J Syst Sci Complex (2024) 37(2): 729–758

Product Form Solution of a Queuing-Inventory System with Lost Sales and Server Vacation*

YUE Dequan · ZHANG Yuying · XU Xiuli · YUE Wuyi

DOI: 10.1007/s11424-024-1207-7
Received: 21 June 2021

Abstract In this study, the authors consider an M/M/1 queuing system with attached inventory under an (s, S) control policy. The server takes multiple vacations whenever the inventory is depleted. It is assumed that the lead time and the vacation time follow exponential distributions. The authors formulate the model as a quasi-birth-and-dearth (QBD) process and derive the stability condition of the system. Then, the stationary distribution in product form for the joint process of the queue length, the inventory level, and the server's status is obtained. Furthermore, the conditional distributions of the inventory level when the server is on and operational, and when it is off due to a vacation, are derived. Using the stationary distribution, the authors obtain some performance measures of the system. The authors investigate analytically the effect of the server's vacation on the performance measures. Finally, several numerical examples are presented to investigate the effects of some parameters on the performance measures, the optimal policy, and the optimal cost.

Keywords Cost function, multiple vacations, product form solution, queuing-inventory system, (s, S) control policy.

1 Introduction

A queueing system with attached inventory is called a queueing-inventory system (QIS) in the literature. In this system, in order to satisfy each customer, items from the inventory are

YUE Dequan (Corresponding author)
School of Science, Yanshan University, Qinhuangdao 066004*, China*. Email: ydq@ysu.edu.cn.
ZHANG Yuying
School of Economics and Management, Yanshan University, Qinhuangdao 066004*, China*.
Email: zyybjzj@163.com.
XU Xiuli
School of Science, Yanshan University, Qinhuangdao 066004*, China*. Email: xxl-ysu@163.com.
YUE Wuyi
The Kyoto College of Graduate Studies for Informatics, Kyoto City 606-8225*, Japan*. Email: w_yue@kcg.edu.
*This work was supported in part by the Natural Science Foundation of China under Grant No. 71971189, the Natural Science Foundation of Hebei Province under Grant No. A2019203313, the Key Project of Scientific Research in Higher Education of Hebei Province of China under Grant No. ZD2018042, and in part by MEXT, Japan.
◇ *This paper was recommended for publication by Editor CHAI Jian.*

required to be on-hand and enough time to complete each service. For example, before items in the inventory are out of the warehouse, the items require some time for retrieval, preparation, packing, and loading, see [1]. As pointed by Zhao and Lian[2], the QIS is different from the traditional queueing system because the attached inventory influences the service, if there is no inventory on-hand, the service will be interrupted. Also, it is different from traditional inventory management because the inventory is consumed at the service rate rather than at the demand rate when there are customers queued up for service. Research on QIS and its explicit analytical solution has attracted considerable attention over recent decades.

The early research work on the QIS was done by Sigman and Simchi-Levi[3] and Melikov and Molchanov[4]. Sigman and Simchi-Levi[3] studied an M/G/1 queuing system with an attached inventory, where it was assumed that customers arriving at the system during an out-of-stock period were backlogged. They developed a light traffic heuristic for finding performance descriptions for their model. Melikov and Molchanov[4] considered a QIS in a transportation/storage system, where a user request was lost if the request arrived when the system already contained the maximum number N of user requests. The exact and approximate solution methods were proposed.

Most of the existing literature on the QIS assumes that any demands that face a zero inventory are lost. These are called lost sales. The paper by Schwarz, et al.[5] requires a special mention since it studied an M/M/1 QIS with the lost sales considered under three different inventory management policies, and it obtained a product form solution for the joint stationary distribution of the number of customers and the inventory level. This was a noteworthy outcome due to the strong correlation that exists between the number of customers joining the system during the lead time and the number of items in the inventory over that period. Subsequently, there have been several research papers on the product form solution for various QIS models.

Saffari, et al.[1] extended the M/M/1 QIS with lost sales and the (r, Q) policy studied by Schwarz, et al.[5] for a case where the lead times followed a mixed exponential distribution. They derived the product form solution for the joint stationary distribution of the queue length and the inventory level. The optimal order size was derived in exact form when the reorder point is predetermined. Saffari, et al.[6] further extended the M/M/1 QIS studied by Saffari, et al.[1] to include the case where the lead time followed a general distribution. Baek and Moon[7] studied an M/M/1 production-inventory system, where the inventory was controlled by either an internal production or an external order. They proved that the joint stationary distribution of the queue length and the inventory level has a product form solution. Krenzler and Daduna[8] studied a QIS in a random environment. They proved a necessary and sufficient condition that the stationary distribution of the joint process of the queueing length and the environment has a product form.

Krishnamoorthy, et al.[9] proposed an M/M/1 QIS with lost sales, where the item was given with a probability to a customer at his service completion epoch. For either (s, Q) or (s, S) control policy, they obtained the product form solution for the joint stationary distribution of the queue length and the inventory level. Krishnamoorthy, et al.[10] studied a supply chain model with a production centre and a distribution centre. The production inventory system adopts a

(rQ, KQ) policy, and the distribution centre adopts (s, Q) policy. The product form solution of the steady-state distribution was obtained. The effect of various performance measures is investigated. Yue, et al.[11] analyzed an M/M/1 QIS with (s, S) policy and batch demands where the size of the batch demand was assumed to follow a geometric distribution. They obtained the product form solution for the joint stationary distribution of the queue length and the inventory level. In addition, they obtained some important performance measures and the average cost functions by using these stationary distributions. More research work on QIS with lost sales can be found in a survey paper by Krishnamoorthy, et al.[12].

Utilization of a server's idle time has been discussed in the context of vacation queueing models. From an economical point of view, these vacation models are more profitable than the classical queueing models since the idle server has been utilized for performing secondary jobs. For more details on this topic, readers may refer to Doshi[13], Takagi[14], Tian and Zhang[15] and Ke, et al.[16]. However, there has been very limited research on QIS that consider a server's vacations.

Narayanan, et al.[17] were first to introduce a server vacation into a QIS with an (s, S) inventory policy. They considered a very general model where the customer demands constituted a Markovian arrival process (MAP), the service times and the vacation times all had phase type (PH) distributions, and the lead time followed a correlated process similar to the customer arrival process. The customers waiting for service were able to renege after a random time. They formulated the model as a level-dependent quasi-birth-and-dearth process and computed the steady state probabilities. Sivakumar[18] studied an M/M/1 QIS with retrial demands and multiple vacations for a server, where (s, S) policy was considered. The lead time and the vacation time were all assumed to be exponentially distributed. Demands that occurred during stock-out periods and/or during server vacation periods entered the orbit of infinite size. The stationary distribution of joint process of the inventory level and the number of customers in the orbit. They also calculate some performance measures and the long-run total expected cost rate. Padmavathi, et al.[19] investigated an (s, S) finite-source inventory system with postponed demands and a modified vacation policy. A demand that occurs during a stock out period or a server inactive period was entered into the pool, and the demands in the pool were selected if the inventory level was above s. They obtained the stationary distribution of the joint process of the mode of the server, the server status, the inventory level, and the number of demands in the pool.

Melikov, et al.[20] proposed a model for a servicing system with perishable inventory and a finite queue of impatient claims where an (s, S) inventory policy was considered, and the server could be in one of three states: Operational, early, and delaying vacations. They developed a method for approximate computation of the system's characteristics. Koroliuk, et al.[21, 22] proposed Markov QIS models with perishable inventory and an (s, S) inventory policy. In [21], it was assumed that the server took vacations if either the inventory level was zero, the queue was empty, or both. Unlike in [21], it was assumed in [22] that the server took a vacation only if there were no customers in the system at the moment its operation completed, and the server returned to operating mode only when the number of customers in the system exceeded some

thresholds. In these studies, they developed an exact method and an approximate method to find its characteristics. Manikandan and Nair[23] proposed a QIS model under (s, Q) policy with working vacation and lost sales. They computed the steady-state probability vector and various performance measures. Jeganathan and Abdul Reiyas[24] studied a QIS model under (s, Q) policy with two heterogeneous servers and working vacation, where one server is exclusively used for high priority customers and another for low priority customers. They computed the steady state distribution of the system and analyzed the distributions of the waiting times of the types of customers. Zhang, et al.[25] studied an M/M/1 QIS with lost sales and server's vacation under a random order size policy. The lead time and the vacation time were all assumed to be distributed exponentially. They obtained the product form solution for the joint stationary distribution of the queue length, the inventory level, and the status of the server under the assumption that the server takes multiple vacations once the inventory is depleted. They obtained some important performance measures and investigated the effect of the server's vacation on the performance measures of the system.

In this paper, we extend the M/M/1 queueing-inventory model with lost sales and (s, S) inventory policy studied by Schwarz, et al.[5] to include the case where the server takes multiple vacations. When the server finishes service of a customer and finds that the inventory is empty, the server leaves for a vacation. If the server finds that the inventory is not empty at the end of a vacation, it returns from the vacation and serves any customer waiting for service. Otherwise, the server takes another vacation immediately and continues in the same manner until it finds that the inventory is not empty. The vacation time and the lead time are assumed to be exponentially distributed. In contrast to the research work in [20–22], in this paper, we have produced a tractable product form solution for the stationary distribution.

We summarize the main contributions of this study as follows: (i) We obtain the stability condition of the system and show that it is independent of both the vacation time and the lead time. (ii) The product form solution for the joint stationary distribution of the queue length, the inventory level, and the status of the server are obtained. (iii) We find that the conditional distribution of the inventory level when the server is off due to a vacation is independent of the arrival rate, and that the conditional distribution of the inventory level when the server is on and operational is independent of the vacation rate. (iv) Some very simple expressions of some performance measures of our system model by means of the corresponding performance measures of its classical inventory system (CIS) model are developed. (v) We investigate the effect of the vacation parameter and some other parameters on performance measures of the system, the optimal policy, and the optimal cost.

The rest of this paper is organized as follows. The system model is described in Section 2. In Section 3, we derive the stationary condition of the system by using QBD process theory. Then, we obtain the product form solution for the stationary distribution of the system in Section 4. Furthermore, the conditional distributions of the inventory level when the server is on and operating, and off due to a vacation are derived in this section. In Section 5, some performance measures are computed. We also obtain some very simple formulas that relate the performance measures of our model and those of the corresponding CIS model. In Section 6,

we give an average cost function and consider optimization of the average cost function under the constraint of service level. We also carry out numerical analysis to investigate the effect of some parameters on the performance measures, the optimal policy, and the optimal average cost. Conclusions are given in Section 7.

2 System Model

Consider a continuous review QIS with server's vacations and lost sales. The arrival process of customers is a Poisson process with rate $\lambda > 0$. There is a single server to serve the customers one by one under a First-Come, First-Served discipline. Each customer requires exactly one item in the inventory for service. The service time is assumed to be exponentially distributed with parameter $\mu > 0$.

Multiple vacations are considered for the server. If the server finds that the inventory is empty at a service completion epoch, the server goes on vacation. On return from this vacation, if the server finds that the inventory is still empty, the server takes another vacation immediately and continues in this fashion until the server finds that the inventory is not empty. It is assumed that the vacation time follows an exponential distribution with parameter $\theta > 0$.

A continuous review (s, S) inventory policy is adopted. Each time the inventory level reaches the reorder point $s \geq 0$ an order is placed for replenishment. Upon replenishment, the inventory level is restocked to level S with $s < S$ no matter how many items are still present in the inventory.

The replenishment lead time is exponentially distributed with parameter $\eta > 0$. Customers arriving during a period when inventory is depleted or during a vacation period are rejected and lost to the system (lost sales). If the inventory is empty at the epoch that the server is ready to serve a customer that is at the head of the line, the service of the customers waiting in the queue starts the moment the next replenishment arrives.

3 Stability Condition

In this section, we develop a QBD process for the system described in Section 2 and derive the stability condition of the system.

Let $X(t)$ be the number of customers at time t, $Y(t)$ be the inventory level at time t, and $Z(t)$ be the status of the server at time t, where $Z(t)$ is 0 if the server is off due to a vacation or 1 if the server is on and operational. Then, the process $\{\Phi(t), t \geq 0\}$=$\{(X(t), Y(t), Z(t)), t \geq 0\}$ forms a continuous time Markov process with state space $\Omega = \cup_{n=0}^{\infty}\{n\}$, where

$$\{n\} = \{(n, 0, 0), (n, 1, 1), (n, 2, 1), \cdots, (n, S, 1), (n, S, 0)\}, \quad n = 0, 1, \cdots$$

is the collection of states with $X(t) = n$, called the level n. The state-transition diagram of the process $\{\Phi(t), t \geq 0\}$ is presented in Figure 1.

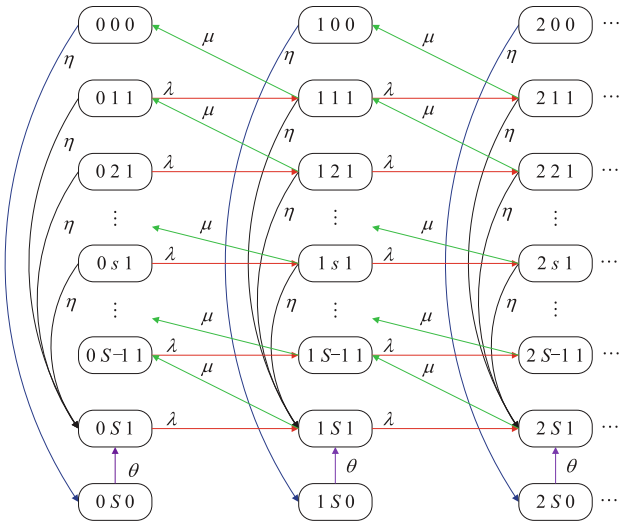


Figure 1 State-transition diagram of the process $\{\Phi(t), t \geq 0\}$

The Markov process $\{\Phi(t), t \geq 0\}$ is a QBD process by noting the assumptions made in Section 2. We have the infinitesimal generator of the process $\{\Phi(t), t \geq 0\}$ as follows:

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{A}_0 & \boldsymbol{B}_0 & & & \\ \boldsymbol{B}_2 & \boldsymbol{B}_1 & \boldsymbol{B}_0 & & \\ & \boldsymbol{B}_2 & \boldsymbol{B}_1 & \boldsymbol{B}_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

where $\boldsymbol{A}_0$, $\boldsymbol{B}_0$, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are all $(S+2) \times (S+2)$ matrices, and they are given as follows:

$$\boldsymbol{A}_0 = \begin{pmatrix} -\eta & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \eta \\ 0 & -(\lambda+\eta) & 0 & \cdots & 0 & 0 & \cdots & 0 & \eta & 0 \\ 0 & 0 & -(\lambda+\eta) & \cdots & 0 & 0 & \cdots & 0 & \eta & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(\lambda+\eta) & 0 & \cdots & 0 & \eta & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\lambda & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & -\lambda & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & -\lambda & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & \theta & -\theta \end{pmatrix},$$

$$\boldsymbol{B}_0 = \text{diag}\{0, \lambda, \lambda, \cdots, \lambda, 0\},$$
$$\boldsymbol{B_1} = \boldsymbol{A}_0 - \frac{\mu}{\lambda}\boldsymbol{B}_0,$$
$$\boldsymbol{B}_2 = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \mu & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \mu & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mu & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{pmatrix}.$$

Consider the matrix $\boldsymbol{B} = \boldsymbol{B}_0 + \boldsymbol{B}_1 + \boldsymbol{B}_2$, which is given by

$$\boldsymbol{B} = \begin{pmatrix} -\eta & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & \eta \\ \mu & -(\mu+\eta) & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \eta & 0 \\ 0 & \mu & -(\mu+\eta) & \cdots & 0 & 0 & \cdots & 0 & 0 & \eta & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(\mu+\eta) & 0 & \cdots & 0 & 0 & \eta & 0 \\ 0 & 0 & 0 & \cdots & \mu & -\mu & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & \mu & -\mu & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & \mu & -\mu & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \theta & -\theta \end{pmatrix}.$$

It is easy to see that $\boldsymbol{B}$ is an infinitesimal generator of a Markov process. Let $\boldsymbol{\tau} = (\tau(0), \tau(1), \cdots, \tau(S+1))$ denote the stationary probability vector of the Markov process with infinitesimal generator $\boldsymbol{B}$. Then, $\boldsymbol{\tau}$ satisfies the following set of equations:

$$\begin{cases} \boldsymbol{\tau B} = \boldsymbol{0}, \\ \boldsymbol{\tau e} = 1, \end{cases} \tag{1}$$

where $\boldsymbol{e}$ is a column vector of 1's of appropriate dimension. In order to derive the stability of the process $\{\Phi(t), t \geq 0\}$, we first solve Equation (1). The solution is given by the following lemma.

Lemma 3.1 *The components of the stationary probability vector $\boldsymbol{\tau}$ of are given by*

$$\tau(i) = \begin{cases} \frac{\mu}{\eta}\left(\frac{\mu}{\mu+\eta}\right)^s \kappa^{-1}, & i = 0, \\ \left(\frac{\mu}{\mu+\eta}\right)^{s-i+1} \kappa^{-1}, & i = 1, 2, \cdots, s, \\ \kappa^{-1}, & i = s+1, s+2, \cdots, S, \\ \frac{\mu}{\theta}\left(\frac{\mu}{\mu+\eta}\right)^s \kappa^{-1}, & i = S+1, \end{cases} \tag{2}$$

where

$$\kappa = \frac{\mu}{\eta} + S - s + \frac{\mu}{\theta}\left(\frac{\mu}{\mu+\eta}\right)^s. \tag{3}$$

Proof Equation (1) can be rewritten as follows:

$$-\eta\tau(0) + \mu\tau(1) = 0, \tag{4}$$

$$-(\mu+\eta)\tau(i) + \lambda\tau(i+1) = 0, \quad i = 1, 2, \cdots, s, \tag{5}$$

$$-\mu\tau(i) + \mu\tau(i+1) = 0, \quad i = s+1, s+2, \cdots, S-1, \tag{6}$$

$$-\mu\tau(S) + \eta\left(\tau(0) + \sum_{i=1}^{s}\tau(i)\right) + \theta\tau(S) = 0, \tag{7}$$

$$-\lambda\tau(0) - \theta\tau(S+1) = 0, \tag{8}$$

$$\tau(0) + \sum_{i=1}^{S+1}\tau(i) + \tau(S+1) = 1. \tag{9}$$

Solving Equation (5) recursively, we get

$$\tau(i) = \left(\frac{\mu}{\mu+\eta}\right)^{s-i+1}\tau(s+1), \quad i = 1, 2, \cdots, s. \tag{10}$$

From Equation (6), we get

$$\tau(i) = \tau(S), \quad i = s+1, s+2, \cdots, S-1. \tag{11}$$

Substituting Equation (11) with $i = s+1$ into Equation (10), we have

$$\tau(i) = \left(\frac{\mu}{\mu+\eta}\right)^{s-i+1}\tau(S), \quad i = 1, 2, \cdots, s. \tag{12}$$

From Equation (4), using Equation (12) with $i = 1$, we have

$$\tau(0) = \frac{\mu}{\eta}\tau(1) = \frac{\mu}{\eta}\left(\frac{\mu}{\mu+\eta}\right)^s\tau(S). \tag{13}$$

From Equation (8), using Equation (13), we have

$$\tau(S+1) = \frac{\eta}{\theta}\tau(0) = \frac{\mu}{\theta}\left(\frac{\mu}{\mu+\eta}\right)^s\tau(S). \tag{14}$$

Substituting Equations (12)–(14) into Equation (9), we obtain $\tau(S)$ as follows:

$$\tau(S) = \kappa^{-1}, \tag{15}$$

where κ is defined by Equation (3). This completes the proof of Lemma 3.1. ∎

Using the stationary probability vector of the infinitesimal generator $\boldsymbol{B}$ given by Lemma 3.1, we can derive the stability condition of the process $\{\Phi(t), t \geq 0\}$.

Theorem 3.2 *The process* $\{\Phi(t), t \geq 0\}$ *is positive recurrent if and only if* $\lambda < \mu$.

Proof From the matrices $\boldsymbol{B}_0$ and $\boldsymbol{B}_2$, we have

$$\boldsymbol{\tau}\boldsymbol{B}_0\boldsymbol{e} = \lambda \sum_{i=1}^{S} \tau(i) = \lambda \left(1 - \tau(0) - \tau(S+1)\right)$$

and

$$\boldsymbol{\tau}\boldsymbol{B}_2\boldsymbol{e} = \mu \sum_{i=1}^{S} \tau(i) = \mu \left(1 - \tau(0) - \tau(S+1)\right).$$

From Neuts[26], the process $\{\Phi(t), t \geq 0\}$ is positive recurrent if and only if

$$\boldsymbol{\tau}\boldsymbol{B}_0\boldsymbol{e} < \boldsymbol{\tau}\boldsymbol{B}_2\boldsymbol{e}. \tag{16}$$

Thus, Equation (16) is equivalent to the following inequality:

$$\lambda \left(1 - \tau(0) - \tau(S+1)\right) < \mu \left(1 - \tau(0) - \tau(S+1)\right).$$

From Lemma 3.1, it is easy to see that

$$1 - \tau(0) - \tau(S+1) = \frac{\frac{\mu}{\eta} + S - s - \frac{\mu}{\eta}\left(\frac{\mu}{\mu+\eta}\right)^s}{\frac{\mu}{\eta} + S - s + \frac{\mu}{\theta}\left(\frac{\mu}{\mu+\eta}\right)^s} > 0.$$

Thus, we have $\lambda < \mu$. This proves Theorem 3.2. ∎

Remark 3.3 We find from Theorem 3.2 that the stability condition for the present model agrees with the stability condition of the classical queueing system (CQS) M/M/1 with the arrival rate λ and the service rate μ. Thus, this stability condition is independent of both the vacation parameter θ and the lead time parameter η.

4 Stationary Distribution

In this section, we shall find the stationary distribution of the process $\{\Phi(t), t \geq 0\}$ by using a similar idea used in [27]. Firstly, we consider a special case of our model: $\mu \to \infty$, i.e., the case of negligible service time. This specific case of the model is denoted as Model I. Then, we use this distribution to derive the stationary distribution for our original system model described in Section 2, which is denoted as Model II.

4.1 Stationary Distribution of Model I

The corresponding Markov process for Model I is defined as $\{\widehat{\Phi}(t), t \geq 0\} = \{(\widehat{Y}(t), \widehat{Z}(t)), t \geq 0\}$, where $\widehat{Y}(t)$ and $\widehat{Z}(t)$ have the same definitions as $Y(t)$ and $Z(t)$ that were defined earlier, respectively. The state space of the process $\{\widehat{\Phi}(t), t \geq 0\}$ for Model I is given as follows:

$$\widehat{\Omega} = \{(0,0)\} \cup \{(i,1),\ i = 1, 2, \cdots, S\} \cup \{(S,0)\}.$$

The state-transition diagram of the process $\{\widehat{\Phi}(t), t \geq 0\}$ is shown by Figure 2.

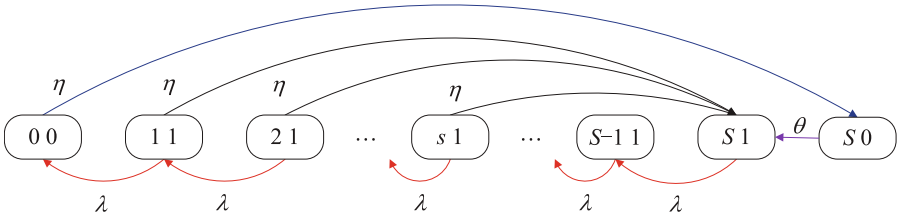


Figure 2 State-transition diagram of the state process $\{\widehat{\Phi}(t), t \geq 0\}$

The infinitesimal generator of the process $\{\widehat{\Phi}(t), t \geq 0\} = \{(\widehat{Y}(t), \widehat{Z}(t)), t \geq 0\}$ is given by

$$\widehat{\boldsymbol{Q}} = \begin{pmatrix}
-\eta & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & \eta \\
\lambda & -(\lambda+\eta) & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \eta & 0 \\
0 & \lambda & -(\lambda+\eta) & \cdots & 0 & 0 & \cdots & 0 & 0 & \eta & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -(\lambda+\eta) & 0 & \cdots & 0 & 0 & \eta & 0 \\
0 & 0 & 0 & \cdots & \lambda & -\lambda & \cdots & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & 0 & \cdots & \lambda & -\lambda & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & \lambda & -\lambda & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \theta & -\theta
\end{pmatrix}.$$

Let $\boldsymbol{\pi_v} = (\pi_v(0,0), \pi_v(1,1), \pi_v(2,1), \cdots, \pi_v(S,1), \pi_v(S,0))$ be the stationary probability vector of the process $\{\widehat{\Phi}(t), t \geq 0\}$. Then $\boldsymbol{\pi_v}$ satisfies the following equations:

$$\begin{cases} \boldsymbol{\pi_v}\widehat{\boldsymbol{Q}} = \boldsymbol{0}, \\ \boldsymbol{\pi_v}\boldsymbol{e} = 1, \end{cases} \tag{17}$$

where $\boldsymbol{e}$ is a column vector of 1's of appropriate dimension.

It is easy to see that matrix $\widehat{\boldsymbol{Q}}$ can be obtained if we change all μ in matrix $\boldsymbol{B}$ by λ. Thus, we can directly get the stationary distribution of the process $\{\widehat{\Phi}(t), t \geq 0\}$ from Lemma 3.1. Therefore, the solution of Equation (17) is given by

$$\pi_v(0,0) = \frac{\lambda}{\eta}\left(\frac{\lambda}{\lambda+\eta}\right)^s K^{-1}, \tag{18}$$

$$\pi_v(i,1) = \begin{cases} \left(\dfrac{\lambda}{\lambda+\eta}\right)^{s-i+1} K^{-1}, & i = 1, 2, \cdots, s, \\ K^{-1}, & i = s+1, s+2, \cdots, S, \end{cases} \tag{19}$$

$$\pi_v(S,0) = \frac{\lambda}{\theta}\left(\frac{\lambda}{\lambda+\eta}\right)^s K^{-1}, \tag{20}$$

where

$$K = \frac{\lambda}{\eta} + S - s + \frac{\lambda}{\theta}\left(\frac{\lambda}{\lambda+\eta}\right)^s. \tag{21}$$

Let $\theta \to \infty$, i.e., the server does not take any vacations. Thus, this special case of Model I corresponds to a classical inventory system (CIS) model. Let $\widetilde{I}$ be the inventory level in the steady state for this CIS model, and let $\theta \to \infty$ in Equations (18)–(21), then we obtain the stationary distribution of the inventory level $\widetilde{I}$ as follows:

$$P(\widetilde{I}=0) = \frac{\lambda}{\eta}\left(\frac{\lambda}{\lambda+\eta}\right)^s K_0^{-1}, \tag{22}$$

$$P(\widetilde{I}=i) = \begin{cases} \left(\dfrac{\lambda}{\lambda+\eta}\right)^{s-i+1} K_0^{-1}, & i = 1,2,\cdots,s, \\ K_0^{-1}, & i = s+1, s+2, \cdots, S, \end{cases} \tag{23}$$

where

$$K_0 = \frac{\lambda}{\eta} + S - s. \tag{24}$$

Remark 4.1 It is easy to verify that the constant K given by Equation (21) can be rewritten as the sum of the constant K_0 of the CIS model and an additional constant K_1 due to the server's vacation as follows:

$$K = K_0 + K_1,$$

where

$$K_1 = \frac{\lambda}{\theta}\left(\frac{\lambda}{\lambda+\eta}\right)^s. \tag{25}$$

4.2 Stationary Distribution of Model II

In this subsection, we derive the stationary distribution of Model II by using the stationary distribution of Model I given Subsection 4.1.

Let $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_0, \boldsymbol{\varphi}_1, \cdots)$ be the stationary probability vector of the process $\{\Phi(t), t \geq 0\}$, where $\boldsymbol{\varphi}_n = (\varphi(n,0,0), \varphi(n,1,1), \varphi(n,2,1), \cdots, \varphi(n,S,1), \varphi(n,S,0))$ is a row vector of $S+2$ dimension. Then, the vector $\boldsymbol{\varphi}$ satisfies the set of the following equations:

$$\begin{cases} \boldsymbol{\varphi}\boldsymbol{Q} = \boldsymbol{0}, \\ \boldsymbol{\varphi}\boldsymbol{e} = 1, \end{cases} \tag{26}$$

where $\boldsymbol{e}$ is a column vector of 1's of appropriate dimension.

Theorem 4.2 *If $\frac{\lambda}{\mu} < 1$, the stationary probability vector of the process $\{\Phi(t), t \geq 0\}$ is given by*

$$\boldsymbol{\varphi} = (\boldsymbol{\varphi}_0, \boldsymbol{\varphi}_1, \cdots),$$

where

$$\boldsymbol{\varphi}_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \boldsymbol{\pi_v}, \quad n = 0, 1, \cdots, \tag{27}$$

and the components of the vector

$$\boldsymbol{\pi_v} = (\pi_v(0,0), \pi_v(1,1), \pi_v(2,1), \cdots, \pi_v(S,1), \pi_v(S,0))$$

are given by Equations (18)–(20).

Proof The first equation of the set of Equation (26) can be rewritten as follows:

$$\boldsymbol{\varphi}_0 \boldsymbol{A}_0 + \boldsymbol{\varphi}_1 \boldsymbol{B}_2 = \boldsymbol{0}, \tag{28}$$

$$\boldsymbol{\varphi}_i \boldsymbol{B}_0 + \boldsymbol{\varphi}_{i+1} \boldsymbol{B}_1 + \boldsymbol{\varphi}_{i+2} \boldsymbol{B}_2 = \boldsymbol{0}, \quad i = 1, 2, \cdots. \tag{29}$$

Let

$$\boldsymbol{\varphi}_n = \gamma \left(\frac{\lambda}{\mu}\right)^n \boldsymbol{\pi_v}, \quad n = 0, 1, \cdots, \tag{30}$$

where γ is a constant.

Now, we need to verify Equation (30) satisfies Equations (28) and (29). Substituting Equation (30) into the left sides of Equations (28) and (29), we have

$$\boldsymbol{\varphi}_0 \boldsymbol{A}_0 + \boldsymbol{\varphi}_1 \boldsymbol{B}_2 = \gamma \boldsymbol{\pi_v} \left(\boldsymbol{A}_0 + \frac{\lambda}{\mu} \boldsymbol{B}_2\right)$$

and

$$\begin{aligned}
\boldsymbol{\varphi}_i \boldsymbol{B}_0 + \boldsymbol{\varphi}_{i+1} \boldsymbol{B}_1 + \boldsymbol{\varphi}_{i+2} \boldsymbol{B}_2 &= \gamma \left(\frac{\lambda}{\mu}\right)^i \boldsymbol{\pi_v} \left[\boldsymbol{B}_0 + \frac{\lambda}{\mu} \boldsymbol{B}_1 + \left(\frac{\lambda}{\mu}\right)^2 \boldsymbol{B}_2\right] \\
&= \gamma \left(\frac{\lambda}{\mu}\right)^i \boldsymbol{\pi_v} \left[\boldsymbol{B}_0 + \frac{\lambda}{\mu}\left(\boldsymbol{A}_0 - \frac{\mu}{\lambda} \boldsymbol{B}_0\right) + \left(\frac{\lambda}{\mu}\right)^2 \boldsymbol{B}_2\right] \\
&= \gamma \left(\frac{\lambda}{\mu}\right)^{i+1} \boldsymbol{\pi_v} \left(\boldsymbol{A}_0 + \frac{\lambda}{\mu} \boldsymbol{B}_2\right), \quad i = 0, 1, \cdots.
\end{aligned}$$

From the structure of the matrices $\boldsymbol{A}_0$, $\boldsymbol{B}_2$ and $\widehat{\boldsymbol{Q}}$, it is easy to verify that

$$\boldsymbol{A}_0 + \frac{\lambda}{\mu} \boldsymbol{B}_2 = \widehat{\boldsymbol{Q}}. \tag{31}$$

From Equations (17) and (31), we have $\boldsymbol{\pi_v}(\boldsymbol{A}_0 + \frac{\lambda}{\mu}\boldsymbol{B}_2) = \boldsymbol{\pi_v}\widehat{\boldsymbol{Q}} = \boldsymbol{0}$. Hence, the right sides of Equations (28) and (29) are zero. Thus, Equations (28) and (29) are satisfied with the assumption given by Equation (30). Applying the normalizing condition $\boldsymbol{\varphi e} = 1$ and noting that $\boldsymbol{\pi_v e} = 1$, we get

$$\gamma \sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^i = 1. \tag{32}$$

Hence, if $\frac{\lambda}{\mu} < 1$ we get from Equation (32) that $\gamma = 1 - \frac{\lambda}{\mu}$. Thus, we complete the proof of Theorem 4.2. ∎

Remark 4.3 Theorem 4.2 shows that the stationary distribution of Model II has a product form solution for the two distributions: One distribution is the stationary distribution of the queue length in the M/M/1 CQS, and the other distribution is the stationary distribution of the inventory level in Model I.

We denote X, Y and Z to be the corresponding variables of $X(t)$, $Y(t)$ and $Z(t)$ under the steady state, respectively. From Theorem 4.2, we can obtain the marginal distributions of theses random variables.

Corollary 4.4 (i) *The marginal stationary distribution of the queue length X is given by*

$$P(X = n) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n, \quad n = 0, 1, \cdots, \tag{33}$$

which is equal to the stationary distribution of the queue length in M/M/1 CQS with arrival rate λ and service rate μ.

(ii) *The marginal stationary distribution of the inventory level Y is given by*

$$P(Y = i) = \begin{cases} \dfrac{\lambda}{\eta}\left(\dfrac{\lambda}{\lambda+\eta}\right)^s K^{-1}, & i = 0, \\ \left(\dfrac{\lambda}{\lambda+\eta}\right)^{s-i+1} K^{-1}, & i = 1, 2, \cdots, s, \\ K^{-1}, & i = s+1, s+2, \cdots, S-1, \\ \left[1 + \dfrac{\lambda}{\theta}\left(\dfrac{\lambda}{\lambda+\eta}\right)^s\right] K^{-1}, & i = S, \end{cases} \tag{34}$$

where K is given by Equation (21).

(iii) *The marginal stationary distribution of the server's status Z is given by*

$$P(Z = j) = \begin{cases} \left(\dfrac{\lambda}{\theta} + \dfrac{\lambda}{\eta}\right)\left(\dfrac{\lambda}{\lambda+\eta}\right)^s K^{-1}, & j = 0, \\ \left[\dfrac{\lambda}{\eta} + S - s - \dfrac{\lambda}{\eta}\left(\dfrac{\lambda}{\lambda+\eta}\right)^s\right] K^{-1}, & j = 1, \end{cases} \tag{35}$$

where K is given by Equation (21).

Proof The results are directly obtained by using Theorem 4.2. ∎

Remark 4.5 (i) Equation (33) shows that the stationary distribution of the queue length is only dependent on the parameters λ and μ, and it is not dependent on the other system parameters. (ii) Equations (34) and (35) show that both the marginal stationary distribution of the inventory level and the marginal stationary distribution of the server's status are not dependent on the service rate μ.

4.3 Comparison of the Distributions of the Inventory Level Between Model II and Its Corresponding CIS Model

In the following, we compare the marginal stationary distribution of the inventory level Y for Model II with that for its corresponding CIS model.

Comparing Equations (18)–(20) and Equations (22) and (23), it is easy to have the following relations between the stationary distributions for Model I and its corresponding CIS model:

$$\pi_v(0,0) = P(\widetilde{I}=0)K_v, \tag{36}$$

$$\pi_v(i,1) = P(\widetilde{I}=i)K_v, \quad i=1,2,\cdots,S, \tag{37}$$

$$\pi_v(S,0) = 1-K_v, \tag{38}$$

where

$$K_v = \frac{\frac{\lambda}{\eta}+S-s}{\frac{\lambda}{\eta}+S-s+\frac{\lambda}{\theta}\left(\frac{\lambda}{\lambda+\eta}\right)^s}. \tag{39}$$

Using relations given by Equations (36)–(38), we have from Theorem 4.2 that

$$P(Y=0) = \pi_v(0,0) = P(\widetilde{I}=0)K_v, \tag{40}$$

$$P(Y=i) = \pi_v(i,1) = P(\widetilde{I}=i)K_v, \quad i=1,2,\cdots,S-1, \tag{41}$$

$$P(Y=S) = \pi_v(S,1)+\pi_v(S,0) = P(\widetilde{I}=S)K_v+1-K_v. \tag{42}$$

Using Equations (40) and (41) and noting $0<K_v<1$, we have

$$P(Y=i) < P(\widetilde{I}=i), \quad i=0,1,\cdots,S-1. \tag{43}$$

Using Equation (42), we have

$$\begin{aligned} P(Y=S) &= P(\widetilde{I}=S)K_v+1-K_v \\ &= P(\widetilde{I}=S)+(1-K_v)(1-P(\widetilde{I}=S)) \\ &> P(\widetilde{I}=S). \end{aligned} \tag{44}$$

Remark 4.6 From Equations (43) and (44), we have the following observations: (i) The marginal probability of the inventory level Y at any level, except the maximum inventory level S for Model II, is less than the probability of the inventory level $\widetilde{I}$ for its corresponding CIS model. (ii) The marginal probability of the inventory level Y at the maximum inventory level S for Model II is larger than the probability of the inventory level $\widetilde{I}$ for its corresponding CIS model.

4.4 Conditional Distribution of the Inventory Level

We consider the conditional distributions of the inventory level when the server is off due to a vacation or is on and operational.

Theorem 4.7 (i) *The conditional distribution of the inventory level when the server is off due to a vacation is given by*

$$P(Y=i|Z=0)=\begin{cases}\dfrac{\theta}{\eta+\theta}, & i=0,\\ \dfrac{\eta}{\eta+\theta}, & i=S.\end{cases} \tag{45}$$

(ii) *The conditional distribution of the inventory level when the server is on and operational is given by*

$$P(Y=i|Z=1)=\begin{cases}\left(\dfrac{\lambda}{\lambda+\eta}\right)^{s}K_c^{-1}, & i=1,2,\cdots,s,\\ K_c^{-1}, & i=s+1,s+2,\cdots,S,\end{cases} \tag{46}$$

where

$$K_c=S-s+\frac{\lambda}{\eta}\left[1-\left(\frac{\lambda}{\lambda+\eta}\right)^{s}\right]. \tag{47}$$

Proof Using Equations (18)–(20) and Equation (35), it is easy to obtain the conditional distributions of the on-hand inventory when the server is off due to a vacation and when the server is on and operational as presented by Equations (45) and (46), respectively. ∎

Remark 4.8 (i) From Equation (45), it is observed that the conditional inventory level when the server is off due to vacation is either 0 or S, and it follows a Bernoulli distribution. Also, the conditional distribution of the inventory level when the server is off due to a vacation is independent of the arrival rate λ, and that it is not dependent on parameters η and θ individually but only on their proportions η/θ. So, the arrival process of customers does not influence the conditional inventory level when the server is off due to a vacation. (ii) From Equations (46) and (47), it is observed that the conditional distribution of the inventory level when the server is on and operational is independent of the vacation rate θ. So, the vacation parameter does not influence the conditional inventory level when the server is on and operational.

From Equation (45), we obtain the conditional mean inventory level when the server is off due to a vacation which is given by

$$E(Y|Z=0)=\frac{\eta S}{\eta+\theta}. \tag{48}$$

From Equation (46), we obtain the conditional mean inventory level when the server is on and operational which is given by

$$E(Y|Z=1)=\frac{\left(S-s+\frac{\lambda}{\eta}\right)I}{S-s+\frac{\lambda}{\eta}\left[1-\left(\frac{\lambda}{\lambda+\eta}\right)^{s}\right]}, \tag{49}$$

where I is the mean inventory level for the CIS model.

Remark 4.9 From Equations (48) and (49), we have the following observations: (i) The conditional mean inventory level when the server is off due to a vacation $E(Y|Z=0)$ is less than the maximum inventory level S. (ii) The conditional mean inventory level when the server is on and operational $E(Y|Z=1)$ is larger than the mean inventory I for the CIS model.

5 Performance Measures of the System

In this section, we are interested in the stationary characteristics of the system. Having determined the stationary distribution given by Theorem4.2, we can compute several performance measures of the operating characteristics for the system explicitly.

We introduce some notations used in the rest of our paper. Notations used for Model II and its corresponding CIS model or its corresponding M/M/1 CQS model are listed in Table 1, where notations with subscript 'v' are for Model II, and notations without any subscripts are for the CIS model or the M/M/1 CQS model.

Table 1 Notations for the performance measures

Symbol	Description
$I,\ I_v$	The mean inventory level
$L,\ L_v$	The mean number of lost sales per unit of time
$\widetilde{L},\ \widetilde{L}_v$	The mean number of lost sales per cycle
$A,\ A_v$	The mean arrival rate of customers who are admitted to the system per unit of time
$R,\ R_v$	The mean reorder rate per unit of time
$\beta,\ \beta_v$	The service level
$N,\ N_v$	The mean number of the customers in the system
$\widetilde{N},\ \widetilde{N}_v$	The mean number of the waiting customers in the queue
$W,\ W_v$	The mean sojourn time of the customers in the system
$\widetilde{W},\ \widetilde{W}_v$	The mean waiting time of the customers in the queue

5.1 Performance Measures Related to Inventory

According to Schwarz, et al.[5] (pp. 60–61), definitions for a cycle and a service level are given by the following. The definitions of the other performance measures in Table 1 are readily known.

Definition 5.1 A cycle and a service level are defined as follows:

(i) A cycle is the time between the placing of two successive orders.

(ii) A service level is defined by

$$\beta = \frac{E(\text{demand satisfied per unit of time})}{E(\text{total demand per unit of time})}.$$

The above definition of the service level is called β-service level by Schwarz, et al.[5], and it is also called the fill rate in [28]. As pointed out by Schwarz, et al.[5], the β-service level is a quantity-originated service measure describing the proportion of demands that are met from stock without accounting for the duration of a stock out.

Firstly, we present some performance measures which have been given by Schwarz, et al.[5] for the CIS model by the following lemma. These performance measures can be also derived from Equations (22) and (23).

Lemma 5.2 *For the CIS Model, we have the following performance measures:*

(i) *The mean inventory level is*

$$I = \left\{ \frac{\lambda}{\eta} \left[s - \frac{\lambda}{\eta} + \frac{\lambda}{\eta} \left(\frac{\lambda}{\lambda+\eta} \right)^s \right] + \frac{1}{2}(S-s)(S+s+1) \right\} K_0^{-1}. \tag{50}$$

(ii) *The mean number of lost sales per unit of time is*

$$L = \frac{\lambda^2}{\eta} \left(\frac{\lambda}{\lambda+\eta} \right)^s K_0^{-1}. \tag{51}$$

(iii) *The mean arrival rate of customers who are admitted to the system per unit of time is*

$$A = \lambda \left[1 - \frac{\lambda}{\eta} \left(\frac{\lambda}{\lambda+\eta} \right)^s K_0^{-1} \right]. \tag{52}$$

(iv) *The mean number of replenishment (reorder rate) per unit of time is*

$$R = \lambda K_0^{-1}. \tag{53}$$

(v) *The mean number of lost sales per cycle is*

$$\widetilde{L} = \frac{\lambda}{\eta} \left(\frac{\lambda}{\lambda+\eta} \right)^s. \tag{54}$$

(vi) *The service level is*

$$\beta = 1 - \frac{\lambda}{\eta} \left(\frac{\lambda}{\lambda+\eta} \right)^s K_0^{-1}. \tag{55}$$

Proof The proof of Lemma 5.2 can be found in [5] (see Theorem 4.3, pp. 66–67). It is easy to see that all these performance measures are independent of the service rate μ. So, all these performance measures holds for the CIS model. ∎

Secondly, we compute some performance measures from view point of inventory for Model II. From Remark 4.5, we know that both the marginal stationary distribution of the inventory level and the marginal stationary distribution of the server's status do not depend on the service rate μ. Thus, we can compute those performance measures that relate to inventory for Model II by using the stationary distribution for Model I that has been given in Subsection 4.1. Furthermore, using the relations given by Equations (36)–(38), we can obtain some very simple formula for the performance measures of Model II by means of the performance measures of the corresponding CIS model that were given by Lemma 5.2.

Theorem 5.3 *For Model II, the performance measures related to inventory are given as follows:*

(i) *The mean inventory level is*

$$I_v = K_v I + (1 - K_v)S. \tag{56}$$

(ii) *The mean number of lost sales per unit of time is*

$$L_v = K_v L + (1 - K_v)\lambda. \tag{57}$$

(iii) *The mean arrival rate of customers who are admitted to the system per unit of time is*

$$A_v = K_v A. \tag{58}$$

(iv) *The mean number of replenishment (reorder rate) per unit of time is*

$$R_v = K_v R. \tag{59}$$

(v) *The mean number of lost sales per cycle is*

$$\widetilde{L}_v = \widetilde{L} + K_1. \tag{60}$$

(vi) *The service level is*

$$\beta_v = K_v \beta. \tag{61}$$

Proof (i) Using Equations (36)–(38), one can easily show that the mean inventory level

$$I_v = \sum_{i=0}^{S} i\pi_v(i,1) + S\pi_v(S,0) = K_v I + (1 - K_v)S.$$

(ii) Lost sales occurs when a demand arrives during a vacation period. Hence, the mean number of lost sales incurred per unit of time is

$$L_v = \lambda(\pi_v(0,0) + \pi_v(S,0)). \tag{62}$$

Substituting Equations (36) and (38) into Equation (62) with some algebra, Equation (57) can be obtained.

(iii) Using Equation (57), the mean arrival rate of customers who are admitted to the system per unit of time is

$$A_v = \lambda - L_v = K_v A.$$

(iv) The mean reorder rate per unit of time is

$$R_v = \eta \left(\sum_{i=0}^{s} \pi_v(i,1) + \pi_v(0,0) \right) = \eta \sum_{i=0}^{s} K_v \pi(i) = K_v R.$$

(v) Obviously, the mean cycle time is R_v^{-1}. Thus, the mean number of lost sales per cycle is

$$\widetilde{L}_v = R_v^{-1} L_v.$$

Using Equations (57) and (59) with some algebra, Equation (60) can be obtained.

(vi) Using Equation (57), one can easily show that the service level is

$$\beta_v = \frac{\lambda - L_v}{\lambda} = K_v \beta.$$

The proof of Theorem 5.3 is completed. ∎

Remark 5.4 From Theorem 5.3, we observed the following relations between Model II and its corresponding CIS model: (i) The mean inventory level I_v is the weighted average summation of I and S, and the mean number of lost sales incurred per unit of time L_v is the weighted average summation of L and the arrival rate λ. (ii) The mean number $\widetilde{L}_v$ of lost sales per cycle can be decomposed into the summation of $\widetilde{L}$ and additional constant K_1 due to server's vacations. (iii) The performance measures R_v, A_v and β_v for Model II are K_v times of their corresponding performance measures R, A and β for the CIS model, respectively. In other words, each of the ratios $\frac{A_v}{A}$, $\frac{R_v}{R}$ and $\frac{\beta_v}{\beta}$ is equal to a constant K_v. All these relationships can be classified into three types of relations: weighted average summation, proportion, and summation. They are summarized in Table 2.

Table 2 Relations of performance measures between Model II and its corresponding CIS model

Performance measures	Relationships	Types of relation
$I,\ I_v$	$I_v = K_v I + (1 - K_v)S$	Weighted average summation
$L,\ L_v$	$L_v = K_v L + (1 - K_v)\lambda$	Weighted average summation
$\widetilde{L},\ \widetilde{L}_v$	$\widetilde{L}_v = \widetilde{L} + K_1$	Summation
$A,\ A_v$	$A_v = K_v A$	Proportion
$R,\ R_v$	$R_v = K_v R$	Proportion
$\beta,\ \beta_v$	$\beta_v = K_v \beta$	Proportion

5.2 Performance Measures Related to Queue

Now, we compute some performance measures related to queue for Model II. Let N_v and $\widetilde{N}_v$ be the mean number of customers in the system and the mean number of waiting customers in the queue, respectively.

From Corollary 4.4 (i), the marginal stationary distribution of the number of customers of the system for Model II is the same as that of the number of customers for the M/M/1/∞ CQS. Thus, the mean number of customers in the system for Model II is given by

$$N_v = \frac{\lambda}{\mu - \lambda}. \tag{63}$$

From Theorem 4.2, the mean number of waiting customers in the queue is given by

$$\begin{aligned}\widetilde{N}_v=&\sum_{n=0}^{\infty}\sum_{i=1}^{S}(n-1)\varphi(n,i,1)+\sum_{n=0}^{\infty}n[\varphi(n,0,0)+\varphi(n,S,0)]\\ =&\sum_{n=0}^{\infty}(n-1)\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n P(Z=1)\\ &+\sum_{n=0}^{\infty}n\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n P(Z=0).\end{aligned}\tag{64}$$

Substituting Equation (35) into Equation (64), we have

$$\widetilde{N}_v=\frac{\lambda}{\mu-\lambda}-\frac{\lambda}{\mu}K_q^{-1},\tag{65}$$

where

$$K_q=\frac{\frac{\lambda}{\eta}+S-s+\frac{\lambda}{\theta}\left(\frac{\lambda}{\lambda+\eta}\right)^s}{\frac{\lambda}{\eta}+S-s-\frac{\lambda}{\eta}\left(\frac{\lambda}{\lambda+\eta}\right)^s}.\tag{66}$$

Let W_v and $\widetilde{W}_v$ be the mean sojourn time of the customers in the system and the mean waiting time of the customers in the queue, respectively. Using Little's formula and Equations (57), (63) and (65), the mean sojourn time of the customers in the system is given by

$$W_v=\frac{N_v}{A_v}=\frac{K_q}{\mu-\lambda},\tag{67}$$

and the mean waiting time of the customers in the queue is given by

$$\widetilde{W}_v=\frac{\widetilde{N}_v}{A_v}=\frac{K_q}{\mu-\lambda}-\frac{1}{\mu},\tag{68}$$

where K_q is given by Equation (66).

Remark 5.5 From Equations (63), (65), (67) and (68), taking the first and the second derivatives, it is easy to see that the performance measures N_v, $\widetilde{N}_v$, W_v and $\widetilde{W}_v$ are decreasing and convex in μ.

5.3 The Effect of the Vacation Rate on Some Performance Measures

In this subsection, we investigate the effect of the vacation rate θ on some performance measures by using the results obtained in last two subsections.

From Table 2 and Equations (65)–(68), it is observed that the constants K_v and K_q play a key role in relating the performance measures of Model II with those of its corresponding CIS model or CQS model. We have the following properties for the two constants K_v and K_q.

Proposition 5.6 (i) *The constant K_v is strictly increasing and concave in θ.*

(ii) *The constant K_q is strictly decreasing and convex in θ.*

Proof (i) Taking the first and the second derivative for K_v with respect to the parameter θ, we have

$$\frac{dK_v}{d\theta} = \lambda \left(\frac{\lambda}{\lambda+\eta}\right)^s \frac{K_0}{(\theta K)^2} > 0$$

and

$$\frac{d^2K_v}{d\theta^2} = -2\lambda \left(\frac{\lambda}{\lambda+\eta}\right)^s \frac{K_0^2}{(\theta K)^3} < 0.$$

Hence, the constant K_v is strictly increasing and concave in θ.

(ii) Taking the first and the second derivative for K_q with respect to the parameter θ, we have

$$\frac{dK_q}{d\theta} = -\frac{\lambda}{\theta^2} \left(\frac{\lambda}{\lambda+\eta}\right)^s K_c^{-1} < 0 \tag{69}$$

and

$$\frac{d^2K_q}{d\theta^2} = \frac{2\lambda}{\theta^3} \left(\frac{\lambda}{\lambda+\eta}\right)^s K_c^{-1} > 0. \tag{70}$$

Hence, the constant K_q is strictly decreasing and convex in θ. ∎

Now, we have the following monotonicity of some performance measures with respect to the vacation rate.

Theorem 5.7 *For the performance measures of Model II, we have the following properties:*
(i) *I_v, L_v and $\widetilde{L}_v$ are strictly decreasing and convex in θ.*
(ii) *A_v, R_v and β_v are strictly increasing and concave in θ.*
(iii) *$\widetilde{N}_v$, W_v and $\widetilde{W}_v$ are strictly decreasing and convex in θ.*

Proof (i) Using the relations for I_v and L_v presented in Table 2, we have

$$\frac{dI_v}{d\theta} = (I-S)\frac{dK_v}{d\theta}, \quad \frac{d^2I_v}{d\theta^2} = (I-S)\frac{d^2K_v}{d\theta^2},$$

and

$$\frac{dL_v}{d\theta} = (L-\lambda)\frac{dK_v}{d\theta}, \quad \frac{d^2L_v}{d\theta^2} = (L-\lambda)\frac{d^2K_v}{d\theta^2}.$$

Obviously, $I - S < 0$ and $L - \lambda < 0$. Thus, using Proposition 5.6 (i), we prove that I_v and L_v are strictly decreasing and convex in θ. Using the relation for $\widetilde{L}_v$ presented in Table 2, we have

$$\frac{d\widetilde{L}_v}{d\theta} = \frac{dK_1}{d\theta} = -\frac{\lambda}{\theta^2}\left(\frac{\lambda}{\lambda+\eta}\right)^s < 0,$$

$$\frac{d^2\widetilde{L}_v}{d\theta^2} = \frac{d^2K_1}{d\theta^2} = \frac{2\lambda}{\theta^3}\left(\frac{\lambda}{\lambda+\eta}\right)^s > 0.$$

Hence, $\widetilde{L}_v$ is strictly decreasing and convex in θ.

(ii) From Table 2, A_v, R_v and β_v are K_v times of the corresponding performance measures A, R and β, respectively. Note that A, R and β are positive and independent of θ, and the result of Theorem 5.7 (ii) obviously holds.

(iii) Using Equations (69) and (70), from Equation (65), we have

$$\frac{d\widetilde{N}_v}{d\theta}=\frac{\lambda}{\mu}K_q^{-2}\frac{dK_q}{d\theta}<0,$$

$$\frac{d^2\widetilde{N}_v}{d\theta^2}=\frac{\lambda}{\mu}K_q^{-3}\left[K_q\frac{d^2K_q}{d\theta^2}-2\left(\frac{dK_q}{d\theta}\right)^2\right]$$

$$=\frac{2\lambda^2K_v}{\mu\theta^3K_q}\left(\frac{\lambda}{\lambda+\eta}\right)^s>0.$$

Thus, we prove that $\widetilde{N}_v$ is strictly decreasing and convex in θ. From Equations (67) and (68), using Proposition 5.6 (ii), it is clear that W_v and $\widetilde{W}_v$ are strictly decreasing and convex in θ. ▌

6 Numerical Analysis

In this section, we develop a cost function by using the performance measures obtained in Section 5 and present some numerical analyses to investigate the effect of some parameters on the performance measures, the optimal policy, and the optimal cost.

6.1 Effect of the System Parameters on Some Performance Measures

In this subsection, we study the effect of the system parameters such as the arrival rate λ, the replenishment rate η and the vacation rate θ on the performance measures since the service rate μ only affects the performance measures such as N_v, $\widetilde{N}_v$, W_v and $\widetilde{W}_v$ and its effects are well detailed in Remark 5.5. The results are given in Tables 3, 4 and 5. We fix the reorder point $s = 6$, and the maximum inventory level $S = 15$, and fix the cost parameter values as $C_1 = 30$, $C_2 = 100$, $C_3 = 500$ and $C_4 = 50$.

As is to be expected, it is observed from Table 3 that L_v, $\widetilde{L}_v$, $\widetilde{N}_v$, W_v and $\widetilde{W}_v$ increase significantly as λ increases. We also note that the mean number of lost sales per cycle $\widetilde{L}_v$ is much larger than the mean number of lost sales per unit of time L_v when λ becomes increasingly larger. This is due to the very small value of the mean reorder rate per unit of time R_v and comparatively high value of L_v, which explains the observation by noting the relation $\widetilde{L}_v = L_vR_v^{-1}$. From Table 3, we observe that I_v, A_v and R_v firstly increase and then decrease with an increase in the arrival rate λ. However, the values of I_v, A_v and R_v vary only slightly even when λ varies significantly. This means that I_v, A_v and R_v are less sensitive than other performance measures. It is observed that the service level β_v decreases significantly with the arrival rate λ. For instance, the service level β_v is below 51% even when λ is larger than 4.

Table 3 The effect of the arrival rate λ on the performance measures for $\eta = 1$, $\theta = 0.1$ and $\mu = 30$

λ	I_v	L_v	$\widetilde{L}_v$	A_v	R_v	β_v	$\widetilde{N}_v$	W_v	$\widetilde{W}_v$
1	10.4805	0.0169	0.1719	0.9831	0.0985	0.9831	0.0006	0.0351	0.0006
4	11.0274	1.9645	11.5343	2.0355	0.1703	0.5089	0.0756	0.0756	0.0371
7	11.2242	5.1017	34.5572	1.8983	0.1476	0.2712	0.2218	0.1603	0.1168
10	11.0378	8.2299	62.0921	1.7701	0.1325	0.1770	0.4115	0.2825	0.2325
13	10.7901	11.3134	91.6701	1.6866	0.1234	0.1297	0.6655	0.4534	0.3946
16	10.5560	14.3697	122.3317	1.6303	0.1175	0.1019	1.0264	0.7010	0.6296
19	10.3500	17.4098	153.6342	1.5902	0.1133	0.0837	1.5827	1.0862	0.9953
21	10.2284	19.4305	174.7397	1.5695	0.1112	0.0747	2.1589	1.4867	1.3756
24	10.0672	22.4555	206.6481	1.5445	0.1087	0.0644	3.7426	2.5898	2.4231
27	9.9277	25.4751	238.7761	1.5249	0.1067	0.0565	8.4917	5.9019	5.5686

Table 4 The effect of the replenishment rate η on the performance measures with $\lambda = 10$, $\theta = 0.1$ and $\mu = 30$

η	I_v	L_v	$\widetilde{L}_v$	A_v	R_v	β_v	$\widetilde{N}_v$	W_v	$\widetilde{W}_v$
1	11.0378	8.2299	62.0921	1.7701	0.1325	0.1770	0.4115	0.2825	0.2325
2	12.1067	7.4046	35.1643	2.5954	0.2106	0.2595	0.3702	0.1926	0.1426
3	12.2888	6.4773	21.4082	3.5227	0.3026	0.3523	0.3229	0.1419	0.0919
4	12.1680	5.4933	13.6131	4.5067	0.4035	0.4507	0.2747	0.1109	0.0609
5	11.9381	4.5274	8.9547	5.4726	0.5056	0.5473	0.2264	0.0914	0.0414
6	11.6878	3.6445	6.0598	6.3555	0.6014	0.6355	0.1822	0.0787	0.0287
7	11.4595	2.8838	4.2021	7.1162	0.6863	0.7116	0.1442	0.0703	0.0203
8	11.2698	2.2569	2.9769	7.7431	0.7581	0.7743	0.1128	0.0646	0.0146
9	11.1213	1.7564	2.1492	8.2436	0.8172	0.8244	0.0878	0.0607	0.0107
10	11.0097	1.3649	1.5781	8.6351	0.8649	0.8635	0.0682	0.0579	0.0079

Table 5 The effect of the vacation rate θ on the performance measures with $\lambda = 10$, $\eta = 1$ and $\mu = 30$

θ	I_v	L_v	$\widetilde{L}_v$	A_v	R_v	β_v	$\widetilde{N}_v$	W_v	$\widetilde{W}_v$
0.5	8.5479	5.5908	16.9342	4.4092	0.3301	0.4409	0.2795	0.1134	0.0634
1.0	7.5951	4.5809	11.2895	5.4191	0.4058	0.5419	0.2290	0.0922	0.0423
1.5	7.1725	4.1329	9.4079	5.8671	0.4393	0.5867	0.2066	0.0852	0.0352
2.0	6.9339	3.8800	8.4671	6.1200	0.4582	0.6120	0.1940	0.0817	0.0317
2.5	6.7806	3.7175	7.9026	6.2825	0.4704	0.6282	0.1859	0.0796	0.0296
3.0	6.6737	3.6043	7.5263	6.3957	0.4789	0.6396	0.1802	0.0782	0.0282
3.5	6.5950	3.5209	7.2575	6.4791	0.4851	0.6479	0.1760	0.0772	0.0272
4.0	6.5347	3.4569	7.0559	6.5431	0.4899	0.6543	0.1728	0.0764	0.0264
4.5	6.4869	3.4062	6.8991	6.5938	0.4937	0.6594	0.1703	0.0758	0.0258
5.0	6.4481	3.3651	6.7737	6.6349	0.4968	0.6635	0.1683	0.0754	0.0254

From Table 4, we observe that an increase in the replenishment rate η results in a decrease in the performance measures L_v, $\widetilde{L}_v$, $\widetilde{N}_v$, W_v and $\widetilde{W}_v$. This agrees with our intuitive expectation. Moreover, as η increases, L_v and $\widetilde{L}_v$ become closer. We also observe that as η increases, there is a slight increase in A_v and R_v as expected. This is a consequence of a decrease in the mean number of lost sales L_v. However, as η increases, there is a comparatively high increase in β_v. For instance, when $\eta > 9$ the service level β_v is higher than 80%. It is observed that a concave of I_v. It initially increases firstly and then decreases with an increase in the replenishment rate η. However, the values of I_v vary only slightly even when η varies significantly. This indicates that I_v is less sensitive than the other performance measures.

We observe the following monotonicity from Table 5: (i) I_v, L_v and $\widetilde{L}_v$ decrease as the vacation rate θ increases; (ii) A_v, R_v and β_v increase with an increase in the vacation rate θ; (iii) $\widetilde{N}_v$, W_v and $\widetilde{W}_v$ decrease with an increase in the vacation rate θ. These results correspond to a certain extent with the analytical results given by Theorem 5.7. It is seen that all the performance measures are more sensitive in the case where θ is a small value compared to the case where θ is a large value.

6.2 Effect of the System Parameters on the Optimal Policy and the Optimal Cost

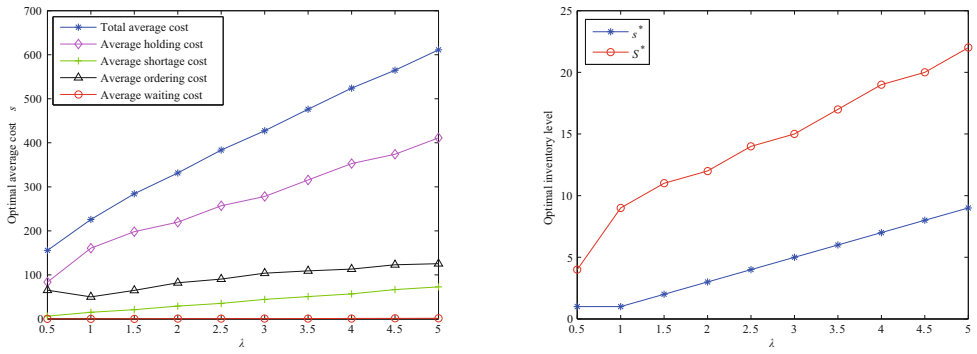
According to Schwarz, et al.[5], we consider the following cost parameters connected with operating the system: A holding cost C_1 of inventory per unit time, a shortage cost C_2 per unit of time for each lost sale demand, a fixed cost C_3 for placing an order, and a waiting cost C_4 per unit of time for each waiting customer in queue.

Let $F(s, S)$ be the total average cost per unit of time for our original model, i.e., Model II. Then, the total average cost per unit of time for Model II is given by

$$F(s, S) = C_1 I_v + C_2 L_v + C_3 R_v + C_4 \widetilde{N}_v, \tag{71}$$

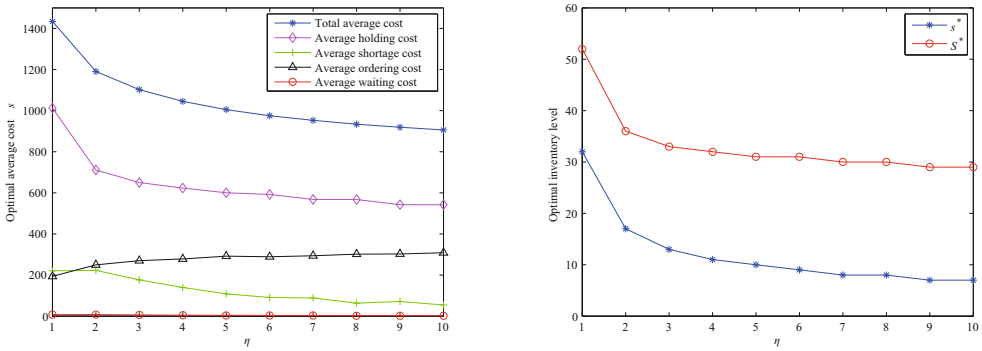
where C_1I_v, C_2L_v, C_3R_v and $C_4\widetilde{N}_v$ are the average holding cost, the average shortage cost, the average ordering cost, and the average waiting cost, respectively. It is difficult to discuss the structural properties of the cost function $F(s,S)$ analytically because of the complexity of the cost function. Hence, we carry out a detailed numerical study for the optimal policy and the optimal cost.

In the following, we numerically investigate the effect of system parameters on the optimal policy and the optimal cost under the constraint of the service level. Hence, the genetic algorithm (see Feng, et al.[29]) is used to find the optimal policy to minimize the total average cost $F(s,S)$ under the constraint of the service level $\beta_v \geq \beta_0$, where β_0 the predetermined service level. We fix the predetermined service level $\beta_0 = 85\%$ and the cost parameters $C_1 = 30$, $C_2 = 100$, $C_3 = 500$ and $C_4 = 50$, and vary any one of the parameters λ, η, θ and μ, and fix all other parameters to be constant. The results are presented in Figures 3–6.



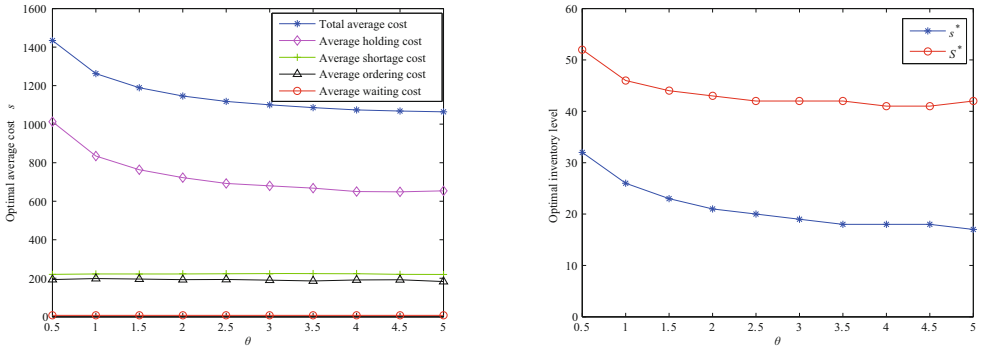
(a) Optimal average costs per unit time (b) Optimal values s^* and S^*

Figure 3 The effect of the arrive rate λ on the optimal cost and the optimal policy with $\eta = 1$, $\theta = 0.5$ and $\mu = 30$



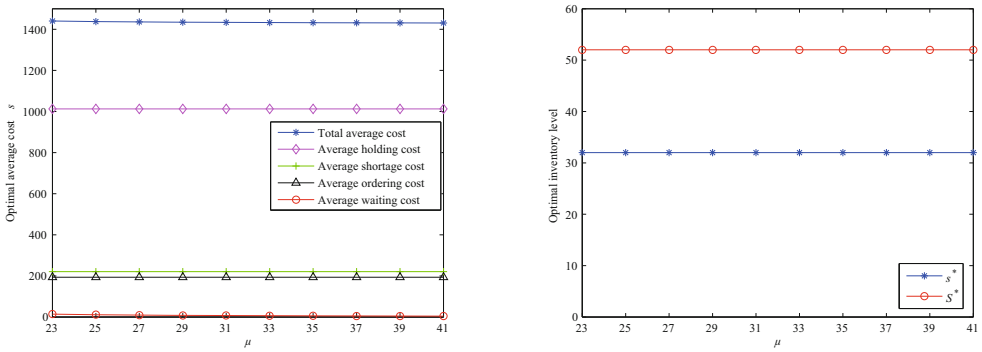
(a) Optimal average costs per unit time (b) Optimal values s^* and S^*

Figure 4 The effect of the replenishment rate η on the optimal cost and the optimal policy with $\lambda = 15$, $\theta = 0.5$ and $\mu = 30$



(a) Optimal average costs per unit time (b) Optimal values s^* and S^*

Figure 5 The effect of the vacation rate θ on the optimal cost and the optimal policy with $\lambda = 15$, $\eta = 1$, and $\mu = 30$



(a) Optimal average costs per unit time (b) Optimal values s^* and S^*

Figure 6 The effect of the service rate μ on the optimal cost and the optimal policy with $\lambda = 15$, $\eta = 1$ and $\theta = 0.5$

From Figures 3–6, we have the following observations:

(i) The optimal average cost $F(s^*, S^*)$ and the optimal policy (s^*, S^*) increase significantly with an increase in the arrival rate λ. Figure 3(a) shows the increasing tendencies of all the components of the total average cost $F(s^*, S^*)$. Especially, the holding cost increases significantly with λ. This leads to the increasing of the optimal policy (s^*, S^*) and the optimal average cost $F(s^*, S^*)$.

(ii) The optimal average cost $F(s^*, S^*)$ and the optimal policy (s^*, S^*) decrease with an increase in the replenishment rate η, and their decreasing rates get slower as the replenishment rate η increases. This can be explained by the following observation from Figure 4(a) that both the holding cost and the shortage cost decrease with the replenishment rate η even though the ordering cost increases slightly with η.

(iii) The optimal average cost $F(s^*, S^*)$ and the optimal policy (s^*, S^*) decrease with an increase in the vacation rate θ. Figure 5(a) shows that the vacation rate mainly affects the holding cost, while hardly affecting the other three components. This indicates a vacation has

a negative effect on the optimal average cost.

(iv) The optimal policy (s^*, S^*) does not vary with an increase in service rate μ, but the optimal average cost decreases slightly with an increase in the service rate μ. This is because the service rate μ mainly affects the waiting cost, and the waiting cost is smaller than the other three cost components since the mean number of waiting customers in queue $\widetilde{N}_v$ is very small, which can be observed from Table 3.

(v) The parameters λ, η and θ have a significant effect on the holding cost, while only having a slight effect on the shortage cost, the ordering cost and the waiting cost. The parameter μ only has a slight effect on the waiting cost and does not affect the other three cost components.

6.3 Effect of the Service Level on Some Performance Measures, the Optimal Policy and the Optimal Cost

In this subsection, we study the effect of the predetermined service level β_0 on the optimal policy, the optimal cost and some performance measures. The numerical results for the optimal policy (s^*, S^*), the optimal cost $F(s^*, S^*)$, and the performance measures β_v, I_v, L_v, R_v and $\widetilde{N}_v$ are displayed in Table 6. In Table 6, we fix the parameters as $\lambda = 7$, $\eta = 1$, $\theta = 0.5$, $\mu = 30$, $C_1 = 30$, $C_2 = 100$, $C_3 = 500$, $C_4 = 50$.

Table 6 Optimal policy and performance measures under service level constraint

$\beta_0(\%)$	(s^*, S^*)	$F(s^*, S^*)$	$\beta_v(\%)$	I_v	L_v	R_v	$\widetilde{N}_v$
65.00	(8, 18)	722.1563	66.92	10.8354	2.3159	0.3209	0.1007
70.00	(9, 20)	725.5027	71.57	12.1528	1.9900	0.3152	0.0865
75.00	(10, 23)	736.3839	76.67	14.0586	1.6329	0.2956	0.0710
80.00	(11, 26)	754.6255	80.83	15.9596	1.3416	0.2775	0.0583
85.00	(14, 27)	781.5079	85.39	17.3011	1.0230	0.3159	0.0445
90.00	(17, 31)	839.6837	90.33	20.4876	0.6766	0.3119	0.0294
95.00	(23, 35)	960.7663	95.04	24.9067	0.3469	0.3563	0.0151

From Table 6, we have the following observations:

(i) As expected, the mean inventory level I_v increases as β_0 increases, and the mean number of lost sales per unit of time L_v and the mean number of the waiting customers in queue $\widetilde{N}_v$ decrease with an increase in β_0.

(ii) The optimal policy (s^*, S^*) and the optimal cost $F(s^*, S^*)$ increase with the increasing of β_0. This is because the inventory is to be maintained with more stocks with an increase in the service level. This leads to a much higher holding inventory cost. This may be the main reason of the increase in the optimal cost.

(iii) The mean reorder rate per unit of time R_v does not display any monotonic behaviour as the predetermined service level β_0 increases. It seems that R_v is slightly sensitive to the changes of the predetermined service level β_0.

(iv) When the predetermined service level is raised some levels from a lower service level, the optimal cost may increase slightly. However, when the predetermined service level is raised by the same number of levels from a higher service level, the optimal cost may increase significantly. For instance, when β_0 is raised 5% from 65% to 70%, the optimal cost only increases around three units of the cost. However, when β_0 is raised the same 5% from 90% to 95%, the optimal cost increases by around 121 units of the cost. This observation suggests that the decision maker should improve the service level appropriately so as to balance the service level and the optimal average cost.

7 Conclusions

In this article we studied a queuing-inventory system with multiple vacations with lost sales under (s, S) policy. We obtained the product form solution for the stationary distribution of the system. We found that the conditional distribution of the inventory level when the server is off due to a vacation is independent of the arrival rate, and that the conditional distribution of the inventory level when the server is on and operational is independent of the vacation rate. We obtained some very simple expressions of some performance measures of our system model by means of the corresponding performance measures of its CIS model. The effect of various parameters on performance measures, the optimal policy and the optimal average cost were investigated. From the numerical study, we found that improving the service level could not only reduce the number of lost sales, but also reduce the number of the waiting customers. It was also found that improving the service level can more effectively manage the optimal inventory policy. The queuing-inventory system with multiple vacations with lost sales under (r, Q) policy could be solved in a similar way as the case of (s, S) policy. The more general model with general distributions of the service time, the lead time or the vacation time could be the direction of the future research.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Saffari M, Haji R, and Hassanzadeh F, A queueing system with inventory and mixed exponentially distributed lead times, *International Journal of Advanced Manufacturing Technology*, 2011, **53**(9): 1231–1237.

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

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

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

[5] Schwarz M, Sauer C, Daduna H, et al., M/M/1 queueing systems with inventory, *Queueing Systems*, 2006, **54**(1): 55–78.

[6] Saffari M, Asmussen S, and Haji R, The M/M/1 queue with inventory, lost sale, and general lead times, *Queueing Systems*, 2013, **75**(1): 65–77.

[7] Baek J W and Moon S K, The M/M/1 queue with a production-inventory system and lost sales, *Applied Mathematics and Computation*, 2014, **233**: 534–544.

[8] Krenzler R and Daduna H, Loss systems in a random environment: Steady state analysis, *Queueing Systems*, 2015, **80**(1): 127–153.

[9] Krishnamoorthy A, Manikandan R, and Lakshmy B, A revisit to queueing-inventory system with positive service time, *Annals of Operations Research*, 2015, **233**: 221–236.

[10] Krishnamoorthy A, Shajin D, and Lakshmy B, Product form solution for some queueing-inventory supply chain problem, *OPSEARCH*, 2016, **53**(1): 85–102.

[11] Yue D, Zhao G, and Qin Y, An M/M/1 queueing-inventory system with geometric batch demands and lost sales, *Journal of Systems Science and Complexity*, 2018, **31**(4): 1024–1041.

[12] Krishnamoorthy A, Shajin D, and Viswanath C N, *Inventory with Positive Service Time: A Survey*, *Advanced Trends in Queueing Theory*, Eds. by Anisimov V and Limnios N, Science, ISTE&J. Wily, London, 2019.

[13] Doshi B T, Queueing systems with vacations: A survey, *Queueing Systems*, 1986, **1**(1): 29–66.

[14] Takagi H, *Queueing Analysis — A Foundation of Performance Evaluation*, Elsevier, Amsterdam, 1991, **1**.

[15] Tian N and Zhang Z G, *Vacation Queueing Models: Theory and Applications*, Springer-Verlag, New York, 2006.

[16] Ke J C, Wu C H, and Zhang Z G, Recent developments in vacation queueing models: A short survey, *International Journal of Operation Research*, 2010, **7**(4): 3–8.

[17] Narayanan V C, Deepak T G, Krishnamoorthy A, et al., On an (s, S) inventory policy with service time, vacation to server and correlated lead time, *Quality Technology and Quantitative Management*, 2008, **5**(2): 129–143.

[18] Sivakumar B, An inventory system with retrial demands and multiple server vacation, *Quality Technology and Quantitative Management*, 2011, **8**(2): 125–146.

[19] Padmavathi I, Lawrence A S, and Sivakumar B, A finite-source inventory system with postponed demands and modified M vacation policy, *OPSEARCH*, 2016, **53**(1): 41–62.

[20] Melikov A Z, Rustamov A M, and Ponomarenko L A, Approximate analysis of a queueing-inventory system with early and delayed server vacations, *Automation and Remote Control*, 2017, **78**(11): 1991–2003.

[21] Koroliuk V S, Melikov A Z, Ponomarenko L A, et al., Asymptotic analysis of the system with server vacation and perishable inventory, *Cybernetics and Systems Analysis*, 2017, **53**(4): 543–553.

[22] Koroliuk V S, Melikov A Z, Ponomarenko L A, et al., Models of perishable queueing-inventory systems with server vacations, *Cybernetics and Systems Analysis*, 2018, **54**(1): 31–44.

[23] Manikandan R and Nair S S, An M/M/1 queueing-inventory system with working vacations,

vacation interruptions and lost sales, *Automation and Remote Control*, 2020, **81**(4): 746–759.

[24] Jeganathan K and Abdul Reiyas M, Two parallel heterogeneous servers Markovian inventory system with modified and delayed working vacations, *Mathematics and Computers in Simulation*, 2020, **172**: 273–304.

[25] Zhang Y, Yue D, and Yue W, A queueing-inventory system with random order size policy and server vacations, *Annals of Operations Research*, 2022, **310**(2): 595–620.

[26] Neuts M F, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, John Hopkins Press, Baltimore, 1981.

[27] Krishnamoorthy A and Narayanan V C, Stochastic decomposition in production inventory with service time, *European Journal of Operational Research*, 2013, **228**(2): 358–366.

[28] Marand A J, Li H, and Thorstenson A, Joint inventory control and pricing in a service-inventory system, *International Journal of Production Economics*, 2019, **209**: 78–91.

[29] Feng J R, Liu Z H, and Liu Z H, A mixed integer genetic algorithms for solving the mixed integer programming problems and simulation implementing, *Journal of System Simulation*, 2004, **16**(4): 845–848.