Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Queueing Inventory With Routine and Expedited Replenishments

Abdul Rof V
Korambayil Ahamed Haji Memorial Unity Women's College
Achyutha Krishnamoorthy
achyuthacusat@gmail.com

CMS College
Dmitry Kozyrev
V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences

Research Article

Keywords:

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11446929/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Queueing Inventory With Routine and Expedited Replenishments

Abdul Rof V[1†], Achyutha Krishnamoorthy[2,3*†], Dmitry Kozyrev[4,5†]

[1*]Department of Mathematics, Korambayil Ahamed Haji Memorial Unity Women's College, Manjeri, Malappuram, 676122, Kerala, India.
[2]Centre for Research in Mathematics, CMS College, Kottayam, 686001, Kerala, India.
[3]Department of Mathematics, Central University of Kerala, Kasaragod, 671320, Kerala, India.
[4]Department of Mathematics, V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznaya Street, 117997, Moscow, Russia.
[5]Probability Theory and Cyber Security Department, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, 117198, Moscow, Russia.

*Corresponding author(s). E-mail(s): achyuthacusat@gmail.com;
Contributing authors: abdulrof@unitywomenscollege.ac.in;
kozyrev-dv@rudn.ru;
†These authors contributed equally to this work.

Abstract

A queueing inventory system with positive lead time is considered. The system operates with two types of replenishment policies: routine replenishment and expedited replenishment. Customers arrive according to a Poisson process with a rate $\boldsymbol{\lambda}$, and the inventory is served based on an exponential distribution with a parameter $\boldsymbol{\mu}$. When the inventory level drops to the threshold $\boldsymbol{s}$, a replenishment order (routine replenishment) is placed. The time it takes for routine replenishment to be realized follows an exponential distribution with a parameter $\boldsymbol{\beta_s}$. However,if the routine replenishment is not realized when the inventory level reaches a lower threshold $\boldsymbol{r}$ (where $\boldsymbol{r < s}$ and $\boldsymbol{r}$ is close to zero), an expedited replenishment order is placed. The lead time for an expedited replenishment

order follows an exponential distribution with a rate parameter $\boldsymbol{\beta_r, \beta_r >> \beta_s}$. Whenever a replenishment is realized, a specific quantity, denoted as $\boldsymbol{Q_s}$ or $\boldsymbol{Q_r}$, is added to the inventory level, corresponding to routine replenishment/ expedited replenishment, respectively. The system has been studied in detail. Its stability condition is established. The steady-state probabilities of the system is obtained in product-form under the assumption that customers are blocked when there is no inventory available. Various performance measures associated with the system are derived. A cost function is formulated, and optimal reorder levels and reorder quantities for both routine and expedited replenishment policies are estimated using numerical illustrations. Additionally, upper and lower bounds for the cycle time distribution and the distribution of the number of consecutive routine/expedited replenishment between two expedited/routine replenishments have been computed.

Keywords: Inventory; Positive-Lead time; Routine and Expedited Replenishment; Product form solution

1 Introduction

In this paper, we investigate a queueing inventory system with positive service time. The system operates under two types of replenishment policies: routine replenishment and expedited replenishment. Customers arrive according to a Poisson process with a rate λ, and upon arrival, if the system is idle, they collect one inventory item before leaving the system. If the system is busy, the arriving customer joins an infinite capacity queue and waits for service.

When the inventory level drops to a designated threshold value s, a routine replenishment order is placed. Upon realization of this order, the inventory level increases by a specified quantity denoted as Q_s. However, in the case where the routine replenishment is not realized in time and the inventory level further drops to a lower threshold value r (where $r < s$ and r is close to zero), an expedited replenishment order is triggered. This ensures the availability of items in the system. When the expedited order is realized, the inventory level increases by an amount of Q_r.

Both routine replenishment and expedited replenishment follow exponential distributions, each with their respective rates: β_s for routine replenishment and β_r for expedited replenishment. It is to be noted that, under the assumption made in this study, customers are not allowed to enter the system when no inventory is available. This assumption is employed to explore the possibility of finding a product-form solution.

To illustrate the relevance of this study, we consider the scenario of a bulk purchase of medicines from a foreign country. In the routine replenishment case, the items are transported by ship, while an expedited situation may require an air transshipment, which is faster but incurs relatively higher costs.

The objective of this study is to determine the optimal order policy for both routine replenishment (s, Q_s) and expedited replenishment (r, Q_r). This will enable us to effectively manage inventory levels, ensuring a balance between cost efficiency and

maintaining necessary inventory availability.
The rest of this paper is structured as follows:
Section 2 provides a comprehensive review of the related literature, presenting the existing knowledge in this specific area. In Section 3, the model is described in detail. This includes an explanation of the stability condition, steady-state analysis, and various performance measures. Section 4 discusses the upper and lower bounds for cycle time distributions. This section explores the limits or extreme values that the cycle time distributions can take. Section 5 investigates the distributions of the number of consecutive replenishments. In section 6, a cost function is constructed using the performance measures. This section involves the development of a function that quantifies the costs associated with the performance measures discussed in section 3. Finally, in section 7, the results of numerical computations are presented. This section showcases the outcomes of the simulations or calculations performed based on the proposed approach.

Highlights of the present work

- Long run product form solution of the system state probability has been obtained.
- The reorder quantity for expedited replenishment and routine replenishment are assumed to be different, distinguishing it from Boxma *et al* [**?**], where the same reorder quantity was assigned for both types of replenishments.
- Optimum values for both the pairs (r, Q_r) and (s, Q_s) have been determined.
- Even though it is difficult to determine the cycle time distribution (time duration between two successive routine replenishment epochs), upper and lower bounds for the same have been determined.
- The distributions of the time duration between two consecutive routine/expedited replenishments have been discussed.

Notations and Abbreviations

Table 1 Notation

Symbol	Description
$\exp(\mu)$	Exponential distribution with parameter μ.
$E_k(\mu)$	Erlang distribution of order k with parameter μ.
$GE_k(\mu_1, \mu_2, \ldots, \mu_k)$	Generalized Erlang distribution of order k with parameters $\mu_1, \mu_2, \ldots, \mu_k$.
(r, Q)	Queueing–inventory policy with reorder level r and fixed order quantity Q.

2 Literature Review

Inventory systems with emergency replenishment policies have been widely studied in the literature. One of the earliest discussions on emergency ordering was presented by Friend [1], where the concept of emergency replenishment and optimal control policies were examined. Later, Tagaras *et al.* [2] introduced the concept of emergency transshipments to address item shortages and proposed an algorithm to determine optimal control policies. Similarly, Moinzadeh and Aggarwal [3] analyzed an $(S-1, S)$ type inventory system where stocking locations may replenish inventory through either a normal or an expensive emergency supply channel, and a numerical study was conducted to determine the optimal system cost. Johansen and Thorstenson [4] also studied emergency replenishment for a single item and developed an algorithm to minimize long-run average inventory costs.

Several researchers have investigated periodic review inventory systems that incorporate emergency ordering options. Tagaras and Vlachos [5] introduced a periodic review system where regular orders are placed periodically while emergency orders with shorter lead times are allowed to prevent stock-outs and reduce costs. Chiang [6, 7] further developed programming models for periodic review systems in which regular orders are placed at the beginning of an order cycle while emergency orders can be placed during each period. Efficient computational procedures were proposed to determine optimal policy parameters.

Another important line of research considers emergency transshipments, substitution, and collaborative inventory systems. Axsäter [8] examined a single-echelon inventory system with multiple warehouses where emergency lateral transshipments and product substitution are permitted. Netessine *et al.* [9] studied dynamic cross-selling in e-commerce under stochastic demand and analyzed models with and without inventory replenishment, including emergency replenishment and lost-sales scenarios. Ji *et al.* [10] investigated emergency replenishment in collaborative commerce, focusing on the benefits of demand information sharing among a warehouse and multiple retailers in reducing system costs.

Research has also addressed emergency replenishment rules and supply chain coordination. Axsäter [11] proposed an emergency replenishment rule for a single-echelon inventory system with compound Poisson demand, aiming to minimize expected costs while accounting for emergency orders with shorter lead times. Gaukler *et al.* [12] studied inventory systems with uncertain lead times and proposed improved replenishment decisions using real-time tracking technologies such as RFID and GPS. Axsäter [13] further analyzed batch ordering policies with compound Poisson demand and developed decision rules that ensure optimal cost performance under emergency replenishment. Johansson and Olsson [14] considered a two-echelon spare parts inventory system in which emergency replenishment from an external supplier is used when regular supply from a central warehouse cannot satisfy demand.

Several studies have also explored emergency replenishment in specialized applications. Poormoaied *et al.* [15] investigated emergency shipments in a retailer's inventory system and determined optimal base-stock levels, review periods, and emergency shipment timing to minimize total operating costs. Li and Ou [16] studied a firm facing uncertain demand for products composed of multiple components, where shortages

are handled through emergency purchases at higher costs while unsatisfied demand is backordered. Luo and Chen [17] developed a model for blood supply chain management considering both regular and emergency demand for blood products and derived optimal ordering and collection policies for hospitals and blood centers.

In addition to classical inventory models, queueing–inventory systems incorporating emergency replenishment have attracted considerable attention. Jacob *et al.* [18] studied a queueing inventory system with one essential and multiple optional items operating in random environments and analyzed system stability and optimal control policies. Melikov *et al.* [19] proposed a double-sourced hybrid replenishment policy in which regular orders are placed when inventory reaches a reorder level and emergency orders are triggered when the level drops below a threshold. Shajin *et al.* [20] analyzed a multi-server production–inventory system with emergency replenishment occurring when inventory reaches zero and derived stationary distributions and performance measures. Melikov and Ozkar [21] further studied an (s, S) queueing inventory system with two types of replenishment depending on the queue length and developed optimal policies using the matrix-geometric method. Melikov *et al.* [22] also proposed a hybrid replenishment policy in a double-source queueing–inventory system where emergency replenishment restores inventory to the maximum level. Shajin and Melikov [23] examined a single-server queueing–inventory model with feedback customers and analyzed system performance under an (s, S) policy.

A comprehensive survey of queueing inventory models up to 2019 was provided by Krishnamoorthy *et al.* [24], covering various replenishment policies, retrial queues, and production–inventory systems. More recently, Preethi and Sivakumar [25] studied a continuous-review inventory system with stock decay for perishable items where demand follows a compound Poisson process. Using the level-crossing method, they derived steady-state distributions and evaluated important system performance measures.

Boxma *et al* [26] suggest a classical inventory model with normal and emergency orders issued at two different inventory levels. The system follows an (r, Q) policy, where r represents the reorder point and Q represents the order quantity. Their focus is on the stationary analysis of the model. The demand for the inventory item is state dependent; emergency orders are triggered as and when the inventory level drops to zero. Average inventory level, average backlog level, and average waiting time are a few of the performance measures analyzed in the paper.

Next we focus on the contribution in our paper. First we give a detailed description of the model.

3 Model Description

We consider a single server queueing inventory system in which customers arrive according to a Poisson process of rate λ. The inventory is served according to an exponentially distributed time with rate μ. When the stock depletes to the level s, an order referred to as routine order, is placed. The lead time (time required for replenishment of the inventory) follows an exponential distribution with parameter

β_s. If the order is not realized, when the stock level drops to r ($r < s$) a more expensive replenishment order, called an expedited order is placed. Expedited order is also exponential with parameter β_r. When the replenishments realize, an amount of Q_r and Q_s are added to the inventory level for expedited replenishment and routine replenishment respectively. So the maximum level of the inventory is $S = r + Q_r + Q_s$.

Key Assumptions

- Customers do not join the system in the absence of inventory. This assumption is made to explore the possibility of a product-form solution.
- When the replenishment order for a specific type is already pending, it will be ordered the next time s is reached cancelling the existing order.
- Without loss of generality, the reorder levels and replenishment quantities are assumed to satisfy: $0 \leq r < s < Q_r < Q_s$.

3.1 The System State Process

Let $N(t)$ be the number of customers and $I(t)$, the inventory level at time t. Then the process $\{(N(t), I(t)) : t \geq 0\}$ is a continuous time Markov chain with state space $\{(i, j) : i \in \mathbb{N}, j = 0, 1, \cdots S\}$, where $\mathbb{N}$ is the set of non negative integers.
We enumerate the system states of this Markov process in lexicographic order of components. We consider i as the level and j as the phase of the process. The infinitesimal generator matrix Q contains all possible transition rates of the Markov Chain during the infinitesimally short interval $[t, t + h)$. By analyzing the system state transitions, we prove the following theorem:

Theorem 1 *The infinitesimal generator matrix Q of the Markov Chain $\{(N(t), I(t)) : t \geq 0\}$ has the block-tridiagonal structure*

$$Q = \begin{bmatrix} B_1 & A_0 & & & \\ A_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}$$

where

$$B_1 = (b_{ij})_{(S+1)\times(S+1)}$$

with

$$b_{ij} = \begin{cases} -(\beta_s + \beta_r) & ; i = j = 1 \\ -(\lambda + \beta_s + \beta_r) & ; j = i, i = 2, 3, \cdots, r + Q_r + 1 \\ -(\lambda + \beta_r) & ; j = i, i = r + Q_r + 2, \cdots, r + Q_s + 1 \\ -\lambda & ; j = i, i = r + Q_s + 2, \cdots, r + Q_r + Q_s + 1 \\ \beta_r & ; i = 1, 2, \cdots, r + Q_s + 1, j = i + Q_r \\ \beta_s & ; i = 1, 2, \cdots, r + Q_r + 1, j = i + Q_s \end{cases}$$

With

$$A_1 = (c_{ij})_{(S+1)\times(S+1)}$$

, the elements are

$$c_{ij} = \begin{cases} -(\beta_s + \beta_r) & ; i = j = 1 \\ -(\lambda + \mu + \beta_s + \beta_r) & ; j = i, i = 2, 3, \cdots, r + Q_r + 1 \\ -(\lambda + \mu + \beta_r) & ; j = i, i = r + Q_r + 2, \cdots, r + Q_s + 1 \\ -(\lambda + \mu) & ; j = i, i = r + Q_s + 2, \cdots, r + Q_r + Q_s + 1 \\ \beta_r & ; i = 1, 2, \cdots, r + Q_s + 1, j = i + Q_r \\ \beta_s & ; i = 1, 2, \cdots, r + Q_r + 1, j = i + Q_s \end{cases}$$

We have

$$A_0 = (a_{ij})_{(S+1)\times(S+1)}$$

with elements

$$a_{ij} = \begin{cases} \lambda & ; j = i, i = 2, 3, \cdots, S \\ 0 & ; otherwise, \end{cases}$$

and

$$A_2 = (d_{ij})_{(S+1)\times(S+1)}$$

with

$$d_{ij} = \begin{cases} \mu & ; j = i - 1, i = 2, 3, \cdots, S + 1 \\ 0 & ; otherwise \end{cases}$$

Proof First we note that the transition from level i to level j is impossible in the infinitesimal duration of time for $|i - j| > 1$. Further, since the customers are blocked when the inventory is not in stock, the first rows B_1, A_0, A_1, and A_2 are all zeros. The diagonal entries of B_1 and A_1 are all negative. The absolute values of such entries are the sums of the rates of all possible transitions from the corresponding state. The transition from each state occurs as a consequence of;
1. An arriving customer joins the system when at least one unit of inventory is available. The intensity of such transition is λ.
2. A customer already in the system joins for service when the inventory level is positive. The intensity of a service is μ. Thus a customer leaves after service at the rate μ
3. When the inventory level drops to s, a routine replenishment order is placed. The intensity of this order realization is β_s.
4. When the inventory drops to the level r, an expedited order is placed. The intensity of which is β_r
The non-diagonal entries of B_1 and A_1 represent transitions in the Markov chain $\{(N(t), I(t)) : t \geq 0\}$ that do not alter the level i which means that these matrices have entries representing phase changes without consequent level changes. These transitions are possible when either a routine or an expedited replenishment is fulfilled.
The entries of the matrix A_0 define the intensities of transition in the Markov Chain $\{(N(t), I(t)) : t \geq 0\}$ that correspond to an increase in the number of customers in the system due to an arrival. It is a diagonal matrix where all diagonal entries, except the first, are λ.
The entries of A_2 define the intensities of transition in the Markov chain $\{(N(t), I(t)) : t \geq 0\}$ that corresponds to a decrease in the number of customers in the system due to a service completion. The lower diagonal of this matrix has entries μ while all other entries are zeros. This completes the proof of the theorem.

□

3.2 Analysis of the System and Stability Condition

Let $A = A_0 + A_1 + A_2 = (A_{ij})$, where,

$$A_{ij} = \begin{cases} -(\beta_s + \beta_r) & ; i = j = 1 \\ -(\mu + \beta_s + \beta_r) & ; j = i, 2 \le i \le r + Q_r + 1 \\ -(\mu + \beta_s) & ; j = i, r + Q_r + 2 \le i \le r + Q_s + 1 \\ -\mu & ; j = i, r + Q_s + 2 \le i \le r + Q_r + Q_s \\ \mu & ; j = i - 1, 2 \le i \le S + 1 \\ \beta_r & ; 1 \le i \le r + Q_s + 1, j = i + Q_r \\ \beta_s & ; 1 \le i \le r + Q_r + 1, j = i + Q_s \end{cases}$$

Let $\phi = (\phi_0, \phi_1, \cdots, \phi_S)$ be the steady state probability vector of A. We solve the equations $\phi A = 0$ and $\phi e = 1$.
To simplify the expressions let $\beta = \frac{\beta_s + \beta_r}{\mu}$, $\beta_1 = \frac{\beta_s}{\mu}$, $\beta_2 = \frac{\beta_r}{\mu}$.
Also write, $L_1 = (1+\beta)^{Q_r+r}\beta - \beta_1(1+\beta)^r - \beta_1\beta r(1+\beta)^{r-1}$;
$L_2 = (1+\beta_1)^{Q_s-Q_r-r-1}L_1 - \frac{\beta_1}{\beta_2}[(1+\beta)^{Q_s-Q_r-r-1} - (1+\beta_1)^{Q_s-Q_r-r-1}](1+\beta)^r\beta$;
$L_3 = \frac{\beta_1}{\beta_2}\phi_{Q_s-Q_r}/\phi_0$;
$L_4 = (1+\beta_1)^r[L_2(1+\beta_1) + \beta_1 + (1+\beta_1)L_3] - (1+\beta_1)^r[(1+\beta)L_3 + \beta]$;
$L_5 = \phi_{Q_s-Q_r+r+1}/\phi_0$;
$L_6 = L_4 + L_5 - \frac{(1+\beta)^r}{\beta_1\beta_2}[\beta_1{\beta_2}^2 - {\beta_1}^2 + \beta(1+\beta_1)^2]$.

Now we have,

$$\phi_i = \begin{cases} (1+\beta)^{i-1}\beta\phi_0 \; ; 1 \le i \le Q_r \\ [(1+\beta)^{i-1}\beta - \beta_1(1+\beta)^{i-Q_r-1} - \beta_1\beta(i-Q_r-1)(1+\beta)^{i-Q_r-2}\phi_0 \; ; Qe+1 \le i \le r+Q_r+1 \\ \{(1+\beta_1)^{i-Q_r-r-1}L_1 - \frac{\beta_1}{\beta_2}[(1+\beta)^{i-Q_r-r-1} - (1+\beta_1)^{i-Q_r-r-1}](1+\beta)^r\beta\}\phi_0 \; ; r+Q_r+2 \le i \le Q_s \\ \{(1+\beta_1)^{i-Q_s-1}[L_2(1+\beta_1) + \beta_1 + (1+\beta_1)L_3] - (1+\beta_1)^{i-Q_s-1}[(1+\beta)L_3 + \beta]\}\phi_0 \; ; Q_s+1 \le i \le Q_s+r+1 \\ [(1+\beta_1)^{Q_r}(1+\beta)^r\beta - \beta(1+\beta)^r + \beta_1(1+\beta_1)^r]\phi_0 \; ; Q_s+r+2 \le i \le Q_s+Q_r+r \\ \{L_6 + \frac{(1+\beta)^r}{\beta_1\beta_2}[\beta_1{\beta_2}^2 + {\beta_1}^3(1+\beta)](1+\beta)^{i-Q_s-r-1} - [L_5 + \beta(1+\beta_1)](1+\beta_1)^{i-Q_s-r-1}\}\phi_0 \; ; Q_n+1 \le i \le Q_s+Q_r+1 \end{cases}$$

where

$$\phi_0 = \Big\{ (1+\beta)^{Q_r+r+1} + (r\beta_1 - 2 - \beta_2)(1+\beta)^r + \frac{\beta_1}{\beta_2}(1+\beta)^r(1+\beta_1)^{Q_s-Q_r-r} - \frac{\beta_1}{\beta_2}(1+\beta_1)^{r+1} - 1 - \frac{\beta_1}{\beta} + \frac{(1+\beta)^r}{\beta\beta_2}(\beta_2^2 + \beta_1^2 + \beta_1^2\beta)[(1+\beta)^{Q_r} - 1] + L_1(1+\beta_1)\left[\frac{(1+\beta)^{Q_s-Q_r-r-1} - 1}{\beta}\right] + (L_2 + L_3)(1+\beta_1)\left[\frac{(1+\beta_1)^{r+1} - 1}{\beta_1}\right] - L_3(1+\beta)\left[\frac{(1+\beta)^{r+1} - 1}{\beta}\right] - (L_5 + \beta + \beta\beta_1)\left[\frac{(1+\beta)^{Q_r} - 1}{\beta_1}\right] \Big\}^{-1}$$

Next we derive condition for stability of the system.

3.3 Stability Condition

Theorem 2 *The system is stable if and only if* $\lambda < \mu$.

Proof The system is stable if and only if $\phi A_0 e < \phi A_2 e$
(the left drift rate is larger than that for the right drift) ie: $(\phi_1 + \phi_2 + \cdots + \phi_S)\lambda < (\phi_1 +$

$\phi_2 + \cdots + \phi_S)\mu$
which implies $\lambda < \mu$

□

Having established the stability condition, we proceed to derive the system state probability in the long run. In this process we establish that the system admits a product form solution. This means that, despite strong correlation between the number of customers arriving during the lead time and the length of the lead time, the random variables $N(t)$ and $I(t)$ are independent as $t \to \infty$

3.4 Steady State Analysis

To find the steady-state probability vector of the process $\{I(t) : t \geq 0\}$, we examine the queueing inventory system with negligible service time in which no backlog is permitted. The infinitesimal generator of the process $\tilde{W}$ is given by $\tilde{W} = W_{ij}$ where

$$W_{ij} = \begin{cases} -(\beta_s + \beta_r) & ; i = j = 1 \\ -(\lambda + \beta_s + \beta_r) & ; j = i, 2 \leq i \leq r + Q_r + 1 \\ -(\lambda + \beta_s) & ; j = i, r + Q_r + 2 \leq i \leq r + Q_s + 1 \\ -\lambda & ; j = i, r + Q_s + 2 \leq i \leq r + Q_r + Q_s \\ \lambda & ; j = i - 1, 2 \leq i \leq S + 1 \\ \beta_r & ; 1 \leq i \leq r + Q_s + 1, j = i + Q_r \\ \beta_s & ; 1 \leq i \leq r + Q_r + 1, j = i + Q_s \end{cases}$$

Let $\pi = (\pi_0, \pi_1, \cdots, \pi_S)$ be the steady state probability vector of W. To compute π we solve the equations $\pi\tilde{W} = 0$ under the normalizing condition $\pi e = 1$.
To simplify the expressions let $\alpha = \frac{\beta_s + \beta_r}{\lambda}$, $\alpha_1 = \frac{\beta_s}{\lambda}$, $\alpha_2 = \frac{\beta_r}{\lambda}$.
Also let $M_1 = (1+\alpha)^{Q_r+r}\alpha - \alpha_1(1+\alpha)^r - \alpha_1\alpha r(1+\alpha)^{r-1}$;
$M_2 = (1+\alpha_1)^{Q_s-Q_r-r-1}M_1 - \frac{\alpha_1}{\alpha_2}[(1+\alpha)^{Q_s-Q_r-r-1} - (1+\alpha_1)^{Q_s-Q_r-r-1}](1+\alpha)^r\alpha$;
$M_3 = \frac{\alpha_1}{\alpha_2}\pi_{Q_s-Q_r}/\pi_0$;
$M_4 = (1+\alpha_1)^r[M_2(1+\alpha_1) + \alpha_1 + (1+\alpha_1)M_3] - (1+\alpha_1)^r[(1+\alpha)M_3 + \alpha]$;
$M_5 = \pi_{Q_s-Q_r+r+1}/\pi_0$;
$M_6 = M_4 + M_5 - \frac{(1+\alpha)^r}{\alpha_1\alpha_2}[\alpha_1{\alpha_2}^2 - {\alpha_1}^2 + \alpha(1+\alpha_1)^2]$.

Then we have,

$$\pi_i = \begin{cases} (1+\alpha)^{i-1}\alpha\pi_0 \; ; 1 \leq i \leq Q_r \\ [(1+\alpha)^{i-1}\alpha - \alpha_1(1+\alpha)^{i-Q_r-1} - \alpha_1\alpha(i-Q_r-1)(1+\alpha)^{i-Q_r-2}\pi_0 \; ; Qe+1 \leq i \leq r+Q_r+1 \\ \{(1+\alpha_1)^{i-Q_r-r-1}M_1 - \frac{\alpha_1}{\alpha_2}[(1+\alpha)^{i-Q_r-r-1} - (1+\alpha_1)^{i-Q_r-r-1}](1+\alpha)^r\alpha\}\pi_0 \; ; r+Q_r+2 \leq i \leq Q_s \\ \{(1+\alpha_1)^{i-Q_s-1}[M_2(1+\alpha_1) + \alpha_1 + (1+\alpha_1)M_3] - (1+\alpha_1)^{i-Q_s-1}[(1+\alpha)M_3 + \alpha]\}\pi_0 \; ; Q_s+1 \leq i \leq Q_s+r+1 \\ [(1+\alpha_1)^{Q_r}(1+\alpha)^r\alpha - \alpha(1+\alpha)^r + \alpha_1(1+\alpha_1)^r]\pi_0 \; ; Q_s+r+2 \leq i \leq Q_s+Q_r+r \\ \{M_6 + \frac{(1+\alpha)^r}{\alpha_1\alpha_2}[\alpha_1{\alpha_2}^2 + {\alpha_1}^3(1+\alpha)](1+\alpha)^{i-Q_s-r-1} - [M_5 + \alpha(1+\alpha_1)](1+\alpha_1)^{i-Q_s-r-1}\}\pi_0 \; ; Qn+1 \leq i \leq Q_s+Q_r+1 \end{cases}$$

In the above π_0 is given by

$$\pi_0 = \Big\{ (1+\alpha)^{Q_r+r+1} + (r\alpha_1 - 2 - \alpha_2)(1+\alpha)^r + \frac{\alpha_1}{\alpha_2}(1+\alpha)^r(1+\alpha_1)^{Q_s-Q_r-r} - \frac{\alpha_1}{\alpha_2}(1+\alpha_1)^{r+1}$$
$$- 1 - \frac{\alpha_1}{\alpha} + \frac{(1+\alpha)^r}{\alpha\alpha_2}(\alpha_2^2 + \alpha_1^2 + \alpha_1^2\alpha)[(1+\alpha)^{Q_r} - 1] + L_1(1+\alpha_1)\left[\frac{(1+\alpha)^{Q_s-Q_r-r-1} - 1}{\alpha}\right]$$
$$+ (L_2 + L_3)(1+\alpha_1)\left[\frac{(1+\alpha_1)^{r+1} - 1}{\alpha_1}\right] - L_3(1+\alpha)\left[\frac{(1+\alpha)^{r+1} - 1}{\alpha}\right] - (L_5 + \alpha + \alpha\alpha_1)\left[\frac{(1+\alpha)^{Q_r} - 1}{\alpha_1}\right]\Big\}^{-1}$$

Using the components of the probability vector π, we compute the probability vector of the original system. Let $\bar{x} = (x_0, x_1, \cdots)$ be the steady-state probability vector of the system $\{(N(t), I(t)) : t \geq 0\}$ where $x_i = (x_i(0), x_i(1), \cdots, x_i(S)), i \geq 1$
Substituting in the equations, $\bar{x}Q = 0$ and $\bar{x}\bar{e} = 1$ we get
$x_0B_1 + x_1A_2 = 0$
$x_iA_0 + x_{i+1}A_1 + x_{i+2}A_2 = 0, i \geq 0.$
Write $x_0 = \zeta\pi$; $x_i = \zeta(\frac{\lambda}{\mu})^i\pi, i \geq 1$ where ζ is a constant to be determined.
$x_0B_1 + x_1A_2 = \zeta\pi(B_1 + \frac{\lambda}{\mu}A_2) = \zeta\pi\tilde{W} = O;$
$x_iA_0 + x_{i+1}A_1 + x_{i+2}A_2 = \zeta(\frac{\lambda}{\mu})^{i+1}\pi(B_1 + \frac{\lambda}{\mu}A_2) = \zeta\pi(\frac{\lambda}{\mu})^{i+1}\tilde{W} = O.$
Applying the normalizing condition $\bar{x}\bar{e} = 1$, we have
$\zeta + \sum_{i=1}^{\infty}\zeta(\frac{\lambda}{\mu})^i = 1$ which implies
$\zeta = 1 - \frac{\lambda}{\mu}$, since $\lambda < \mu$ by stability condition.

Theorem 3 *Under the necessary and sufficient condition $\rho = \frac{\lambda}{\mu} < 1$, the steady state probability vector $\bar{x} = (x_1, x_2, \cdots, x_n, \cdots)$ of the process $\{(N(t), I(t)) : t \geq 0\}$ is given by,*
$x_0 = (1-\rho)\pi$;
$x_i = \rho^i(1-\rho)\pi, i = 1, 2, \cdots$ where π is the steady state probability vector of the process $\{I(t) : t \geq 0\}$.

Thus we are able to obtain product form solution for the system state distribution except that a compact expression is not available for the inventory level distribution.

With the availability of the system state distribution we can get fairly simple expressions for the system performance matrices.

3.5 Performance Measures

- Average number of customers in the system, $l_s = \frac{\lambda}{\mu-\lambda}$
- Average number of customers in the queue, $L_q = \frac{\lambda^2}{\mu(\mu-\lambda)}$
- Average number of inventory in the system, $I_a = \sum_{j=1}^{r+Q_s+Q_r} j\pi_j$
- Average normal replenishment orders in unit time,
  $R_N = \frac{\lambda\mu}{\lambda+\mu}x_0(s+1) + \sum_{i=1}^{\infty}x_i(s+1)\mu$
  $= \frac{2\lambda\mu}{\lambda+\mu}\alpha(1+\alpha)^s\pi_0$
- Average emergency replenishment orders in unit time,
  $R_E = \frac{\lambda\mu}{\lambda+\mu}x_0(r+1) + \sum_{i=1}^{\infty}x_i(r+1)\mu$
  $= \frac{2\lambda\mu}{\lambda+\mu}\alpha(1+\alpha)^r\pi_0$

- Average number of customers lost in unit time, $A_L = \lambda\pi_0$
- Average number of customers joining the system in unit time, $A_s = \lambda(1 - \pi_0)$
- Average number of customers in the system when the inventory is not in stock $W_0 = L_s\pi_0$

We turn to the derivation of a "cycle time duration" distribution. *Define* the cycle length as the time elapsed between two consecutive hits of the inventory level to s.

4 Upper and Lower Bounds for the Cycle Time Distribution

Since there are two probable types of replenishments in a cycle time, its exact distribution is hard to compute. Therefore, we derive the upper and lower bounds for the distribution. In fact, these bounds turn out to be the LUB and GLB of the distribution. If the routine order isn't realized before the stock level drops to r, an expedited order for replenishment is placed at that epoch. To find the distribution of the time duration elapsed, starting from the inventory level dropping to s, consequent to arrivals and service completions, replenishment of either or both types and finally the inventory level going down to s, we have to consider the following four cases:

Case 1:s to s Transition with both types of replenishments

Since the time elapsed between the replenishments depends on the number of customers in the system, at the epoch at which the stock level falls to s, we have the following sub-cases to be considered. If there are $N(\geq s + Q_s + Q_r)$ customers in the system, future arrivals need not be considered. Otherwise, we have to consider the customers arriving in between. These cases are discussed below separately in a and b .

a) When the number of customers $N \geq s + Q_s + Q_r$

In this scenario, we have to consider the following two cases.

Case (i): Replenishment occurs before the inventory level hits zero.

Consider the transitions of the inventory level,

$s \to s-1 \to s-2 \to \cdots \to r-k+1 \to r-k$; where $0 \leq k \leq r-1$.

The transition occurs due to a service completion with rate μ. At this level, an expedited replenishment occurs with rate β_r. The time until replenishment is Erlang Distribution of order $s-r+k$ with parameter μ, denoted by $E_{s-r+k}(\mu)$. Next, we consider the transitions, $Q_r + r - k \to Q_r + r - k - 1 \to Q_r + r - k - 2 \to \cdots \to Q_r + r - c + 1 \to Q_r + r - c$, where $0 \leq c \leq Q_r + r - s - 1$. At this level, routine replenishment is realized. The time elapsed is Erlang Distribution with parameter μ and order $c-k$, denoted by $E_{c-k}(mu)$. The next transitions are $Q_s + Q_r + r - c \to Q_s + Q_r + r - c - 1 \to Q_s + Q_r + r - c - 2 \to \cdots \to s + 1 \to s$.The time elapsed until it reaches s follows Erlang Distribution with parameter μ of order $Q_s + Q_r + r - c - s$ denoted by $E_{Q_s+Q_r+r-c-s}(\mu)$. So the elapsed time between two consecutive hits of the inventory level s is the convolution of the above three Erlang Distributions. So the cycle length distribution is

$E_{s-r+k}(\mu) * E_{c-k}(\mu) * E_{Q_s+Q_r+r-c-s}(\mu) \sim E_{Q_s+Q_r}(\mu)$

Case (ii): Replenishment occurs after the inventory level hits zero.

First, we consider the transitions of the inventory levels,

$s \to s-1 \to s-2 \to \cdots \to 1 \to 0$.

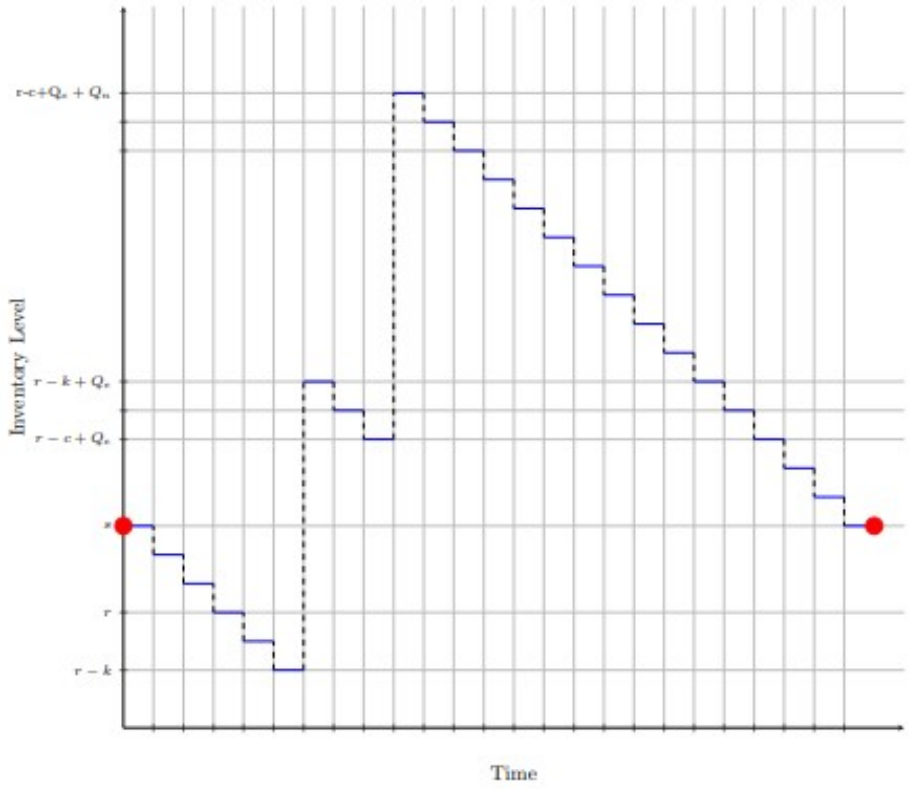


Fig. 1 The Cycle Time Distribution with both types of replenishments in which routine replenishment is realized first.

The time until it reaches zero is Erlang distributed of order s and parameter μ , denoted by $E_s(\mu)$. Now the system becomes empty and stays in the same state for an exponentially distributed duration of time which is exponentially distributed with parameter β_r after which the expedited replenishment is realized. Next, we consider the transitions, $Q_r \rightarrow Q_r - 1; Q_r - 1 \rightarrow Q_r - 2; \cdots Q_r - c + 1 \rightarrow Q_r - c$; where $0 \leq c \leq Q_r - s - 1$. At this time, routine replenishment is realized. This follows Erlang Distribution of order c with parameter μ, denoted by $E_c(\mu)$. The next transitions are, $Q_s + Q_r - c \rightarrow Q_r - c - 1; Q_r - c - 1 \rightarrow Q_r - c - 2; \cdots s + 1 \rightarrow s$. This is Erlang Distribution with parameter μ and of order $Q_s + Q_r - c - s$, denoted by, $E_{Q_s+Q_r-c-s}(\mu)$. So the cycle length has the distribution,
$E_s(\mu) * exp(\beta_r) * E_c(\mu) * E_{Q_s+Q_r-c-s}(\mu) \sim GE_{Q_s+Q_r+1}(\mu, \beta_r)$.
b) When the Number of Customers $N < s + Q_s + Q_r$
In this case, we may have to consider the future arrivals of customers into the system. Once the system has enough customers for continuous service, we stop looking for

future arrivals. For convenience we interchange the level and phase variables in this discussion. We start from the state (s, N). We can see that the cycle time distribution is of the phase type. The initial probability vector of the phase-type distribution is one at the position corresponding to (s, N) and zero elsewhere. The state space of the Markov chain $\{(I(t), N(t)) : t \geq 0\}$ arranged in lexicographic order is
$\{(s, N), (s-1, N-1), \cdots, (0, N-s), (s, N+1), (s-1, N), \cdots, (0, N-s+1), \cdots, (r, N), (r-1, N-1), \cdots, (0, N-r), \cdots, (r+Q_r, N), \cdots (r+Q_r-N, 0), \cdots, (r+Q_r, N+1), \cdots (r+Q_r-N-1, 0), \cdots, (r+Q_r+Q_s, N), \cdots (r+Q_r+Q_s-N, 0), \cdots, (r+Q_r+Q_s, N+1), \cdots (r+Q_s+Q_r-N-1, 0), \cdots, (s+k, N), \cdots, (s, N-k)\}$.
The absorbing states of the Markov Chain are $(s, *)$ where $*$ belongs to $\{0, 1, \cdots, Q_s + Q_r + N - s\}$ which represents a departure epoch. If T is the block matrix of the transitions between the transient states, the distribution of the cycle length is $1 - \alpha e^{Tt} e$ where α is the initial probability vector, which has the entry 1 corresponding to (s, N) and 0 elsewhere.
Note: In the above discussion, it is not essential that expedited replenishment takes place first, followed by routine replenishment; it can be the other way also. However, in order for expedited order placement, the routine replenishment should not take place before reaches the level r.

Case 2: s to s transition without an expedited replenishment in between
First we assume that there are $N > s + Q_s$ customers in the system. The transition from s to s can occur through only a routine replenishment. If the inventory is replenished before reaches the level r, an expedited order is not relevant. Further, when the routine order placed at s is not realized up to the level r, an expedited order may be placed. We consider the case when the routine replenishment order is fulfilled, and before the expedited order is realized, the inventory level reaches the level s again. We discuss both cases in detail.
a) Routine replenishment takes place before reaches the inventory level r
We consider the transitions in the inventory levels:
$s \to s-1 \to s-2 \to \cdots \to s-k+1 \to s-k$, where $0 \leq k \leq s-r+1$.
At this level the routine replenishment is realized. This is an Erlang distribution of order k with parameter μ, denoted by $E_k(\mu)$. The absorbing state of the distribution is $\Delta_5 = \{(s-k+Q_s, N-s+k)\}$. The next transitions are,
$Q_s + s - k \to Q_s + s - k - 1 \to Q_s + s - k - 2 \to \cdots \to s + 1 \to s$, where $0 \leq k \leq s - r + 1$. This is an Erlang Distribution of order $Q_s - k$ with parameter μ, denoted by $E_{Q_s-k}(\mu)$. So the cycle length distribution is
$E_k(\mu) * E_{Q_s-k}(\mu) \sim E_{Q_s}(\mu)$
b) Routine replenishment takes place after the inventory level reaches r
In this case, we consider the following situation. An order for routine replenishment is placed; the replenishment takes place only after the level r is reached (thus expedited order is placed). However, expedited replenishment does not take place until the inventory level again drops to s. Here we have to consider the sub-cases: i) the routine replenishment realizes before reaches the level zero and ii) after reaches level zero.
i) The routine replenishment is realized before the inventory level reaches zero:

The transitions are,
$s \to s-1 \to s-2 \to \cdots r-k+1 \to r-k$, where $0 \le k \le r-1$.
At this level the routine replenishment is realized. The corresponding time duration follows Erlang distribution of order $s-r+k$ with parameter μ, denoted by $E_{s-r+k}(\mu)$. The routine replenishment is realized at this point. This can be denoted by the absorbing state $\Delta_6 = \{(r-k+Q_s, N-r+k)\}$. The next transitions are,
$Q_s + r - k \to Q_s + r - k - 1, Q_s + r - k - 1 \to Q_s + r - k - 2 \to \cdots \to s+1 \to s$, where $0 \le k \le s-r+1$. The consequent time duration follows Erlang distribution of order $Q_s + r - s - k$ with parameter μ, denoted by $E_{Q_s+r-s-k}(\mu)$. So the cycle length distribution is,
$E_{s-r+k}(\mu) * E_{Q_s+r-s-k}(\mu) \sim E_{Q_s}(\mu)$
ii) The routine replenishment is realized after the inventory level reaches zero
In this case, the transitions are,
$s \to s-1 \to s-2 \to \cdots \to 0$.
At this level the system stays at the level zero for an exponentially distributed duration with parameter β_s, after which the routine replenishment is realized. Then the system state reaches the state $(Q_s, N-s)$. From this state the following transitions take place:
$Q_s \to Q_s - 1 \to Q_s - 2 \to \cdots \to s+1 \to s$. The cumulative time on account of these transitions follows the Erlang distribution of order $Q_s - s$ with parameter μ, denoted by $E_{Q_s-s}(\mu)$. The transitions follows the Erlang distribution of order $s-r+k$ with parameter μ, denoted by $E_{s-r+k}(\mu)$.So the cycle length distribution is
$E_s(\mu) * exp(\beta_s) * E_{Q_s-s}(\mu) \sim GE_{Q_s+1}(\mu, \beta_s)$.

Case 3: s to s transition only with expedited replenishment

First, we assume that the number of customers in the system, $N > Q_r + s$. When the routine order placed at s is not realized up to the level r, an expedited order is placed. We consider the case when the expedited replenishment order is fulfilled, and before the routine order is realized, the inventory level reaches s again. By considering all cases similar to *Case 2* we can see that the cycle time has the following distributions.
i) If the expedited replenishment is realized before the inventory level hits zero,the duration of the time has distribution $E_{Q_r}(\mu)$
ii) If the expedited replenishment is realized after the inventory level hits zero:
Then the distribution of the time duration is
$E_s(\mu) * exp(\beta_r) * E_{Q_r-s}(\mu) GE_{Q_r+1}(\mu, \beta_r)$

Case 4: The extreme case: Each transition is followed by an idle time

We shall consider the following cases: i) the inventory level does not reach zero and ii) the inventory level reaches zero during the lead times.
i) Consider the case when the queue is empty in the absence of customers waiting for service when the inventory level drops to s. After an exponential duration of time with parameter λ, a customer arrives. Even after he is served out no customer arrives for some time. So at the completion of his service, the queue becomes empty; the server waits for the next arrival. The process continues until the inventory level reaches $r-k$, where $0 \le k \le r-1$ an expedited replenishment order is realized. The transition s to $s-1$ to $s-2 \cdots$ to $r-k+1$ to $r-k$ follows $GE_2(\lambda, \mu)$. So the distribution of the transition from s to $r-k$ is

$GE_2(\lambda,\mu) * GE_2(\lambda,\mu), \cdots, GE_2(\lambda,\mu)$
which is a generalized Erlang of order $2(s-r+k)$. If there are no customers in the system, the arrival and service continue till the inventory level reaches $Q_r + r - c$. The transitions follow a Generalized Erlang of order $2(c-k)$. At this level, the routine replenishment is realized and the inventory level reaches $r-c+Q_r+Q_s$. From this level, by a sequence of arrivals and services the inventory level returns to s. The transition time duration follows the Generalized Erlang distribution of order $2(r-c+Q_r+Q_s-s)$. So the cycle time has distribution:
$GE_{2s-2r+2k}(\lambda,\mu) * GE_{2c-2k}(\lambda,\mu) * GE_{2r-2c+2Q_r+2Q_s-2s}(\lambda,\mu)$
$\sim GE_{2Q_r+2Q_s}(\lambda,\mu)$
In case ii), when the inventory level hits zero, the system remains empty until an expedited replenishment (or routine replenishment) is realized. So the cycle time has the distribution:
$GE_{2Q_r+2Q_s}(\lambda,\mu) * exp(\beta_r) \sim GE_{2Q_r+2Q_s+1}(\lambda,\mu,\beta_r)$
It is a matter worth mentioning that the *Note* in case 1 is applicable here. The case is illustrated in the below diagram.

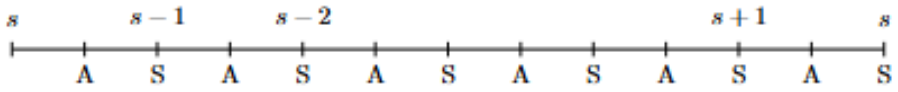


Fig. 2 The extreme case of cycle time distribution. Here A represents a single arrival and S represents the service completion epoch of that arrival

5 Distribution of the Number of Replenishments of Routine/Expedited Type in between two Expedited/Routine Replenishments

It is of independent interest to study the number of consecutive routine replenishments that occur without an expedited replenishment in between as well as the number of consecutive expedited replenishments between two routine replenishments. It is enough to derive the distribution for the first one since the derivation of the other is on the same lines. We determine the distribution of the number of routine replenishments that are placed without any expedited order placements in between.

i) **Distribution of the number of consecutive routine replenishments without an expedited replenishment in between:**

We consider the case when the inventory level drops to s at which epoch a routine order is placed. If the routine order is realized before the inventory level reaches r, the expedited order will not be placed. We look for a situation in which only routine replenishments are realized in consecutive cycles. We define the random variables;

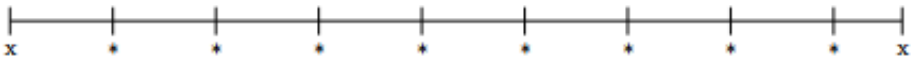


Fig. 3 Diagram showing consecutive routine replenishment epochs between two expedited replenishment epochs. 'x' represents epochs of expedited replenishment, '*' represents epochs of routine replenishment.

$$I = \begin{cases} 0 & \textit{if there is at least one customer in the system waiting for service,} \\ 1 & \textit{otherwise.} \end{cases}$$

$X =$ *The number of service completions before the realization of a routine replenishment*, and U is *the inter arrival time random variables*, which follows the exponential distribution with parameter λ.

Now, consider the random variable $IU + X$. If there is at least one customer waiting for service, then $IU + X = X$, and has distribution $P\{N \geq 1\}exp(\mu)$ where $N = lim_{t\to\infty}N(t)$.
If there is no customer in the system, then, $IU + X > X$ and has distribution, $GE_2(\lambda, \mu)P\{N = 0\}$. Let $Z =$*Number of Routine replenishments before level r is reached in a cycle.* Then Z takes the values $0, 1, 2, \ldots$.
$P\{Z = 0\} =$*Probability that routine replenishment occurs only after inventory level reaches r* $=$*Probability of at least* $(s - r)$ *service completions before routine replenishment* $=$*Probability of k service completions without an arrival and* $(s - r - k)$ *services with an arrival* $= E_k(\mu)(P\{N = k\}) * GE_{2(s-r-k)}(\lambda, \mu)(P\{N = 0\})$ $= E_k(\mu)\rho * GE_{2(s-r-k)}(\lambda, \mu)(1 - \rho) = \rho^k(1 - \rho)^{2s-2r-2k}GE_{2s-2r-k}(\lambda, \mu)$
$P\{Z = 1\} =$*Probability that one routine replenishment occurs before inventory level reaches r* $=$*Probability of at most* $(s - r - 1)$ *service completions before a routine replenishment realization, from the time when the order for the same is placed.* $=$*Probability of k service completions without an arrival,* $(s - r - 1 - k)$ *services with an arrival and a routine Replenishment*
$= (P\{N = k\})E_k(\mu) * (P\{N = 0\}GE_{2(s-r-k-1)}(\lambda, \mu)) * exp(\beta_s)$
$= \rho^k(1 - \rho)^{2s-2r-2k-2}GE_{2s-2r-k-1}(\lambda, \mu, \beta_s)$

$P\{Z = l\}$ is the $l - fold$ convolution of the distribution of $Z = 1$.

□

The distribution of the number of consecutive expedited replenishments without a routine replenishment in between can be computed in a similar way.

6 Optimization Problem

Based on the performance measures discussed in *section 3.5*, we construct the following cost function:
$F(r, s, Q_s, Q_r) = K_n.R_N + C_n.Q_s + K_e.R_E + C_e.Q_r + H_c.I_a + W_0.W_c + A_L.L_c$ where,
K_n is the fixed cost and C_n price per unit of items for routine replenishment,
K_e is the fixed cost and C_e price per unit of items for expedited replenishment,
H_c is the holding cost per unit,
W_c is the waiting cost per customer per unit time, and
L_c is the revenue loss per customer loss.

7 Numerical Illustration

To showcase the system's behaviour concerning the parameters, we will provide numerical examples while assuming that the stability conditions are met. We will be able to analyse and understand how the system operates under different parameter configurations and observe the impact of these parameters on various performance measures. Further, we can determine the optimal pairs (r, Q_r) and (s, Q_s) numerically.

7.1 Effect of Parameters

The arrival rate, service rate, and replenishment rates have profound effects on system efficiency, inventory management, customer experience, and system stability. Managing these parameters effectively is crucial in ensuring a well-functioning inventory queueing system that meets customer expectations and minimizes stock-outs. We check the effect of the parameters λ, μ, β_s and β_r on various performance measures of the system.

7.1.1 Effect of arrival rate λ

Table 1 presents the changes in performance measures as the arrival rate λ varies.The number of customers both in the system and in the queue increases as the arrival rate increases. However, the arrival of customers into the system does not impact the inventory level. As the number of customers increases, the demand for inventory also increases, leading to more frequent replenishments of both routine and expedited stocks. Consequently, the rate of customer loss and waiting time for customers will likely increase due to the higher inventory needs.

7.1.2 Effect of service rate μ

*Table 2*presents the changes in performance measures as the arrival rate μ varies. As the service rate increases, the number of customers both in the system and the queue naturally decreases. Conversely, as the on-hand inventory decreases, the average number of both regular and urgent replenishments increases. The loss of customers also increases with a higher service rate. Moreover, the waiting time of customers in the absence of inventory tends to increase slightly.

Table 2 Effect of the parameter λ. Fix $\mu = 7$, $\beta_s = 2$, $\beta_r = 5$, $r = 3$, $s = 8$, $Q_r = 12$, $Q_s = 20$.

λ	L_s	L_q	I_a	R_N	R_E	A_L	W_0
1	0.1667	0.0238	22.1419	0.0154	0.0044	0.0003	0.0001
2	0.4000	0.1143	22.1419	0.0274	0.0078	0.0006	0.0003
3	0.7500	0.3214	22.1419	0.0370	0.0105	0.0009	0.0007
4	1.3333	0.7619	22.1419	0.0449	0.0128	0.0013	0.0017
5	2.5000	1.7857	22.1419	0.0515	0.0146	0.0016	0.0039
6	6.0000	5.1429	22.1419	0.0570	0.0162	0.0019	0.0113

Table 3 Effect of the parameter μ. Fix $\lambda = 3$, $\beta_s = 2$, $\beta_r = 5$, $r = 3$, $s = 8$, $Q_r = 12$, $Q_s = 20$.

μ	L_s	L_q	I_a	R_N	R_E	A_L	W_0
4	3.0000	2.2500	23.5181	0.0101	0.0013	0.0000	0.0002
5	1.5000	0.9000	23.0444	0.0179	0.0033	0.0001	0.0003
6	1.0000	0.5000	22.5856	0.0271	0.0064	0.0004	0.0005
7	0.7500	0.3214	22.1419	0.0370	0.0105	0.0009	0.0007
8	0.6000	0.2250	21.7132	0.0473	0.0155	0.0018	0.0010
9	0.5000	0.1667	21.2986	0.0576	0.0211	0.0032	0.0013

7.1.3 Effect of routine replenishment rate β_s

Table 3 presents the changes in performance measures as the routine replenishment rate β_s varies. When the routine replenishment rate increases, it results in increased inventory, reduced routine replenishments, minimal expedited replenishments, decreased customer loss, and almost eliminated waiting time for customers. Further, The number of customers in the queue and the system depends solely on the arrival rate and service rate, and not on the replenishment rates.

Table 4 Effect of the parameter β_s. Fix $\lambda = 3$, $\mu = 5$, $\beta_r = 8$, $r = 3$, $s = 8$, $Q_r = 12$, $Q_s = 20$.

β_s	L_s	L_q	I_a	R_N	R_E	A_L	W_0
2	1.5000	0.9000	23.0514	0.0179	0.0033	0.0000	0.0001
3	1.5000	0.9000	23.8414	0.0070	0.0007	0.0000	0.0000
4	1.5000	0.9000	24.2517	0.0031	0.0002	0.0000	0.0000
5	1.5000	0.9000	24.5004	0.0015	0.0000	0.0000	0.0000
6	1.5000	0.9000	24.6668	0.0007	0.0000	0.0000	0.0000
7	1.5000	0.9000	24.7858	0.0004	0.0000	0.0000	0.0000

7.1.4 Effect of expedited replenishment rate β_r

Table 4 presents the changes in performance measures as the expedited replenishment rate β_r varies. Increasing the expedited replenishment rate can have positive effects, such as a slight increase in the inventory level, negligible loss of customers, and negligible waiting time.

Table 5 Effect of the parameter β_r. Fix $\lambda = 3$, $\mu = 5$, $\beta_s = 2$, $r = 3$, $s = 8$, $Q_r = 12$, $Q_s = 20$.

β_r	L_s	L_q	I_a	R_N	R_E	A_L	W_0
3	1.5000	0.9000	23.0357	0.0179	0.0033	0.0003	0.0005
4	1.5000	0.9000	23.8414	0.0179	0.0033	0.0002	0.0004
5	1.5000	0.9000	23.0444	0.0179	0.0033	0.0001	0.0003
6	1.5000	0.9000	23.0472	0.0179	0.0033	0.0001	0.0002
7	1.5000	0.9000	23.0495	0.0179	0.0033	0.0001	0.0002
8	1.5000	0.9000	23.0514	0.0179	0.0033	0.0000	0.0001

7.2 Optimal Cost: Illustrative Examples

We compute the pairs (r, Q_r) and (s, Q_s) that minimize the system cost F discussed in *section 3.6*, by assuming values to the relevant costs and parameters.

7.2.1 Optimal pair (r, Q_r)

Fixing the parameters and the values of s and Q_s, we find the optimal pair (r^*, Q_r^*) such that

$$F(r^*, s, Q_s, Q_r^*) = \min_{r,Q_r} F(r, s, Q_s, Q_r).$$

Corresponding to the input values, from *Table 5*, the optimal pair is $(r^*, Q_r^*) = (10, 16)$ and the minimum cost is \$649.10.

Table 5: For
$\lambda = 3,\ \mu = 4,\ \beta_s = 1,\ \beta_r = 2,\ s = 14,\ Q_s = 32$
$(K_n, C_n, K_e, C_e, W_c, L_c, H_c) = (\$1000,\ \$8,\ \$2500,\ \$20,\ \$20,\ \$30,\ \$1.5)$

7.2.2 Optimal pair (s, Q_s)

Fixing the parameters and values of r and Q_r, we find the optimal pair (s^*, Q_s^*) such that

$$F(r, s^*, Q_s^*, Q_r) = \min F(r, s, Q_s, Q_r).$$

From *Table 6*, the optimal pair $(s^*, Q_s^*) = (8, 26)$ and the minimum cost \$749.80.

Table 6: For
$\lambda = 3,\ \mu = 4,\ \beta_s = 1,\ \beta_r = 3,\ r = 6,\ Q_r = 24$
$(K_n, C_n, K_e, C_e, W_c, L_c, H_c) = (\$1000,\ \$8,\ \$2500,\ \$20,\ \$20,\ \$30,\ \$1.5)$

Table 6 Total cost for different values of r and Q_r. The minimum cost occurs at $(r^*, Q_r^*) = (10, 16)$ with cost \$649.10.

$Q_r \rightarrow$ $r \downarrow$	16	18	20	22	24	26	28	30
0	696.04	711.00	736.19	767.84	803.60	841.97	882.01	923.11
2	674.18	698.10	729.02	764.33	802.43	842.30	883.30	925.02
4	661.17	690.90	725.51	763.17	802.77	843.60	885.21	927.32
6	653.89	687.36	724.34	763.51	804.07	845.51	887.51	929.88
8	650.31	686.18	724.67	764.81	805.98	847.81	890.07	932.59
10	649.10	686.50	725.96	766.72	808.28	850.37	892.78	935.41
12	649.40	687.79	727.87	769.02	810.84	853.08	895.60	938.29

Table 7 Total cost for different values of s and Q_s. The minimum cost occurs at $(s^*, Q_s^*) = (8, 26)$ with cost \$749.80.

$Q_s \rightarrow$ $s \downarrow$	26	28	30	32	34	36	38	40
8	749.80	767.13	784.47	801.84	819.22	836.61	854.02	871.43
10	750.35	767.64	784.95	802.28	819.64	837.01	854.39	871.79
12	751.21	768.43	785.69	802.98	820.29	837.63	854.98	872.35
14	752.55	769.67	786.85	804.07	821.32	838.60	855.90	873.22
16	754.64	771.62	788.66	805.77	822.92	840.11	857.33	874.58
18	757.90	774.65	791.50	808.42	825.42	842.47	859.57	876.71
20	763.01	779.39	795.92	812.57	829.32	846.16	863.06	880.03
22	770.99	786.80	802.84	819.06	835.43	851.92	868.53	885.22

8 Conclusions

The present work has provided a comprehensive analysis of a queueing inventory system with two types of replenishment policies: routine replenishment and expedited replenishment. The system operates under a Poisson arrival process for customers and an exponential distribution for serving inventory. Replenishment orders, with positive lead times, are placed when the inventory level drops to specific thresholds. This study has focused on calculating the steady-state probabilities in product form, assuming customer blocking when inventory is unavailable. Additionally, various performance measures have been explored, including the numerical analysis of a cost function. Optimal reorder levels and reorder quantities for both types of replenishment policies, have been estimated using numerical illustrations. Furthermore, the distribution of the time between two consecutive routine/expedited replenishments and the upper and lower bounds for the cycle time distribution, have been analysed. Building upon these findings, the proposed extension of studying fluid inventory will be an invaluable contribution to the field. By considering fluid inventory, where the counting of individual items is no longer meaningful, a more continuous and dynamic perspective

on inventory management can be achieved. This extension aims to explore alternative modelling approaches that accurately capture the behaviour and performance of inventory systems operating in a continuous manner. By studying fluid inventory with both routine and expedited replenishment policies, we can gain deeper insights into the dynamics of inventory, allowing for more flexible and adaptive approaches. This research will contribute to improving the efficiency and effectiveness of inventory management strategies.

References

[1] Friend, J.K.: Stock control with random opportunities for replenishment. J. Oper. Res. Soc. **11**(3), 130–136 (1960)

[2] Tagaras, G., Cohen, M.A.: Pooling in two-location inventory systems with non-negligible replenishment lead times. Manag. Sci. **38**(8), 1067–1083 (1992)

[3] Moinzadeh, K., Aggarwal, P.K.: An information-based multi-echelon inventory system with emergency orders. Oper. Res. **45**(5), 694–701 (1997)

[4] Johansen, S.G., Thorstenson, A.: An inventory model with poisson demands and emergency orders. Int. J. Prod. Econ. **56**, 275–289 (1998)

[5] Tagaras, G., Vlachos, D.: A periodic review inventory system with emergency replenishments. Manag. Sci. **47**(3), 415–429 (2001)

[6] Chiang, C.: A note on optimal policies for a periodic inventory system with emergency orders. Comput. Oper. Res. **28**(2), 93–103 (2001)

[7] Chiang, C.: Optimal replenishment for a periodic review inventory system with two supply modes. Eur. J. Oper. Res. **149**(1), 229–244 (2003)

[8] Axsäter, S.: Evaluation of unidirectional lateral transshipments and substitutions in inventory systems. Eur. J. Oper. Res. **149**(2), 438–447 (2003)

[9] Netessine, S., Savin, S., Xiao, W.: Revenue management through dynamic cross-selling in e-commerce retailing. Oper. Res. **54**(5), 893–913 (2006)

[10] Ji, S.F., Du, X.L., Huang, X.Y.: Emergency replenishment model and analysis of distribution systems in collaborative commerce. J. Northeastern Univ. (Nat. Sci.) **28**(5), 753 (2007)

[11] Axsäter, S.: A heuristic for triggering emergency orders in an inventory system. Eur. J. Oper. Res. **176**(2), 880–891 (2007)

[12] Gaukler, G.M., Özer, Ö., Hausman, W.H.: Order progress information: Improved dynamic emergency ordering policies. Prod. Oper. Manag. **17**(6), 599–613 (2008)

[13] Axsäter, S.: An improved decision rule for emergency replenishment. Int. J. Prod. Econ. **157**, 313–317 (2014)

[14] Johansson, L., Olsson, F.: Age-based inventory control in a multi-echelon system with emergency replenishments. Eur. J. Oper. Res. **265**(3), 951–961 (2018)

[15] Poormoaied, S., Atan, Z., de Kok, T., Van Woensel, T.: Optimal inventory and timing decisions for emergency shipments. IISE Trans. **52**(8), 904–925 (2020)

[16] Li, Y., Ou, J.: Optimal ordering policy for complementary components with partial back-ordering and emergency replenishment under spectral risk measure. Eur. J. Oper. Res. **284**(2), 538–549 (2020)

[17] Luo, Z., Chen, X.: Blood order and collection problems with two demand classes and emergency replenishment. J. Oper. Res. Soc. **72**(3), 501–518 (2021)

[18] Jacob, J., Shajin, D., Krishnamoorthy, A., Vishnevsky, V., Kozyrev, D.: Queueing-inventory with one essential and m optional items with environment change process forming correlated renewal process (mep). Mathematics **10**, 104 (2021)

[19] Melikov, A., Mirzayev, R., Nair, S.S.: Double sources queuing-inventory system with hybrid replenishment policy. Mathematics **10**(14), 2423 (2022)

[20] Shajin, D., Krishnamoorthy, A., Melikov, A.Z., Sztrik, J.: Multi-server queuing production inventory system with emergency replenishment. Mathematics **10**, 3839 (2022)

[21] Melikov, A., Ozkar, S.: Performance analysis and optimization of state-dependent replenishment policy in queuing-inventory system. Oper. Res. **25**(4), 1–36 (2025)

[22] Melikov, A., Mirzayev, R., Nair, S.S.: Double sources queuing-inventory system with hybrid replenishment policy. Mathematics **10**(14), 2423 (2022)

[23] Shajin, D., Melikov, A.: Queueing-inventory system with return of purchased items and customer feedback. RAIRO Oper. Res. **59**(3), 1443–1473 (2025)

[24] Krishnamoorthy, A., Shajin, D., Narayanan, W.: Inventory with positive service time: A survey. Adv. Trends Queueing Theory **2**, 201–238 (2021)

[25] Preethi, K., Sivakumar, B.: System point level crossing approach for lost sales deteriorating inventory systems with positive lead time. Methodol. Comput. Appl. Probab. **27**(3), 1–19 (2025)

[26] Boxma, O., Perry, D., Stadje, W.: Stationary analysis of an (r, q) inventory model with normal and emergency orders. J. Appl. Probab. **60**(1), 106–126 (2023)