Queueing Models and
Service Management
Vol. 6, No. 1, page 1-26, 2023

QMSM
©AU 2023

Analysis of a Queueing-Inventory System with Synchronous Vacation of Multiple Servers

Dequan Yue[1,*], Ziqin Ye[1] and Wuyi Yue[2]

[1]School of Science
Yanshan University, Qinhuangdao 066004, China
[2]School of Applied Information Technology
The Kyoto College of Graduate Studies for Informatics, Kyoto 606-8225, Japan

(*Received January 2023 ; accepted April 2023*)

Abstract: This paper considers a queueing-inventory system with synchronous vacation of multiple servers. The stocks are replenished by an (s, S) policy. When the inventory is empty, all servers synchronize multiple vacations, and the vacation time follows an exponential distribution. In this paper, we first establish a three-dimensional Markov process for the number of customers, the inventory level and the status of the servers in the system. Then, we give the solution of the steady-state probability distribution of the system by using the matrix-geometric solution method, and derive some important performance measures. In order to deal with the case of larger or super larger dimension of the state space, an approximate method to calculate the steady-state probability distribution of the system is developed. We further develop a total average cost function. Finally, we investigate the effect of system parameters on the optimal number of servers, the optimal inventory policy and the optimal average cost function through numerical illustration.

Keywords: Queueing-inventory systems, (s, S) policy, multiple servers, lost sales, synchronous vacation.

1. Introduction

A queueing system with attached inventory is called a queueing-inventory system. In such system, every customer takes on-hand items from the inventory and requires some positive service time. A simple example is a retail market where customers spend time to pay for items that they want to purchase (see Baek and Moon [1]). Another example is a pure inventory system where it requires some time to deal with items for retrieval, preparation, packing, and loading before items in inventory are out of the warehouse (see Saffari et al. [18]). A queueing-inventory system is different from both the traditional inventory system and the traditional queueing system (see Zhao and Lian [29]). During the last decades, queueing-inventory systems have drawn much attention of many researchers because of their differing characteristics from both the classic queueing systems and the classic inventory systems.

* Corresponding author
Email: ydq@ysu.edu.cn

In real life, long queues often consume a lot of time and cost. When there is only one server, it often leads to long waiting time for customers and increases waiting cost. Therefore, it is often to use faster serves or hire additional servers to reduce costs associated with customers sojourn time in the system. Many studies discussed multiple server facilities in queueing-inventory system. Yadavalli et al. [25] studied a continuous review perishable inventory system with multi-server service facility, where an (s, S) policy is considered. The (s, S) policy is a well-known control policy in queueing-inventory system, in which inventory is raised to an order-up-to level S whenever it falls below reorder-level s at a review instant. In [25], the authors assumed that customers follow Markov arrivals, and also considered a second flow of negative customers. They obtained the joint probability distribution of the number of busy servers, the inventory level and the number of customers in the orbit in the steady state of the system. Various measures of stationary system performance were computed and the total expected cost per unit time was calculated. Yadavalli et al. [26] also studied a continuous review retrial inventory system with a finite source of customers and identical multiple servers in parallel. They assumed that all the c servers are homogeneous, and the service times are exponentially distributed. In [26], the Laplace-Stieltjes transform of the waiting time distribution and the moments of the waiting time distribution were calculated. Rajkumar et al. [17] considered a multi-server inventory system at a service facility with an (s, S) policy. In [17], the authors assumed the customers to arrive according to a Poisson process. They calculated the Laplace-Stieltjes transforms of the first passage time and the waiting time of a tagged customer.

Krishnamoorthy et al. [10] studied a M/M/c queueing-inventory system with an (s, Q) policy and lost sales. They obtained a product form solution of the steady-state distribution for the case of two servers, and also obtained a matrix-geometric solution of the steady-state distribution for the case of more than two servers. They have derived conditional distribution of the inventory level, conditioned on the number of customers in the system, and conditional distribution of the number of customers, conditioned on the inventory level. They also computed the optimal number of servers and also computed the optimal (s, Q) pair values and the corresponding minimum cost. Wang[24] further investigated a multi-server retrial queuing-inventory system with two demand classes and an (s, Q) policy. Assume that the demand arrival is Markov arrival process (MAP) and the lower priority customers are impatient. They derived the steady-state probability of the system using the truncation approximation method and investigated the problem to optimize the number of servers, stock and reorder levels, retrial and service rates such that the average operating cost is minimized. Jeganathan et al. [6] studied a Markovian inventory-queueing system with server interruptions of two heterogeneous servers and an (s, Q) policy. They applied a matrix method to obtain the steady-state joint probability distribution of customers level in the queue, retrial group, status of the servers and stock level of heterogeneous system. They also studied the significant effect of a heterogeneous system to compare with a homogeneous system. Jeganathan et al. [5] also studied a perishable inventory model with two dedicated servers and one flexible server, where the service rates are different with respect to stations and servers. By using matrix analysis method, the joint stationary distribution of the number of customers in station 1 and station 2, the status of the servers and the inventory level under steady state are investigated. Various system performance measures are derived and the long-run total expected cost rate is calculated. For more research work on queuing-inventory systems, we refer the reader to survey papers given by Krishnamoorthy et al. [9] and Krishnamoorthy et al. [11].

However, all the research works mentioned above assumed that the servers are always available to work even if the system is out of stock. In practice, it is very common to allow the servers to do some secondary jobs for improving server's utilization when the servers are idle. This period that the servers to do some secondary jobs is called a vacation in the queueing literature. The readers may reference Doshi [4], Takagi [21], Tian and Zhang [22] and Ke et al. [8] for more details on various queuing systems with vacations of multiple servers.

Daniel and Ramanarayanan [3] were the first to study a queueing-inventory system by considering the server's vacation, in which a vacation was granted when the inventory is exhausted. They assumed that the customer arrival times, the lead times and the rest times all follow arbitrary distributions. They obtained the steady-state probabilities of the system by using renewal and convolution techniques. Afterwards, many authors focus on queueing-inventory systems with server's vacation. Viswanath et al. [23] considered a server's vacation in MAP/PH/1 queueing-inventory system in which the server went for a vacation whenever there were no waiting customers, or the inventory level was zero. They derived the steady-state probability distribution by using the matrix-geometric solution method. A finite-source queueing-inventory system with an (s, Q) policy and a modified vacation policy was investigated by Padmavathi et al. [16], in which the server went into an inactive period whenever the inventory level reaches zero. Melikov et al. [14] proposed a model for a queueing-inventory system with perishable inventory and early and delayed vacations of the server under the (s, S) policy where the server went into a vacation if either the level of inventory in the system or the queue length was zero. They developed a method for approximate computation of the system's characteristics. Zhang et al. [28] considered a random order size policy in an M/M/1 queueing-inventory system with multiple server vacation. They derived a product form solution for the stationary distribution under the assumption that the server takes multiple vacations once the inventory is depleted.

Suganya et al. [20] studied a queueing-inventory system with two heterogeneous servers and multiple vacations. They assumed the customers to arrive according to a Markovian arrival process, and two parallel servers to provide heterogeneous phase type services to customers. They obtained the joint probability distribution of the number of customers in the system, inventory level and server status in the steady state of the system. Suganya et al. [19] also studied a queueing-inventory system with a finite number of homogeneous sources of customers and multiple vacations of two heterogeneous servers. They obtained the joint probability distribution and some important performance measures, and investigated the optimality of the expected total cost rate by numerical illustration. A retrial production inventory system with vacation and two heterogenous servers was considered by Jose and Beena [7]. They considered an (s, S) policy for production of items. The stability condition and the steady-state probabilities of the system were calculated by using matrix analysis methods.

All these papers mentioned above considered single server or two heterogeneous servers in the queueing-inventory system models (or production system model) with server vacations. In this paper, we also considered an (s, S) policy as presented in [20] and [19]. However, this paper is different from the papers presented in [20] and [19], in which we consider c homogeneous servers in a queueing-inventory system with synchronous multiple vacations. In this system, when the inventory is empty, all c servers synchronize multiple vacations. Also, in contrast to the M/M/c queueing-inventory model studied by Krishnamoorthy et al. [10], in this paper we consider an M/M/c queueing-inventory model with synchronous vacation of multiple servers and an (s, S) inventory policy. In [10], Krishnamoorthy et al. considered an M/M/c queueing-inventory model with an (s, Q) policy and without server vacations. They obtained a product form solution of the steady-state distribution for the case of $c = 2$, and also obtained the matrix-geometric solution of the steady-state distribution for the case of $c \geq 3$. Another difference from Krishnamoorthy et al. [10] is that in this paper we obtained both an exact solution and an approximated solution to calculate the steady-state probability distribution of the system.

A motivating example of our model comes from online sales of some electronic products. For example, consider an online retailer who sells two types of computers, e.g., Type 1 and Type 2. There are one or more than one servers who serve the customers. If a customer purchases one computer by online from the retailer, it usually needs some service time of the servers for preparation, packing, and mailing. The retailer can adopt an (s, S) policy to manage the inventory of each types of computers. Once the inventory level of Type 1 (or Type 2) computers drop to level s, an order of Type 1 (or Type 2) computers is placed for a variable replenishment quantity such that upon replenishment, the on-hand inventory is restocked to level S. When the inventory of Type 1 (or Type 2) computers is depleted, the retailer often sells Type 1 (or Type 2) computer by online reservation. During this reservation period, servers do not provide service to customers demanding Type 1 (or Type 2) computers until the deadline of the reservation, and they may provide service to customers of demanding Type 2 (or Type 1) computer. This period of online reservation can be looked as synchronous vacations of servers who serve customers demanding Type 1 (or Type 2) computers. Such systems exist in most online sales companies and can be studied with the model presented in this paper.

The main contributions of this paper are as follows: (i) We derive the stability condition of the system and find that it does not depend on the vacation time. (ii) We obtain both the exact solution and the approximate solution of the steady-state distribution of the joint process of the queue length, the on-hand inventory level and the server's status. The approximate solution method can be used to calculate the steady-state probability distribution of the system with larger or super larger dimension of the state space. This is the main difference in methodology adopted between this paper with those aforementioned papers. (iii) We considered not only the optimal problem of finding the number of servers but also the optimal problem of finding the joint (c, s, S) policy. (iv) The effects of the various system parameters on the optimal joint (c, s, S) policy and its corresponding average cost are numerically investigated.

The rest of this paper is organized as follows. Firstly, we describe the system model in Section 2. In Section 3, we derive the stability condition of the system by using the quasi birth-death (QBD) process theory. Then, we give the matrix-geometric solution of the steady-state probability of the system. Based on this, we compute some performance measures. An approximate method to calculate the steady-state probability distribution of the system is developed in Section 4. In Section 5, we derive a total average cost function and present some numerical results. Conclusions are given in Section 6.

2. System Model

We consider a queueing-inventory system with synchronous vacation of multiple servers. Figure 1 shows the schematic diagram of the proposed model.

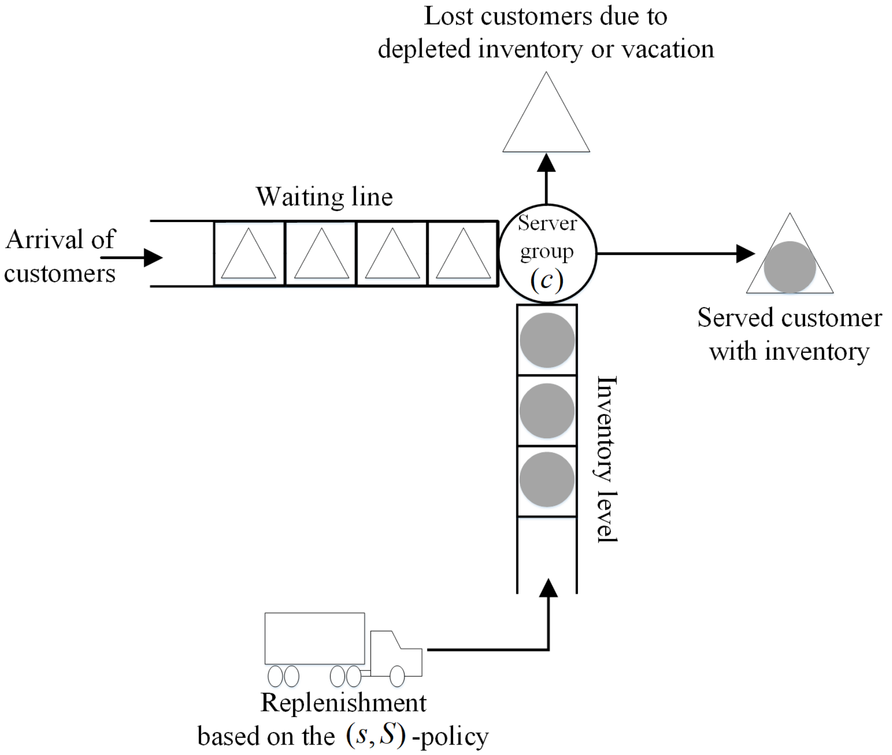

Figure 1. Schematic diagram of the proposed model in this paper.

In this system model, customers arrive according to a Poisson process with rate λ. There are c homogeneous servers in the system. Each customer requires exactly a single item in the inventory for service. Customers are served one by one under a First-Come, First-Served (FCFS) discipline. The service times are exponentially distributed with parameter μ.

All c servers begin a vacation simultaneously whenever the on-hand inventory is empty. At the end of each vacation, if the on-hand inventory in the system is not empty, all servers return back to the system to provide service for customers at any time, otherwise all the severs take another vacation immediately and continue in the same manner until the servers finds the on-hand inventory is not empty. We call this vacation policy a multiple synchronous vacation. The vacation time of the servers follows another exponential distribution with parameter θ.

In this paper, the (s, S) inventory policy with a continuous review is considered. That is to say that, each time the on-hand inventory reaches the reorder point s, an order is placed for a variable replenishment quantity such that upon replenishment, the on-hand inventory is restocked to level S ($s < S$). The replenishment lead time is exponentially distributed with parameter η. It is assumed that $c < s$.

Customers arriving during a period when inventory is depleted or when the servers are off for vacation are rejected and lost. This is a lost sales situation. If a server is ready to serve a customer who is at the head of the line and there is no item in inventory, the service starts at the moment that the next replenishment arrives. It is assumed that the arrival process, the service time, the lead time and the vacation time are independent each other.

3. System Analysis

In this section, we perform a steady-state analysis for the system model. We first formulate a quasi birth-death (QBD) process and derive a stability condition of the system. Then, we compute the stationary distribution of the joint process of the number of customers in the system, the inventory level and the status of the servers. Based on the stationary distribution, we obtain some performance measures of the system.

3.1. Stability Condition

Let $\boldsymbol{\Phi(t)}$ = $\{(M(t), N(t), J(t)),\ t \geq 0\}$ be a state process of the system, where $M(t)$ denotes the number of customers in the system at time t, $N(t)$ denotes the inventory level at time t, and $J(t)$ denotes the status of the servers at time t. $J(t)$ is defined as either 0 or 1 according to whether the servers are off for vacation or on for servicing, respectively. Then, the process $\boldsymbol{\Phi(t)}$ is a QBD process with state space: $\Omega = \cup_{m=0}^{\infty}\{\boldsymbol{m}\}$, where

$$\boldsymbol{m} = \{(m, 0, 0), (m, 1, 1), \ldots, (m, S, 1), (m, S, 0)\}\,,\ m \geq 0$$

is the collection of states with $M(t) = m$, called the level $\boldsymbol{m}$. The number of states in a level $\boldsymbol{m}$ is $S + 2$.

The infinitesimal generator of the process $\boldsymbol{\Phi(t)}$ is given as follows:

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{A_0} & \boldsymbol{C} & & & & & & \\ \boldsymbol{B_1} & \boldsymbol{A_1} & \boldsymbol{C} & & & & & \\ & \boldsymbol{B_2} & \boldsymbol{A_2} & \boldsymbol{C} & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & \boldsymbol{B_{c-1}} & \boldsymbol{A_{c-1}} & \boldsymbol{C} & & \\ & & & & \boldsymbol{B} & \boldsymbol{A} & \boldsymbol{C} & \\ & & & & & \ddots & \ddots & \ddots \end{pmatrix}$$

where each block matrix is $(S+2) \times (S+2)$ matrix.

For $0 \leq m \leq c$, we have

$$\boldsymbol{A_m} = \begin{pmatrix} -\eta & & & & & & & 0 & \eta \\ & f_1^m & & & & & & \eta & 0 \\ & & \ddots & & & & & \vdots & \vdots \\ & & & f_s^m & & & & \eta & 0 \\ & & & & f_{s+1}^m & & & 0 & 0 \\ & & & & & \ddots & & \vdots & \vdots \\ & & & & & & f_{S-1}^m & 0 & 0 \\ & & & & & & & f_S^m & 0 \\ & & & & & & & \theta & -\theta \end{pmatrix}$$

where

$$f_k^0 = \begin{cases} -(\eta+\lambda), & 1 \leq k \leq s \\ -\lambda, & s+1 \leq k \leq S, \end{cases}$$

and for $1 \leq m \leq c$,

$$f_k^m = \begin{cases} -(\eta+\lambda+k\mu), & 1 \leq k \leq m \\ -(\eta+\lambda+m\mu), & m+1 \leq k \leq s \\ -(\lambda+m\mu), & s+1 \leq k \leq S. \end{cases}$$

For $1 \leq m \leq c$, we have

$$\boldsymbol{B_m} = \begin{pmatrix} 0 & & & & & \\ g_1^m & 0 & & & & \\ & g_2^m & \ddots & & & \\ & & \ddots & 0 & & \\ & & & g_S^m & 0 & \\ & & & & 0 & 0 \end{pmatrix}$$

where

$$g_k^m = \begin{cases} k\mu, & 1 \leq k \leq m \\ m\mu, & m+1 \leq k \leq S, \end{cases}$$

$\boldsymbol{A} = \boldsymbol{A_c}$, $\boldsymbol{B} = \boldsymbol{B_c}$, and $\boldsymbol{C} = \text{diag}\{0, \lambda, \ldots, \lambda, 0\}$ is a diagonal matrix.

Consider the matrix $\boldsymbol{H_c} = \boldsymbol{A} + \boldsymbol{B} + \boldsymbol{C}$, which is given by

$$\boldsymbol{H_c} = \left(\begin{array}{cccccccccccc}
-\eta & & & & & & & & & & 0 & \eta \\
\mu & h_1 & & & & & & & & & \eta & 0 \\
 & 2\mu & h_2 & & & & & & & & \eta & 0 \\
 & & \ddots & \ddots & & & & & & & \vdots & \vdots \\
 & & & c\mu & h_c & & & & & & \eta & 0 \\
 & & & & \ddots & \ddots & & & & & \vdots & \vdots \\
 & & & & & c\mu & h_s & & & & \eta & 0 \\
 & & & & & & c\mu & h_{s+1} & & & 0 & 0 \\
 & & & & & & & \ddots & \ddots & & \vdots & \vdots \\
 & & & & & & & & c\mu & h_{S-1} & 0 & 0 \\
 & & & & & & & & & c\mu & h_S & 0 \\
 & & & & & & & & & & \theta & -\theta
\end{array}\right)$$

where

$$h_k = \begin{cases} -(\eta + k\mu), & 1 \leq k \leq c \\ -(\eta + c\mu), & c+1 \leq k \leq s \\ -c\mu, & s+1 \leq k \leq S. \end{cases}$$

It is easy to see that $\boldsymbol{H_c}$ is an infinitesimal generator of a Markovian process. Let $\boldsymbol{\xi} = (\xi_{0,0}, \xi_{1,1}, \ldots, \xi_{S,1}, \xi_{S,0})$ be the steady-state probability vector of the infinitesimal generator $\boldsymbol{H_c}$. Therefore, the vector $\boldsymbol{\xi}$ satisfies the following equations:

$$\begin{cases} \boldsymbol{\xi H_c} = \boldsymbol{0} \\ \boldsymbol{\xi e} = 1 \end{cases} \tag{1}$$

where $\boldsymbol{e}$ is a column vector of 1's of appropriate dimension. Eq. (1) can be rewritten as follows:

$$\eta\xi_{0,0} = \mu\xi_{1,1}, \tag{2}$$

$$n\mu\xi_{n,1} - [\eta + (n-1)\mu]\,\xi_{n-1,1} = 0,\ 2 \leq n \leq c, \tag{3}$$

$$c\mu\xi_{n,1} - (\eta + c\mu)\xi_{n-1,1} = 0,\ c+1 \leq n \leq s+1, \tag{4}$$

$$\xi_{n,1} = \xi_{n-1,1},\ s+2 \leq n \leq S, \tag{5}$$

$$\eta(\xi_{1,1} + \xi_{2,1} + \cdots + \xi_{s,1}) - c\mu\xi_{S,1} + \theta\xi_{S,0} = 0, \tag{6}$$

$$\eta\xi_{0,0} = \theta\xi_{S,0}, \tag{7}$$

$$\xi_{0,0} + \sum_{i=1}^{S} \xi_{i,1} + \xi_{S,0} = 1. \tag{8}$$

From Eqs. (2)-(5), it is easy to get

$$\xi_{n,1} = \begin{cases} \alpha_n \xi_{0,0}, & 1 \le n \le c \\ \left(1+\dfrac{\eta}{c\mu}\right)^{n-c} \alpha_c \xi_{0,0}, & c+1 \le n \le s+1 \\ \left(1+\dfrac{\eta}{c\mu}\right)^{s+1-c} \alpha_c \xi_{0,0}, & s+2 \le n \le S \end{cases} \tag{9}$$

where

$$\alpha_n = \begin{cases} \dfrac{\eta}{\mu}, & n=1 \\ \dfrac{\eta}{n!\mu^n} \prod\limits_{i=1}^{n-1} (\eta + i\mu), & 2 \le n \le c. \end{cases} \tag{10}$$

From Eq. (7), we have

$$\xi_{S,0} = \frac{\eta}{\theta} \xi_{0,0}. \tag{11}$$

Substituting Eqs. (9) and (11) into Eq. (8), we get

$$\xi_{0,0} = \left(1 + \frac{\eta}{\theta} + \sum_{n=1}^{c-1} \alpha_n + \alpha_c \gamma \right)^{-1} \tag{12}$$

where

$$\gamma = \left(1+\frac{\eta}{c\mu}\right)^{s+1-c} \left(S - s + \frac{c\mu}{\eta}\right) - \frac{c\mu}{\eta}, \tag{13}$$

and $\sum_{n=1}^{0}$ is defined to be zero. In the following sections, we always define $\sum_{n=k}^{0} = 0$ for $k \ge 1$.

Using the steady-state probabilities given by Eqs. (9)-(12), we can derive the stability condition of the process $\boldsymbol{\Phi(t)}$.

Theorem 1. The process $\boldsymbol{\Phi(t)}$ with the infinitesimal generator $\boldsymbol{Q}$ is positive recurrent if and only if

$$\rho < 1 - \frac{\sum\limits_{n=1}^{c-1} \left(1 - \dfrac{n}{c}\right) \alpha_n}{\sum\limits_{n=1}^{c-1} \alpha_n + \alpha_c \gamma} \tag{14}$$

where $\rho = \dfrac{\lambda}{c\mu}$, α_n and γ are given by Eqs. (10) and (13), respectively.

Proof. From the well-known result given by Neuts [15], the process $\boldsymbol{\Phi(t)}$ is positive recurrent if and only if $\boldsymbol{\xi C e} < \boldsymbol{\xi B e}$. It is easy to see that

$$\boldsymbol{\xi C e} = \lambda \sum_{n=1}^{S} \xi_{n,1} = \lambda \xi_{0,0} \left(\frac{1}{\xi_{0,0}} - 1 - \frac{\eta}{\theta} \right),$$

and

$$\begin{aligned}\boldsymbol{\xi B e} &= \mu \sum_{n=1}^{c-1} n\xi_{n,1} + c\mu \sum_{n=c}^{S} \xi_{n,1} \\ &= c\mu(1-\xi_{0,0}-\xi_{S,0}) - \mu \sum_{n=1}^{c-1} (c-n)\xi_{n,1} \\ &= c\mu\xi_{0,0}\left[\frac{1}{\xi_{0,0}} - 1 - \frac{\eta}{\theta} - \sum_{n=1}^{c-1}\left(1-\frac{n}{c}\right)\alpha_n\right].\end{aligned}$$

Thus, $\boldsymbol{\xi C e} < \boldsymbol{\xi B e}$ is equivalent to the following inequality:

$$\frac{\lambda}{c\mu} < 1 - \frac{\sum_{n=1}^{c-1}\left(1-\frac{n}{c}\right)\alpha_n}{\frac{1}{\xi_{0,0}} - 1 - \frac{\eta}{\theta}}. \tag{15}$$

Substitute Eq. (12) into Eq. (15), we obtain Eq. (14). This proves Theorem 1.

Remark 1. (i) Eq. (14) shows that the stability condition of this system does not depend on the parameter of the vacation time. Besides, it is stronger than the stability condition of the classical M/M/c queueing system: $\frac{\lambda}{c\mu} < 1$. (ii) For single server case $c = 1$, the stability condition of the system agrees with the stability condition of the classical M/M/1 queueing system: $\frac{\lambda}{\mu} < 1$.

3.2. Steady-State Probability Distribution

In this section, we derive the joint steady-state probability distribution of the number of customers in the system, the inventory level and the servers' status.

3.2.1. Matrix-geometric Solution

Let $\boldsymbol{x} = (\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_2}, \ldots)$ be the steady-state probability vector of the process $\boldsymbol{\Phi(t)}$, where

$$\boldsymbol{x_m} = (x_m(0,0), x_m(1,1), \ldots, x_m(S,1), x_m(S,0)), \ m \geq 0.$$

Then $\boldsymbol{x}$ satisfies the following equations:

$$\begin{cases} \boldsymbol{xQ} = \boldsymbol{0} \\ \boldsymbol{xe} = 1 \end{cases} \tag{16}$$

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

From Neuts [15], under the stability condition of the system given by Theorem 1, the steady-state probability vector $\boldsymbol{x}$ can be expressed as follows:

$$\boldsymbol{x_m} = \boldsymbol{x_c R^{m-c}}, \ m \geq c \tag{17}$$

where $\boldsymbol{R}$ is the minimal nonnegative solution to the matrix quadratic equation

$$\boldsymbol{R^2B} + \boldsymbol{RA} + \boldsymbol{C} = \boldsymbol{0}, \tag{18}$$

and satisfies with the spectral radius sp$(\boldsymbol{R}) < 1$, and the vectors $\boldsymbol{x_0}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_c}$ are the positive solutions of the following equations:

$$(\boldsymbol{x_0}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_c})\, \boldsymbol{B}[\boldsymbol{R}] = \boldsymbol{0}, \tag{19}$$

where

$$\boldsymbol{B}[\boldsymbol{R}] = \begin{pmatrix} \boldsymbol{A_0} & \boldsymbol{C} & & & \\ \boldsymbol{B_1} & \boldsymbol{A_1} & \boldsymbol{C} & & \\ & \ddots & \ddots & \ddots & \\ & & \boldsymbol{B_{c-1}} & \boldsymbol{A_{c-1}} & \boldsymbol{C} \\ & & & \boldsymbol{B_c} & \boldsymbol{RB} + \boldsymbol{A}, \end{pmatrix},$$

and the normalizing condition

$$\sum_{m=0}^{c-1} \boldsymbol{x_m e} + \boldsymbol{x_c} \left(\boldsymbol{I} - \boldsymbol{R}\right)^{-1} \boldsymbol{e} = 1. \tag{20}$$

The Eq. (19) can be rewritten as follows:

$$\boldsymbol{x_0 A_0} + \boldsymbol{x_1 B_1} = \boldsymbol{0}, \tag{21}$$

$$\boldsymbol{x_{i-1} C} + \boldsymbol{x_i A_i} + \boldsymbol{x_{i+1} B}_{i+1} = \boldsymbol{0},\ 1 \le i \le c-1, \tag{22}$$

$$\boldsymbol{x_{c-1} C} + \boldsymbol{x_c}(\boldsymbol{RB} + \boldsymbol{A}) = \boldsymbol{0}. \tag{23}$$

It is easy to solve Eqs. (21)-(23) subject to the normalizing condition Eq. (20). The solutions $\boldsymbol{x_0}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_c}$ are computed iteratively as

$$\boldsymbol{x_i} = \boldsymbol{x_{i+1} F_{i+1}},\ 0 \le i \le c-1 \tag{24}$$

where

$$\begin{cases} \boldsymbol{F_0} & = & \boldsymbol{0} \\ \boldsymbol{F_{i+1}} & = & -\boldsymbol{B_{i+1}} \left(\boldsymbol{F_i C} + \boldsymbol{A_i}\right)^{-1},\ 0 \le i \le c-1, \end{cases} \tag{25}$$

and $\boldsymbol{x_c}$ can be determined by the following equations:

$$\begin{cases} \boldsymbol{x_c}(\boldsymbol{F_c C} + \boldsymbol{RB} + \boldsymbol{A}) = \boldsymbol{0} \\ \boldsymbol{x_c} \left[\sum_{m=0}^{c-1} \prod_{i=0}^{c-m-1} \boldsymbol{F_{c-i}} + \left(\boldsymbol{I} - \boldsymbol{R}\right)^{-1}\right] \boldsymbol{e} = 1. \end{cases} \tag{26}$$

3.2.2. Algorithmic Computation of the Rate Matrix $\boldsymbol{R}$

In order to calculate the steady-state probability of the system, it is necessary to solve the rate matrix $\boldsymbol{R}$. Unfortunately, it is not feasible to find the analytical solution from Eq. (18). However, several quadratically-convergent algorithms such as Logarithmic Reduction algorithm and Cyclic Reduction algorithm for computing the matrix $\boldsymbol{R}$ have been proposed. We refer to Bean [2] for physical interpretations of these algorithms. We use the Logarithmic Reduction algorithm given by Latouche and Ramaswami [12] to compute the rate matrix $\boldsymbol{R}$. The main steps involved in the logarithmic reduction algorithm for computation of $\boldsymbol{R}$ are listed here as follows:

Step 1. $\boldsymbol{C}^{(0)} = (-\boldsymbol{A})^{-1}\boldsymbol{C},\ \boldsymbol{B}^{(0)} = (-\boldsymbol{A})^{-1}\boldsymbol{B},\ \boldsymbol{G}^{(0)} = \boldsymbol{B}^{(0)},\ \boldsymbol{H}^{(0)} = \boldsymbol{C}^{(0)}$.

Step 2. Consider

$$\boldsymbol{T_1}^{(j)} = \left(\boldsymbol{C}^{(j)}\right)^2,\ \boldsymbol{T_2}^{(j)} = \left(\boldsymbol{B}^{(j)}\right)^2,\ \boldsymbol{U}^{(j)} = \boldsymbol{C}^{(j)}\boldsymbol{B}^{(j)} + \boldsymbol{B}^{(j)}\boldsymbol{C}^{(j)},$$

$$\boldsymbol{C}^{(j+1)} = \left(\boldsymbol{I} - \boldsymbol{U}^{(j)}\right)^{-1}\boldsymbol{T_1}^{(j)},\ \boldsymbol{B}^{(j+1)} = \left(\boldsymbol{I} - \boldsymbol{U}^{(j)}\right)^{-1}\boldsymbol{T_2}^{(j)},$$

$$\boldsymbol{G}^{(j+1)} = \boldsymbol{G}^{(j)} + \boldsymbol{H}^{(j)}\boldsymbol{B}^{(j+1)},\ \boldsymbol{H}^{(j+1)} = \boldsymbol{H}^{(j)}\boldsymbol{C}^{(j+1)}.$$

Continue Step 2 until $\left\|\boldsymbol{e} - \boldsymbol{G}^{(j+1)}\boldsymbol{e}\right\|_\infty < \varepsilon$.

Step 3. $\boldsymbol{R} = -\boldsymbol{C}\left(\boldsymbol{A} + \boldsymbol{C}\boldsymbol{G}^{(j+1)}\right)^{-1}$.

3.3. Performance Measures

Based on the steady-state probability distribution obtained in Subsection 3.2, we can derive the performance measures of the system in steady state. Here, we list some of performance measures of interest.

3.3.1. Mean inventory level

Let I denote the mean inventory level. We note that the probability that the inventory level is n is $\sum\limits_{m=0}^{\infty} x_m(n,1)$ for $1 \le n \le S-1$, and the probability that the inventory level is S is $\sum\limits_{m=0}^{\infty} [x_m(S,1) + x_m(S,0)]$. Hence, the mean inventory level is given by

$$\begin{aligned} I &= \sum_{m=0}^{\infty}\sum_{n=1}^{S} n x_m(n,1) + \sum_{m=0}^{\infty} S x_m(S,0) \\ &= \left[\sum_{m=0}^{c-1} \boldsymbol{x_m} + \boldsymbol{x_c}(\boldsymbol{I} - \boldsymbol{R})^{-1}\right]\boldsymbol{\nu} \end{aligned}$$

where $\boldsymbol{\nu} = (0,1,2,\ldots,S,S)^T$ is a column vector with dimension of $S+2$.

3.3.2. Mean number of busy servers

Let N_b denote the mean number of busy servers. We consider the following two cases: (i) For $1 \leq k \leq c-1$, the number of busy servers is k if and only if the inventory level is k and the number of customers in the system is larger than or equal to k, or the number of customers in the system is k and the inventory level is larger than k; (ii) The number of busy servers is c if and only if both the number of customers in the system and the inventory level are larger than or equal to c. Thus, the mean number of busy servers is given by

$$\begin{aligned} N_b &= \sum_{k=1}^{c-1} k \left[\sum_{m=k}^{\infty} x_m(k,1) + \sum_{n=k+1}^{S} x_k(n,1) \right] + c \sum_{m=c}^{\infty} \sum_{n=c}^{S} x_m(n,1) \\ &= \sum_{m=1}^{c-1} \boldsymbol{x_m} \boldsymbol{\delta_m} + \boldsymbol{x_c} (\boldsymbol{I} - \boldsymbol{R})^{-1} \boldsymbol{\delta_c} + \sum_{m=1}^{c-1} m \boldsymbol{x_m} \boldsymbol{\varepsilon_m} \end{aligned}$$

where $\boldsymbol{\delta_m} = (0,1,\ldots,m,0,\ldots,0)^T$ for $1 \leq m \leq c-1$, $\boldsymbol{\delta_c} = (0,1,\ldots,c,c,\ldots,c,0)^T$, and $\boldsymbol{\varepsilon_m} = (\underbrace{0,\ldots,0}_{m+1},1,1,\ldots,1,0)^T$ for $1 \leq m \leq c-1$ are column vectors of dimension of $S+2$.

3.3.3. Mean reorder rate

Let E_r be the mean reorder rate, i.e., the mean number of replenishments per unit of time. We note that a reorder is triggered when the inventory level drops to n $(0 \leq n \leq s)$, and if there are m customers in the system the probability when a reorder is triggered is $\sum_{n=1}^{s} x_m(n,1) + x_m(0,0)$. Thus, the mean reorder rate is given by

$$\begin{aligned} E_r &= \eta \sum_{m=0}^{\infty} \sum_{n=1}^{s} x_m(n,1) + \eta \sum_{m=0}^{\infty} x_m(0,0) \\ &= \eta \left[\sum_{m=0}^{c-1} \boldsymbol{x_m} + \boldsymbol{x_c} (\boldsymbol{I} - \boldsymbol{R})^{-1} \right] \boldsymbol{\chi} \end{aligned}$$

where $\boldsymbol{\chi} = (\underbrace{1,1,\ldots,1}_{s+1},0,\ldots,0)^T$ is a column vector with dimension of $S+2$.

3.3.4. Mean order size

Let E_o be the mean order size. We note that a reorder is triggered when the inventory level drops to $n(0 \leq n \leq s)$, and the order size is $S-n$. Hence, the mean order size is given by

$$\begin{aligned} E_o &= \sum_{m=0}^{\infty} \sum_{n=1}^{s} (S-n) x_m(n,1) + S \sum_{m=0}^{\infty} x_m(0,0) \\ &= \left[\sum_{m=0}^{c-1} \boldsymbol{x_m} + \boldsymbol{x_c} (\boldsymbol{I} - \boldsymbol{R})^{-1} \right] \boldsymbol{\sigma} \end{aligned}$$

where $\boldsymbol{\sigma} = (S, S-1, \ldots, S-s, 0, \ldots, 0)^T$ is a column vector with dimension of $S+2$.

3.3.5. Mean loss rate of customers

Let E_l be the mean loss rate of customers. Since the customer who arrives at epoch when the inventory is zero or when the server is off for vacation is lost, the mean loss rate of customers is given by

$$\begin{aligned} E_l &= \lambda \sum_{m=0}^{\infty} \left[x_m(0,0) + x_m(S,0) \right] \\ &= \lambda \left[\sum_{m=0}^{c-1} \boldsymbol{x_m} + \boldsymbol{x_c} \left(\boldsymbol{I} - \boldsymbol{R} \right)^{-1} \right] \boldsymbol{\tau_1} \end{aligned}$$

where $\boldsymbol{\tau_1} = (1, 0, \dots, 0, 1)^T$ is a column vector with dimension of $S + 2$.

3.3.6. Mean number of waiting customers in the queue

Let L_q be the mean number of waiting customers in the queue. We consider the following two cases: (i) The servers are off for vacation. If there are m customers in the system, the probability that there are m customers waiting in the queue is $x_m(0,0) + x_m(S,0)$. (ii) The servers are on for servicing. We assume that there are m customers in the system and n items in the inventory. For this case, if $c \le n \le S$ and $m \ge c + 1$, the probability that there are $m - c$ customers waiting in the queue is $\sum_{n=c}^{S} x_m(n,1)$; otherwise, if $1 \le n \le c - 1$ and $m \ge n + 1$, the probability that there are $m - n$ customers waiting in the queue is $x_m(n,1)$. Hence, the mean number of waiting customers in the queue is given by

$$\begin{aligned} L_q &= \sum_{m=1}^{\infty} m \left[x_m(0,0) + x_m(S,0) \right] + \sum_{m=c+1}^{\infty} \sum_{n=c}^{S} (m - c) \, x_m(n,1) \\ &\quad + \sum_{n=1}^{c-1} \sum_{m=n+1}^{\infty} (m-n) x_m(n,1) \\ &= \sum_{m=1}^{c-1} \boldsymbol{x_m} \boldsymbol{\tau_1} + \sum_{m=2}^{c-1} \boldsymbol{x_m} \boldsymbol{\tau_m} + \boldsymbol{x_c} \left(\boldsymbol{I} - \boldsymbol{R} \right)^{-1} \varsigma + \boldsymbol{x_c} \boldsymbol{R} \left(\boldsymbol{I} - \boldsymbol{R} \right)^{-2} \boldsymbol{\omega} \end{aligned}$$

where $\boldsymbol{\tau_1} = (1, 0, \dots, 0, 1)^T$, $\boldsymbol{\tau_m} = (0, m-1, m-2, \dots, 1, 0, \dots, 0)^T$ for $2 \le m \le c - 1$, $\varsigma = (1, c-1, c-2, \dots, 1, 0, \dots, 0)^T$, and $\boldsymbol{\omega} = (\underbrace{1, 1, \dots, 1}_{c+1}, 0, \dots, 0, 1)^T$ are column vectors with dimension of $S + 2$.

3.3.7. Other performance measures

The mean number of customers who are admitted to the system per unit time is given by

$$\lambda_A = \lambda - E_l.$$

Using Little's formula, the mean waiting time of a customer in the queue is given by

$$W_q = \frac{L_q}{\lambda_A}.$$

The mean number of vacations per time unit is given by

$$E_v = \theta \left[\sum_{m=0}^{\infty} (x_m(0,0) + x_m(S,0)) \right] = \frac{\theta}{\lambda} E_l.$$

4. Approximate Analysis

In Section 3, we computed the steady-state probability distribution of the system model by using the matrix-geometric solution technique. However, this method is efficient only for models of moderate dimension, and are not efficient for models with large or super larger dimensions. Therefore, below we propose an approximate method to compute efficiently the steady-state probability distribution of the system model so that we can perform asymptotic analysis of the system model with the larger or super larger dimension of the state space.

Firstly, we consider the conditional joint probability distribution of the inventory level and the servers' status given the number of customers in the system. Let $\boldsymbol{\zeta}^m = \big(\zeta^m(0,0), \zeta^m(1,1), \zeta^m(2,1), \ldots, \zeta^m(S,1), \zeta^m(S,0)\big)$ be the steady-state conditional probability distribution of the inventory level and the servers' status conditioned on that there are m $(m \geq 1)$ customers in the system, where $\zeta^m(n,j)$ is the conditional probability that the number of items in the inventory is n and the servers' status is j, and j is either 0 or 1 according to whether the servers are off for vacation or on for servicing. The explicit expressions for the conditional probability distribution $\boldsymbol{\zeta}^m$ are given in the following theorem.

Theorem 2. (1) When $1 \leq m < c$, the steady-state conditional probability distribution is given by

$$\zeta^m(0,0) = \left(1 + \frac{\eta}{\theta} + \sum_{n=1}^{m-1} \beta_n + \beta_m \kappa \right)^{-1}, \tag{27}$$

$$\zeta^m(n,1) = \begin{cases} \beta_n \zeta^m(0,0), & 1 \leq n \leq m \\ \left(1 + \dfrac{\eta}{m\mu}\right)^{n-m} \beta_m \zeta^m(0,0), & m+1 \leq n \leq s+1 \\ \left(1 + \dfrac{\eta}{m\mu}\right)^{s+1-m} \beta_m \zeta^m(0,0), & s+2 \leq n \leq S, \end{cases} \tag{28}$$

$$\zeta^m(S,0) = \frac{\eta}{\theta} \zeta^m(0,0) \tag{29}$$

where

$$\beta_n = \begin{cases} \dfrac{\eta}{\mu}, & n = 1 \\ \dfrac{\eta}{n!\mu^n} \displaystyle\prod_{i=1}^{n-1} (\eta + i\mu), & 2 \leq n \leq m, \end{cases} \tag{30}$$

$$\kappa = \left(1 + \frac{\eta}{m\mu}\right)^{s+1-m} \left(S - s + \frac{m\mu}{\eta}\right) - \frac{m\mu}{\eta}. \tag{31}$$

(2) When $m \geq c$, the steady-state conditional probability distribution is given by

$$\zeta^m(0,0) = \left(1 + \frac{\eta}{\theta} + \sum_{n=1}^{c-1} \alpha_n + \alpha_c \gamma\right)^{-1}, \tag{32}$$

$$\zeta^m(n,1) = \begin{cases} \alpha_n \zeta^m(0,0), & 1 \leq n \leq c \\ \left(1 + \frac{\eta}{c\mu}\right)^{n-c} \alpha_c \zeta^m(0,0), & c+1 \leq n \leq s+1 \\ \left(1 + \frac{\eta}{c\mu}\right)^{s+1-c} \alpha_c \zeta^m(0,0), & s+2 \leq n \leq S, \end{cases} \tag{33}$$

$$\zeta^m(S,0) = \frac{\eta}{\theta} \zeta^m(0,0) \tag{34}$$

where α_n and γ are defined by Eqs. (10) and (13).

Proof. (1) For the case of $1 \leq m < c$, we get the infinitesimal generator $\boldsymbol{Q_1}$ of the process $\{(N(t), J(t)),\ t \geq 0\}$ under the condition that $M(t) = m$ as follows:

$$\boldsymbol{Q_1} = \begin{pmatrix}
-\eta & & & & & & & & & & 0 & \eta \\
\mu & l_1 & & & & & & & & & \eta & 0 \\
 & 2\mu & l_2 & & & & & & & & \eta & 0 \\
 & & \ddots & \ddots & & & & & & & \vdots & \vdots \\
 & & & c\mu & l_m & & & & & & \eta & 0 \\
 & & & & \ddots & \ddots & & & & & \vdots & \vdots \\
 & & & & & c\mu & l_s & & & & \eta & 0 \\
 & & & & & & c\mu & l_{s+1} & & & 0 & 0 \\
 & & & & & & & \ddots & \ddots & & \vdots & \vdots \\
 & & & & & & & & c\mu & l_{S-1} & 0 & 0 \\
 & & & & & & & & & c\mu & l_S & 0 \\
 & & & & & & & & & & \theta & -\theta
\end{pmatrix}$$

where

$$l_k = \begin{cases} -(\eta + k\mu), & 1 \leq k \leq m \\ -(\eta + m\mu), & m+1 \leq k \leq s \\ -m\mu, & s+1 \leq k \leq S. \end{cases}$$

The conditional probability distribution $\boldsymbol{\zeta^m}$ satisfies the following equations:

$$\begin{cases} \boldsymbol{\zeta^m Q_1} = 0 \\ \boldsymbol{\zeta^m e} = 1 \end{cases} \tag{35}$$

where $\boldsymbol{e}$ is a column vector of 1's of appropriate dimension. If we compare the matrix $\boldsymbol{Q_1}$ and the matrix $\boldsymbol{H_c}$, it is easy to see that $\boldsymbol{Q_1} = \boldsymbol{H_m}$. Thus, we get the solution of Eq. (35) by using the solution of Eq. (1) as given by Eqs. (27)-(29).

(2) For the case of $m \geq c$, we get the infinitesimal generator $\boldsymbol{Q_2}$ of the process $\{(N(t), J(t)),\ t \geq 0\}$ under the condition that $M(t) = m$ as follows:

$$\boldsymbol{Q_2} = \left(\begin{array}{cccccccccccc}
-\eta & & & & & & & & & & 0 & \eta \\
\mu & h_1 & & & & & & & & & \eta & 0 \\
 & 2\mu & h_2 & & & & & & & & \eta & 0 \\
 & & \ddots & \ddots & & & & & & & \vdots & \vdots \\
 & & & c\mu & h_c & & & & & & \eta & 0 \\
 & & & & \ddots & \ddots & & & & & \vdots & \vdots \\
 & & & & & c\mu & h_s & & & & \eta & 0 \\
 & & & & & & c\mu & h_{s+1} & & & 0 & 0 \\
 & & & & & & & \ddots & \ddots & & \vdots & \vdots \\
 & & & & & & & & c\mu & h_{S-1} & 0 & 0 \\
 & & & & & & & & & c\mu & h_S & 0 \\
 & & & & & & & & & & \theta & -\theta
\end{array}\right)$$

where

$$h_k = \begin{cases} -(\eta + k\mu), & 1 \leq k \leq c \\ -(\eta + c\mu), & c+1 \leq k \leq s \\ -c\mu, & s+1 \leq k \leq S. \end{cases}$$

The conditional probability distribution $\boldsymbol{\zeta^m}$ satisfies the following equations:

$$\begin{cases} \boldsymbol{\zeta^m Q_2} = 0 \\ \boldsymbol{\zeta^m e} = 1 \end{cases} \tag{36}$$

where $\boldsymbol{e}$ is a column vector of 1's of appropriate dimension. If we compare the matrix $\boldsymbol{Q_2}$ and the matrix $\boldsymbol{H_c}$, we observe an interest fact that $\boldsymbol{Q_2} = \boldsymbol{H_c}$. Thus, we get the solution of Eq. (36) by using the solution of Eq. (1) as given by Eqs. (32)-(34).

Next, we consider an inventory system with negligible service time, i.e., $\mu \to \infty$. The other assumptions are the same as given in Section 2. This inventory system becomes a single-server inventory system with an (s, S) policy and lost sales. We call this system a modified system. Let $\{(\hat{N}(t), \hat{J}(t)),\ t \geq 0\}$ be the state process of this modified system, where $\hat{N}(t)$ is the inventory level at time t and $\hat{J}(t)$ is the status of the server which is defined as either 0 or 1 according to whether the server is off for vacation or on for servicing, respectively. The state space of the process is given by $\hat{\Omega} = \{(0,0), (1,1), (2,1), \ldots, (S,1), (S,0)\}$. Let $\boldsymbol{\pi} = (\pi_{0,0}, \pi_{1,1}, \ldots, \pi_{S,1}, \pi_{S,0})$ be the steady-state probability vector of the process $\{(\hat{N}(t), \hat{J}(t)),\ t \geq 0\}$. From the known result given by Zhang [27] (Eqs. (3-21)-(3-24), p. 30), we get the components of the probability vector $\boldsymbol{\pi}$ as follows:

$$\pi_{0,0} = \frac{\lambda}{\eta}\left(\frac{\lambda}{\lambda + \eta}\right)^s K, \tag{37}$$

$$\pi_{i,1} \;=\; \begin{cases} \left(\dfrac{\lambda}{\lambda+\eta}\right)^{s-i+1} K, & 1 \le i \le s \\ K, & s+1 \le i \le S, \end{cases} \tag{38}$$

$$\pi_{S,0} = \frac{\lambda}{\theta}\left(\frac{\lambda}{\lambda+\eta}\right)^{s} K \tag{39}$$

where

$$K = \frac{1}{\dfrac{\lambda}{\eta} + S - s + \dfrac{\lambda}{\theta}\left(\dfrac{\lambda}{\lambda+\eta}\right)^{s}}. \tag{40}$$

Further, using the conditional joint probability distribution given in Theorem 2 and the probability distribution given by Eqs. (37) and (38), we compute approximately the steady-state probability distribution of the original system model described in Section 2.

Let M, N and J be the corresponding variables of $M(t)$, $N(t)$ and $J(t)$ in the steady state, respectively. Then, using conditional probability formula, we have

$$\begin{aligned} x_m(n,j) &= P(M=m, N=n, J=j) \\ &= P(M=m)P(N=n, J=j|M=m),\ (m,n,j) \in \Omega,\ m \ge 1, \end{aligned}$$

or equivalently,

$$\boldsymbol{x_m} = P(M=m)\boldsymbol{\zeta^m},\ m \ge 1$$

where $\boldsymbol{\zeta^m}$ is determined by Theorem 2.

In general, the queue length M depends on the inventory level N and the servers' status J. So, its probability distribution is different from the probability distribution of the queue length of the classical M/M/c queue. Therefore, we compute approximately the probability distribution of M by the probability distribution of the queue length of the classical M/M/c queue with arrival rate λ and variable service rates that depend on the number of customers in the system and the inventory level, which is denoted by ψ_m, $m \ge 0$.

Let $[m]$ be a enlarged state that unites all the states in level $\boldsymbol{m}$. The state space of all the enlarged states is denoted by $\Theta = \{[m],\ m \ge 0\}$. Denote $\psi_m = P(M = [m]),\ m \ge 0$. Then, similar to the approximate method proposed by Melikov et al. [13], the steady-state probability distribution of the original system model is approximately determined by

$$\boldsymbol{x_m} \approx \psi_m \boldsymbol{\zeta^m},\ m \ge 1, \tag{41}$$

or equivalently,

$$x_m(n,j) \approx \psi_m\, \zeta^m(n,j),\ (m,n,j) \in \Omega,\ m \ge 1$$

where $\zeta^m(n,j)$ is given by Theorem 2. We note from Eq. (24) that the vector $\boldsymbol{x_0} = \boldsymbol{x_1 F_1}$. Thus, $\boldsymbol{x_0}$ can be approximately determined by

$$\boldsymbol{x_0} \approx \psi_1 \boldsymbol{\zeta^1 F_1} \tag{42}$$

where $\boldsymbol{F}_1$ is given by Eq. (25).

Now, we compute approximately the probability distribution $\psi_m = P(M = [m]),\ m \geq 0$. Denote $\{\hat{M(t)},\ t \geq 0\}$ to be a Markov process defined in state space Θ. Let $q([i],[j])$ be the intensity of the transition from the enlarged state $[i]$ to the enlarged state $[j]$. We note that the transition from state $[j+1]$ to state $[j]$ only when the servers are on working and the inventory level is positive. For simplicity, we only consider the case that the number of on-hand inventory items is larger than or equal to the number of busy servers. This means that the servers will not be idle due to shortage of inventory. For this, in what follows we assume that the intensity of replenishment arrival significantly exceeds the intensity of the server leaving for vacation. Under this asymptotical condition, the intensity of the transition from state $[i]$ to state $[i-1]$ can be approximately determined by $i\mu \sum\limits_{k=i}^{S} \pi_{k,1}$ for $1 \leq i \leq c$, and by $c\mu \sum\limits_{k=c}^{S} \pi_{k,1}$ for $i \geq c$, where $\pi_{k,1}$, $1 \leq k \leq S$, is given by Eq. (38). Hence, we obtain

$$q([i],[j]) = \begin{cases} \lambda, & \text{if } [j] = [i+1],\ i \geq 0 \\ \mu_i, & \text{if } [j] = [i-1],\ 1 \leq i \leq c \\ \mu_c, & \text{if } [j] = [i-1],\ i \geq c+1 \\ 0, & \text{in other cases} \end{cases} \tag{43}$$

where

$$\mu_i = i\mu \sum_{k=i}^{S} \pi_{k,1}, 1 \leq i \leq c.$$

Using Eq. (38), we get

$$\mu_i = i\mu \left[1 - \frac{\lambda}{\theta} \left(\frac{\lambda}{\lambda + \eta} \right)^{s} K - \frac{\lambda}{\eta} \left(\frac{\lambda}{\lambda + \eta} \right)^{s-i+1} K \right],\ 1 \leq i \leq c \tag{44}$$

where K is given by Eq. (40).

We note that $\{\hat{M(t)},\ t \geq 0\}$ is a birth-and-death process on that state space Θ. The transition diagram of the states inside the enlarged states in Θ is illustrated in Figure 2. If $\varrho = \dfrac{\lambda}{\mu_c} < 1$, the steady-state probability distribution of the queue length M is given by

$$\psi_m = \begin{cases} \prod\limits_{i=1}^{m} \dfrac{\lambda}{\mu_i} \psi_0, & 1 \leq m \leq c \\ \prod\limits_{i=1}^{c} \dfrac{\lambda}{\mu_i} \varrho^{m-c} \psi_0, & m \geq c+1 \end{cases} \tag{45}$$

where

$$\psi_0 = \left\{ 1 + \sum_{k=1}^{c} \prod_{i=1}^{k} \frac{\lambda}{\mu_i} + \prod_{i=1}^{c} \frac{\lambda}{\mu_i} (1 - \varrho)^{-1} \right\}^{-1}. \tag{46}$$

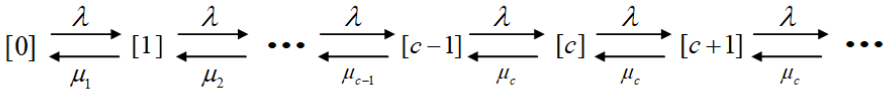

Figure 2. The transition diagram of the states inside the enlarged states in Θ.

5. Numerical Analysis

In this section, we develop a total average cost function by using the performance measures obtained in Section 4 and present some numerical analysis.

A total average cost function $F(c, s, S)$ is defined by

$$F\left(c,s,S\right)=C_1L_q+C_2I+C_3E_l+C_4E_r+C_5E_oE_r+C_6N_b+C_7E_vc \qquad (47)$$

where C_1 is a waiting cost per unit time per customer in the queue, C_2 is a holding cost of inventory per unit time, C_3 is a cost incurred due to loss of per customer, C_4 is a fixed cost for placing an order, C_5 is a replenishment cost per item, C_6 is a cost incurred per unit time per busy server , and C_7 is a vacation cost per unit time per server.

The cost function $F(c, s, S)$ is a nonlinear function of the decision variables, and all the decision variables are discrete integer variables. It is difficult to analyze the convexity of the cost function due to its complexity. For this, we use the traversal search method to find the optimal inventory policy that minimizes the cost function $F(c, s, S)$.

Firstly, we compute the optimal average cost with the approximation method presented in Section 4 and with the exact method presented in Section 3 by numerical examples.

Example 1. We compute the optimal reorder point s^* and its corresponding optimal average cost F by using the exact method presented in Section 3 and the approximation method presented in Section 4, respectively for various values of parameter c, keeping the maximum inventory capacity S fixed. Let $F_1(s^*)$ and $F_2(s^*)$ be the optimal average costs computed with the exact method and the approximation method, respectively, and let Δ_1 be the relative error of the optimal average cost $F_2(s^*)$ on the optimal average cost $F_1(s^*)$, then we can define Δ_1 as follows:

$$\Delta_1=\left|\frac{F_1(s^*)-F_2(s^*)}{F_1(s^*)}\right|.$$

In Table 1, we show the numerical results, where we set the cost parameters as: $C_1 = 10$, $C_2 = 5$, $C_3 = 55$, $C_4 = 25$, $C_5 = 15$, $C_6 = 5$ and $C_7 = 45$, and the other system parameters as: $\lambda = 4$, $\mu = 6$, $\theta = 0.8$, $\eta = 6$, $S = 20$ as an example.

Table 1. The optimal order point and the optimal average cost for various parameter c.

c	4	5	6	7	8	9	10
s^*	5	6	7	8	9	10	11
$F_1(s^*)$	90.5923	96.4501	103.2159	110.3754	117.8357	125.6048	133.7158
$F_2(s^*)$	96.8432	99.1778	101.3243	104.0954	107.4976	111.4794	116.0445
Δ_1	0.0690	0.0283	0.0183	0.0569	0.0877	0.1125	0.1322

Table 1 shows that the optimal reorder points s^* computed by using the exact method and the approximation method are the same for different values of parameter c. However, its corresponding optimal average costs $F_1(s^*)$ and $F_2(s^*)$ are different, and the relative error of the optimal average cost $F_2(s^*)$ varies from 0.0183 to 0.1322. From Table 1, we also observe that the optimal reorder point s^* and its corresponding optimal costs $F_1(s^*)$ and $F_2(s^*)$ increase with the increasing of the number of servers. This agrees with our intuitive expectation.

Example 2. We compute the optimal inventory policy (s, S) and the optimal average cost F by using the exact method presented in Section 3 and the approximation method presented in Section 4, respectively for various values of parameter c. The numerical results are shown in Table 2, where $F_1(s^*, S^*)$ and $F_2(s^*, S^*)$ are the optimal average costs computed with the exact method and the approximation method, respectively. Let Δ_2 be the relative error of optimal total average costs $F_2(s^*, S^*)$, which is defined by

$$\Delta_2 = \left| \frac{F_1(s^*, S^*) - F_2(s^*, S^*)}{F_1(s^*, S^*)} \right|.$$

In Table 2, the cost parameters and the other system parameters are assumed as the same as in Table 1.

Table 2. The optimal inventory policy and the optimal average cost for various parameter c.

c	4	5	6	7	8	9	10
(s^*, S^*)	$(5,13)$	$(6,14)$	$(7,15)$	$(8,16)$	$(9,17)$	$(10,18)$	$(11,19)$
$F_1(s^*, S^*)$	83.2335	90.8513	99.1771	107.7274	116.3475	124.9820	133.6133
(s^*, S^*)	(5, 16)	(6, 16)	(7, 17)	(8, 18)	(9, 18)	(10, 19)	(11, 20)
$F_2(s^*, S^*)$	94.1229	96.9054	99.6108	103.0473	107.0579	111.4098	116.0445
Δ_2	0.1308	0.0666	0.0044	0.0434	0.0798	0.1086	0.1315

From Table 2, we observe that the optimal reorder points s^* computed by using the exact method and the approximation method for different values of parameter c are the same, but the optimal maximum inventory capacity S^* are little different. However, its corresponding optimal average costs $F_1(s^*, S^*)$ and $F_2(s^*, S^*)$ are different, and the relative errors of the optimal average costs F_2 varies from 0.0044 to 0.1315.

Next, we consider an optimal problem from the servers' perspective. When the inventory policy (s, S) is given, we consider the optimal problem of finding the number of servers to minimize the average cost $F(c, s, S)$ which is denoted here by $F(c)$.

Example 3. Assume the inventory policy $(s, S) = (20, 50)$ to be given, we can find the optimal number of servers c and the optimal average cost $F(c)$ for various service rates. The numerical results are shown in Table 3. The system parameters are set as: $\lambda = 5$, $\theta = 4$ and $\eta = 1.1$, and the cost parameters are set as: $C_1 = 10$, $C_2 = 2$, $C_3 = 100$, $C_4 = 25$, $C_5 = 15$, $C_6 = 5$ and $C_7 = 45$.

It is obvious that increasing the service rate μ can reduce the number of servers. This can be also observed from Table 3. However, we observe from Table 3 that when service rate μ

Table 3. The optimal number of servers and the optimal average cost for various service rates.

μ	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2
c	15	14	12	11	10	9	9	8	8
$F(c)$	183.4039	99.6534	104.8701	97.0346	97.9742	114.2536	93.8314	111.2735	90.3845

increases from 0.8 to 1.2, the optimal number of servers c decreases slightly, and the optimal average cost $F(c)$ exhibits small fluctuations. This shows that higher service rates do little help to reduce the optimal average costs, so the service rate should be increased moderately.

Finally, we conduct a numerical analysis by using the approximating method to consider the effect of parameters λ, μ, θ and η on the optimal policy (c, s, S) and the optimal average cost $F(c, s, S)$. The numerical results are shown in Tables 4-7. We assume the system parameters as: $\lambda = 8$, $\mu = 10$, $\theta = 4$ and $\eta = 7$, unless their values are mentioned in the respective tables as for each case. In Tables 4-7, we set the cost parameters as: $C_1 = 10$, $C_2 = 5$, $C_3 = 35$, $C_4 = 80$, $C_5 = 100$, $C_6 = 50$ and $C_7 = 45$.

Example 4. We consider the effect of the arrival rate λ on the optimal policy (c, s, S) and the optimal cost $F(c, s, S)$ in Table 4.

Table 4. The effect of the arrival rate λ on the optimal policy and the optimal cost.

λ	2	3	4	5	6	7	8
(c, s, S)	$(2, 3, 13)$	$(2, 3, 14)$	$(3, 4, 18)$	$(3, 4, 20)$	$(3, 4, 21)$	$(3, 4, 22)$	$(4, 5, 28)$
$F(c, s, S)$	121.2146	152.8183	183.7741	217.7202	255.8373	298.2934	344.3292

From Table 4, we observe that the optimal number of servers c and the optimal reorder point s increase slightly with the increasing of the arrival rate λ, and the optimal maximum inventory capacity S and the optimal average cost $F(c, s, S)$ increase significantly with the increasing of the parameter λ. This is because that the number of customers arriving in the system per unit time increases with the increasing of the parameter λ. Therefore, more servers may be needed to reduce the customers' waiting cost. As a result, the optimal maximum inventory capacity S and the optimal average cost $F(c, s, S)$ increase.

Example 5. We consider the effect of the service rate μ on the optimal policy (c, s, S) and the optimal average cost $F(c, s, S)$ in Table 5.

Table 5. The effect of the service rate μ on the optimal policy and the optimal cost.

μ	10	15	20	25	30	35	40
(c, s, S)	$(4, 5, 28)$	$(3, 4, 26)$	$(3, 4, 27)$	$(3, 4, 28)$	$(3, 4, 28)$	$(3, 4, 29)$	$(3, 4, 29)$
$F(c, s, S)$	344.3292	273.1697	240.7529	223.5737	213.3170	206.5791	201.8599

From Table 5, it is found that the optimal number of servers c decreases slightly and the optimal total average cost $F(c, s, S)$ decreases significantly with the increasing of the parameter μ. This is because that more items are taken by customers from the inventory with the increasing of the service rate μ. Therefore, the number of servers in the system can be decreased to reduce the vacation costs of servers.

Example 6. We consider the effect of the parameter θ on the optimal policy (c, s, S) and the optimal cost $F(c, s, S)$ in Table 6.

It is observed from Table 6 that the optimal number of servers c and the optimal reorder point s decrease slightly with the increasing of the parameter θ, while the optimal average cost $F(c, s, S)$ increases slowly with the increasing of the parameter θ. This is because that the mean vacation time of the servers decreases as the increase of the parameter θ. Therefore, the number of waiting customers in the queue and the probability of the servers' vacation decrease. As a result, the number of servers in the system can be decreased to reduce the vacation costs of servers.

Table 6. The effect of the parameter θ on the optimal policy and the optimal cost.

θ	2	3	4	5	6	7	8
(c, s, S)	$(4, 5, 26)$	$(4, 5, 27)$	$(4, 5, 28)$	$(3, 4, 23)$	$(3, 4, 24)$	$(3, 4, 24)$	$(3, 4, 25)$
$F(c, s, S)$	313.2859	328.7768	344.3292	355.1125	365.1191	374.9571	384.5179

Example 7. We consider the effect of the parameter η on the optimal policy (c, s, S) and the optimal cost $F(c, s, S)$ in Table 7.

Table 7. The effect of the parameter η on the optimal policy and the optimal average cost.

η	0.8	2	5	7	9	11	13	17	19
(c, s, S)	$(4, 5, 33)$	$(5, 6, 37)$	$(5, 6, 30)$	$(4, 5, 28)$	$(3, 4, 18)$	$(3, 4, 21)$	$(3, 4, 19)$	$(3, 4, 19)$	$(3, 4, 19)$
$F(c, s, S)$	672.6195	478.4653	370.3118	344.3292	327.7637	315.8212	307.0550	295.0992	290.7479

From Table 7, it is observed that the optimal policy (c, s, S) exhibits small fluctuations when η increases from 0.8 to 11 and then tend to be stable. When the parameter η increases to a certain extent, e.g., when η increases from 13 to 19, the optimal policy does not vary and keeps to be $(3, 4, 19)$. This is because that the average replenishment time decreases slightly when η increases from 13 to 19. Thus, the optimal policy (c, s, S) tends to be stable. We observed from Table 7 that the optimal average cost $F(c, s, S)$ first decreases significantly and then decreases slightly with the increasing of the parameter η. This is because that the average replenishment time is reduced when the increase of the parameter η increases. Therefore, the lost cost due to lost customers during the stock-out period is reduced. This may be the main reason that results in the decreasing of the average cost $F(c, s, S)$. However, when the parameter η increases to a certain extent, the average replenishment time decreases slightly. This result in the slight decrease of the average cost $F(c, s, S)$.

6. Conclusions

In this paper, we analyzed a queueing-inventory system with vacations of multiple servers and an (s, S) replenishment policy. The steady-state probability vector of the system was obtained by using the matrix-geometric solution method. An approximate method to calculate the steady-state probability distribution of the system was developed to deal with larger or super larger dimension of the state space. Various performance measures of the system were derived. Numerical results showing the effect of the system parameters on the optimal number of servers, the optimal policy and the optimal total average cost were obtained. In this system model, we assumed that the service rates of all the servers are identical, which is appropriate when the service process is electronically or mechanically controlled. However, in an inventory system with human servers, it may be appropriate to assume that the servers have the different service rates. Therefore, queueing-inventory systems with heterogeneous servers and servers' vacations should be worthy of further study. However, this extension would be more challenging due to their analytical complexity.

Acknowledgments

This research was supported in part by National Natural Science Foundation of China (No. 71971189), and the Research Project of Science and Technology of University of Hebei Province, China (No. ZD2018042), and is supported in part by MEXT, Japan.

References

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

[2] Bean, N., Latouche, G., & Taylor, P. (2018). Physical interpretations for quasi-birth-and-death process algorithms. *Queueing Models and Service Management*, 1, 59-78.

[3] Daniel, J. K., & Ramanarayanan, R. (1988). An (s, S) inventory system with rest periods to the server. *Naval Research Logistics*, 35, 119-123.

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

[5] Jeganathan, K., Melikov, A. Z., Padmasekaran, S., Kingsly, S. J., & Lakshmi, K. P. (2019). A stochastic inventory model with two queues and a flexible server. *International Journal of Applied and Computational Mathematics*, 5, 1-27.

[6] Jeganathan, K., Reiyas, M. A., Lakshmi, K. P., & Saravanan, S. (2019). Two server Markovian inventory systems with server interruptions: Heterogeneous vs. homogeneous servers. *Mathematics and Computers in Simulation*, 155, 177-200.

[7] Jose, K. P., & Beena, P. (2020). On a retrial production inventory system with vacation and multiple servers. *International Journal of Applied and Computational Mathematics*, 6, 1-17.

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

[9] Krishnamoorthy, A., Lakshmy, B., & Manikandam, R. (2011). A survey on inventory models with positive service time, *OPSEARCH*, 48, 153-169.

[10] Krishnamoorthy, V., Manikandan, R., & Shajin, D. (2015). Analysis of a multiserver queueing-inventory system. *Advances in Operations Research*, 2015, 1-16.

[11] Krishnamoorthy, A., Shajin, D., & Viswanath, C. N. (2019). *Inventory with positive service time: A survey*, In V. Anisimov & N. Limnios (Eds), Advanced trends in queueing theory: Series of books "Mathematics and Statistics". Science, ISTE & J. Wily, London.

[12] Latouche, G., & Ramaswami, V. (1993). A logarithmic reduction algorithm for quasi-birth-death processes. *Journal of Applied Probability*, 30, 650-674.

[13] Melikov, A. Z., Ponimarenko, L. A., & Bagirova, S. A. (2016). Models of queuein-inventory systems with randomized lead policy. *Journal of Automation and Information Sciences*, 48, 23-35.

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

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

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

[17] Rajkumar, M., Sivakumar, B., & Arivarignan, G. (2014). An infinite waiting hall at a multi-server inventory system. *International Journal of Inventory Research*, 2, 189-221.

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

[19] Suganya, C., Lawrence, A. S., & Sivakumar, B. (2018). A finite-source inventory system with service facility, multiple vacations of two heterogeneous servers. *International Journal of Information and Management Sciences*, 29, 257-277.

[20] Suganya, C., Sivakumar, B., & Arivarignan, G. (2017). Numerical investigation on MAP/PH(1), PH(2)/2 inventory system with multiple server vacations. *International Journal of Operational Research*, 29, 1-33.

[21] Takagi, H. (1991). *Queueing Analysis-A Foundation of Performance Evaluation*, Vol. 1. Elsevier, Amusterdam.

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

[23] Viswanath, C. N., Deepak, T. G., Krishnamoorthy, A., & Krishkumar, B. (2008). On $(s,\ S)$ inventory policy with service time, vacation to server and correlated lead time. *Quality Technology and Quantitative Management*, 5, 129-144.

[24] Wang, F. F. (2015). Approximation and optimization of a multi-server impatient retrial queueing-inventory system with two demand classes. *Quality Technology & Quantitative Management*, 12, 269-292.

[25] Yadavalli, V. S. S., Sivakumar, B., Arivarignan, G., & Adetunji, O. (2011). A multi-server perishable inventory system with negative customer. *Computers and Industrial Engineering*, 61, 254-273.

[26] Yadavalli, V. S. S., Sivakumar, B., Arivarignan, G., & Adetunji, O. (2012). A finite source multi-server inventory system with service facility. *Computers and Industrial Engineering*, 63, 739-753.

[27] Zhang, Y. (2022). Performance analysis and optimal control of queueing inventory systems with vacations. *Doctorial Dissertation* (in Chinese). Yanshan University, Qinhuangdao.

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

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