*}-\pi_{0^*}\right)$. Applying the condition $\pi A_0\mathbf{e}<\pi A_2\mathbf{e}$ and simplify, we obtain the condition as stated in the theorem.

### 3.2. Steady state Probability vector of the system

Suppose that the system is stable. Let $\mathbf{x}$ denote the steady-state probability vector of the generator $Q$. Then we have

$$\mathbf{x}Q=0\quad\text{and}\quad\mathbf{x}e=1$$

Partitioning $\mathbf{x}$ according to the level of the state space as : $\mathbf{x}=(\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_2,...)$. Then, the above set of equations reduce to :

$$x_0A_{00}+x_1A_{10}=0$$

$$x_0A_{01}+x_1A_1+x_2A_2=0$$

$$x_{n-1}A_0 + x_nA_1 + x_{n+1}A_2 = 0 \quad \text{for } n = 2, 3, 4, \ldots$$

Under the assumption that the stability condition holds, $\mathbf{x}$ is obtained as in [3] $\mathbf{x_n} = x_1R^{n-1};\ n = 2, 3, 4, \ldots$, where $R$ is the minimal non-negative solution to the matrix quadratic equation $R^2A_2 + RA_1 + A_0 = 0$ and the boundary equations are given by $x_0A_{00} + x_1A_{10} = 0$, $x_0A_{01} + x_1(A_1 + RA_2) = 0$. The normalizing condition gives

$$\mathbf{x_0}\left[I + K(I - R)^{-1}\right]\mathbf{e} = 1$$

where $K = -A_{01}\left(A_1 + RA_2\right)^{-1}$

## 4. Some performance measures

1. Expected number of Type 1 customers in the system

$$E_N = \sum_{n=0}^{\infty} n\mathbf{x}_n\mathbf{e}$$

2. Expected number of backlogged Type 2 demands

$$E_B = \sum_{n=1}^{\infty}\sum_{j=1}^{b} j\left(\mathbf{x}_n(0_j) + x_n(0_j^*)\right) + \sum_{j=1}^{b} jx_0(0_J^*)$$

3. Expected inventory level in the system

$$E_I = \sum_{n=0}^{\infty}\sum_{j=1}^{S} j\mathbf{x}_n(j) + \sum_{j=1}^{S} j\mathbf{x_0}(j^*)$$

4. Expected reorder rate

$$E_R = \mu\sum_{n=1}^{\infty}\mathbf{x}_n(s+1) + \lambda_2\sum_{n=1}^{\infty}\mathbf{x}_n(s+1)$$

5. Expected loss rate of Type 1 customers

$$E_L^{\text{Type 1}} = \lambda_1\left(\sum_{n=1}^{\infty}\sum_{j=1}^{b}\left[x_n(0_j) + x_n(0_j^*\right] + \sum_{j=1}^{b} x_0(0_j^*)\right)$$

6. Expected loss rate of Type 2 customers

$$E_L^{\text{Type 2}} = \lambda_2\left(\sum_{n=1}^{\infty}\left[x_n(0_b) + x_n(0_b^*\right] + x_0(0_b^*)\right)$$

## 5. Distribution of waiting time of a Type 1 customer

In this section, we derive the waiting time distribution for a Type 1 customer. To this end, we consider a tagged Type 1 customer who joins the queue as the $r^{th}$ customer at the time of arrival. We refer to this position as the rank of the tagged customer. The rank progressively decreases as customers ahead of the tagged customer leave the system after receiving service.

We consider the Markov process $\mathbf{W} = \{(X(t), I(t)) ; t \geq 0 \}$, where $X(t)$ denotes the rank of the tagged customer at time $t$, and $I(t)$ denotes the inventory level at time $t$. We define $X(t) = r$ to mean that there are $r-1$ customers ahead of the tagged customer in the system; thus, the tagged customer is said to have rank $r$. The state space of the process $\mathbf{W}$ is given by

$$\{(n,i);\ n \in \{r, r-1, r-2, \ldots, 1\};\ i \in \{0_b^*, 0_b, 0_{(b-1)^*}, 0_{b-1}, \cdots, 0^*, 0, 1, 2, \cdots, S\}\} \cup \{\Delta\}$$

where $\Delta$ is the absorbing state, indicating that the tagged customer has been selected for service. The possible transitions and corresponding rates are given in the Table 1

Table 1: Transitions and corresponding rates for $\mathbf{W}$

| From | To | Rate | Conditions |
|---|---|---|---|
| $(n,i)$ | $(n-1, i-1)$ | $\mu$ | $n \in \{r, r-1, r-2, \ldots, 2\},\ \ i \in \{1, 2, 3, \ldots, S\}$ |
| $(1,i)$ | $\Delta$ | $\mu$ | $i \in \{1, 2, 3, \ldots, S\}$ |
| $(n,i)$ | $(n,S)$ | $\beta$ | $n \in \{r, r-1, r-2, \ldots, 1\},\ i \in \{0_b^*, 0_b, 0_{(b-1)^*}, 0_{b-1}, \cdots, 0^*, 0, 1, 2, \cdots, S\}$ |
| $(n,i)$ | $(n, i-1)$ | $\lambda_2$ | $n \in \{r, r-1, r-2, \ldots, 1\},\ \ i \in \{0_b^*, 0_b, 0_{(b-1)^*}, 0_{b-1}, \cdots, 0^*, 0, 1, 2, \cdots, S\}$ |

The infinitesimal generator $\mathcal{Q}_{\mathbf{W}}$ of $\mathbf{W}$ is of the form :

$$\mathcal{Q}_{\mathbf{W}} = \begin{pmatrix} T_r & T_r^0 \\ \mathbf{0} & 0 \end{pmatrix}$$

where $T_r$ is a square matrix of size $r(S+2b+2)$, $T^0$ is a column vector of size $r(S+2b+2)$, and $\mathbf{0}$ is a row vector of zeros of size $r(S+b+2)$. We have :

$$T_r = \begin{pmatrix} D & M & & & & \\ & D & M & & & \\ & & \ddots & \ddots & & \\ & & & & D & M \\ & & & & & D \end{pmatrix}.$$

where the sub-blocks $D$ and $M$ are square matrices of size $S+2b+2$, defined as follows:

$$(M_{ij})_{ij} = \begin{cases} \mu & \text{for } j = i-1;\ i = 2,4,6,\ldots,2b+2 \\ \mu & \text{for } j = i-1;\ i = 2b+3, 2b+4,\ldots,S+2b+2 \\ 0 & \text{otherwise} \end{cases}$$

$$(D_{ij})_{ij} = \begin{cases} \lambda_2 & \text{for } j = i-2;\ i = 3,4,\ldots,2b+2 \\ \lambda_2 & \text{for } j = i-1;\ i = 2b+3, 2b+4,\ldots,S+2b+2 \\ \beta & \text{for } i = 1,2,\ldots,s+2b+2;\ j = S+2b+2 \\ -\beta & \text{for } i = j = 1 \\ -(\mu+\beta) & \text{for } i = j = 2 \\ -(\lambda_2+\beta) & \text{for } i = j = 3,5,7,\ldots,2b+1 \\ -(\mu+\lambda_2+\beta) & \text{for } i = j = 4,6,8,\ldots,2b+2 \\ -(\lambda_1+\lambda_2+\mu+\beta) & \text{for } i = j = 2b+3, 2b+4,\ldots,s+2b+2 \\ -(\lambda_1+\lambda_2+\mu) & \text{for } i = j = s+2b+3, s+2b+4,\ldots,S+2b+2 \\ 0 & \text{otherwise} \end{cases}$$

and $T_r^0$ is a column vector of size $r(S+2b+2)$ whose $j^{th}$ entry is given by

$$(T^0)_j = \begin{cases} \mu & \text{for } \ j = (r-1)(2b+S+2)+(b+2),\ (r-1)(2b+S+2)+(b+3),\cdots,r(2b+S+2) \\ 0 & \text{otherwise} \end{cases}$$

Let $p_r$ denote the probability that a tagged customer has rank $r$, which is given by $p_r = x_r\mathbf{e}$, where $\mathbf{e}$ is a column vector of ones of size $b+S+1$. Define $y_r = \frac{x_r}{p_r}$. Now, the probability vector $\alpha_r = e_1 \otimes y_r$, ( where the vector $e_1$ is an r-tuple with a 1 in the first entry and 0 in all other entries ) defines the initial distribution for the Markov chain $\mathbf{W}$.

**Theorem 2.** *The expected waiting time of a general Type 1 customer is given by*

$$E(\mathbf{W}_{\text{Type 1}}) = \sum_{r=1}^{\infty} p_r E_r = \sum_{r=1}^{\infty} p_r \left(-\alpha_r T_r^{-1}\mathbf{e}\right).$$

*Proof.* For a given rank $r$, the waiting time of the tagged customer until it reaches the absorbing state $\Delta$ is the time to absorption in the continuous-time Markov chain $\mathbf{W}$ with generator $T_r$. The expected absorption time for initial distribution $\alpha_r$ is

$$E_r = -\alpha_r T_r^{-1}\mathbf{e}.$$

Since the overall waiting time distribution is a mixture over possible ranks $r$ (with weights $p_r$), the expected waiting time is

$$E(\mathbf{W}_{\text{Type 1}}) = \sum_{r=1}^{\infty} p_r E_r = \sum_{r=1}^{\infty} p_r \left(-\alpha_r T_r^{-1}\mathbf{e}\right).$$

This completes the proof.

## 6. Distribution of Type 2 demands (backlogs)

When there is no inventory, Type 2 demands accumulate as backlogs, though these backlogs cannot exceed the finite capacity $b$. Consider the Markov process $\{B(t); t \geq 0\}$ , where $B(t)$ is the number of unmet Type 2 demands at time $t$. The state space of the process is $\{0, 1, 2, \ldots, b\}$.

The infinitesimal generator $\mathcal{Q}_{\mathcal{B}}$ of the process is given by

$$\mathcal{Q}_B = \begin{pmatrix} -\lambda_2 & \lambda_2 & & & \\ \beta & -(\lambda_2+\beta) & \lambda_2 & & \\ \vdots & & \ddots & \ddots & \\ \beta & & & -(\lambda_2+\beta) & \lambda_2 \\ \beta & & & & -\beta \end{pmatrix}.$$

Let $\xi = (\xi_0, \xi_1, \xi_2, \ldots, \xi_b)$ be the steady-state probability vector associated with the generator matrix $\mathcal{Q}_{\mathcal{B}}$. This vector must satisfy the set of equations

$$\xi \mathcal{Q}_{\mathcal{B}} = 0 \text{ and } \xi e = 1.$$

The stationary condition $\xi \mathcal{Q}_{\mathcal{B}} = 0$ leads to the following set of equations:

1. **For $\xi_0$ :**
$$-\lambda_2 \xi_0 + \beta \sum_{j=1}^{b} \xi_j = 0$$

2. **For $\xi_j$; $j = 1, 2, \ldots, b-1$:**
$$\lambda_2 \xi_{j-1} - (\lambda_2 + \beta)\, \xi_j = 0$$

3. **For $\xi_b$:**
$$\lambda_2 \xi_{b-1} - \beta \xi_b = 0$$

Using the normalizing condition $\xi e = 1$, we derive:

$$\xi_0 = \frac{\beta}{\lambda_2 + \beta}$$

Solving the recurrence relation for $\xi_j$ gives the general form:

$$\xi_j = \left(\frac{\lambda_2}{\lambda_2 + \beta}\right)^j \xi_0 \text{ for } j = 1, 2, \ldots, b-1$$

The final component $\xi_b$ is determined as:

$$\xi_b = \frac{\lambda_2}{\beta}\left(\frac{\lambda_2}{\lambda_2 + \beta}\right)^{b-1}$$

**Theorem 3.** *Under the condition of stability, the steady-state probabilities $\xi_j$ of having $j$ Type 2 backlogs $(j = 0, 1, 2, \ldots, b)$ are given by:*

$$\xi_j = \left(\frac{\beta}{\lambda_2 + \beta}\right)\left(\frac{\lambda_2}{\lambda_2 + \beta}\right)^j \text{ for } j = 0, 1, 2, \ldots, b-1$$

*and*

$$\xi_b = \frac{\lambda_2}{\beta}\left(\frac{\lambda_2}{\lambda_2 + \beta}\right)^{b-1}$$

## 7. Cost function

Based on the performance measure, we define a cost function as follows:

$$K = C_1 E_N + C_2 E_I + C_3 E_R + C_4 E_L^{\text{Type 1}} + C_5 E_L^{\text{Type 2}}$$

, where:

1. $C_1$ = Holding cost of Type 1 customers per unit time
2. $C2$ = Holding cost of inventory
3. $C_3$ = Reordering cost
4. $C_4$ = Cost due to loss of Type 1 customers
5. $C_5$ = Cost due to loss of Type 2 customers

## 8. Numerical Illustrations

In this section, we present numerical results to illustrate the behavior of the proposed queueing-inventory model under various parameter settings. The performance measures examined include the expected number of Type 1 customers in the system ($E_N$), the expected inventory level ($E_I$), the expected reordering rate ($E_R$), and the expected loss rates for both types of customers ($E_L^{\text{Type 1}}$ and $E_L^{\text{Type 2}}$).

Figure 1 shows how $\lambda_1$ affects the system performance measures $E_N, E_R, E_I, E_L^{\text{Type 1}}$ and $E_L^{\text{Type 2}}$. As $\lambda_1$ increases, the expected number of Type 1 customers, $E_N$, increases rapidly, indicating increased system congestion. Both $E_I$ and $E_R$ steadily rise with increasing values of $\lambda_1$, indicating increased stock and replenishment frequency. The expected loss rate for Type 1 customers, $E_L^{Type1}$, rises roughly linearly with $\lambda_1$. For Type 2 customers, on the other hand, it increases at first but then decreases as $\lambda_1$ increases. This trend could be because the faster replenishment effectively addresses their demand.

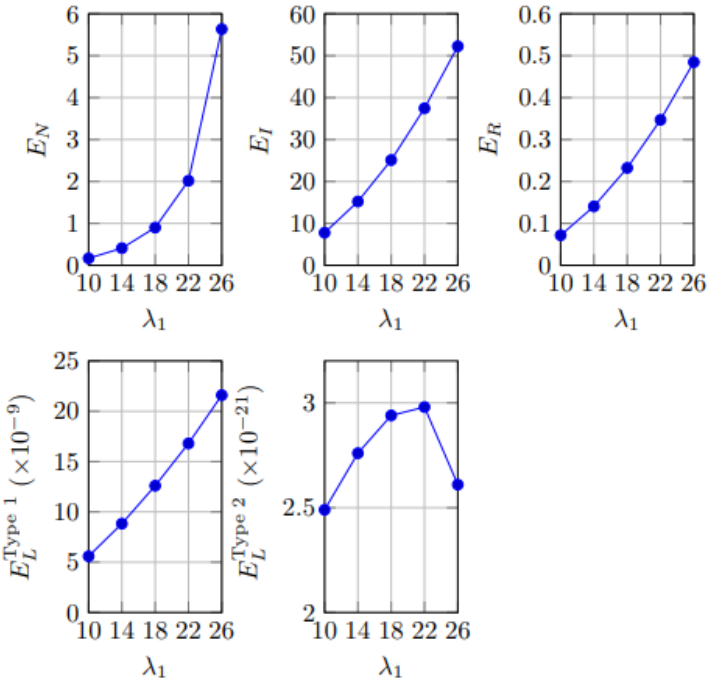


Figure 1: Effect of $\lambda_1$ on performance measures (Fix values: $S = 100, s = 40, b = 30, \lambda_2 = 10, \beta = 20, \mu = 30$ ).

Figure 2 illustrates the effect of the service rate $\mu$ on system performance

measures. With an increase in $\mu$, the performance measures $E_N, E_I$ and $E_R$ show a decreasing trend, which aligns with expectations. However the loss rates $E_L^{\text{Type 1}}$ and $E_L^{\text{Type 2}}$ increase gradually. This suggests that a higher service rate $\mu$ accelerates service and reduces congestion in the system.

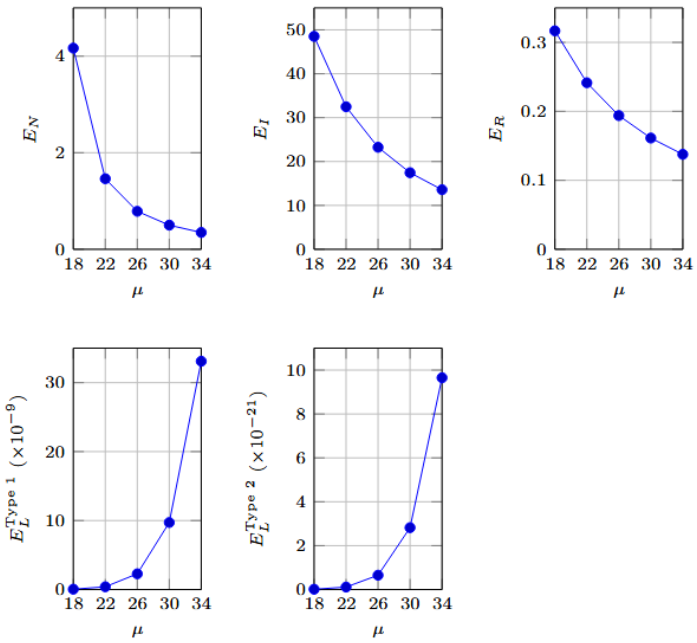


Figure 2: Effect of $\mu$ on performance measures (Fix values: $S = 100, s = 40, b = 30, \lambda_1 = 15, \lambda_2 = 10, \beta = 20$ ).

Figure 3 displays the effect of $\lambda_2$ on $E_N, E_I, E_R, E_L^{\text{Type 1}}$ and $E_L^{\text{Type 2}}$. Figure shows that $\lambda_2$ does not affect queue length. The reorder rate $E_R$ increases gradually, while the average inventory $E_I$ decreases slightly, reflecting a small rise in consumption as $\lambda_2$ grows. The expected loss of Type 1 customers $E_L^{\text{Type 1}}$ increases steadily, though at a very small scale, while the expected loss of Type 2 customers $E_L^{\text{Type 2}}$ grows more sharply, indicating their sensitivity to rising demand. Overall, higher $\lambda_2$ mainly increases reorder activity and loss probabilities, with minimal effect on queue size.

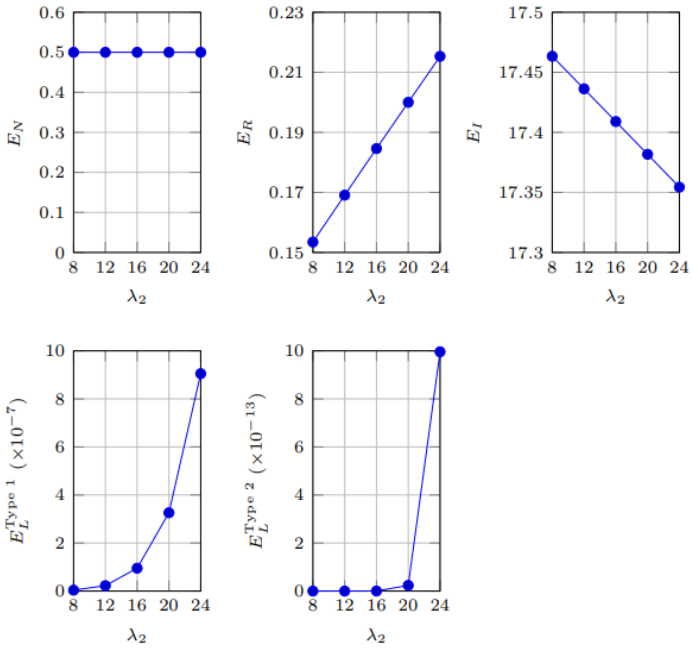


Figure 3: Effect of $\lambda_2$ on performance measures (Fix values: $S = 100, s = 40, b = 30, \lambda_1 = 15, \beta = 20, \mu = 30$ ).

From Figure 4, it is clear that the replenishment rate, $\beta$, has a strong influence on the loss measures while leaving other performance measures largely unaffected. The expected number of customers $E_N$ remains constant, showing that replenishment speed does not change queue length. The average inventory level $E_I$ and reorder rate $E_R$ increase only slightly with $\beta$, indicating a stable inventory position. In contrast, the expected loss of Type 1 customers $E_L^{\text{Type 1}}$ decreases rapidly on a logarithmic scale, as faster replenishment ensures stock availability for their demand. A similar but even sharper decline is seen for Type 2 losses $E_L^{\text{Type 2}}$, which fall drastically as $\beta$ increases. Overall, higher replenishment rates greatly reduce customer losses while keeping other system measures almost unchanged.

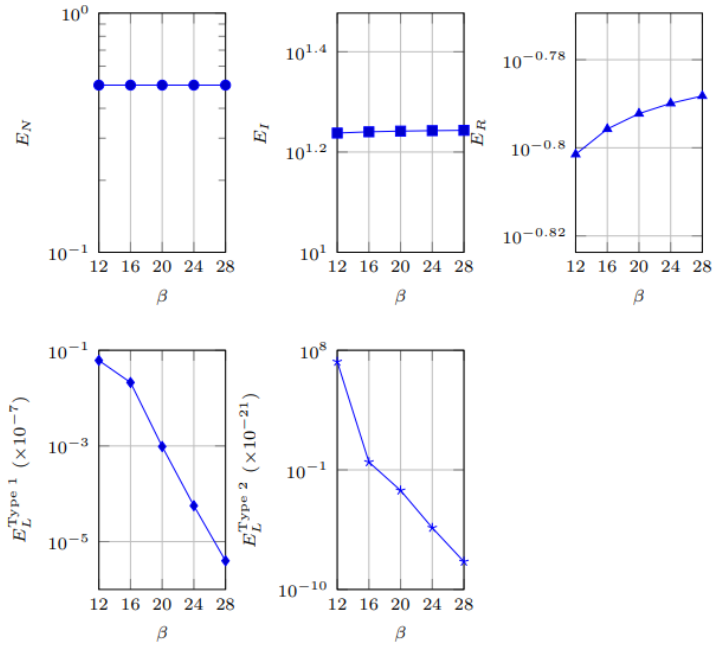


Figure 4: Effect of $\beta$ on performance measures (Fix values: $S = 100, s = 40, b = 30, \lambda_1 = 15, \lambda_2 = 10, \mu = 30$ ).

From Figure 5, we observe that as $b$ increases, the expected backlog $E_B$ rises slowly since more Type 2 customers are allowed to wait. However, the expected losses of both customer types, especially $E_L^{\text{Type 2}}$, decrease sharply because a larger backlog capacity reduces the chance of lost demand. The inventory level $E_I$ and reorder rate $E_R$ remain almost unchanged, showing that backlog mainly influences customer-related measures rather than stock movement. Overall, increasing $b$ improves service for Type 2 customers by reducing losses, but at the cost of slightly higher backlog accumulation.

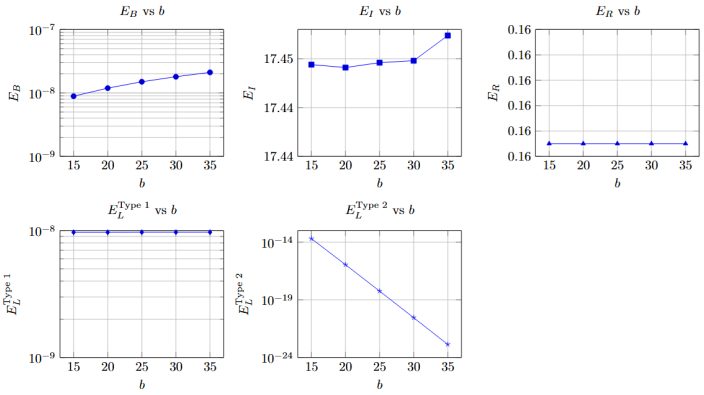


Figure 5: Effect of $b$ on performance measures (Fix values: $S = 100, s = 40, \lambda_1 = 15, \lambda_2 = 10, \beta = 20, \mu = 30$ ).

## 9. Cost Analysis

The cost function $K = C_1E_N + C_2E_I + C_3E_R + C_4E_L^{\text{Type 1}} + C_5E_L^{\text{Type 2}}$ combines contributions from expected queue length, inventory level, reorder rate, and customer loss measures. The sensitivity plots (Figure 6) reveal how the system parameters influence this aggregate cost. As the service rate $\mu$ increases, $K$ decreases sharply due to reduced congestion and losses, highlighting the efficiency gains from faster service. In contrast, variations in the backlog cap $b$ have only a marginal effect on $K$, showing that enlarging the backlog limit beyond a moderate threshold yields diminishing returns. Increasing the arrival rate of Type 1 customers, $\lambda_1$, causes $K$ to grow rapidly, reflecting the higher congestion and customer losses in the system. On the other hand, the cost $K$ decreases slightly with higher Type 2 arrival rate $\lambda_2$, since instantaneous service for these customers avoids significant buildup in the system. Finally, as the replenishment rate $\beta$ rises, $K$ grows gradually with a concave trend, suggesting that while faster replenishment improves availability, it also increases holding and ordering costs, leading to a net moderate increase in $K$.

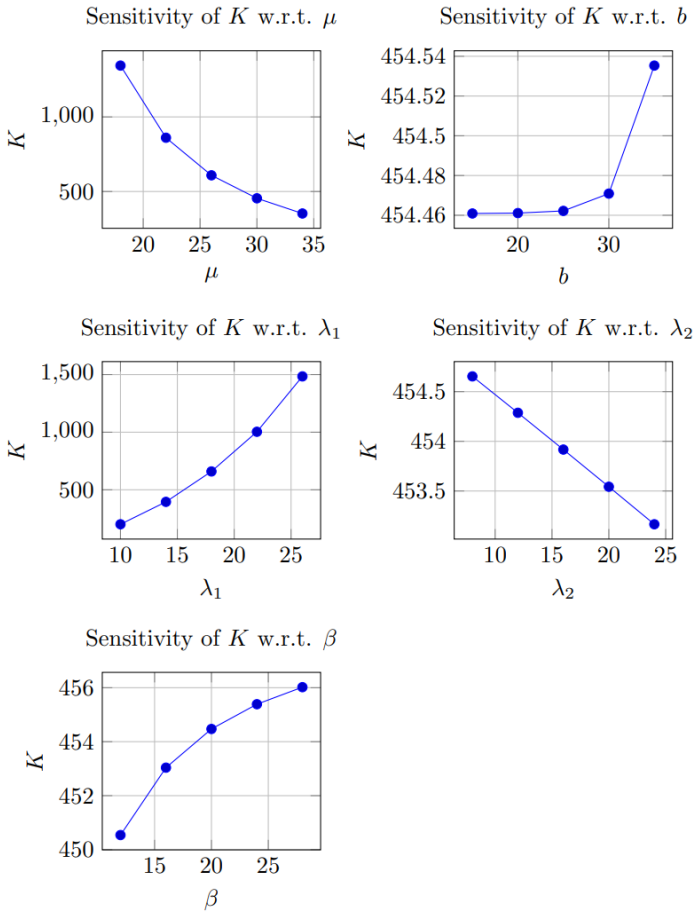


Figure 6: Sensitivity of the cost function $K$ with respect to the parameters $\lambda_1, \lambda_2, \mu, \beta, b$ (Fixed values: $S = 100, s = 40$).

Figure 7 illustrates the variation in total cost as a function of $S$, across various values of $s$. The results show that for a fixed $s$, the cost increases steadily with the values of $S$. This is due to increasing $S$ generally raises the inventory holding expenses. The minimum cost is observed when both $s$ and $S$ are at their lowest tested values, suggesting that adopting a smaller reorder level $s$ in combination with a moderate value of $S$ yields the most cost-efficient policy.

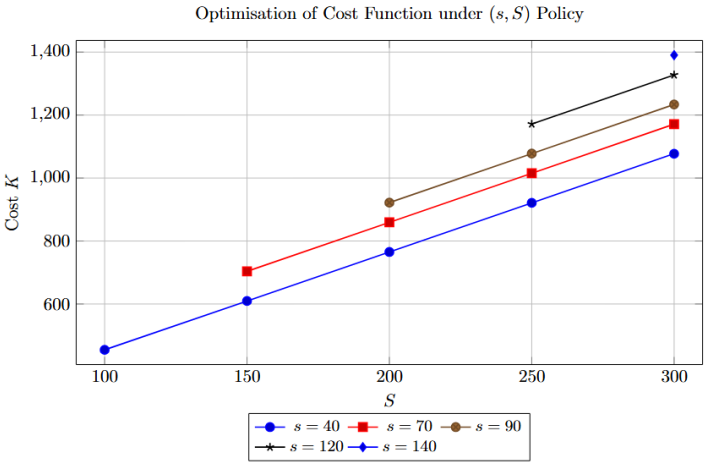


Figure 7: Cost function for different values of $s$ and $S$

These numerical results validate the model and provide practical guidance for selecting optimal inventory control parameters in real-world systems.

## 10. Conclusion

This study examined a queueing-inventory system with two types of customers—those requiring processed items with service time, and those requesting unprocessed items that may be backlogged if inventory is unavailable. The system follows an $(s, S)$ inventory policy, with restocking delays modeled as exponentially distributed lead times. We used a continuous-time Markov chain and the matrix-geometric method to analyze the system's long-term behavior and computed key performance measures such as the expected queue length, inventory level, the waiting time distribution for Type 1 customers, and the backlog distribution for Type 2 customers. A cost function was also developed to combine these measures and illustrated its sensitivity over various parameters of the system.

## References

[1] Simchi-Levi, D., & Sigman, K. (1992). Light traffic heuristic for an M/G/1 queue with limited inventory. *Annals of Operations Research*, 40(1), 371–380.

[2] Melikov, A. Z., & Molchanov, A. A. (1992). Stock optimization in transportation/storage systems. *Cybernetics and Systems Analysis*, 28(3), 484–487.

[3] Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models: An algorithmic approach*. Johns Hopkins University Press.

[4] Neuts, M. F. (1984). Matrix-analytic methods in queuing theory. *European Journal of Operational Research*, 15(1), 2–12.

[5] Bolch, G., Greiner, S., de Meer, H., & Trivedi, K. S. (1998). *Queueing networks and Markov chains: Modeling and performance evaluation with computer science applications*. Wiley.

[6] Chakravarthy, S. R., & Alfa, A. S. (Eds.). (1996). *Matrix-analytic methods in stochastic models*. CRC Press.

[7] Asmussen, S. (2003). *Applied probability and queues* (2nd ed.). Springer.

[8] Gross, D., Shortle, J. F., Thompson, J. M., & Harris, C. M. (2018). *Fundamentals of queueing theory* (5th ed.). Wiley.

[9] Schwarz, M., & Daduna, H. (2006). Queueing systems with inventory management under random lead times and with backordering. *Mathematical Methods of Operations Research*, 64(3), 383–414. https://doi.org/10.1007/s00186-006-0085-1

[10] Berman, O., & Kim, E. (1999). Stochastic models for inventory management at service facilities. *Communications in Statistics: Part C, Stochastic Models*, 15(4), 695–718. https://doi.org/10.1080/15326349908807558

[11] Berman, O., & Sapna, K. P. (2002). Optimal service rates of a service facility with perishable inventory items. *Naval Research Logistics*, 49(6), 464–482.

[12] Kocer, U. U., & Ozkar, S. (2023). A production queueing-inventory system with two customer classes and a server subject to breakdown. *Annals of Operations Research*, 331(2), 1089–1117. https://doi.org/10.1007/s10479-023-05275-9

[13] AlMaqbali, K. A. K., Joshua, V. C., & Krishnamoorthy, A. (2023). Multi-class, multi-server queueing inventory system with batch service. *Mathematics*, 11(4), 830. https://doi.org/10.3390/math11040830

[14] Melikov, A. Z., & Fatalieva, M. R. (1998). Situational inventory in counter-stream serving systems. *Engineering Simulation*, 15(6), 839–848.

[15] Krishnamoorthy, A., Shajin, D., & Narayanan, W. (2021). Inventory with positive service time: A survey. *Queueing Theory*, 2, 201–237. Wiley.

[16] Zhao, N., & Lian, Z. (2011). A queueing-inventory system with two classes of customers. *International Journal of Production Economics*, 129(1), 225–231.

Shajeeb P U
*Department of Mathematics, Cochin University of Science and Technology, Kochi, Kerala-682022, India.*
*Department of Mathematics, Govt. Engineering College, Thrissur, Kerala-680009, India.*
*E-mail:* shajeeb.usman@gmail.com

Jaison Jacob, *
*Department of Mathematics, St. Aloysius College, Elthuruth, Thrissur, Kerala-680611, India.*
*E-mail:* jaisonjacob@staloysiuselt.edu.in

Achyutha Krishnamoorthy
*Centre for Research in Mathematics, CMS College, Kottayam, Kerala-686001, India.*
*Department of Mathematics, Central University of Kerala, Kasaragod-671320, India.*
*E-mail:* achyuthacusat@gmail.com

Received 12 March 2025
Accepted 22 August 2025

*Corresponding author.