Queueing Models and
Service Management
Vol. 9, No. 1, page 45–73, 2026

QMSM

Queuing-Inventory System with Attraction-Retention Mechanisms Under a Partial Synchronous Vacation Policy: The Case of Ethio Telecom Service Center in Arba Minch, Ethiopia

Berhanu Mekonen Alemu[1,*], Natesan Thillaigovindan[2] and Getinet Alemayehu Wole[3]

[1]Department of Mathematics
College of Natural and Computional Sciences, Arba Minch University, Arba Minch, Ethiopia

[2]Department of Mathematics
College of Natural and Computational Sciences, Arba Minch University, Arba Minch, Ethiopia

[3]Department of Mathematics
College of Natural and Computational Sciences, Haramaya University, Harar, Ethiopia

(*Received April 2025; accepted February 2026*)

Abstract: Quality of service (QoS) is a critical factor for customer satisfaction and operational efficiency, particularly in service-driven organizations such as Ethio Telecom in Ethiopia. To address congestion and customer impatience, this study investigates a finite-capacity multi-server Markovian queuing-inventory system (MQIS) that explicitly incorporates attraction-retention mechanisms alongside C removable servers operating under a partial synchronous vacation policy. The attraction-retention strategies are modeled to influence customer arrival rates and patience levels by encouraging customers to remain in the system through incentives or improved service quality, thereby mitigating balking and reneging behaviors. In this setting, any D servers ($0 < D < C$) may take simultaneous vacations as a group when no customers are waiting at a service completion epoch, while the remaining $C - D$ servers continue to operate, either actively serving or idling depending on the inventory status. Both service and vacation times are exponentially distributed, and inventory is managed using a continuous-review (q, Q) policy that replenishes stock to level Q once it drops to q. A continuous-time Markov process is formulated to analyze the system and steady-state probabilities are derived to evaluate performance measures. To minimize the total cost, a cost-loss model is proposed and solved using a genetic algorithm to determine the optimal service rate and the number of servers allocated for vacation. Numerical experiments based on primary data collected from Ethio Telecom's Arba Minch branch demonstrate how attraction-retention mechanisms, along with other system parameters, impact optimal policies and cost metrics. The proposed model is applicable to a wide range of service environments, including supermarkets, telecom centers, hospitals, production systems and restaurants, and can be extended to incorporate batch service, customer retrials, or catastrophic events.

* Corresponding author
Email: berhanumekonen6@gmail.com

Keywords: Attraction-retention mechanisms, partial synchronous vacation, removable servers, steady-state distribution, stochastic queuing-inventory system.

1. Introduction

Stochastic modeling serves as a vital interface between theoretical frameworks and practical applications. Queuing theory, pioneered by A. K. Erlang in his seminal work "The Theory of Probabilities and Telephone Conversations," provides essential tools for modeling systems involving waiting lines. In real-world scenarios, customers are influenced by inventory systems, leading to queuing-inventory systems (QIS), which are widely applied in healthcare, logistics, retail and telecommunications.

Customer impatience, manifested as balking or reneging, can significantly diminish service efficiency and revenue, particularly in high-demand settings like Ethio Telecom. In QIS contexts, customers are served on a first-come, first-served (FCFS) basis and each service reduces the inventory by one unit. Replenishment is managed externally, often via a continuous-review policy. Unlike standalone queuing or inventory models, QIS captures the interdependence between service capacity and stock dynamics (Zhao and Lian [23]).

In the system we examine, customers are served and depart immediately upon service completion, with inventory decreasing by one unit at each service completion epoch. Inventory is replenished by an external vendor. Specifically, Ethio Telecom service stations experience customer queues, with items sold on a First-Come-First-Serve (FCFS) basis when available. Unlike traditional inventory or queuing models, a QIS involves the interaction between service processes and inventory levels. Consequently, QIS has garnered significant research interest, with applications in integrated supply chain management (Schwarz *et al.* [14]), vehicle maintenance (Krishnamoorthy *et al.* [10]) and medical services (Arun [4]).

The foundational work on QIS by Melikov and Molchanov [11] and by Sigman and Simchi-Levi [16] focused on idealized models such as the $M/G/1$ system with limited inventory. Later developments considered server unavailability due to vacations, repairs or supplementary tasks. Traditional multi-server vacation models typically assume full synchronization, where all servers take a break simultaneously.

Classical queuing theory assumes that servers are always available for service; however, servers may become unavailable due to repairs, maintenance, or breaks. These periods of unavailability, known as vacations, have prompted the study of QIS models that incorporate server vacations. In multi-server systems, most vacation models assume simultaneous vacations for all servers (synchronous vacation). Tian *et al.* [19], have explored partial server vacations and their impact on waiting time. In practice, however, some servers may remain on duty while others take a vacation.

Daniel and Ramanarayanan [7] introduced the vacation concept in QIS, while Divya *et al.* [8] analyzed a single-server QIS with impatient customers, where the server takes vacations to replenish inventory. Nithya *et al.* [12] found that breaks, such as tea breaks, enhance performance and reduce stress. More recently, Yue *et al.* [22] studied synchronous vacation models in multi-server QIS.

In some QIS studies, customers are always served after service completion. However, in certain scenarios, customers may leave without service, such as in retail settings where they may listen to a pitch but not make a purchase. Inventory issues can be categorized into two types: (1) customers are never forced to leave, and (2) customers leave due to impatience when the system is full or inventory is unavailable. Baccelli and Bremaud [5] discussed factors influencing customer impatience, such as queue length, waiting time, and busy periods.

Attraction mechanisms, such as rewards and service quality improvements, can encourage customers to join the queue but may increase waiting times and impatience. Retaining customers who might otherwise leave (renege) is crucial for minimizing costs. The goal is to balance service costs, inventory management and waiting times to optimize the system's total cost.

Recently, Sun *et al.* [17] investigated customer balking behavior in a multi-server queue with synchronous multiple uninterrupted vacations under $N-$policy. They examined equilibrium and socially optimal balking strategies under different information levels (observable Vs unobservant) and demonstrated that pricing strategies vary with information transparency. Their findings reveal that regardless of information level, the $N-$policy should be designed for social optimization and disclosing server status to customers is beneficial when the threshold N predetermined. These insights on balking behavior and information disclosure complement this study, where balking is influenced by both system congestion and attraction mechanisms.

Most QIS studies focus on customer loss and synchronous vacation policies. However, these policies may not always be applicable. This study aims to analyze a QIS with removable servers under an asynchronous vacation policy, where service continues even during vacation periods and customers are not lost. An example is telecommunication service centers, where customers queue to purchase electronic devices, and a specified number of servers may take synchronous vacations when inventory runs low.

Choi *et al.* [6] and Rangaswamy *et al.* [13] reviewed QIS theory and applications. A few studies have incorporated customer attraction and retention mechanisms for impatient customers, particularly under a partial synchronous vacation policies. This study addresses this gap by examining customer impatience, which can lead to balking or reneging in a finite capacity multi-server QIS and explores strategies where servers adopt attraction and retention mechanisms.

Recent advancements in queuing-inventory modeling have addressed increasingly complex system behaviors. In a related development, Yue *et al.* [22] analyzed a multi-server QIS under a synchronous vacation policy, where all servers go on vacation simultaneously when inventory is depleted. Verma *et al.* [20] analyzed an $M/M/1$ queuing-inventory system with batch demands under an (S, Q) policy, deriving steady-state distributions and optimal parameters using iterative methods and ARENA simulation. Their work on batch demands is relevant for future extensions of this study.

Jegannathan *et al.* [9] investigated retrial queue with server-level a partial synchronous vacation and (s, S) inventory policy. Sharma *et al.* [15] presented a comprehensive review of queuing systems that incorporate customer impatience, retention strategies, and vacation mechanisms. Ahmadi *et al.* [1] studied inventory systems with preorder strategies under waiting-time risk, using a mean–Conditional Value at Risk (mean–CVaR) approach to quantify and manage service delays. Their work highlights the importance of integrating risk-aware decision-making into inventory control, particularly in service systems facing uncertainty in lead times and demand.

Previous research on queuing-inventory systems examined customer attraction-retention mechanisms and server vacation policies independently. For instance, studies on customer retention primarily focus on how promotional incentives or service improvements affect customer impatience and arrival rates [15], whereas vacation models typically analyze server availability assuming either full synchronization or fully independent asynchronous breaks [9]. However, the combined modeling of attraction-retention effects alongside partial or asynchronous server vacations remains scarce. Our work bridges this gap by simultaneously incorporating congestion-sensitive customer behavior and a partial synchronous vacation framework, where a subset of servers take joint breaks while others remain active. This integration captures more realistic operational dynamics, reflecting how customer decisions and server scheduling interact in practice, providing a more comprehensive and novel analytical framework compared to existing literature.

Despite such progress, there remains a gap in the literature. Few models integrate attraction-retention strategies for impatient customers with partial synchronous server vacations, where only a subset of servers take coordinated breaks. Most existing models assume either full synchronization or independent vacations. Moreover, they rarely capture the influence of customer behavior based on congestion, queue length or promotional strategies.

This paper addresses this gap by proposing a finite-capacity multi-server Markovian QIS model that incorporates both attraction-retention mechanisms and partial synchronous vacation policy. In the proposed system, any D servers, with $0 < D < C$, may take a joint vacation when no customers are waiting at a service completion epoch, while the remaining $C-D$ servers continue to operate, either serving customers or idling depending on inventory availability. At vacation completion, if the number of customers in the system is not greater than $C - D$, the same D servers extend their vacation for another period. Otherwise, they return to serve.

Our model differs significantly from those proposed by Yue *et al.* [22] and Jegannathan *et al.* [9]. Yue *et al.* assumed complete synchronization, where all servers simultaneously go on vacation if the inventory is depleted. In contrast, Jegannathan *et al.* considered asynchronous vacations at the server level, allowing each server to decide independently. Our model introduces a hybrid approach: a subset of D servers jointly take vacations, while the remaining servers continue to provide uninterrupted service, representing a partial synchronous vacation scheme.

Customer decisions in our model are influenced by both system congestion and promotional incentives. Those who find the system crowded may balk, while others may be retained using attraction strategies such as reward schemes or enhanced service quality. These mechanisms affect both arrival rate and patience level.

We develop a continuous-time Markov process to model the system dynamics, derive the steady-state probabilities and compute system performance measures. A cost-loss function is formulated to optimize the number of servers on vacation and service rate. This optimization problem is solved using a genetic algorithm. Numerical experiments using data from the Ethio Telecom service center in Arba Minch validate the proposed model and provide managerial insights.

This study contributes to the growing body of QIS literature by introducing a novel framework that integrates congestion-sensitive customer behavior, inventory management and coordinated server vacations. The model is applicable to a wide range of service environments, including telecommunications, supermarkets, hospitals and manufacturing. Extensions to batch service, retrials, and catastrophic inventory loss scenarios are identified as promising areas for future research.

The remainder of this paper is organized as follows. Section 2 presents the model description and notations. Section 3 discusses the methodology. Section 4 presents the analytical results. Section 5 provides the data and assumptions. Section 6 contains numerical results and discussion. Section 7 concludes the paper.

2. Description of the Model

We consider a finite-capacity multi-server queuing-inventory system (MQIS) with attraction-retention mechanisms for impatient customers, dedicated C-removable servers, a partial synchronous vacation policy and stochastic lead times. The following notations and assumptions are used throughout the paper:

i. Customers arrive individually according to a Poisson process with rate λ ($\lambda \geq 0$). Upon arrival, a customer either joins the queue or balks. Let b_n denote the probability that a customer joins the queue when n customers are already in the system. This state-dependent probability that models customer behavior based on system congestion is defined as:

$$b_n = \begin{cases} 1, & 0 \leq n \leq C - D - 1 \\ \frac{N-n}{N}, & C - D \leq n \leq N \end{cases}$$

This form, inspired by Ancker and Gafarian [3], reflects full willingness to join when the system is relatively empty and a decreasing inclination as congestion increases. The probability of balking is $1 - b_n$.

Due to attraction mechanisms (e.g., rewards, gift coupons, promotions), the effective arrival rate can increase by a factor β, which represents the proportional increase in customer inflow. As arrival rate increases, delays may occur, leading to customer impatience. Retention strategies (e.g., high-quality service, loyalty programs) are employed to reduce this impatience, with effectiveness modeled by a retention rate r.

ii. The system has C removable servers, a finite waiting room of size N, and a maximum inventory level Q. Each service consumes exactly one inventory item, thus inventory depletes by one unit per service completion.

iii. Among the C servers, a subset of D servers ($D < C$) can take synchronous vacations. That is, when triggered, D servers go on vacation as a group, while the remaining $C-D$ servers continue providing service. This policy is referred to as a *partial synchronous vacation*, as discussed in Tian and Xu [19].

iv. Customers form a single queue and are served on a first-come, first-served (FCFS) basis. Service times are independent and identically distributed exponential random variables with density $s(t) = \mu e^{-\mu t}$, $t \geq 0$, where $\mu > 0$ is the service rate. Once service begins, it is completed without interruption.

v. After joining the queue, a customer waits for a random amount of time T before potentially abandoning the system. If not served within time T, the customer leaves (reneges). The reneging time follows an exponential distribution with density $f(t) = \alpha e^{-\alpha t}$, $t \geq 0$, where $\alpha > 0$ is the reneging rate.
If $n \leq C-D$ and inventory is available, customers are served immediately and reneging does not occur. If $n > C - D$, then $(n - C + D)$ customers wait in the queue. The average reneging rate in this state is $(n - C + D)(1 - r)\alpha$, accounting for retention mechanisms. Hence the state-dependent reneging rate is:

$$R(n) = \begin{cases} (n - C + D)(1 - r)\alpha, & C - D < n < N \\ 0, & 0 \leq n \leq C - D \end{cases}$$

vi. When there are no customers waiting in the queue or on-hand inventory is depleted at a service completion epoch, D servers begin a synchronous vacation for a random duration V. If no customers are waiting in the queue upon a vacation completion, then the same or a different set of D servers begin another vacation. The vacation time V follows an exponential distribution with density $v(t) = \xi e^{-\xi t}$, $t \geq 0$, where $\xi > 0$ is the vacation rate.

vii. The system adopts an (q, Q) inventory replenishment policy. When the on-hand inventory level drops to the reorder point q, a replenishment order is triggered to restore the inventory up to level Q ($q < Q$). The lead time for replenishment follows an exponential distribution with rate parameter η ($\eta > 0$).
We assume that $D < q$ to ensure proper inventory control. If $D \geq q$, there is a possibility that some of the active servers are on vacation before the inventory level drops to q, thereby preventing inventory consumption and halting the reorder process. This could cause a deadlock state in the inventory system. The condition $D < q$ guarantees that at least some servers remain active to continue inventory depletion and maintain system responsiveness.

3. Research Methodology

3.1. Research design

The aim of this study is to reduce cost-loss due to impatient customers often noticed at Ethio Telecom service centers. For such QIS models, encouraging arrivals and retention of reneging customers are significantly important to attain suitable optimal total cost. This has motivated us to analyze queuing systems attached to inventory. Attraction-retention mechanisms for impatient customers are adopted to minimize cost loss, and for dedicated C-removable servers, a partial synchronous vacation policy is employed to serve the customers without break. The analysis is based on building a mathematical model for such a QIS with the assumption of Markov process. The effect of study parameters on various performance measures are discussed using numerical data collected from Ethio Telecom service center, Arba Minch.

3.2. Method of data collection

The data used in this study are primary data collected through direct observations from Ethio Telecom, Arba Minch district head office by the researcher for the number of customers and measurements of time recorded on the data sheet for a period of two weeks, non-inclusive of Sundays. The service time and arrival time of customers are determined every day by a stopwatch using mobile phone. The data is collected based on the research design and research approach within the standard workweek in Ethiopia for 48 hours. The data collection is typically performed eight hours a day, starting from $8:30$ am to $12:30$ pm and $1:30$ pm to $5:30$ pm for 12 days.

3.3. Time measurements and data extraction

From time measurements the required data is extracted using the process below. Customers are registered for identification when entering Ethio Telecom service center, Arba Minch district head office using indices, $i = 1, 2, 3, \cdots, N$, each of which is linked to the following time measures:

(i) customer's arrival time, (ii) customer's service completion time, (iii) inter-arrival times between consecutive arrivals, (iv) number of customers in the queue, (v) number of servers on vacation, (vi) number of customers reneged, (vii) waiting time of customer in the queue, (viii) service time of customer, (ix) waiting time of customer in the system.

The number of customers in the system N, the number of servers on vacation D, and the inter-arrival and service times were averaged across each of the eight working hours per day. Utilizing the associated $M/M/C/N$ MQIS model equations, we compute the mean operating characteristics of the QIS in Arba Minch district head office of Ethio Telecom service center. Mathematical software MATLAB is used for numerical analysis.

3.4. Method of data analysis

Mathematical models are prominently tractable even for complicated systems. Therefore both analytical and numerical analyses of the model are carried out. The governing differential-difference equations of the QIS in terms of steady-state probabilities and study parameters are formulated. Recursive method is employed to obtain closed-form expressions for steady-state probabilities of the system size. After completion of data collection, it is checked for completeness and exported to SPSS version 20 to get expected values for further analysis. Based on recorded data a numerical analysis is performed to study the impact of parameters on the effectiveness of the service center.

4. Analysis of the Model

In this section we carry out analytical analysis of the finite capacity multi-server $MQIS$ with attraction-retention mechanisms for impatient customers and dedicated $C-$removable servers operating under a partial synchronous vacation policy and random lead time. Let the state of the system at time t be described by the random variables $X(t)$ which denotes the system size, $Y(t)$ which describes the inventory level, and $Z(t)$ denotes the status of the server, defined as

$$Z(t) = \begin{cases} 0, & D \ servers \ are \ on \ vacation \ at \ time \ t \\ 1, & All \ C \ servers \ are \ available \ for \ service \ at \ time \ t \end{cases}$$

The stochastic process

$$\Phi(t) = \{(X(t), Y(t), Z(t));\ t \geq 0\}$$

is a three-dimensional Markov process with state space

$$\Omega = \{(n, s, 0) : 0 \leq n \leq C - D, 0 \leq s \leq Q\} \bigcup \{(n, s, 1) : C - D + 1 \leq n \leq N; 0 \leq s \leq Q\}$$

For the process $\Phi(t)$, the steady-state probability distribution is defined as:

$$p_{i,n,s} = \lim_{t\to\infty} Pr\{X(t) = n, Y(t) = s;\ 0 \leq s \leq Q,\ Z(t) = i;\ i = 0, 1\}$$

We assume that there is no customer in the system and the inventory level is Q at time $t = 0$. Then by applying Markov process and using state-transition diagram of QIS presented in Figure 1, the Chapman Kolmogorov equations (1 - 11) governing the model are given below. In the state transition diagram λ_* and α_* represent, $\lambda(1 + \beta)b_{C-D}$ and $(N - C + D)\alpha$ respectively.

$$-\lambda(1 + \beta)p_{0,0,0} + (\mu + \eta)p_{0,1,s} = 0;\ s = 0, 1, 2, \cdots, q,\ n = 0 \tag{1}$$
$$-\lambda(1 + \beta)p_{0,0,0} + \mu p_{0,1,s} = 0;\ s = q + 1, q + 2, \cdots, Q - 1,\ n = 0$$

$$\begin{aligned} &\lambda(1 + \beta)p_{0,n-1,s} - (\lambda(1 + \beta) + n\mu + \eta)p_{0,n,s} + (n + 1)(\mu + \eta)p_{0,n+1,s} = 0; \\ &s = 1, 2, \cdots, q,\ 1 \leq n < C - D \end{aligned} \tag{2}$$

$$\begin{aligned} &\lambda(1 + \beta)p_{0,n-1,s} - (\lambda(1 + \beta) + n\mu)p_{0,n,s} + (n + 1)\mu p_{0,n+1,s} = 0; \\ &s = q + 1, q + 2, \cdots, Q - 1,\ 1 \leq n < C - D \end{aligned}$$

$$\begin{aligned} &\lambda(1 + \beta)p_{0,C-D-1,s} - [\lambda(1 + \beta)b_{C-D,s} + (C - D)(\mu + \eta)]p_{0,C-D,s} + [(C - D)(\mu + \eta) \\ &+ alpha]p_{0,C-D+1,s} + (C - D + 1)(\mu + \eta)p_{1,C-D+1,s} = 0;\ s = 1, 2, \cdots, q,\ n = C - D \end{aligned} \tag{3}$$

$$\begin{aligned} &\lambda(1 + \beta)p_{0,C-D-1,s} - [\lambda(1 + \beta)b_{C-D,s} + (C - D)\mu]p_{0,C-D,s} + [(C - D)\mu + \alpha]p_{0,C-D+1,s} \\ &+ (C - D + 1)\mu p_{1,C-D+1,s} = 0;\ s = q + 1, q + 2, \cdots, Q - 1,\ n = C - D \end{aligned}$$

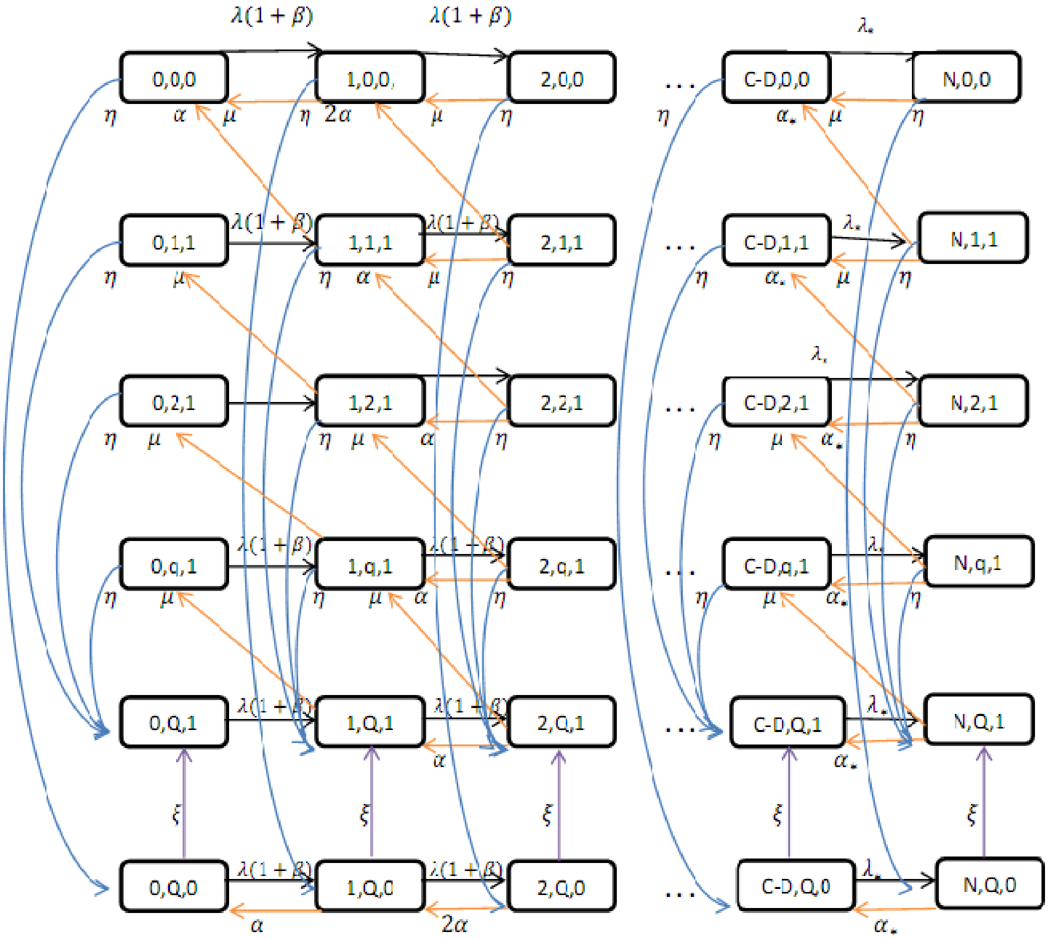


Figure 1. State-transition diagram of the QIS model

$$\lambda(1+\beta)b_{n-1}p_{0,n-1,s} - [\xi + \lambda(1+\beta)b_n + (C-D)(\mu+\eta) + (n+D-C)\alpha(1-r)] \\ p_{0,n,s} + [(C-D)(\mu+\eta) + (n+1+D-C)\alpha(1-r)]p_{0,n+1,s} = 0;\ s = 1,2,\cdots,q, \\ C-D < n < N \tag{4}$$

$$\lambda(1+\beta)b_{n-1}p_{0,n-1,s} - [\xi + \lambda(1+\beta)b_n + (C-D)\mu + (n+D-C)\alpha(1-r)]p_{0,n,s} + \\ [(C-D)\mu + (n+1+D-C)\alpha(1-r)]p_{0,n+1,s} = 0; \\ s = q+1, q+2, \cdots, Q-1,\ C-D < n < N$$

$$\lambda(1+\beta)b_{N-1}p_{0,N-1,s} - [\xi + (C-D)(\mu+\eta) + (N+D-C)\alpha(1-r)]p_{0,N,s} = 0; \\ s = 1,2,\cdots,q,\ n = N \tag{5}$$

$$\lambda(1+\beta)b_{N-1}p_{0,N-1,s} - [\xi + (C-D)\mu + (N+D-C)\alpha(1-r)]p_{0,N,s} = 0; \\ s = q+1, q+2, \cdots, Q-1,\ n = N$$

$$\xi p_{0,C-D+1,s}(t) - [\lambda(1+\beta) + (C-D+1)(\mu+\eta)]p_{1,C-D+1,s} + (C-D+2)(\mu+\eta) \\ p_{1,C-D+2,s} = 0;\ s = 1,2,\cdots,q\ n = C-D+1 \tag{6}$$

$$\xi p_{0,C-D+1,s}(t) - [\lambda(1+\beta) + (C-D+1)\mu]p_{1,C-D+1,s} + (C-D+2)\mu \\ p_{1,C-D+2,s} = 0;\ s = q+1, q+2, \cdots, Q-1,\ n = C-D+1$$

$$\xi p_{0,n,s} + \lambda(1+\beta)p_{1,n-1,s} - [\lambda(1+\beta) + n(\mu+\eta))p_{1,n,s} + [(n-1)(\mu+\eta)p_{1,n+1,s} \\ = 0; s = 1, 2, \cdots, q,\ C-D+2 \leq n \leq C-1 \tag{7}$$

$$\xi p_{0,n,s} + \lambda(1+\beta)p_{1,n-1,s} - [\lambda(1+\beta) + n\mu)p_{1,n,s} + [(n-1)\mu p_{1,n+1,s} = 0; \\ s = q+1, q+2, \cdots, Q-1,\ C-D+2 \leq n \leq C-1$$

$$\xi p_{0,C,s} + \lambda(1+\beta)p_{1,C-1,s} - [\lambda(1+\beta)b_{C,s} + C(\mu+\eta)]p_{1,C,s} + (C(\mu+\eta)+ \\ \alpha(1-r))p_{1,C+1,s} = 0;\ s = 1, 2, \cdots, q,\ n = C \tag{8}$$

$$\xi p_{0,C,s} + \lambda(1+\beta)p_{1,C-1,s} - [\lambda(1+\beta)b_{C,s} + C\mu]p_{1,C,s} + (C\mu+ \\ \alpha)p_{1,C+1,s} = 0;\ s = q+1, q+2, \cdots, Q-1,\ n = C$$

$$\xi p_{0,n,s} + \lambda(1+\beta)b_{n-1}p_{1,n-1,s} - [\lambda(1+\beta)b_n + C(\mu+\eta) + (n-C)\alpha(1-r)]p_{1,n,s} \\ + [C(\mu+\eta) + (n+1-C)\alpha(1-r)p_{1,n+1,s} = 0;\ s = 1, 2, \cdots, q,\ C < n < N \tag{9}$$

$$\xi p_{0,n,s} + \lambda(1+\beta)b_{n-1}p_{1,n-1,s} - [\lambda(1+\beta)b_n + C\mu + (n-C)\alpha(1-r)]p_{1,n,s} + [C\mu \\ + (n+1-C)\alpha(1-r)p_{1,n+1,s} = 0;\ s = q+1, q+2, \cdots, Q-1,\ C < n < N$$

$$\xi p_{0,N,s} + \lambda(1+\beta)b_{N-1}p_{1,N-1,s} - [C(\mu+\eta) + (N-c)\alpha(1-r)]p_{1,N,s} = 0; s = 1, 2, \cdots, q \tag{10}$$

$$\xi p_{0,N,s} + \lambda(1+\beta)b_{N-1}p_{1,N-1,s} - [C\mu + (N-c)\alpha(1-r)]p_{1,N,s} = 0;\ s = q+1, q+2, \cdots, Q-1$$

Normalization condition is:

$$\sum_{s=0}^{Q} \left(\sum_{n=0}^{C-D} p_{0,n,s} + \sum_{n=C-D+1}^{N} p_{1,n,s} \right) = 1 \tag{11}$$

4.1. Steady-state solution

To acquire neat and closed-form solutions, recursive technique is utilized to find all steady-state probabilities $p_{0,n,s}$ and $p_{1,n,s}$ in terms of $p_{0,0,0}$, λ, μ, β, b_n, r, α, ξ and η. The solution approach of the system is stated as follows:

Theorem: If the steady-state difference equations of an $M/M/C/N$ queuing-inventory system (MQIS) with attraction-retention mechanisms for impatient customers and dedicated C-removable servers operating under a partial synchronous vacation policy and (q, Q) inventory policy are as given in (1) - (11), then the steady-state probabilities of the system size are given by:

$$p_{0,0,0} = \left[L + \sum_{n=C-D}^{N-1} \left[\frac{(\lambda(1+\beta))^n}{\mu+\eta+\xi} \prod_{i=C-D+1}^{N-1} b_i \right] + p_{1,N,s} \right]^{-1} \tag{12}$$

$$p_{0,1,s} = \rho p_{0,0,0} \tag{13}$$

$$p_{0,n,s} = \frac{1}{n!}\rho^n p_{0,0,0};\ 2 \leq n \leq C - D, \tag{14}$$

$$p_{1,1,s} = \frac{\lambda\left(1+\beta\right)}{\mu+\eta+\xi}\prod_{i=2}^{N-1}\left[b_i\right]p_{0,0,0} \tag{15}$$

$$p_{1,n,s} = \frac{\left(\lambda\left(1+\beta\right)\right)^n}{\mu+\eta+\xi}\prod_{i=C-D+1}^{N-1}\left[b_i\right]p_{0,0,0};\ C-D+1 \leq n < N \tag{16}$$

$$p_{1,N,s} = \frac{\left(\lambda\left(1+\beta\right)\right)^N}{\mu+\eta+\xi}\prod_{i=C-D+1}^{N-1}\left[b_i\right]p_{0,0,0} \tag{17}$$

where

$$L = 1+\rho+\sum_{n=2}^{C-D-1}\left[(n!)^{-1}\left(\rho\right)^n\right]$$

$$\rho = \rho^* + \rho^{**}$$

$$\rho^* = \left(\frac{\lambda\left(1+\beta\right)}{\mu+\eta}\right)$$

$$\rho^{**} = \left(\frac{\lambda\left(1+\beta\right)}{\mu}\right)$$

$$b_i = b_i^* + b_i^{**},\ i = C-D+1, C-D+2, \cdots, N-1$$

$$b_i^* = \frac{b_{i-1}}{\mu+\eta+(i-(C-D))\alpha(1-r)+\xi},\ i = C-D+1, C-D+2, \cdots, N-1$$

$$b_i^{**} = \frac{b_{i-1}}{\mu+(i-(C-D))\alpha(1-r)+\xi},\ i = C-D+1, C-D+2, \cdots, N-1$$

Proof of the theorem (see Appendix A).

4.2. Special cases

In this section results pertaining to some models which are found in the existing literature have been deduced from our model by taking specific values of the parameters b_n, α, ξ, β and r

Case 1: In the absence of balking, multiple-servers, server vacation and attraction-retention mechanisms for impatient customers (i.e. for $b_n = 1, \forall n = 0, 1, 2, \cdots, N, C = 1, \xi = \beta = r = 0$), the model at hand reduces to $M/M/1/N$ queuing scheme with inventory for impatient customers under a random order size, studied by Alnowibet *et al.* [2].

Case 2: For fixed balking rate (i.e. $b_n = b, \forall n = 0, 1, \cdots, N$), in the absence of reneging and attraction-retention mechanisms for impatient customers (i.e. $\alpha = \beta = r = 0$), the study reduces to a multi-server retrial QIS with a partial synchronous vacations, studied by Jeganathan *et al.* [9].

4.3. Performance measures

In this section various system performance metrics are defined explicitly and presented in the sequel:

1. **The mean inventory level** (E_I)**:** The expected inventory level (E_I) of the multi-server QIS with vacation in steady-state is given by

$$E_I = \sum_{s=0}^{Q} \sum_{n=0}^{C-D} s p_{0,n,s} + \sum_{s=0}^{Q} \sum_{n=C-D+1}^{N} s p_{1,n,s} \tag{18}$$

   When average inventory level is very high, this means we may have to change the order quantity. If it is very low we will lose more customers.

2. **The average number of new orders** (E_r)**:** The expected number of inventory units replenished per unit time (i.e., reorder rate) is called the average number of new orders or simply expected reorder rate (E_r). Under the (q, Q) policy, a reorder is triggered when the inventory level drops to s $(0 \leq s \leq q)$, and the order size is $(Q - s)$ at the end of service. Thus, the expected reorder rate is given by:

$$E_r = \eta \sum_{s=0}^{q} (Q - s) \left(\sum_{n=0}^{C-D} p_{0,n,s} + \sum_{n=C-D+1}^{N} p_{1,n,s} \right) \tag{19}$$

3. **Number of cycles** (E_{r_k})**:** A cycle is defined as the time between placing two successive orders (see Schwarz *et al.* [14]). Here, E_{r_k} is the number of cycles for the order size k. So, the mean cycle time is given by

$$E_{r_k} = E_r^{-1} \tag{20}$$

4. **The average order size** (E_0)**:** We note that a reorder is triggered when inventory level drops to s $(0 \leq s \leq q)$, and the order size is $(Q - s)$. Hence, mean order size is given by

$$E_0 = \sum_{n=0}^{C-D} Q p_{0,n,0} + \sum_{s=0}^{q} \sum_{n=C-D+1}^{N} (Q - s) p_{1,n,s} \tag{21}$$

5. **The average lost sales rate** (LS)**:** The expected lost sales per unit time for new customers is given by:

$$LS = \lambda(1 + \beta) \left[\sum_{n=0}^{C-D} p_{0,n,0} + \sum_{s=1}^{Q} p_{1,N,s} \right] \tag{22}$$

   and the expected lost sales per cycle is defined by

$$LS_c = \frac{LS}{E_r} \tag{23}$$

Lost sales for new customers happen for two reasons. First, it may happen when the customer finds waiting space, but there is no on-hand inventory. Second, it may happen when the customer finds waiting space full and is unable to enter, but there is enough on-hand inventory.

6. **Safety Stock (SS):** Inventory can be divided into working stock and safety stock. Working stock is inventory acquired and held in advance of requirements so that expected demand can be satisfied and ordering can be done on a lot size rather than on an as-needed basis. Lot sizing is done to minimize ordering and holding costs. Safety stock is held in reserve to protect against uncertainties of supply and demand (Tempelmeier [18]). The safety stock is defined by Schwartz *et al.* [14] as the palm stationary mean value given by

$$SS = E(NIP\ just\ before\ arrival\ of\ a\ replenishment) \tag{24}$$

where NIP is Net Inventory Position. The net inventory position is defined as inventory on-hand $(Y_{net} = Y)$. The net inventory position is usually defined as on-hand inventory minus back-orders. In order to determine safety stock required to guarantee a specified $\alpha-$service-level, stationary probability distribution of stock on hand must be known.

7. $\alpha-$**service-level:** The $\alpha-$service-level is an event-oriented service measure. It measures the probability that all customer orders arriving within a given time interval will be completely delivered from stock on hand, i.e. without delay. With reference to a demand period, α denotes the probability that an arbitrarily arriving customer will be served from stock on hand, i.e. without an inventory-related delay. Generally, in inventory management, $\alpha-$service-level is often used to denote the desired level of service. It specifically measures the probability that a customer demand can be fulfilled entirely from the existing stock on hand, without any stock-outs or delays during a replenishment cycle. It is denoted by Greek letter α and given by:

$$\alpha = 1 - Pro(Stockout)$$

where α is the $\alpha-$service-level, $Pro(Stockout)$ is the probability of stocking-out during a replenishment cycle. Let DL be the random variable representing demand during lead time and η be the random variable representing lead time. The memoryless property of exponential distribution states that:

$$Pro(DL > s + t|DL > s) = Pro(DL > t)$$

This property allows us to simplify the calculation of α-service-level for the continuous review inventory system using exponential distribution. Thus $\alpha-$service-level can be expressed as:
$\alpha = Pro(DL \leq Q - I|DL > 0)$ which is equivalent to:

$$\alpha = 1 - e^{-\lambda_{DL}(Q-I)} \tag{25}$$

where λ_{DL} is rate parameter for the demand distribution, Q is order quantity and I is current inventory level. If an order cycle is considered as the standard period of reference, then α denotes the probability of restock within an order cycle, which is equal to the proportion of all order cycles with no stock-out situation.

8. **Quality of service for new customers:** Let β_1 be the quality of service measure of the new customer or β_1-service level. According to Schwarz *et al.* [14], β_1-service level is defined by

$$\beta_1 = \frac{E(satisfied\ demand\ per\ unit\ time)}{E(total\ demand\ per\ unit\ time)}$$

The β_1-service level is given by

$$\beta_1 = \frac{\lambda(1+\beta) - LS}{\lambda(1+\beta)} \tag{26}$$

When service quality is closer to one, new customers receive better service, and when it is closer to zero, new customers receive worst service.

9. **Effective mean arrival rate** (λ_{eff})**:** The expected arrival rate of customers who are admitted to the system per unit time is given by

$$\lambda_{eff} = \lambda(1+\beta) - LS = \lambda(1+\beta)\beta_1 \tag{27}$$

Note that, $\alpha-$ and $\beta-$service levels do not provide any information on length of waiting time that a customer may experience.

10. **System length** (L_s)**:** The expected number of customers in the system denotes customers in line plus those who are being served. Let N_s be the random variable that describes the number of customers in the system at steady-state, then

$$L_s = E(N_s)$$

Thus

$$L_s = \sum_{s=0}^{Q}\sum_{n=1}^{C-D} np_{0,n,s} + \sum_{s=1}^{Q}\sum_{n=C-D+1}^{N} np_{1,n,s} \tag{28}$$

11. **Queue length** (L_q)**:** The expected number of customers in the waiting room denotes the number of customers in the system excluding those who are being served. Let N_q be the random variable that describes the number of customers awaiting in the line, and $(L_q) = E(N_q)$. Then

$$L_q = \sum_{s=0}^{Q}\sum_{n=C-D}^{N} \left[n - (C-D)\right] p_{1,n,s}$$

which is equivalent to

$$L_q = \sum_{s=0}^{Q}\sum_{n=C-D+1}^{N} \left[n - (C-D)\right] p_{1,n,s} \tag{29}$$

12. **Expected waiting times in the system** (W_s)**, and in the queue** (W_q)**:** The performance measures, average waiting time in the queue (W_q) and average waiting time in the system (W_s) are obtained using well-known Little's formula, $W_q = \frac{L_q}{\lambda}$, $W_s = \frac{L_s}{\lambda}$. Now when we are triggered to apply Little's formula in the proposed model, we know customers arrive into the system at a rate of $\lambda(1+\beta)$. However, all the customers who arrive do not join the system in view of finite capacity limitation. In this connection, effective arrival rate λ_{eff} (27) is crucial in our model which is different from the overall arrival rate $\lambda(1+\beta)$. As a result various performance measures based on effective average arrival rate (λ_{eff}) are defined in the sequel:

$$W_s = \frac{L_s}{\lambda_{eff}} \tag{30}$$

$$W_q = \frac{L_q}{\lambda_{eff}} \tag{31}$$

13. **Expected busy probability:** Busy probability that there are no idle servers in the system (P_B) is given by:

$$P_B = \sum_{s=1}^{Q} \sum_{n=C-D}^{N} p_{1,n,s} \tag{32}$$

On the other side, if the number of customers in the system is less than the number of servers who are on duty it is not difficult to observe idle severs in the system and this scenario depicted by idle probability of the server.

14. **Idle probability (**P_I**):** The uncertainty that there are servers who are idle in the system is given by:

$$P_I = \sum_{s=0}^{Q} \sum_{n=0}^{C-D-1} p_{0,n,s} + \sum_{n=C-D}^{N} p_{1,n,0} \tag{33}$$

15. **Expected system lost** (LR)**:** The average number of customers who are not getting service in the system. It is the sum of average balking (BR) and reneging (RR). *Average balking in the system* (BR) is the expected number of customers who are not entering the queue, *average system reneging* (RR) is the expected number of customers who leave the queue without getting service. The corresponding mathematical expressions are given as follows:

$$BR = \sum_{s=0}^{Q} \sum_{n=C-D}^{N} \lambda(1+\beta)(1-b_n)p_{1,n,s}$$

which simplifies to

$$BR = \lambda(1+\beta) \sum_{s=0}^{Q} \sum_{n=C-D}^{N} (1-b_n)p_{1,n,s} \tag{34}$$

$$RR = \sum_{s=0}^{Q} \sum_{C-D}^{N} (n - C + D)\alpha(1 - r)p_{1,n,s}$$

which is equivalent to

$$RR = \alpha(1 - r) \sum_{s=0}^{Q} \sum_{C-D+1}^{N} (n - C + D)p_{1,n,s} \tag{35}$$

In (35) $[(n - C + D)(1 - r)\alpha]$ is the instantaneous reneging rate of customers.

$$LR = BR + RR \tag{36}$$

Even if there are customers who are not served, it is customary to observe that there are patient customers who wait in the queue and get served. In this regard, the average number of customers who are serviced (GR) is considered.

16. **Expected gain (GR):** The average number of customers who join the queue and leave the system after completion of service is given by:

$$GR = L_s - LR \tag{37}$$

Lastly, from our daily experience, there are customers who are lost due to server vacations.

17. **Expected vacation loss (VL_r):** The average number of customers who are lost due to servers vacation is given by:

$$VL_r = \xi \sum_{s=0}^{Q} \sum_{n=0}^{C-D} np_{0,n,s}, \tag{38}$$

where ξ is reneging rate triggered by server vacation.

4.4. Cost-loss model

Economically, there are costs connected with operating the system originating from both queuing of customers and holding inventory. Once we have measures of performance, we develop cost model as the total expected cost per unit time, where the number of servers on vacation D and the service rate μ are decision variables and we are focused on the optimization part of the model. Our objective is to determine optimum number of servers to go on vacation D^* and optimum service rate μ^* that minimizes the total expected cost loss. Let
C_1 = holding cost of inventory per unit time,
C_2 = fixed cost for placing an order,
C_3 = replenishment cost per item,
C_4 = cost per unit time when a server is on vacation,
C_5 = cost per unit time when a server is busy,
C_6 = cost per unit time when a server is idle,

C_7 = cost per unit time when a customer joins in the queue and waits for service (cost of waiting),
C_8 = cost per unit time when a customer joins the system and is served (cost of service),
C_9 = cost per unit time when a customer balks or reneges.
Using the definition of each cost element listed above and its corresponding system characteristics the total expected cost loss function per unit time is given by:

$$F(\mu, D) = C_1E_I + C_2E_r + C_3E_0E_r + C_4VL_r + C_5P_B + C_6P_I + C_7L_q + C_8GR + C_9LR \quad (39)$$

In equation (39), first term C_1E_I is inventory cost incurred for inventory holding, second term C_2E_r is order cost incurred when placing an order, third term $C_3E_0E_r$ is cost incurred when a replenishment is performed, 4^{th} term C_4VL_r is the cost incurred when servers go for vacation, 5^{th} term C_5P_B is the cost incurred when the servers are busy, 6^{th} term C_6P_I is the cost incurred by idle servers, 7^{th} term C_7L_q is the cost incurred when customers waiting in line for service, 8^{th} term C_8GR is the cost incurred by customers who are served and last term C_9LR is the cost incurred by customers who are lost.

Thus the problem reduces to:

$$Minimize\ F(\mu, D)$$

Subject to the constraints

$$0 \leq \mu \leq M_1$$

$$0 \leq D \leq M_2$$

where M_1 is a real number and M_2 ia an integer. The cost function $F(\mu, D)$ is nonlinear with μ, and D as decision variables. It is difficult to analyze the convexity of the cost function due to its complexity. To treat such types of problems meta-heuristic algorithms, like genetic algorithm is employed to solve the problem using the computer software MATLAB. The appropriate results are obtained using genetic algorithm developed (Appendix B).

5. Data Presentation

NB: Time: $12:30$ pm - $1:30$ pm Data collection is not considered due to lunch break.

6. Numerical Results and Discussion

In this section, we conduct a numerical investigation to assess how the probability p_n, customer impatience rates ($(1-b_n)$, $(1-b_N)$, α and ξ), and other key system parameters influence the performance measures of the telecommunication service system. The numerical study uses real data collected from the Ethio Telecom service center located at the Arba Minch district office.

The operational environment at the service center represents a dynamic balance between fluctuating customer demand and available service capacity. Currently the center employs 13 dedicated vendors who serve on an average 89 customers per day. However, maintaining

Table 1. Summary of Number of Arrivals $(A(t))$, reneges of customers $(R(t))$, and number of servers on vacation $(D(t))$ at Ethio Telecom, Arba Minch district head office from February $04 - 09, 2024$ and February $11 - 16, 2024$ at any time (t)

Week	Day	Time (t)	$A(t)$	$D(t)$	$R(t)$
Week (I)	Monday	$8:30$ am $- 5:30$ pm	15.125000	9.875000	4.750000
	Tuesday	$8:30$ am $- 5:30$ pm	9.875000	7.600000	1.750000
	Wednesday	$8:30$ am $- 5:30$ pm	10.500000	4.760000	2.250000
	Thursday	$8:30$ am $- 5:30$ pm	5.500000	7.940000	0.500000
	Friday	$8:30$ am $- 5:30$ pm	17.500000	1.780000	4.000000
	Saturday	$8:30$ am $- 5:30$ pm	8.500000	4.750000	2.250000
Week (II)	Monday	$8:30$ am $- 5:30$ pm	11.125000	5.450000	1.250000
	Tuesday	$8:30$ am $- 5:30$ pm	13.875000	9.250000	5.250000
	Wednesday	$8:30$ am $- 5:30$ pm	7.500000	4.850000	2.250000
	Thursday	$8:30$ am $- 5:30$ pm	6.500000	7.950000	1.500000
	Friday	$8:30$ am $- 5:30$ pm	18.500000	5.650000	3.000000
	Saturday	$8:30$ am $- 5:30$ pm	9.500000	1.750000	2.250000

consistent service quality is challenged by resource fluctuations, notably server availability. The overall customers reneging rate is estimated to be 0.029 and on the average 6 servers are observed to be on vacation at any given time. The reneging rate increases to 0.044 during vacation periods due to reduced service capacity, modeled by an incremental impatience effect $\xi = 0.015$.

Customers arrive and join the queue to primarily purchase mobile phones or accessories and are served on a first-come, first-served basis. The system operates under a steady-state arrival rate of $\lambda = 11.17$ customers per hour, with an observed average service rate of $\mu = 7.25$ customers per hour. Using our model, we determine the optimal number of servers to go on vacation, denoted by D^*, as well as the optimal service rate μ^*, with the goal of minimizing the expected cost loss while simultaneously maintaining acceptable levels of customer satisfaction and efficient inventory turnover.

These findings provide significant practical implications for Ethio Telecom's operational management. The optimal vacation scheduling enables more balanced workload distribution and efficient server scheduling, allowing a limited number of servers to rest without compromising continuous service delivery. Additionally, identifying the optimal service rate μ^* serves as a practical benchmark for staff performance evaluation, supporting decisions related to staff allocation, targeted training, and shift planning. Given the capacity-constraint of the system, it is noteworthy that the utilization factor $\rho = \frac{\lambda}{\mu}$ need not be strictly less than one; instead, it should be optimized jointly with cost considerations and service-level targets to achieve overall system efficiency.

The future of Ethio Telecom service center is ripe with possibilities. The current service level of 0.988 at the Arba Minch district head office, indicating a high probability of satisfactory service, suggests that newly arriving customers can be effectively retained by maintaining and improving service quality. This underscores the promising potential of the telecommunication service center to achieve enhanced operational efficiency, improved cus-

Table 2. The ordering and receiving of items at Ethio Telecom, Arba Minch district head office from February $04-09$, 2024 and February $11-16$, 2024 at any time (t)

Week	Day	Time (t)	Ordering Time	Receiving Time
Week (I)	Monday	$8:30$ am $-$ $5:30$ pm	$9:10$ am	–
	Tuesday	$8:30$ am $-$ $5:30$ pm	–	–
	Wednesday	$8:30$ am $-$ $5:30$ pm	–	$10:20$ am
	Thursday	$8:30$ am $-$ $5:30$ pm	–	–
	Friday	$8:30$ am $-$ $5:30$ pm	–	–
	Saturday	$8:30$ am $-$ $5:30$ pm	–	–
Week (II)	Monday	$8:30$ am $-$ $5:30$ pm	–	–
	Tuesday	$8:30$ am $-$ $5:30$ pm	–	–
	Wednesday	$8:30$ am $-$ $5:30$ pm	$9:30$ am	–
	Thursday	$8:30$ am $-$ $5:30$ pm	–	–
	Friday	$8:30$ am $-$ $5:30$ pm	–	$11:00$ am
	Saturday	$8:30$ am $-$ $5:30$ pm	–	–

tomer satisfaction and greater empowerment of service agents.

To model the system dynamics at any time t, we define the probability that an arriving customer joins the queue as

$$b_n = \begin{cases} 1, & 0 \leq n \leq C-D-1 \\ \frac{N-n}{N}, & C-D \leq n \leq N \end{cases}$$

following the approach of Ancker and Gafarian [3].

The cost parameters used in this study are given as:

$$C_1 = 100,\ C_2 = 110,\ C_3 = 120,\ C_4 = 150,\ C_5 = 130,$$
$$C_6 = 140,\ C_7 = 120,\ C_8 = 56,\ C_9 = 80,$$

as reported in Wang and Chang [21] and Yue *et al.* [22].

Using a genetic algorithm, we evaluate the optimal values of the number of servers on vacation (D^*) and the service rate (μ^*), along with the minimum expected cost F. Additionally, performance measures including expected inventory level, expected order size, expected system and queue lengths, and system loss or gain are computed at these optimal values to provide comprehensive operational insights.

Next, we perform a sensitivity analysis on the optimal number of vendors to go on vacation (D^*), the optimal service rate (μ^*) and the minimum expected cost loss (F^*) with respect to variations in key system parameters. These parameters include the reordering rate η, attraction rate β, retention rate r, arrival rate λ, reneging rate α and vacation rate ξ.

Since the decision variable D is integer-valued and the expected cost function is nonlinear and complex, deriving closed-form analytical expressions for the optimal values D^* and μ^* is a challenging task. Therefore, a heuristic approach is employed to obtain the optimal values D^* and μ^*, which satisfy the following local optimality conditions:

$$F(D^*-1) > F(D^*) < F(D^*+1), \tag{40}$$

$$F(\mu^* - d) > F(\mu^*) < F(\mu^* + d), \tag{41}$$

where d is a positive constant representing a small perturbation in the service rate.

The numerical results for the optimal values D^*, μ^*, the optimal cost F^* and other key performance measures are summarized in Tables 3 to 5 and depicted graphically in Figures 2 to 5.

Table 3. System performance measures for $\alpha = 0.029$, $\beta = 0.5$, $r = 0.9$, $\xi = 0.015$ and $\eta = 0.0588$

D^*	1	2	3	4	5	6
λ	11.17	11.2	11.23	11.26	11.29	11.32
μ^*	7.25	7	6.75	6.5	6.25	6
E_I	205.4764	114.0650	55.0347	22.7978	8.1025	2.5199
E_r	0.0034	0.0032	0.0035	0.0047	0.0071	0.0112
E_0	2.9041	2.7501	2.9919	3.9741	6.0793	9.5530
LS	0.0790	0.0756	0.0758	0.0760	0.0762	0.0764
β_1	0.9900	0.9916	0.9930	0.9941	0.9950	0.9955
λ_{eff}	16.5870	16.6584	16.7264	16.7905	16.8498	16.9033
L_s	0.2130	0.5934	1.6032	3.8243	7.8958	14.2035
L_q	0.0254	0.0271	0.0305	0.0350	0.0403	0.0462
W_s	0.0128	0.0356	0.0958	0.2278	0.4686	0.8403
W_q	0.0015	0.0016	0.0018	0.0021	0.0024	0.0028
GR	0.1900	0.5177	1.3913	3.3185	6.8622	12.3721
LR	0.0231	0.0756	0.2119	0.5059	1.0336	1.8314
P_B	0.0019	0.0063	0.0178	0.0431	0.0894	0.1611
P_I	0.0562	0.0487	0.0420	0.0364	0.0322	0.0299
VL_r	0.0032	0.0089	0.0240	0.0574	0.1184	0.2131
F^*	280.1642	860.5207	2392.2963	5755.4699	11916.6169	21461.2462

First, we fix the estimated parameter values as follows: reneging rate $\alpha = 0.029$, customer encouragement rate $\beta = 0.5$, retention rate $r = 0.9$, reneging rate due to server vacation $\xi = 0.015$ and replenishment lead time rate $\eta = 0.0588$. We then vary the arrival rate λ, service rate μ and the number of servers on vacation D. The corresponding numerical results are presented in Table 3. From this table the following observations are made:

(i) The idle probability of vendors P_I decreases as the arrival rate λ increases, while the expected lost sales LS declines slightly. However the total expected cost F rises significantly with an increase in λ.

(ii) Performance measures such as expected inventory level E_0, lost revenue LR, system length L_s, queue length L_q, system waiting time W_s, queue waiting time W_q, gain rate GR, blocking probability P_B, vendor loss rate VL_r and expected reneging E_r all increase as λ increases.

Similarly, fixing $\lambda = 11.17$, $\eta = 0.0588$, $\xi = 0.015$, and $D = 6.0625$, we vary α, β, r and μ. The results are summarized in Table 4, and the key implications are:

Table 4. System performance measures for $\lambda = 11.17$, $\eta = 0.0588$, $\xi = 0.015$ and $D^* = 6.0625$

α	0.024	0.025	0.026	0.027	0.028	0.029
β	0	0.1	0.2	0.3	0.4	0.5
r	0.9	0.8	0.7	0.6	0.5	0.4
μ^*	7.25	7	6.75	6.5	6.25	6
E_I	199.2917	111.2516	77.2427	52.2170	50.5371	30.8915
E_r	0.0127	0.0106	0.0094	0.0092	0.0098	0.0110
E_0	10.7812	8.9741	8.0128	7.8317	8.3305	9.3894
LS	0.0842	0.0806	0.0809	0.0811	0.0813	0.0815
β_1	0.9529	0.9689	0.9799	0.9873	0.9921	0.9952
λ_{eff}	10.6439	11.9043	13.1340	14.3358	15.5143	16.6745
L_s	2.2535	3.5798	5.3985	7.7266	10.5394	13.7777
L_q	0.0267	0.0394	0.0567	0.0797	0.1089	0.1446
W_s	0.2117	0.3007	0.4110	0.5390	0.6793	0.8263
W_q	0.0016	0.0025	0.0040	0.0061	0.0091	0.0136
GR	2.0591	3.2324	4.8261	6.8462	9.2634	12.0207
LR	0.1944	0.3473	0.5724	0.8804	1.2760	1.7570
P_B	0.0240	0.0394	0.0604	0.0872	0.1193	0.1563
P_I	0.1917	0.1400	0.0998	0.0695	0.0473	0.0315
VL_r	0.0338	0.0537	0.0810	0.1159	0.1581	0.2067
F^*	3290.1746	5320.8875	8093.9643	11633.8971	15903.9669	20815.4862

(i) The expected cost F increases as the service rate μ decreases, with higher values of α and β, and lower values of r.

(ii) Performance measures including system length L_s, queue length L_q, system waiting time W_s, queue waiting time W_q, gain rate GR, lost revenue LR, blocking probability P_B and vendor loss rate VL_r increase with decreasing μ and r and increasing α and β. Conversely, idle probability E_I, expected reneging E_r, expected inventory level E_0 and lost sales LS decrease under these parameter changes.

Finally, fixing the service rate $\mu = 5.9101$, replenishment lead time rate $\eta = 0.0588$, and number of servers on vacation $D = 6.0625$, we vary the parameters α, β, λ, r and ξ. The results are presented in Table 5, and the key findings are summarized as follows:

(i) The idle probability E_I, lost sales LS and vendor idle probability P_I all decrease with increasing values of α, β and λ, as well as with decreasing values of r and ξ. Conversely, the system length L_s, queue length L_q, system waiting time W_s, queue waiting time W_q, gain rate GR, lost revenue LR, blocking probability P_B, vendor loss rate VL_r and total expected cost F increase with increasing α, β and λ, and decreasing r and ξ.

(ii) The expected reneging E_r and expected inventory level E_0 do not exhibit consistent trends with changes in α, β, λ, r and ξ.

For the parameter values $\beta = 0.5$, $\lambda = 11.17$, $\alpha = 0.029$, $r = 0.9$, $\eta = 0.0588$, $\xi = 0.015$ and $D = 6$, two graphs are plotted: Figure 2 and Figure 3.

Table 5. System performance measures for $\mu^* = 5.9101$, $\eta = 0.0588$ and $D = 6.0625$

β	0	0.1	0.2	0.3	0.4	0.5
λ	11.17	11.2	11.23	11.26	11.29	
α	0.024	0.025	0.026	0.027	0.028	0.029
r	0.9	0.8	0.7	0.6	0.5	0.4
ξ	0.019	0.018	0.017	0.016	0.015	0.016
E_I	212.5517	179.8840	148.1061	117.9075	90.0582	65.3475
E_r	0.0098	0.0092	0.0093	0.0097	0.0105	0.0115
E_0	8.3099	7.8511	7.8729	8.2581	8.9086	9.7454
LS	0.0895	0.0857	0.0859	0.0861	0.0864	0.0866
β_1	0.9758	0.9832	0.9883	0.9917	0.9941	0.9958
λ_{eff}	10.9000	12.1130	13.3176	14.5169	15.7132	16.9086
L_s	4.5870	6.2747	8.1831	10.2568	12.4421	14.6907
L_q	0.0240	0.0313	0.0408	0.0533	0.0697	0.0909
W_s	0.4208	0.5180	0.6145	0.7066	0.7918	0.8688
W_q	0.0014	0.0020	0.0028	0.0040	0.0058	0.0083
GR	4.2076	5.6970	7.3549	9.1259	10.9579	12.8054
LR	0.3793	0.5777	0.8282	1.1309	1.4841	1.8853
P_B	0.0511	0.0705	0.0924	0.1161	0.1410	0.1667
P_I	0.1151	0.0865	0.0651	0.0491	0.0371	0.0282
VL_r	0.0872	0.1129	0.1391	0.1641	0.1866	0.2351
F^*	6858.2855	9427.3651	12327.0269	15474.3029	18788.4109	22201.5213

Figure 2 shows that the waiting time in the queue decreases as the service rate μ increases.

Figure 3 shows that the average number of customers in the queue decreases as the number of customers leaving the system after service increases. This phenomenon is directly related to the case of lost sales, as the average number of lost sales decreases with the number of customers leaving the system after receiving service becomes larger. The system's efficiency increases, indicating that it offered more services and thus consumed the inventory faster. The lost sales concerning the number of customers leaving drops as the average number of customers waiting in line decreases, the expected lost sales for new customers declines as the number of customers leaving after service from the system becomes larger. This is practically true because, under uncertain inventory order sizes, when the speed of service with respect to the number of customers in a queuing system increases, the expected loss of sales decreases.

From Figure 4 we observe that E_I decreases as arrival rate λ increases.

Figures 5 (a) and (b) demonstrate the impacts of expected service rate μ and arrival rate λ on total expected cost F. Figure 5 (a) shows that higher service rates reduce the total expected cost. Figure 5 (b) indicates that total expected cost F increases with increasing arrival rate λ.

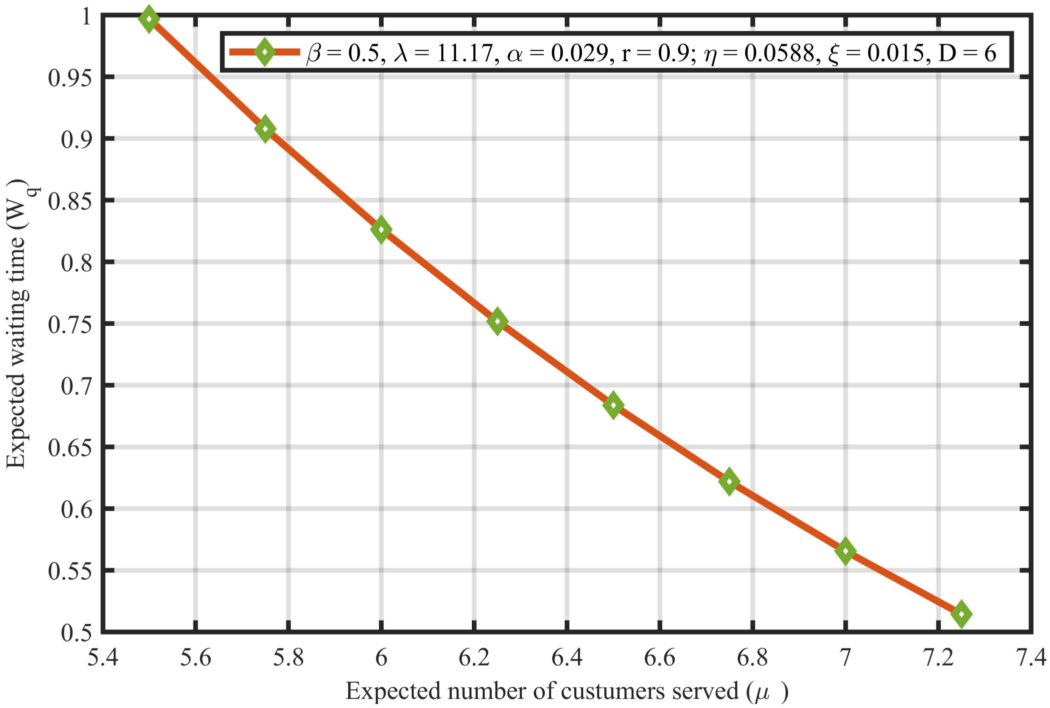


Figure 2. The effect of service rate μ on the waiting time in the queue W_q

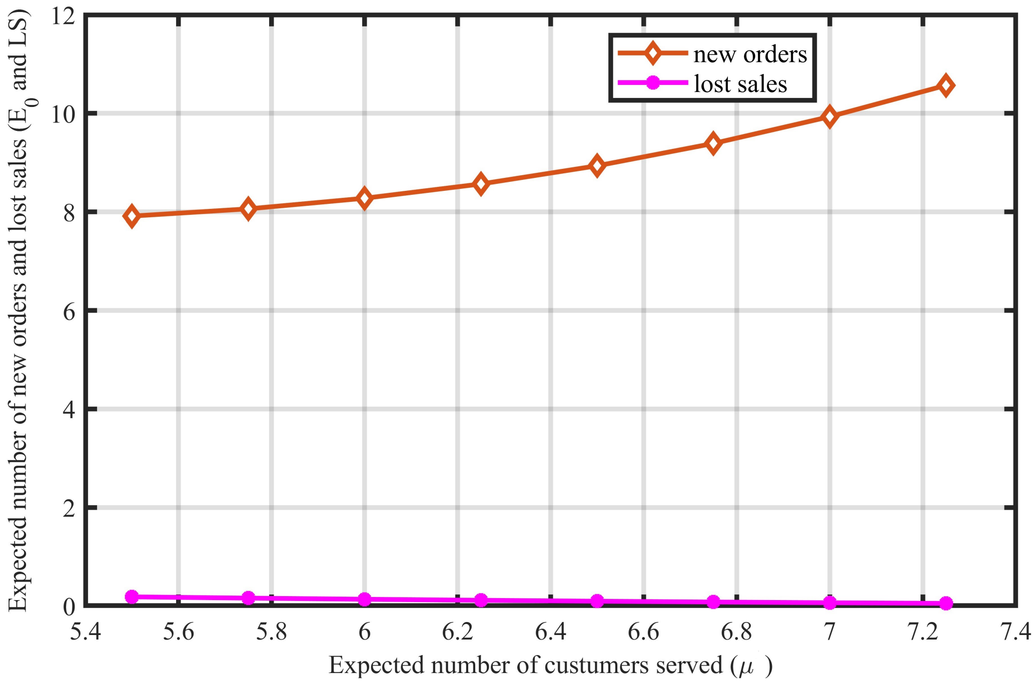


Figure 3. The effect of service rate μ on the mean number of replenishment E_0 and mean lost sales LS

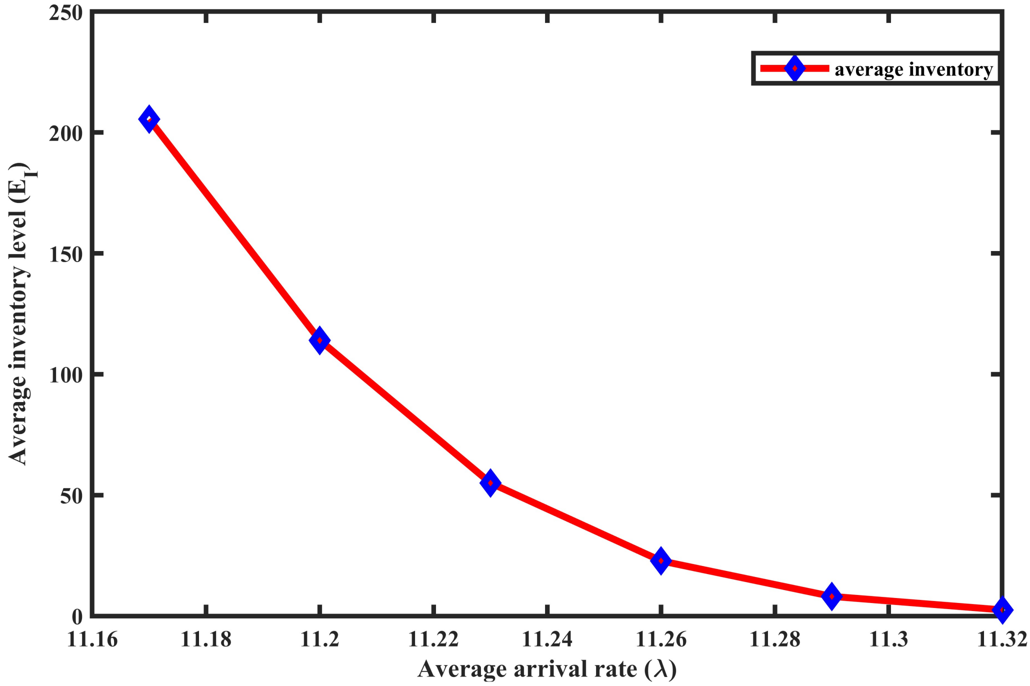
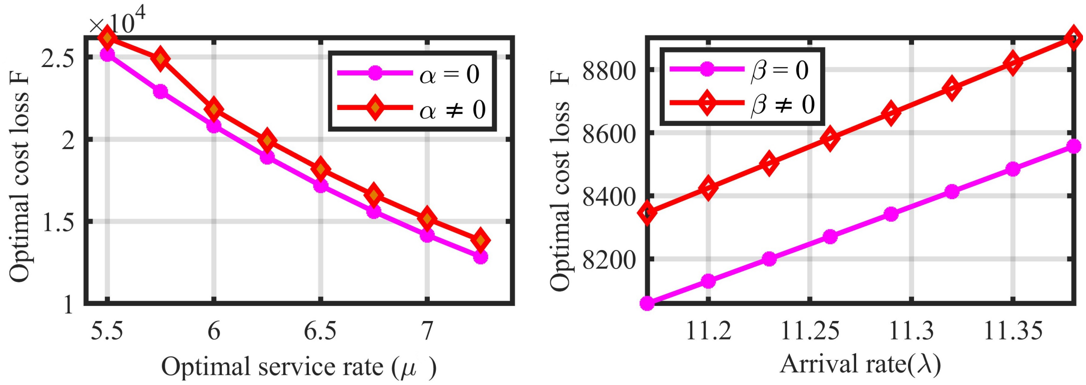


Figure 4. The effect of arrival rate λ on the mean inventory level E_I



(a) (b)

Figure 5. The effect of service rate μ and arrival rate λ on the cost function F^*

7. Conclusions

This study reveals that while key performance measures such as the mean number of customers in the system (L_s), in the queue (L_q), mean waiting times (W_s, W_q) and mean loss rate (LR) depend on the service rate, while many other metrics do not. Therefore Ethio Telecom should avoid blindly increasing service capacity since it does not guarantee a reduction in customer loss rate, although it can reduce waiting times and improve customer satisfaction. This trade-off must be carefully balanced against increased operating costs. Effective management requires an integrated approach to inventory and service processes, focusing on minimizing customer loss by maintaining favorable arrival and purchase rates, which ultimately drives profitability and growth. Hence strategic decisions should optimize service capacity improvements alongside inventory control and customer satisfaction to enhance overall system performance and business sustainability.

Future Directions

To extend the present research, several promising avenues may be explored:

- **Batch Arrivals and Retrials**: Modeling group arrivals and customer retrials to capture scenarios common in public services and healthcare.
- **Server Interruptions**: Incorporating unexpected server failures and operational disruptions for enhanced realism.
- **Advanced Arrival and Service Processes**: Using Markov Arrival Processes (MAPs) and Phase-Type (Ph) distributions to better capture variability in demand and service.
- **Neutrosophic Logic Integration**: Applying neutrosophic logic for managing uncertainty and ambiguity in customer behavior and system optimization, enabling more robust decision-making.

Acknowledgments

The authors would like to thank the editor-in-chief and the anonymous referees for their comments and suggestions which helped us to bring this manuscript in the current form.

References

[1] Ahmadi, T., Hesaraki, A. F., Mahmoodi, A., & Marandi, A. (2025). Managing customer waiting times in an inventory system using Conditional Value-at-Risk measure. *Annals of Operations Research*, 344(1), 1–45.

[2] Alnowibet, K. A., Alrasheedi, A. F., & Alqahtani, F. S. (2022). Queuing models for analyzing the steady-state distribution of stochastic inventory systems with random lead time and impatient customers. *Processes*, 10(4), 624.

[3] Ancker Jr, C. J., & Gafarian, A. V. (1963). Some queuing problems with balking and reneging. *Operations Research*, 11(1), 88–100.

[4] Arun, C. P. (2000). Queueing and inventory theory in clinical practice: application to clinical toxicology. *Annals of the New York Academy of Sciences*, 919(1), 284–287.

[5] Baccelli, F., & Bremaud, P. (2003). *Elements of queueing theory: Palm Martingale calculus and stochastic recurrences* (Vol. 26). Berlin: Springer.

[6] Choi, K. H., Yoon, B. K., & Am Moon, S. (2019). Queueing inventory systems with phase-type service distributions: a literature review. *Industrial Engineering & Management Systems*, 18(3), 330–339.

[7] Daniel J. K., & Ramanarayanan R. (1987). An inventory system with two servers and rest periods. *Journal of the Center for Operational Research Studies*, 29(1-2), 95–100.

[8] Divya, V., Krishnamoorthy, A., Vishnevsky, V. M., & Kozyrev, D. V. (2020). On a queueing system with processing of service items under vacation and $N-$policy with impatient customers. *Queueing Models and Service Management*, 3(2), 167–201.

[9] Jeganathan, K., Harikrishnan, T., Lakshmi, K. P., & Nagarajan, D. (2023). A multi-server retrial queuing-inventory system with a partial synchronous multiple vacations. *Decision Analytics Journal*, 9, 100333.

[10] Krishnamoorthy, A., Shajin, D., & Lakshmy, B. (2016). On a queuing-inventory with reservation, cancellation, common life time and retrial. *Annals of Operations Research*, 247(1), 365–389.

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

[12] Nithya, M., Joshi, G. P., Sugapriya, C., Selvakumar, S., Anbazhagan, N., Yang, E., & Doo, I. C. (2022). Analysis of stochastic state-dependent arrivals in a queuing-iventory system with multiple server vacation and retrial facility. *Mathematics*, 10(17), 3041.

[13] Rangaswamy, M. (2023). Queueing-inventory systems: A survey. *arXiv preprint*, arXiv:2308.06518.

[14] Schwarz, M., Sauer, C., Daduna, H., Kulik, R., & Szekli, R. (2006). $M/M/1$ queuing systems with inventory. *Queueing Systems*, 54, 55–78.

[15] Sharma, R., Kumar, N., & Patel, A. (2023). A review of queueing models with impatience, vacation, and retention strategies. *Journal of Queueing Systems and Applications*, 45(2), 101-–118.

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

[17] Sun, W., Xie, X., & Zhang, Z. (2025). Customer balking behaviors in a multi-server queue with synchronous multiple uninterrupted vacations under N-policy. *Queueing Models and Service Management*, 8(4), 1–26.

[18] Tempelmeier, H. (2020). *Inventory analytics: prescriptive analytics in supply chains*. BoD–Books on Demand.

[19] Tian, N., & Xu, X., (2005). The waiting time of an M/M/c queue with partial servers vacations. *OR Transactions*, 9(2), 1–8.

[20] Verma, A., Samanta, S. K., & Isotupa, K. S. (2025). A batch demands queueing-inventory model with positive replenishment time. *Queueing Models and Service Management*, 8(1), 1–20.

[21] Wang, H., & Chang, Y. C. (2002). Cost analysis of a finite capacity Markovian queue $(M/M/N)$ with balking, reneging and server breakdowns. *Mathematical Methods of Operations Research*, 56(3), 169–180.

[22] Yue, D., Ye, Z., & Yue, W. (2023). Analysis of a queuing-inventory system with synchronous vacation of multiple servers. *Queuing Models and Service Management*, 6(1), 1–26.

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

Appendix

Appendix A: Proof of Theorem 4.1

Proof. We obtain the steady-state probabilities using the recursive method. Re-arranging (1) and we get the value of $p_{0,1,s}$ as

$$p_{0,1,s} = \rho p_{0,0,0},\ s = 0, 1, 2, \cdots, q, \cdots, Q$$

which is (13). From (2),

$$\lambda(1+\beta)p_{0,0,s} - (\lambda(1+\beta)+\mu+\eta)p_{0,1,s} + (2)(\mu+\eta)p_{0,2,s} = 0;\ n = 1$$

$$\lambda(1+\beta)p_{0,0,s} - (\lambda(1+\beta)+\mu)p_{0,1,s} + (2)\mu p_{0,2,s} = 0;\ n = 1$$

Using (1), we have

$$-\lambda(1+\beta)p_{0,1,s} + 2(\mu+\eta)p_{0,2,s} = 0;\ s = 0, 1, 2, \cdots, q$$

$$-\lambda(1+\beta)p_{0,1,s} + 2\mu p_{0,2,s} = 0;\ s = q+1, q+2, \cdots, Q$$

$$\implies p_{0,2,s} = \left(\frac{(\lambda(1+\beta))^2}{2(\mu+\eta)^2} + \frac{(\lambda(1+\beta))^2}{2\mu^2}\right) p_{0,0}$$

$$= \frac{1}{2}\rho^2 p_{0,0,0};\ s = 0, 1, 2, \cdots, Q$$

where $\rho = \rho^* + \rho^{**}$, $\rho^* = \frac{\lambda(1+\beta)}{\mu+\eta}$, $\rho^* = \frac{\lambda(1+\beta)}{\mu}$
Similarly, for $n = 2$ the above procedure gives

$$p_{0,3,s} = \frac{1}{6}\rho^3 p_{0,0,0};\ s = 0, 1, 2, \cdots, Q$$

Therefore, by the inductive principle we can conclude for any n $(3 < n \leq C-D)$ as follows

$$p_{0,n,s} = \frac{1}{n!}\rho^n p_{0,0,0};\ s = 0, 1, 2, \cdots, Q$$

which is (14). From (3) for $n = C-D+1$ we can obtain

$$p_{1,C-D+1} = \frac{(\lambda(1+\beta))^{C-D+1}}{\mu+\eta+\xi}\left(\frac{b_{C-D}}{\mu+\eta+\alpha(1-r)+\xi}\right)\left(\frac{b_{C-D+1}}{\mu+\eta+2(\alpha(1-r))+\xi}\right)\cdots \left(\frac{b_{N-2}}{\mu+\eta+(N-1-(C-D))\,\alpha(1-r)+\xi}\right)p_{0,0,0};\ s = 0, 1, 2, \cdots, q.$$

$$p_{1,C-D+1} = \frac{\lambda^{C-D+1}(1-\eta_I)}{\mu+\xi}\left[\left(\frac{b_{C-D}}{\mu+\alpha+\xi}\right)\left(\frac{b_{C-D+1}}{\mu+2\alpha+\xi}\right)\right.$$
$$\left.\cdots\left(\frac{b_{N-2}}{\mu+(N-1-(C-D))\,\alpha+\xi}\right)\right]p_{0,0,0}$$

Therefore, by the inductive principle we can conclude for any $n \ (C-D<n<N)$ as follows

$$p_{1,n} = \frac{(\lambda(1+\beta))^n}{\mu+\eta+\xi}\prod_{i=C-D+1}^{N-1}[b_i]\,p_{0,0,0}$$

where

$$b_i = b_i^* + b_i^{**},\ i = C-D+1, C-D+2,\cdots,N-1$$
$$b_i^* = \frac{b_{i-1}}{\mu+\eta+(i-(C-D))\alpha(1-r)+\xi},\ i = C-D+1, C-D+2,\cdots,N-1$$
$$b_i^{**} = \frac{b_{i-1}}{\mu+(i-(C-D))\alpha(1-r)+\xi},\ i = C-D+1, C-D+2,\cdots,N-1$$

which is (16). It is not difficult to obtain (15) and (17) from (16) for $n=1$ and $n=N$ respectively. And hence, equations (14) and (16) completely determine all the steady-state probabilities. Now it only remains to determine $p_{0,0,0}$. To find $p_{0,0,0}$ we apply the normalization condition,

$$p_{0,0,0} + \sum_{s=1}^{Q}\left[\sum_{n=0}^{C-D}p_{0,n,s} + \sum_{n=C-D+1}^{N}p_{1,n,s}\right] = 1$$

$$p_{0,0,0} + Lp_{0,0,0} + \sum_{n=C-D}^{N-1}\left[\frac{(\lambda(1+\beta))^n}{\mu+\eta+\xi}\prod_{i=C-D+1}^{N-1}b_i\right]p_{0,0,0} + p_{1,N,s} = 1,\ s = 0,1,2,\cdots,q,\cdots,Q$$

where

$$b_i = b_i^* + b_i^{**},\ i = C-D+1, C-D+2,\cdots,N-1$$
$$b_i^* = \frac{b_{i-1}}{\mu+\eta+(i-(C-D))\alpha(1-r)+\xi},\ i = C-D+1, C-D+2,\cdots,N-1$$
$$b_i^{**} = \frac{b_{i-1}}{\mu+(i-(C-D))\alpha(1-r)+\xi},\ i = C-D+1, C-D+2,\cdots,N-1$$
$$p_{1,N,s} = \frac{(\lambda\,(1+\beta))^N}{\mu+\eta+\xi}\prod_{i=C-D+1}^{N-1}[b_i]\,p_{0,0,0}$$

$$\Longrightarrow p_{0,0,0} = \left[L + \sum_{n=C-D}^{N-1}\left[\frac{(\lambda(1+\beta))^n}{\mu+\eta+\xi}\prod_{i=C-D+1}^{N-1}b_i\right] + p_{1,N,s}\right]^{-1},\ s = 0,1,2,\cdots,q,\cdots,Q$$

where

$$L = 1 + \rho + \sum_{n=2}^{C-D-1} \left[(n!)^{-1} (\rho)^n \right]$$

$$\rho = \rho^* + \rho^{**}$$

$$\rho^* = \left(\frac{\lambda (1+\beta)}{\mu + \eta} \right)$$

$$\rho^{**} = \left(\frac{\lambda (1+\beta)}{\mu} \right)$$

$$b_i = b_i^* + b_i^{**},\ i = C-D+1, C-D+2, \cdots, N-1$$

$$b_i^* = \frac{b_{i-1}}{\mu + \eta + (i-(C-D))\alpha(1-r) + \xi},\ i = C-D+1, C-D+2, \cdots, N-1$$

which is (12). The system is stable for any value of the utilization factor ρ.
This completes the proof.

Appendix B: Genetic Algorithm for section 4.4

begin
Set
cost function, $F(m)$
population size, n
maximum generation, $MaxGen$
length of chromosome, l_c
minimum value of m, mm
maximum value of m, mM
for $mm \leq m \leq mM$
randomly generate an initial population of n chromosomes, $m_1, m_2, \cdots, m_n$
end for
set an iteration counter $i = 0$
compute the fitness values of each chromosome, $F(m_1), F(m_2), \cdots, F(m_n)$
while ($i \leq MaxGen$)
select a pair of chromosomes from initial population based on fitness,
Apply crossover operation on selected pair with crossover probability, p_c,
Apply mutation on the offspring with mutation probability, p_m,
Replace old population with newly generated population,
Iteration increment, $i = i + 1$
end while
return the best solution, $m - best$ and $F - best$
end