Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 5(5): 4322–4344.
DOI: 10.3934/math.2020276
Received: 17 December 2019
Accepted: 28 April 2020
Published: 09 May 2020

Research article

An unreliable discrete-time retrial queue with probabilistic preemptive priority, balking customers and replacements of repair times

Shaojun Lan[1,*] **and Yinghui Tang**[2]

[1] School of Mathematics and Statistics, Sichuan University of Science and Engineering, Zigong, 643000, Sichuan, China

[2] School of Fundamental Education, Sichuan Normal University, Chengdu, 610066, Sichuan, China

* **Correspondence:** Email: sjlan89@163.com; Tel: +8618280717825.

Abstract: This paper deals with a discrete-time $Geo/G/1$ retrial queueing system with probabilistic preemptive priority and balking customers, in which the server is subject to starting failures and replacements in the repair times may occur with some probability. If the server is found busy at an arrival epoch, the newly arriving customer either interrupts the customer in service to begin its own service with probability p or enters the orbit with probability $1-p$. When an arriving customer (external or repeated) finds the server free, he must turn on the server. If the server is activated successfully, the customer receives service immediately. Otherwise, the server undergoes a repair process. If an external arrival finds that the server is under repair, he decides either to join the orbit with probability q or leaves the system completely (balking) with probability $1-q$. Applying the supplementary variable method and the generating function technique, we analyze the Markov chain underlying the considered queueing model and derive the stationary distributions under different system states, the generating functions for the number of customers in the orbit and in the system, as well as some crucial performance measures in steady state. Especially, some corresponding results under special cases are directly obtained by setting appropriate parameter values. Further, some numerical examples are provided to examine the effect of various system parameters on queueing characteristics. Finally, an operating cost function is formulated to discuss numerically a cost optimization problem.

Keywords: discrete-time retrial queue; probabilistic preemptive priority; balking customers; starting failures; replacement
Mathematics Subject Classification: 60K25, 68M20, 90B22

1. Introduction

In the study of the classical queueing theory, it is always assumed that when an arriving customer finds the server temporarily unavailable (e.g., the server may be busy, on vacation or under repair), the customer either joins the waiting line to get service or leaves the service system forever. However, in day-to-day queueing activities, we often meet some congestion situations where the arriving customers who cannot be served in time leave the service zone and join a retrial group (called orbit) to approach the server for their requests after some random length of time. In literature such queueing phenomena are referred to as retrial queues which arise naturally in the real world. Nowadays, retrial queues have been widely and successfully used to model many realistic congestion scenarios such as telephone switching systems, computer and communication systems, packet switching networks and collision-avoidance star local area networks. Taking call center as an example, when a caller finds the line busy, he/she may decide to enter a virtual retrial group and can retry for service again and again after a random amount of time until the line is found available. During the last few decades, a considerable amount of work has been devoted to investigating retrial queueing model. For comprehensive details on retrial queues we refer the reader to the outstanding survey papers by Yang and Templeton [1], Falin [2], Artalejo [3, 4], and Kim and Kim [5], and the monographs by Falin and Templeton [6], and Artalejo and Gómez-Corral [7].

In the retrial queueing literature, the previous studies were focused mainly on continuous time setting. However, in recent years, there has been an increasing interest in addressing discrete-time retrial queues. The research of discrete-time retrial queueing systems does have important practical significance in daily life. In modern communication technology, there are numerous systems, such as broadband integrated services digital network (BISDN), time division multiple access (TDMA) and asynchronous transfer mode (ATM), operating in discrete-time environment where the events (e.g., arrival of packets and their forward transmissions) can only occur at regularly spaced epochs. This fact indicates that discrete-time queues are more suitable than their continuous-time counterparts in characterizing the behaviors of data communication and computer networks. The pioneer work concerning discrete-time queue can be traced to Meisling [8]. Detailed analysis and applications of discrete-time queues can be found in the books by Hunter [9], Bruneel and Kim [10], and Woodward [11]. Since there are repeated requests for establishing a communication link among various nodes and repeated transmissions of unsuccessfully received packets, retrial phenomenon is inherited in network communication system which operates in discrete time regime. Yang and Li [12] pioneered the study of discrete-time retrial queue. They investigated a $Geo/G/1$ retrial queueing system with geometric retrial times by using generating function technique. Based on the fundamental work of Yang and Li [12], many researchers have concentrated on the study of discrete-time retrial queueing model, see e.g., [13–19], and references therein.

In many real life scenarios such as computer and communication networks, flexible manufacturing system, transportation system and production system, we often encounter the case that the service station may be subject to breakdowns (e.g., arrival of virus to the CPU, hardware breakdowns) when rendering service to customers. The breakdowns of service facilities will result in a temporarily unavailable period of the systems and therefore the performance and the efficiency of these systems will be heavily affected. As a consequence, the managers or manufacturers not only have to face the disruption of production and the loss of products, but also afford the additional repair expense due to the broken machine. Under such circumstances, the research of repairable queueing system is well worth doing from the viewpoint of queueing and reliability theory. Atencia and Moreno [20] considered a discrete-time $Geo/G/1$ retrial queue where the server may fail during busy period (called active breakdowns). Also, Atencia and Moreno [21] applied the supplementary variables method to examine a discrete-time $Geo/G/1$ retrial queueing system associated with starting failures (called passive breakdown) and repairs, in which an arriving customer who finds the server is idle must turn on the server and the server may fail to be activated. Wang and Zhao [22] dealt with a discrete-time $Geo/G/1$ retrial queueing model where all the arriving customers require a first essential service while only some of them may opt for a second optional service. Several other papers on the discrete-time retrial queues with server breakdowns can be referred to [23–25].

It is observed from the existing literature that in almost all the models of repairable discrete-time retrial queues (including the above-mentioned models), the authors always assumed that during repair period the repair times can last, without any interruption, from the beginning of the repair until the completion of the repair. But in real-world situations, the repair of the broken server may be interrupted by some casual random factors such as the limited ability of repairman, the failures of repair facility and the delayed supplies of the required device components. On the basis of this fact, in the present study, we introduce the repair replacement discipline into our model. More specifically speaking, when the repair is subject to interruption, a replacement occurs in the remaining repair times and a new repair whose length is independent of the previous repair times commences from the beginning. The introduction of repair replacements makes the considered system more realistic and flexible. Up to now, the study regarding the repair replacement discipline in discrete-time queue can only be found in [26], but it does not take retrial phenomenon into consideration.

A very important class of queues in congestion situations is retrial queues with priorities. The priority queue has received remarkable attention in the queueing literature because different kinds of customers need different quality of service (QoS). Queues with priority subscribers can be used in a wide range of stochastic service systems like manufacturing and production systems, transportation systems, telecommunication industry, computer and communication systems, etc. In recent years, many fruitful theoretical results have been reported in the area of retrial queues with priorities. Atencia [27] analyzed a $Geo/G/1$ retrial queueing system with priority services in which the arriving customers can decide to go directly to the server expelling out of the system the customer that is currently being served, if any, or to join the retrial orbit. Lan and Tang [28] considered a discrete-time $Geo/G/1$ retrial queueing system with non-preemptive priority, working vacations and vacation interruption, and derived various crucial performance measures in steady state. More studies on retrial queues with priorities can be referred to [29–33].

Another significant feature in queueing literature is queue with balking behavior proposed by Haight [34]. The balking means that when the arriving customers find that the server is unavailable or the queue is long, they may be discouraged and decide to abandon the system without service. Queueing models with balking phenomenon arise in many real-life congestion problems such as customers' impatience at a telephone switchboard, hospital emergency rooms handling critical patients, machine repair models and perishable goods for storage in inventory systems. Additionally, since a customer who balks is definitely considered to be lost, the balking of customers can greatly lead to the economic loss, which is what the system managers do not want to see. Therefore, it is worth investigating the queues with balking customers. Although the incorporation of balking customers in queueing systems is rather significant, there is only little work concerning discrete-time queues with balking customers in the literature, see [35–38].

In this paper, we are interested in analyzing a single-server discrete-time $Geo/G/1$ retrial queueing system with probabilistic preemptive priority and balking customers taking into account starting failures and replacements of repair times. To the best of our knowledge, the present investigation is the first one that deals with such system, and there are some significant differences between our research and the existing works. The contributions and advantages of this paper are stated as follows.

(1) *Model.* A novel discrete-time $Geo/G/1$ retrial queueing system is considered. By incorporating probabilistic preemptive priority, balking customers, starting failures and replacements of repair times, this model has much greater flexibility in characterizing some complicated stochastic phenomena.

(2) *Methodology and Results.* Employing the supplementary variable method and the generating function technique, the probability generating functions of the queue size under different server states are obtained. Various performance measures in steady state are also derived. Some corresponding results for special discrete-time queues are directly obtained by our results.

(3) *Numerical Illustrations.* Sensitivity analysis is carried out to illustrate the impact of different system parameters on the performance characteristics, which can provide insight to the system managers so as to supervise the operation status of this system and reduce the congestion problem. In addition, we establish a cost function to determine the optimal value of replacement probability η so as to help the system managers or decision-makers regulate the system economically.

A practical justification of the considered model is the following flexible manufacturing system that operates in discrete-time environment where the events (e.g., arrival of components and their processing time) can only occur at regularly spaced epochs. Since diverse components have different processing requirements, the situation with priority levels is often encountered in practice. Assume that there are two types of components (the urgent components and the regular components). The raw components arrive at the system according to a Bernoulli process. If the system is busy processing component upon an arrival, the newly arriving component either interrupts the component in service and occupies the system to begin its own processing immediately with probability p (corresponding to the urgent components), or leaves the service zone and enters a group of blocked components with complementary probability $\bar{p}$ (corresponding to the regular components). In order to save energy, the system will be turned off once there is no component needing to be processed, and the new arrival must turn on the server to commence its service. Also, the system may be subjected to starting failures. When the system breaks down, it will be fixed immediately. During a repair period, replacements may take place due to the limited ability of repairman (or the failures of repair facility), and the new arrivals may balk with some probability. In this scenario, the flexible manufacturing system, the raw components, the failures of system, and a group of blocked components correspond to the server, the customers, unreliable server, and retrial customers in the queueing terminology, respectively.

The rest of this paper is as follows. In the next section, a detailed description of the mathematical model under consideration is given. Section 3 is dedicated to the steady-state analysis by the supplementary variable method and the generating function technique. We obtain the closed-form expressions for the generating functions of the stationary distribution of the considered system. In Section 4, various important performance indices of our model are given. Further, some special cases of the proposed model are deduced in Section 5. In Section 6, some numerical results for sensitivity analysis are provided. Section 7 focuses on a cost optimization problem. At last, concluding remarks and suggestions for future work are given in Section 8.

2. Model formulation

Consider a single-server discrete-time $Geo/G/1$ retrial queueing system with probabilistic preemptive priority, starting failures, balking rule and replacements of repair times. It is well known that the probability of two or more queueing activities occurring simultaneously is zero in continuous-time queueing systems while it is not so in discrete-time queueing systems. In discrete-time queues, the time axis is divided into equal time intervals (called slots) and is marked with $0, 1, 2, \cdots, m, \cdots$. All the queueing events (arrivals, departures, beginning and ending of repairs, and retrials) are nonnegative integer-valued random variables and only occur around the slot boundaries. As a result, these queueing activities may take place at the same time and the order of occurrence of these events must be stipulated. In this study, it is assumed that the arrivals, the retrials and the beginning of the repairs occur in time interval (m, m^+), and the departures and the completion of the repairs occur in time interval (m^-, m). That is, we consider an early arrival system (EAS). More details on the EAS policy and related concepts can be referred to [9]. To make it clearer, the various time epochs at which queueing events occur are displayed in Figure 1.

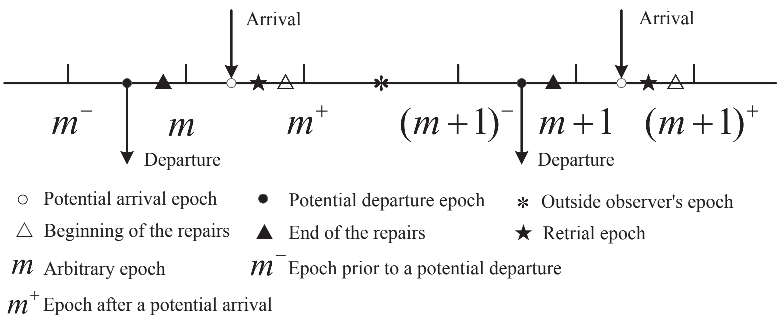

Figure 1. Various time epochs in an early arrival system (EAS).

The detailed formulation of our model is described as follows. Throughout this paper, we have $\bar{x} = 1 - x$ for any real number $x \in [0, 1]$.

Customers arrive at the system according to a Bernoulli process with parameter λ. There is no waiting space in front of the service station. If the server is busy rendering service to a customer upon an arrival, the newly arriving customer either interrupts the customer in service and occupies the server to begin its own service immediately with probability p or leaves the service zone and enters a group of blocked customers (called retrial orbit) with complementary probability $\bar{p}$. The interrupted customer joins the orbit and its service resumes from the beginning after some random length of time. It is supposed that the customers in the orbit form a waiting line in accordance with first-come-first-served (FCFS) regime, that is, only the customer (if any) at the head of the orbit is allowed to make retrials to get service. Successive inter-retrial times of any customer are governed by a general distribution with probability distribution function (pdf) $\{a_i\}_{i=0}^{\infty}$ and probability generating function (pgf) $A(z) = \sum_{i=0}^{\infty} a_i z^i$. When the server is free and both a new external arrival and a retrial take place at the same time, the customer making retrials gives up the attempt for service and the newly arriving customer has access to the server. In this case, the retrial customer goes back to its position in the retrial queue.

If an arriving customer (external or repeated) finds the server free, he/she must turn on the server to commence his/her service. If the server is switched on successfully (with probability β), the customer is served immediately and leaves the system forever after service completion. Otherwise, if the server is activated unsuccessfully (with complementary probability $\bar{\beta}$), the server immediately undergoes a repair process and the customer must enter the orbit. The repair times of the broken server are independent and identically distributed with pdf $\{r_i\}_{i=1}^{\infty}$, pgf $R(z) = \sum_{i=1}^{\infty} r_i z^i$ and n-th factorial moments $\mu_{2,n}$. During a repair period, replacements in the repair times may take place, that is, in every slot of a repair period, a replacement in the remaining repair time occurs with probability η, and does not occur with probability $\bar{\eta}$. The replacement time is assumed to be negligible. After repair completion, the server is as good as new. Moreover, if an external arrival finds that the server is down (under repair), he/she decides either to join the orbit with probability q or leaves the system completely (balking) with probability $\bar{q}$.

The service times follow an arbitrary distribution with pdf $\{s_i\}_{i=1}^{\infty}$, pgf $S(z) = \sum_{i=1}^{\infty} s_i z^i$, and n-th factorial moments $\mu_{1,n}$. It is further assumed that all the random variables involved in our model are independent of each other. Finally, in order to avoid trivial cases and make our system more realistic, we suppose that $0 < \lambda < 1$, $0 < p \leq 1$, $0 < \beta \leq 1$, $0 \leq q \leq 1$, and $0 \leq \eta < 1$.

3. The Markov chain

The state of the system at time m^+ (the epoch immediately after time epoch m) can be designated by the process

$$\{X_m, m \in \mathrm{N}\} = \{(J_m, \xi_{J_m,m}, N_m), m \in \mathrm{N}\}, \tag{3.1}$$

where N_m denotes the number of repeated customers in the orbit, and J_m represents the state of the server as

$$J_m = \begin{cases} 0, & \text{the server is free at time } m^+, \\ 1, & \text{the server is busy at time } m^+, \\ 2, & \text{the server is under repair at time } m^+. \end{cases} \tag{3.2}$$

If $J_m = 0$ and $N_m > 0$, then $\xi_{0,m}$ denotes the remaining retrial time. If $J_m = 1$, then $\xi_{1,m}$ corresponds to the remaining service time of the customer currently being served. If $J_m = 2$, then $\xi_{2,m}$ represents the residual repair time.

After introducing the supplementary variables corresponding to the remaining retrial time, the remaining service time and the remaining repair time, the next state of the system is independent of the past state and only depends on the present state. Therefore, it can be readily shown that $\{X_m, m \in \mathrm{N}\}$ is the Markov chain of our queueing model with state space

$$\Omega = \{(0,0)\} \cup \{(0,i,k) : i \geq 1, k \geq 1\} \cup \{(1,i,k) : i \geq 1, k \geq 0\}$$

$$\cup \{(2,i,k) : i \geq 1, k \geq 1\}. \tag{3.3}$$

The steady-state probability distributions of the Markov chain $\{X_m, m \in \mathrm{N}\}$ are defined as follows:

$$\begin{aligned}
\pi_{0,0} &= \lim_{m\to\infty} \Pr\left[J_m = 0, N_m = 0\right], &(3.4)\\
\pi_{0,i,k} &= \lim_{m\to\infty} \Pr\left[J_m = 0, \xi_{0,m} = i, N_m = k\right], i \geq 1, k \geq 1, &(3.5)\\
\pi_{1,i,k} &= \lim_{m\to\infty} \Pr\left[J_m = 1, \xi_{1,m} = i, N_m = k\right], i \geq 1, k \geq 0, &(3.6)\\
\pi_{2,i,k} &= \lim_{m\to\infty} \Pr\left[J_m = 2, \xi_{2,m} = i, N_m = k\right], i \geq 1, k \geq 1. &(3.7)
\end{aligned}$$

By using the supplementary variable method, we establish the Kolmogorov equations for the stationary distribution of the system under investigation as follows

$$\begin{aligned}
\pi_{0,0} &= \bar{\lambda}\pi_{0,0} + \bar{\lambda}\pi_{1,1,0}, &(3.8)\\
\pi_{0,i,k} &= \bar{\lambda}\pi_{0,i+1,k} + \bar{\lambda}a_i\pi_{1,1,k} + \bar{\lambda}a_i\pi_{2,1,k},\ i \geq 1, k \geq 1, &(3.9)\\
\pi_{1,i,k} &= \delta_{0,k}\lambda\beta s_i\pi_{0,0} + \bar{\lambda}\beta s_i\pi_{0,1,k+1} + (1-\delta_{0,k})\,\lambda\beta s_i\sum_{j=1}^{\infty}\pi_{0,j,k} + \lambda\beta s_i\pi_{1,1,k}\\
&\quad +\bar{\lambda}\beta s_i a_0\pi_{1,1,k+1} + (1-\delta_{0,k})\,\lambda p s_i\sum_{j=2}^{\infty}\pi_{1,j,k-1} + (1-\delta_{0,k})\,\lambda\bar{p}\pi_{1,i+1,k-1}\\
&\quad +\bar{\lambda}\pi_{1,i+1,k} + (1-\delta_{0,k})\,\lambda\beta s_i\pi_{2,1,k} + \bar{\lambda}\beta a_0 s_i\pi_{2,1,k+1},\ i \geq 1, k \geq 0, &(3.10)\\
\pi_{2,i,k} &= \delta_{1,k}\lambda\bar{\beta}r_i\pi_{0,0} + \bar{\lambda}\bar{\beta}r_i\pi_{0,1,k} + (1-\delta_{1,k})\,\lambda\bar{\beta}r_i\sum_{j=1}^{\infty}\pi_{0,j,k-1} + \lambda\bar{\beta}r_i\pi_{1,1,k-1}\\
&\quad +\bar{\lambda}\bar{\beta}r_i a_0\pi_{1,1,k} + (1-\delta_{1,k})\,\lambda\bar{\beta}r_i\pi_{2,1,k-1} + \bar{\lambda}\bar{\beta}r_i a_0\pi_{2,1,k}\\
&\quad + (1-\delta_{1,k})\,\lambda q\bar{\eta}\pi_{2,i+1,k-1} + (1-\delta_{1,k})\,\lambda q\eta r_i\sum_{j=2}^{\infty}\pi_{2,j,k-1}\\
&\quad + \left(\bar{\lambda} + \lambda\bar{q}\right)\bar{\eta}\pi_{2,i+1,k} + \left(\bar{\lambda} + \lambda\bar{q}\right)\eta r_i\sum_{j=2}^{\infty}\pi_{2,j,k},\ i \geq 1, k \geq 1, &(3.11)
\end{aligned}$$

where $\delta_{i,j}$ is the Kronecker delta, i.e., $\delta_{i,j} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$, and the normalization condition is

$$\pi_{0,0} + \sum_{i=1}^{\infty}\sum_{k=1}^{\infty}\pi_{0,i,k} + \sum_{i=1}^{\infty}\sum_{k=0}^{\infty}\pi_{1,i,k} + \sum_{i=1}^{\infty}\sum_{k=1}^{\infty}\pi_{2,i,k} = 1. \tag{3.12}$$

In order to derive the solutions of (3.8)–(3.11), we introduce the following generating functions

$$\begin{aligned}
\varphi_0(x,z) &= \sum_{i=1}^{\infty}\sum_{k=1}^{\infty}\pi_{0,i,k}x^i z^k, &(3.13)\\
\varphi_1(x,z) &= \sum_{i=1}^{\infty}\sum_{k=0}^{\infty}\pi_{1,i,k}x^i z^k, &(3.14)
\end{aligned}$$

$$\varphi_2(x,z) = \sum_{i=1}^{\infty}\sum_{k=1}^{\infty}\pi_{2,i,k}x^i z^k, \tag{3.15}$$

and the auxiliary generating functions

$$\varphi_{0,i}(z) = \sum_{k=1}^{\infty}\pi_{0,i,k}z^k,\ i\geq 1, \tag{3.16}$$

$$\varphi_{1,i}(z) = \sum_{k=0}^{\infty}\pi_{1,i,k}z^k,\ i\geq 1, \tag{3.17}$$

$$\varphi_{2,i}(z) = \sum_{k=1}^{\infty}\pi_{2,i,k}z^k,\ i\geq 1. \tag{3.18}$$

Multiplying Eq (3.9)–(3.11) by z^k, summing over k, and taking into account (3.8), it follows that

$$\varphi_{0,i}(z) = \bar{\lambda}\varphi_{0,i+1}(z) + \bar{\lambda}a_i\varphi_{1,1}(z) + \bar{\lambda}a_i\varphi_{2,1}(z) - \lambda a_i\pi_{0,0},\ i\geq 1, \tag{3.19}$$

$$\begin{aligned}\varphi_{1,i}(z) =\ & \frac{\bar{\lambda}\beta}{z}s_i\varphi_{0,1}(z) + \lambda\beta s_i\varphi_0(1,z) + \left(\frac{\lambda z+\bar{\lambda}a_0}{z}\beta - \lambda pz\right)s_i\varphi_{1,1}(z)\\ & +\tau_1(z)\varphi_{1,i+1}(z) + \lambda pzs_i\varphi_1(1,z) + \frac{\lambda z+\bar{\lambda}a_0}{z}\beta s_i\varphi_{2,1}(z)\\ & +\frac{z-a_0}{z}\lambda\beta s_i\pi_{0,0},\ i\geq 1,\end{aligned} \tag{3.20}$$

$$\begin{aligned}\varphi_{2,i}(z) =\ & \bar{\lambda}\bar{\beta}r_i\varphi_{0,1}(z) + \lambda\bar{\beta}zr_i\varphi_0(1,z) + \left(\lambda z+\bar{\lambda}a_0\right)\bar{\beta}r_i\varphi_{1,1}(z)\\ & +\tau_2(z)\left(\bar{\eta}\varphi_{2,i+1}(z) + \eta r_i\varphi_2(1,z)\right) + (z-a_0)\lambda\bar{\beta}r_i\pi_{0,0}\\ & +\left[\left(\lambda z+\bar{\lambda}a_0\right)\bar{\beta} - \tau_2(z)\eta\right]r_i\varphi_{2,1}(z),\ i\geq 1,\end{aligned} \tag{3.21}$$

where $\tau_1(z) = \bar{\lambda} + \lambda\bar{p}z$, $\tau_2(z) = \bar{\lambda} + \lambda\bar{q} + \lambda qz$.

Multiplying (3.19)–(3.21) by x^i and summing over i gives

$$\frac{x-\bar{\lambda}}{x}\varphi_0(x,z) = \bar{\lambda}\left(A(x)-a_0\right)\left(\varphi_{1,1}(z) + \varphi_{2,1}(z) - \frac{\lambda}{\bar{\lambda}}\pi_{0,0}\right) - \bar{\lambda}\varphi_{0,1}(z), \tag{3.22}$$

$$\begin{aligned}\frac{x-\tau_1(z)}{x}\varphi_1(x,z) =\ & \frac{\bar{\lambda}\beta}{z}S(x)\varphi_{0,1}(z) + \lambda\beta S(x)\varphi_0(1,z) + \lambda pzS(x)\varphi_1(1,z)\\ & +\left[\frac{\left(\lambda z+\bar{\lambda}a_0\right)\beta - \lambda pz^2}{z}S(x) - \tau_1(z)\right]\varphi_{1,1}(z)\\ & +\frac{\lambda z+\bar{\lambda}a_0}{z}\beta S(x)\varphi_{2,1}(z) + \frac{z-a_0}{z}\lambda\beta S(x)\pi_{0,0},\end{aligned} \tag{3.23}$$

$$\begin{aligned}\frac{x-\tau_2(z)\bar{\eta}}{x}\varphi_2(x,z) =\ & \bar{\lambda}\bar{\beta}R(x)\varphi_{0,1}(z) + \lambda\bar{\beta}R(x)\varphi_0(1,z)\\ & +\left(\lambda z+\bar{\lambda}a_0\right)\bar{\beta}R(x)\varphi_{1,1}(z)\\ & +\left\{\left[\left(\lambda z+\bar{\lambda}a_0\right)\bar{\beta} - \tau_2(z)\eta\right]R(x) - \tau_2(z)\bar{\eta}\right\}\varphi_{2,1}(z)\end{aligned}$$

$$+\tau_2(z)\eta R(x)\varphi_2(1,z) + (z-a_0)\,\lambda\bar{\beta}R(x)\pi_{0,0}. \tag{3.24}$$

Setting $x = 1$ in (3.22)–(3.24) and solving for $\varphi_0(1,z)$, $\varphi_1(1,z)$ and $\varphi_2(1,z)$, we have

$$\lambda\varphi_0(1,z) = \bar{\lambda}\,(1-a_0)\left(\varphi_{1,1}(z) + \varphi_{2,1}(z) - \frac{\lambda}{\bar{\lambda}}\pi_{0,0}\right) - \bar{\lambda}\varphi_{0,1}(z), \tag{3.25}$$

$$\begin{aligned}\lambda z(1-z)\varphi_1(1,z) &= \bar{\lambda}\beta(1-z)\varphi_{0,1}(z) + \left[\tau_3(z)\beta - z(\bar{\lambda}+\lambda z)\right]\varphi_{1,1}(z)\\ &\quad+\tau_3(z)\beta\varphi_{2,1}(z) - (1-z)\lambda\beta a_0\pi_{0,0},\end{aligned} \tag{3.26}$$

$$\begin{aligned}\lambda q(1-z)\varphi_2(1,z) &= \bar{\lambda}\bar{\beta}(1-z)\varphi_{0,1}(z) + \tau_3(z)\bar{\beta}\varphi_{1,1}(z)\\ &\quad+\left[\tau_3(z)\bar{\beta} - \tau_2(z)\right]\varphi_{2,1}(z) - (1-z)\,\lambda\bar{\beta}a_0\pi_{0,0},\end{aligned} \tag{3.27}$$

where $\tau_3(z) = z + \bar{\lambda}a_0(1-z)$.

Substituting Eq (3.25)–(3.27) into (3.23) and (3.24), after some tedious algebraic manipulation, we obtain

$$\begin{aligned}&z(1-z)\frac{x-\tau_1(z)}{x}\varphi_1(x,z)\\ &\quad= \bar{\lambda}\beta(1-z)(1-\bar{p}z)S(x)\varphi_{0,1}(z)\\ &\qquad+\left\{\left[\beta(1-\bar{p}z)\tau_3(z) - pz^2\right]S(x) - z(1-z)\tau_1(z)\right\}\varphi_{1,1}(z)\\ &\qquad+\beta(1-\bar{p}z)\tau_3(z)S(x)\varphi_{2,1}(z)\\ &\qquad-(1-z)(1-\bar{p}z)\lambda\beta a_0 S(x)\pi_{0,0},\end{aligned} \tag{3.28}$$

$$\begin{aligned}&\lambda q(1-z)\frac{x-\tau_2(z)\bar{\eta}}{x}\varphi_2(x,z)\\ &\quad= \bar{\lambda}\bar{\beta}(1-z)\tau_4(z)R(x)\varphi_{0,1}(z) + \bar{\beta}\tau_3(z)\tau_4(z)R(x)\varphi_{1,1}(z)\\ &\qquad+\left\{\left[\bar{\beta}\tau_3(z)\tau_4(z) - \tau_2(z)\eta\right]R(x) - \lambda q(1-z)\tau_2(z)\bar{\eta}\right\}\varphi_{2,1}(z)\\ &\qquad-\lambda\bar{\beta}(1-z)\tau_4(z)a_0R(x)\pi_{0,0},\end{aligned} \tag{3.29}$$

where $\tau_4(z) = \lambda q(1-z) + \tau_2(z)\eta$.

Letting $x = \bar{\lambda}$ in (3.22), $x = \tau_1(z)$ in (3.28) and $x = \tau_2(z)\bar{\eta}$ in (3.29) respectively, it finally leads to

$$\begin{aligned}&\lambda\left(A(\bar{\lambda}) - a_0\right)\pi_{0,0}\\ &\quad= -\bar{\lambda}\varphi_{0,1}(z) + \bar{\lambda}\left(A(\bar{\lambda}) - a_0\right)\varphi_{1,1}(z) + \bar{\lambda}\left(A(\bar{\lambda}) - a_0\right)\varphi_{2,1}(z),\end{aligned} \tag{3.30}$$

$$\begin{aligned}&(1-z)(1-\bar{p}z)\lambda\beta a_0 S(\tau_1(z))\pi_{0,0}\\ &\quad= \bar{\lambda}\beta(1-z)(1-\bar{p}z)S(\tau_1(z))\varphi_{0,1}(z)\\ &\qquad+\left\{\left[\beta(1-\bar{p}z)\tau_3(z) - pz^2\right]S(\tau_1(z)) - z(1-z)\tau_1(z)\right\}\varphi_{1,1}(z)\\ &\qquad+\beta(1-\bar{p}z)\tau_3(z)S(\tau_1(z))\varphi_{2,1}(z),\end{aligned} \tag{3.31}$$

$$\lambda\bar{\beta}(1-z)\tau_4(z)a_0R(\tau_2(z)\bar{\eta})\pi_{0,0}$$

$$= \bar{\lambda}\bar{\beta}(1-z)\tau_4(z)R(\tau_2(z)\bar{\eta})\varphi_{0,1}(z) + \bar{\beta}\tau_3(z)\tau_4(z)R(\tau_2(z)\bar{\eta})\varphi_{1,1}(z) + \left\{\left[\bar{\beta}\tau_3(z)\tau_4(z) - \tau_2(z)\eta\right]R(\tau_2(z)\bar{\eta}) - \lambda q(1-z)\tau_2(z)\bar{\eta}\right\}\varphi_{2,1}(z), \tag{3.32}$$

Solving the above system for $\varphi_{0,1}(z)$, $\varphi_{1,1}(z)$ and $\varphi_{2,1}(z)$, we get

$$\varphi_{0,1}(z) = \frac{\left(A(\bar{\lambda}) - a_0\right) z D_0(z)}{D(z)} \frac{\lambda}{\bar{\lambda}} \pi_{0,0}, \tag{3.33}$$

$$\varphi_{1,1}(z) = \frac{\lambda\beta A(\bar{\lambda})(1-\bar{p}z)S(\tau_1(z))\tau_2(z)D_1(z)}{D(z)}\pi_{0,0}, \tag{3.34}$$

$$\varphi_{2,1}(z) = \frac{\lambda\bar{\beta}A(\bar{\lambda})zR(\tau_2(z)\bar{\eta})\tau_4(z)D_2(z)}{D(z)}\pi_{0,0}, \tag{3.35}$$

where

$$\begin{aligned} D(z) &= S(\tau_1(z))\left\{R(\tau_2(z)\bar{\eta})\left[pz^2\bar{\beta}[\lambda qz + \bar{\lambda}A(\bar{\lambda})\tau_4(z)] + \Delta_1(z)\right] \right. \\ &\quad \left. + \lambda q\tau_2(z)\bar{\eta}\left[\beta(1-\bar{p}z)\Delta_2(z) - pz^2\right]\right\} \\ &\quad + z\tau_1(z)\left\{R(\tau_2(z)\bar{\eta})\left[\bar{\beta}\tau_4(z)\Delta_2(z) - \tau_2(z)\eta\right] - \lambda q\bar{\eta}(1-z)\tau_2(z)\right\}, \end{aligned} \tag{3.36}$$

$$\begin{aligned} D_0(z) &= S(\tau_1(z))\{\lambda q\tau_2(z)\bar{\eta}\left[pz - (1-\bar{p}z)\beta\right] \\ &\quad - R(\tau_2(z)\bar{\eta})\left[\lambda qpz^2\bar{\beta} + (\beta - pz\bar{\beta})\tau_2(z)\eta\right]\} \\ &\quad + \tau_1(z)\left\{R(\tau_2(z)\bar{\eta})\left[\tau_2(z)\eta - \bar{\beta}z\tau_4(z)\right] + \lambda q\bar{\eta}(1-z)\tau_2(z)\right\}, \end{aligned} \tag{3.37}$$

$$D_1(z) = \eta R(\tau_2(z)\bar{\eta}) + \lambda q\bar{\eta}(1-z), \tag{3.38}$$

$$D_2(z) = pzS(\tau_1(z)) + (1-z)\tau_1(z), \tag{3.39}$$

$$\Delta_1(z) = \left[(\beta - pz\bar{\beta})z + (1-\bar{p}z)\beta\bar{\lambda}A(\bar{\lambda})\right]\tau_2(z)\eta, \tag{3.40}$$

$$\Delta_2(z) = z + \bar{\lambda}A(\bar{\lambda})(1-z). \tag{3.41}$$

Inserting (3.33)–(3.35) into (3.22), (3.28) and (3.29) gives

$$\varphi_0(x,z) = \frac{A(x) - A(\bar{\lambda})}{x - \bar{\lambda}} \frac{\lambda xzD_0(z)}{D(z)}\pi_{0,0}, \tag{3.42}$$

$$\varphi_1(x,z) = \frac{S(x) - S(\tau_1(z))}{x - \tau_1(z)} \frac{\lambda\beta xA(\bar{\lambda})(1-\bar{p}z)\tau_1(z)\tau_2(z)D_1(z)}{D(z)}\pi_{0,0}, \tag{3.43}$$

$$\varphi_2(x,z) = \frac{R(x) - R(\tau_2(z)\bar{\eta})}{x - \tau_2(z)\bar{\eta}} \frac{\lambda\bar{\beta}\bar{\eta}xA(\bar{\lambda})z\tau_2(z)\tau_4(z)D_2(z)}{D(z)}\pi_{0,0}. \tag{3.44}$$

At this point, the only unknown is $\pi_{0,0}$ and it can be determined by the normalization condition $\pi_{0,0} + \varphi_0(1,1) + \varphi_1(1,1) + \varphi_2(1,1) = 1$. Taking $x = z = 1$ in (3.42)–(3.44) and using L'Hospital rule, we can get

$$\pi_{0,0} = \frac{D(1)}{pA(\bar{\lambda})S\left(\bar{\lambda} + \lambda\bar{p}\right)\left[\beta\eta R(\bar{\eta}) + \lambda\bar{\beta}\bar{q}\left(\bar{\eta} - R(\bar{\eta})\right)\right]}, \tag{3.45}$$

where $D(1) = S\left(\bar{\lambda} + \lambda\bar{p}\right)\left[R(\bar{\eta})[\lambda pq\bar{\beta} + \bar{\lambda}p\eta A(\bar{\lambda}) + (\beta - p\bar{\beta})\eta] - \lambda pq\bar{\beta}\bar{\eta}\right] - (\bar{\lambda} + \lambda\bar{p})\beta\eta R(\bar{\eta})$.

From the representation of $\pi_{0,0}$, as $\pi_{0,0} > 0$, we have that $D(1) > 0$ is a necessary condition for the system to be stable. Also, employing Foster's criterion (see [39]), we can show that $D(1) > 0$ is also a sufficient condition for the stability of the system.

We summarize the above results in the following theorem.

Theorem 1. If $D(1) > 0$, the stationary distribution of the Markov chain $\{X_m, m \in \mathrm{N}\}$ has the following generating functions

$$\varphi_0(x,z) = \frac{A(x)-A(\bar{\lambda})}{x-\bar{\lambda}}\frac{\lambda xzD_0(z)}{D(z)}\pi_{0,0}, \tag{3.46}$$

$$\varphi_1(x,z) = \frac{S(x)-S(\tau_1(z))}{x-\tau_1(z)}\frac{\lambda\beta xA(\bar{\lambda})(1-\bar{p}z)\tau_1(z)\tau_2(z)D_1(z)}{D(z)}\pi_{0,0}, \tag{3.47}$$

$$\varphi_2(x,z) = \frac{R(x)-R(\tau_2(z)\bar{\eta})}{x-\tau_2(z)\bar{\eta}}\frac{\lambda\bar{\beta}\bar{\eta}xA(\bar{\lambda})z\tau_2(z)\tau_4(z)D_2(z)}{D(z)}\pi_{0,0}, \tag{3.48}$$

where $\pi_{0,0}$ is given by (3.45).

Based on Theorem 1, we can directly derive the marginal generating functions for the number of customers in the orbit under different server states and the generating functions for the number of customers in the orbit and in the system.

Corollary 1. (1) The marginal generating function for the number of customers in the orbit when the server is idle is given by

$$\pi_{0,0} + \varphi_0(1,z) = \frac{A(\bar{\lambda})K(z)}{D(z)}\pi_{0,0}, \tag{3.49}$$

where

$$\begin{aligned} K(z) = {} & S(\tau_1(z))\Big\{R(\tau_2(z)\bar{\eta})\Big[pz^2\bar{\beta}[\lambda qz+\bar{\lambda}\tau_4(z)]+[(\beta-pz\bar{\beta})z+(1-\bar{p}z)\beta\bar{\lambda}]\tau_2(z)\eta\Big] \\ & +\lambda q\tau_2(z)\bar{\eta}\Big[\beta(1-\bar{p}z)(\bar{\lambda}+\lambda z)-pz^2\Big]\Big\} \\ & +z\tau_1(z)\Big\{R(\tau_2(z)\bar{\eta})\Big[\bar{\beta}\tau_4(z)(\bar{\lambda}+\lambda z)-\tau_2(z)\eta\Big]-\lambda q\bar{\eta}(1-z)\tau_2(z)\Big\}. \end{aligned} \tag{3.50}$$

(2) The marginal generating function for the number of customers in the orbit when the server is busy is given by

$$\varphi_1(1,z) = \frac{(1-S(\tau_1(z)))\beta A(\bar{\lambda})\tau_1(z)\tau_2(z)D_1(z)}{D(z)}\pi_{0,0}. \tag{3.51}$$

(3) The marginal generating function for the number of customers in the orbit when the server is down (under repair) is given by

$$\varphi_2(1,z) = \frac{1-R(\tau_2(z)\bar{\eta})}{1-\tau_2(z)\bar{\eta}}\frac{\lambda\bar{\beta}\bar{\eta}A(\bar{\lambda})z\tau_2(z)\tau_4(z)D_2(z)}{D(z)}\pi_{0,0}. \tag{3.52}$$

(4) The probability generating function of the number of customers in the orbit is given by

$$\phi(z) = \pi_{0,0} + \varphi_0(1,z) + \varphi_1(1,z) + \varphi_2(1,z). \tag{3.53}$$

(5) The probability generating function of the number of customers in the system is given by

$$\psi(z) = \pi_{0,0} + \varphi_0(1,z) + z\varphi_1(1,z) + \varphi_2(1,z). \tag{3.54}$$

4. Performance measures

In the previous discussion, the analytical results for the probability generating functions of various queue size distributions for different system states are established. Now some important performance measures of the system in steady state are obtained as follows.

(1) The probability that the system is empty is given by

$$\pi_{0,0} = \frac{D(1)}{pA(\bar{\lambda})S\left(\bar{\lambda} + \lambda\bar{p}\right)\left[\beta\eta R(\bar{\eta}) + \lambda\bar{\beta}\bar{q}\left(\bar{\eta} - R(\bar{\eta})\right)\right]}, \tag{4.1}$$

where

$$\begin{aligned} D(1) \;=\; & S\left(\bar{\lambda} + \lambda\bar{p}\right)\left[R(\bar{\eta})[\lambda pq\bar{\beta} + \bar{\lambda}p\eta A(\bar{\lambda}) + (\beta - p\bar{\beta})\eta] - \lambda pq\bar{\beta}\bar{\eta}\right] \\ & -(\bar{\lambda} + \lambda\bar{p})\beta\eta R(\bar{\eta}). \end{aligned} \tag{4.2}$$

(2) The probability that the system is non-empty (occupied) is given by

$$\varphi_0(1,1) + \varphi_1(1,1) + \varphi_2(1,1) = 1 - \pi_{0,0}. \tag{4.3}$$

(3) The probability that the server is idle is given by

$$\pi_{0,0} + \varphi_0(1,1) = \frac{K(1)}{pS\left(\bar{\lambda} + \lambda\bar{p}\right)\left[\beta\eta R(\bar{\eta}) + \lambda\bar{\beta}\bar{q}\left(\bar{\eta} - R(\bar{\eta})\right)\right]}, \tag{4.4}$$

where

$$\begin{aligned} K(1) \;=\; & S\left(\bar{\lambda} + \lambda\bar{p}\right)\left[R(\bar{\eta})[\lambda pq\bar{\beta} + \bar{\lambda}p\eta + (\beta - p\bar{\beta})\eta] - \lambda pq\bar{\beta}\bar{\eta}\right] \\ & -(\bar{\lambda} + \lambda\bar{p})\beta\eta R(\bar{\eta}). \end{aligned} \tag{4.5}$$

(4) The probability that the server is busy is given by

$$\varphi_1(1,1) = \frac{\left(\bar{\lambda} + \lambda\bar{p}\right)\left[1 - S(\bar{\lambda} + \lambda\bar{p})\right]\beta\eta A(\bar{\lambda})R(\bar{\eta})}{D(1)}\pi_{0,0}. \tag{4.6}$$

(5) The probability that the server is down (under repair) is given by

$$\varphi_2(1,1) = \frac{\lambda p\bar{\beta}\bar{\eta}A(\bar{\lambda})S(\bar{\lambda} + \lambda\bar{p})(1 - R(\bar{\eta}))}{D(1)}\pi_{0,0}. \tag{4.7}$$

(6) The expected number of customers in the orbit, denoted by $\mathrm{E}[W]$, is given by

$$\mathrm{E}[N] = \phi'(z)|_{z=1}. \tag{4.8}$$

(7) The expected number of customers in the system, denoted by $\mathrm{E}[L]$, is given by

$$\mathrm{E}[L] = \psi'(z)|_{z=1} = \mathrm{E}[N] + \varphi_1(1,1). \tag{4.9}$$

(8) The arrival rate to the orbit is given by

$$P_{Orbit} = \lambda\bar{p}\varphi_1(1,1) + \lambda q\varphi_2(1,1). \tag{4.10}$$

(9) The arrival rate to the system is given by

$$P_{System} = \lambda\left(\pi_{0,0} + \varphi_0(1,1) + \varphi_1(1,1)\right) + \lambda q\varphi_2(1,1) = \lambda\left(1 - \bar{q}\varphi_2(1,1)\right). \tag{4.11}$$

(10) The loss probability of a customer due to server breakdowns

$$P_{Loss} = \lambda\bar{q}\varphi_2(1,1). \tag{4.12}$$

(11) The equilibrium interruption frequency of the service due to preemptive resume is given by

$$P_{Interruption} = \lambda p\varphi_1(1,1). \tag{4.13}$$

(12) The steady-state availability of the server is given by

$$P_A = 1 - \varphi_2(1,1). \tag{4.14}$$

(13) The steady-state replacement frequency is given by

$$P_{Replacement} = \eta\varphi_2(1,1). \tag{4.15}$$

(14) According to the well-known Little's formula, the expected sojourn time of a customer in the system, denoted by $\mathrm{E}[W]$, is given by

$$\mathrm{E}[W] = \frac{\mathrm{E}[L]}{P_{System}}. \tag{4.16}$$

5. Special cases

In this section, some special cases are directly deduced by choosing appropriate values of some critical parameters in our results.

(1) If $a_0 = 1$ (i.e., without retrials), then $A\left(\bar{\lambda}\right) = 1$ and our model becomes the discrete-time $Geo/G/1$ queue with preemptive resume, starting failures, balking customers and replacements of repair times. In this case, the probability generating function of the number of customers in the system is

$$\begin{aligned}\psi(z) &= \left[\frac{K(z) + (1 - S(\tau_1(z)))\beta z\tau_1(z)\tau_2(z)D_1(z)}{K(z)}\right. \\ &\qquad \left. + \frac{\lambda\bar{\beta}\bar{\eta}z\tau_2(z)\tau_4(z)D_2(z)\left(1 - R(\tau_2(z)\bar{\eta})\right)}{\left(1 - \tau_2(z)\bar{\eta}\right)K(z)}\right]\pi_{0,0},\end{aligned} \tag{5.1}$$

where $\tau_1(z)$, $\tau_2(z)$, $\tau_4(z)$, $K(z)$, $K(1)$, $D_1(z)$ and $D_2(z)$ are given as previously defined, and

$$\pi_{0,0} = \frac{K(1)}{pS\left(\bar{\lambda} + \lambda\bar{p}\right)\left[\beta\eta R(\bar{\eta}) + \lambda\bar{\beta}\bar{q}\left(\bar{\eta} - R(\bar{\eta})\right)\right]}. \tag{5.2}$$

(2) When $p \to 0$, $q = 1$ and $\eta = 0$ in our model, the queueing system under consideration reduces to the discrete-time $Geo/G/1$ retrial queue with starting failures. In this case, the probability generating function of the number of customers in the system is

$$\psi(z) = \frac{\beta A(\bar{\lambda})(1-z)(\bar{\lambda}+\lambda z)S(\bar{\lambda}+\lambda z)}{\left(\beta S(\bar{\lambda}+\lambda z)+\bar{\beta} z R(\bar{\lambda}+\lambda z)\right)\left[z+\bar{\lambda}A(\bar{\lambda})(1-z)\right]-z(\bar{\lambda}+\lambda z)}\pi_{0,0}, \tag{5.3}$$

where $\pi_{0,0} = \frac{\lambda+\bar{\lambda}A(\bar{\lambda})-\bar{\beta}-\lambda\beta\mu_{1,1}-\lambda\bar{\beta}\mu_{2,1}}{\beta A(\bar{\lambda})}$. This result coincides with the corresponding formula presented by Corollary 1(5) of [23] with $\theta = 0$.

(3) If $q = 0$ and $\beta = 1$, then the proposed model becomes the discrete-time $Geo/G/1$ retrial queue with FCFS preemptive resume discipline. In this case, the probability generating function of the number of customers in the system is

$$\psi(z) = \frac{A(\bar{\lambda})(1-\bar{p}z)(\bar{\lambda}+\lambda z)S(\bar{\lambda}+\lambda\bar{p}z)}{S(\bar{\lambda}+\lambda\bar{p}z)\left[z+\bar{\lambda}A(\bar{\lambda})(1-\bar{p}z)\right]-z(\bar{\lambda}+\lambda\bar{p}z)}\pi_{0,0}, \tag{5.4}$$

where $\pi_{0,0} = \frac{S(\bar{\lambda}+\lambda\bar{p})\left(1+\bar{\lambda}pA(\bar{\lambda})\right)-\left(\bar{\lambda}+\lambda\bar{p}\right)}{pA(\bar{\lambda})S(\bar{\lambda}+\lambda\bar{p})}$.

(4) If $p = 1$, $q = 0$ and $\beta = 1$, then the proposed model is reduced to the discrete-time $Geo/G/1$ retrial queue with LCFS preemptive resume discipline. In this case, the probability generating function of the number of customers in the system is

$$\psi(z) = \frac{(\bar{\lambda}+\lambda z)\left[S(\bar{\lambda})\left(1+\bar{\lambda}A(\bar{\lambda})\right)-\bar{\lambda}\right]}{S(\bar{\lambda})\left(z+\bar{\lambda}A(\bar{\lambda})\right)-\bar{\lambda}z}. \tag{5.5}$$

(5) If $p \to 0$, $q = 0$ and $\beta = 1$, then our system can be reduced to the $Geo/G/1$ retrial queue with general retrial times. In this case, the probability generating function of the number of customers in the system is

$$\psi(z) = \frac{(1-z)\,S\left(\bar{\lambda}+\lambda z\right)\left(\bar{\lambda}+\lambda z\right)\left(\lambda+\bar{\lambda}A\left(\bar{\lambda}\right)-\lambda\mu_{1,1}\right)}{\left[z+(1-z)\,\bar{\lambda}A\left(\bar{\lambda}\right)\right]S\left(\bar{\lambda}+\lambda z\right)-z\left(\bar{\lambda}+\lambda z\right)}, \tag{5.6}$$

which is in agreement with the corresponding result given by Corollary 1(4) of [13].

(6) If $a_0 = 1$, $p \to 0$, $q = 0$ and $\beta = 1$, our system becomes the classical discrete-time $Geo/G/1$ queue. In this case, the probability generating function of the number of customers in the system is

$$\psi(z) = \frac{(1-\lambda\mu_{1,1})(1-z)\,S\left(\bar{\lambda}+\lambda z\right)}{S\left(\bar{\lambda}+\lambda z\right)-z}, \tag{5.7}$$

which matches with the result obtained by Hunter [9].

6. Numerical results

In this section, some numerical results are presented to illustrate the effect of various system parameters on the performance measures of our queueing model, such as the probability that the system is empty $\pi_{0,0}$, the probability that the server is down $\varphi_2(1,1)$, the loss probability of a customer due to server breakdowns P_{Loss}, and the expected number of customers in the system $\mathrm{E}[L]$. All the computations are done by developing program in MATLAB software and the values of the parameters involved in these numerical examples are selected to satisfy the stability condition $D(1) > 0$. Throughout this section, it is assumed that the service time is governed by a geometric distribution with parameter χ and the inter-retrial times are geometrically distributed with parameter $\bar{\alpha} = 1 - \alpha$, where α is the probability that a repeated customer does not make a retrial in a slot.

In Figures 2–7, we assume that the repair time follows a geometric distribution with parameter $\sigma = 0.4$. Figures 2 and 3 are depicted to explore the impact of p on $\pi_{0,0}$ and $\mathrm{E}[L]$ for different values of α. We set default parameters for Figures 2 and 3 as $\lambda = 0.2$, $\chi = 0.6$, $\sigma = 0.4$, $q = 0.2$, $\eta = 0.15$, and $\beta = 0.7$. It can be observed from Figure 2 that for fixed α, $\pi_{0,0}$ is insensitive to the change of p, which is due to the Markovian property of geometric service time distribution. On the other hand, as α increases from 0.15 to 0.75, $\pi_{0,0}$ shows a decreasing trend. The reason is that as α increases (i.e., the probability that customer does not make retrial increases), the number of customers in the orbit gradually increases, which undoubtedly leads to the decrease of $\pi_{0,0}$. From Figure 3, one can see that the mean system size $\mathrm{E}[L]$ decreases with the increase of p while increases as α increases, which is in accordance with our expectation. In fact, when α increases, more and more customers accumulate in the retrial orbit, which certainly causes the increase of the mean queue length of system $\mathrm{E}[L]$.

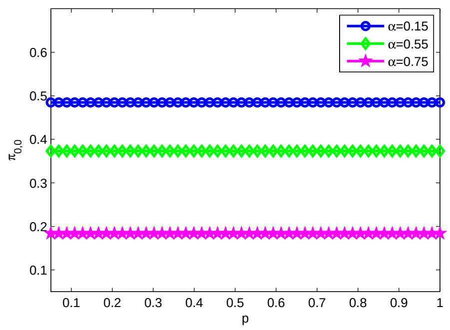

Figure 2. The effect of p on $\pi_{0,0}$ for different values of α.

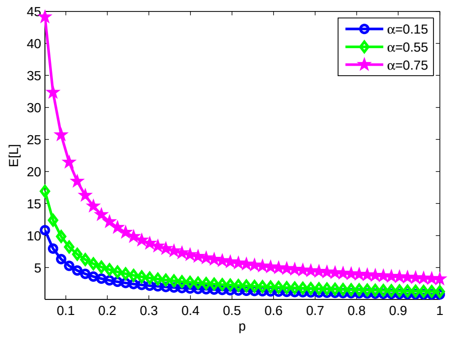

Figure 3. The effect of p on $\mathrm{E}[L]$ for different values of α.

Figures 4–7 describe the influence of q on $\pi_{0,0}$, $\mathrm{E}[L]$, $\varphi_2(1,1)$ and P_{Loss} for different values of β. We set default parameters for Figures 4–7 as $\lambda = 0.2$, $\chi = 0.6$, $\sigma = 0.4$, $p = 0.35$, $\eta = 0.15$, and $\alpha = 0.15$. Figure 4 shows that $\pi_{0,0}$ slightly decreases with the increment of q for $\beta = 0.6$ and $\beta = 0.8$. Actually, as the values of q become larger, the newly arriving customers are more likely to join the system, which in turn reduces the probability that the system is empty. However, when $\beta = 1$, we observe that $\pi_{0,0}$ keeps unchanged with the increase of q. This is due to the fact that for $\beta = 1$, the idle server is activated successfully with probability 1 and our model is reduced to the corresponding queue without starting failures. Therefore, all arrivals must enter the system regardless of the value of q. Additionally, for a fixed q, $\pi_{0,0}$ shows an increasing trend as β increases. This is intuitively true. Actually, as the value of β (i.e., the probability that the idle server is activated successfully) becomes larger, the server can serve more customers, which can obviously lead to the increase of the probability that the system is empty. The reverse effect of q and β on $\mathrm{E}[L]$ is plotted in Figure 5. From Figure 6, there is a gradual growth in $\varphi_2(1,1)$ with the increase of q. This could be due to the fact that the increasing q will lead to the excessive accumulation of customers in the orbit, and the system operates under an overload condition, which makes the idle server be prone to breakdowns. Also, it is clear from Figure 6 that $\varphi_2(1,1)$ decreases with the increasing values of β, which matches with the reality. In particular, it is noted that $\varphi_2(1,1)$ is equal to zero when $\beta = 1$, which matches with our intuition. Moreover, as it is expected in Figure 7, P_{Loss} shows a decreasing trend with the increasing values of q or β. In deed, as q increases, the probability that an arrival joins the queue increases, and hence the probability of loss of customer decreases. Similarly, when β (i.e., the probability that the idle server is activated successfully) increases, the probability that the arrivals enter the system increases, which evidently results in the decrease of the probability of loss of customer P_{Loss}.

In Figure 8, $\varphi_2(1,1)$ is plotted against the parameter η with three types of repair time distributions, i.e., geometric ($\sigma = 1/3$), deterministic ($r_3 = 1$), arbitrary ($r_1 = 0.2$, $r_3 = 0.6$, $r_5 = 0.2$). It is easy to be seen that the three distributions have the same mean value 3. The other default parameters for Figure 8 are taken as $\lambda = 0.2$, $\chi = 0.6$, $p = 0.35$, $q = 0.2$, $\alpha = 0.15$, $\beta = 0.7$. We can observe from Figure 8 that when the repair time follows geometric distribution, the value of $\varphi_2(1,1)$ is the lowest and remains constant with the increase of η, which is due to the Markovian property of geometric distribution. In addition, it is remarkable that η has a significant impact on $\varphi_2(1,1)$ when the repair time is governed by a deterministic distribution. That is, as the replacement probability η increases, the probability that the system is under repair becomes bigger and bigger. This observation suggests that system designers should pay attention to choosing appropriate repair time distribution to avoid the practical congestion situation.

The above sensitivity analysis is highly consistent with the practical situation and our expectation, which not only demonstrates the validity of our queueing model and analytical results, but also can provide managerial insight to the system designers and decision makers so as to reduce the congestion problem encountered in their respective discrete-time queueing systems.

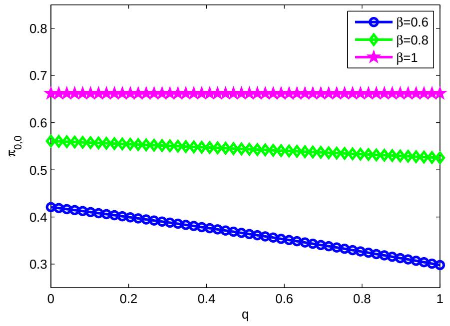

Figure 4. The effect of q on $\pi_{0,0}$ for different values of β

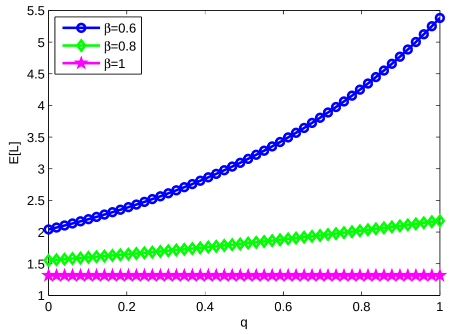

Figure 5. The effect of q on E[L] for different values of β.

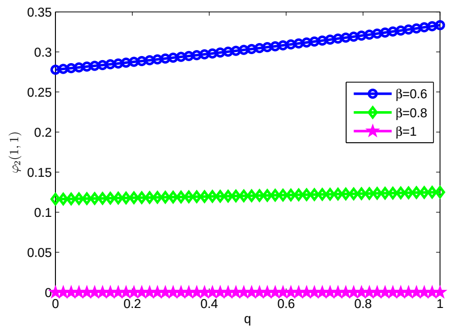

Figure 6. The effect of q on $\varphi_2(1, 1)$ for different values of β.

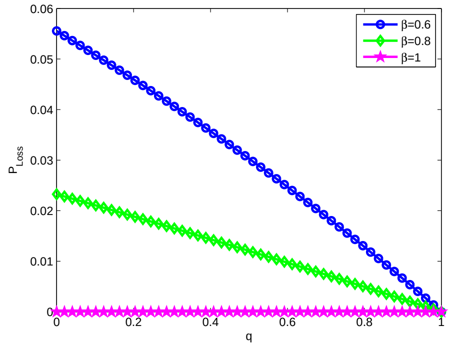

Figure 7. The effect of q on P_{Loss} for different values of β.

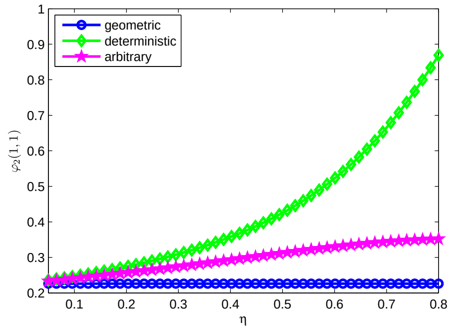

Figure 8. $\varphi_2(1, 1)$ versus η for different repair time distributions.

7. Cost optimization analysis

In practice, the operating cost of the system plays a key role in the analysis of economic revenue. Therefore, in order to make the system more profitable, system designers or system managers are usually interested in minimizing the operating cost of unit time. To demonstrate the applicability of the results obtained in the previous discussion, in this section, we develop an expected operating cost function per unit time for the queueing model under investigation in this paper, in which the replacement probability η is a decision variable. Our objective is to determine the optimum value of η, say η^*, so as to minimize the long-run expected operating cost per unit time.

To this end, we first discuss the expected length of a busy cycle period C. Denote by $\mathrm{E}[B]$ the expected length of the period that begins at the epoch at which a new arrival finds the system is empty and the server is free and ends at a service completion epoch at which the server becomes idle and the system is empty again. Let $\mathrm{E}[I]$ be the expected length of the period that starts at the instant at which the system becomes empty and ends at the instant when the first customer arrives at the empty system. Thus, the expected length of a busy cycle period is given by $\mathrm{E}[C] = \mathrm{E}[B] + \mathrm{E}[I]$. Since the inter-arrival times follow a geometric distribution with parameter λ, $\mathrm{E}[I]$ can be expressed as $\mathrm{E}[I] = \frac{1}{\lambda}$.

Furthermore, by using the argument of an alternating renewal process, we get

$$\pi_{0,0} = \frac{\mathrm{E}[I]}{\mathrm{E}[C]}, \tag{7.1}$$

which gives

$$E[C] = \frac{E[I]}{\pi_{0,0}} = \frac{1}{\lambda \pi_{0,0}}, \tag{7.2}$$

where $\pi_{0,0}$ is determined by (3.45).

We now begin to study the cost optimization problem. Let us define the cost elements as follows.

$C_h \equiv$ per unit time cost of every customer present in the system;

$C_i \equiv$ per unit time cost for keeping the server idle;

$C_b \equiv$ per unit time cost for keeping the server busy;

$C_r \equiv$ per unit time fixed repair cost for broken server.;

$C_s \equiv$ fixed setup cost per busy cycle;

$C_o \equiv$ fixed cost for the every replacement of repair times.

Utilizing the above cost elements and the corresponding performance measures obtained previously, the total expected cost function per unit time is given by

$$\begin{aligned} TC(\eta) &= C_h E[L] + C_i(\pi_{0,0} + \varphi_0(1,1)) + C_b\varphi_1(1,1) + C_r\varphi_2(1,1) \\ &\quad + C_s\frac{1}{E[C]} + C_o\eta\varphi_2(1,1). \end{aligned} \tag{7.3}$$

One may note that it would be a hard task to solve the cost minimization problem (7.3) by using analytic method because $TC(\eta)$ is highly non-linear and complex. Here, we use the parabolic method to find the optimum value η^*. The details about the parabolic method can be referred to [40]. According to the polynomial approximation theory, the unique optimum of the quadratic function agreeing with the objective function $g(x)$ at 3-point pattern $\left\{x^{(l)}, x^{(m)}, x^{(r)}\right\}$ occurs at

$$x^{(q)} = \frac{g(x^{(l)})\big((x^{(m)})^2-(x^{(r)})^2\big)+g(x^{(m)})\big((x^{(r)})^2-(x^{(l)})^2\big)+g(x^{(r)})\big((x^{(l)})^2-(x^{(m)})^2\big)}{2\big[g(x^{(l)})\big(x^{(m)}-x^{(r)}\big)+g(x^{(m)})\big(x^{(r)}-x^{(l)}\big)+g(x^{(r)})\big(x^{(l)}-x^{(m)}\big)\big]}. \tag{7.4}$$

The parabolic method uses this approximation to improve the current 3-point pattern by replacing one of its points with an approximate optimum $x^{(q)}$. For the purpose of clarity, the procedures of the parabolic method are described as follows.

Step 1 (Initialization). Choose a starting 3-point pattern $\{x^{(l)}, x^{(m)}, x^{(r)}\}$ along with a stopping tolerance $\varepsilon = 10^{-6}$, and initialize the iteration counter $i = 0$.

Step 2 (Stopping). If $\left|x^{(q)} - x^{(m)}\right| \le \varepsilon$, stop and report approximate optimum solution $x^{(m)}$.

Step 3 (Quadratic fit). Compute a quadratic fit optimum $x^{(q)}$ according to (7.3) and (7.4). If $x^{(q)} \le x^{(m)}$, go to Step 4 . If $x^{(q)} > x^{(m)}$, go to Step 5.

Step 4 (Left). If $g\left(x^{(m)}\right)$ is less than $g\left(x^{(q)}\right)$, then update $x^{(q)} \to x^{(l)}$. Otherwise, replace $x^{(m)} \to x^{(r)}$, $x^{(q)} \to x^{(m)}$. Either way, advance $i = i + 1$, and return to Step 2.

Step 5 (Right). If $g\left(x^{(m)}\right)$ is less than $g\left(x^{(q)}\right)$, then update $x^{(q)} \to x^{(r)}$. Otherwise, replace $x^{(m)} \to x^{(l)}$, $x^{(q)} \to x^{(m)}$. Either way, advance $i = i + 1$, and return to Step 2.

In the following numerical experiment, we employ the parabolic method to determine the optimum replacement probability η^*.

Example 1. We consider a practical problem concerning the flexible manufacturing system mentioned in Introduction (Section 1). It is assumed that the raw components arrive at the system according to a Bernoulli process with parameter λ. The processing times per component, the inter-retrial times and the repair times are geometrically distributed with parameter χ, $\bar{\alpha}$ and σ, respectively. The default values of the system parameters and the cost elements are taken as $\lambda = 0.2$, $\chi = 0.6$, $\sigma = 0.4$, $p = 0.35$, $q = 0.2$, $\alpha = 0.25$, $\beta = 0.7$, $C_h = \$35$, $C_i = \$20$, $C_b = \$55$, $C_r = \$15$, $C_s = \$10$ and $C_o = \$25$. The effect of η on the system operating cost is illustrated in Figure 9. From the information of Figure 9, we observe that there is an optimum value of η to minimize the system cost and we choose the initial 3-point pattern $\eta^{(l)} = 0.2$, $\eta^{(m)} = 0.25$ and $\eta^{(r)} = 0.3$. Applying the parabolic method as mentioned above with the stopping tolerance $\varepsilon = 10^{-6}$, after six iterations, one can see from Table 1 that the minimum expected operating cost per unit time converges to the solution $\eta^* = 0.259841$ with value \$102.899274.

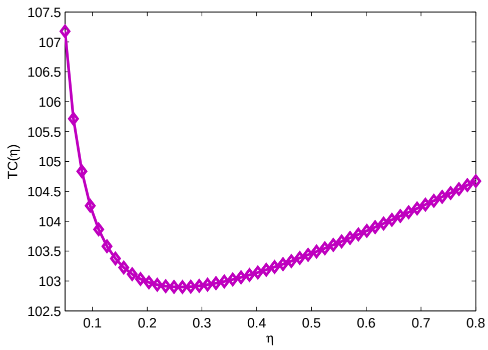

Figure 9. The effect of η on $TC(\eta)$.

Table 1. The parabolic method in searching for the optimum solution.

No. of iterations	$\eta^{(l)}$	$\eta^{(m)}$	$\eta^{(r)}$	$TC(\eta^{(l)})$	$TC(\eta^{(m)})$	$TC(\eta^{(r)})$	η^*	$TC(\eta^*)$	Tolerance
0	0.200000	0.250000	0.300000	102.986247	102.901156	102.925387	0.263918	102.899580	0.013918
1	0.250000	0.263918	0.300000	102.901156	102.899580	102.925387	0.260376	102.899279	0.003542
2	0.250000	0.260376	0.263918	102.901156	102.899279	102.899580	0.259925	102.899274	4.513242×10^{-4}
3	0.250000	0.259925	0.260376	102.901156	102.899274	102.899279	0.259853	102.899274	7.166558×10^{-5}
4	0.250000	0.259853	0.259925	102.901156	102.899274	102.899274	0.259843	102.899274	9.823264×10^{-6}
5	0.250000	0.259843	0.259853	102.901156	102.899274	102.899274	0.259842	102.899274	1.541255×10^{-6}
6	0.250000	0.259842	0.259843	102.901156	102.899274	102.899274	0.259841	102.899274	2.152343×10^{-7}

8. Conclusions

In the foregoing study, we considered a discrete-time $Geo/G/1$ retrial queue with probabilistic preemptive priority, balking customers, starting failures and replacements of repair times. The introduction of replacements in the repair times is an interesting novelty of this investigation. By employing the supplementary variable method and the generating function technique, the closed-form expressions for the probability generating functions of the stationary distributions of different system states, orbit size and system size were obtained. Some performance measures such as the probabilities that the server is free, busy period or down, the mean steady-state system queue length, the mean sojourn time were also derived. Furthermore, sensitivity analysis was carried out by some numerical examples, which can help the system managers better understand the operating characteristics of the system. Finally, we established a cost structure to search for optimal replacement probability for minimizing the system cost. The incorporation of retrial policy, preemptive resume priority, balking behavior, starting failures and replacements of repair times makes our system closer to real-life congestion scenarios and the analysis of this paper can provide potentially practical application in telecommunication systems, flexible manufacturing systems, transportation system, inventory problems, etc.

For future research, one could extend the present work by considering server vacations, multi-optional services, Markovian arrival process (MAP) of customers. It is also interesting to consider the case that the replacement time cannot be negligible.

Acknowledgments

The authors would like to sincerely thank two anonymous reviewers for their valuable comments and constructive suggestions which are very helpful in improving the presentation of this paper. This research is supported by the Scientific Research Project of Sichuan University of Science and Engineering (No. 2017RCL56), the Opening Project of Sichuan Province University Key Laboratory of Bridge Non-destruction Detecting and Engineering Computing (No. 2019QYJ03) , and the National Natural Science Foundation of China (No. 71571127).

Conflict of interest

The authors declare no conflicts of interest in this paper.

References

1. T. Yang, J. G. C. Templeton, *A survey on retrial queues*, Queueing Syst., **2** (1987), 201–233.
2. G. I. Falin, *A survey of retrial queues*, Queueing Syst., **7** (1990), 127–167.
3. J. R. Artalejo, *A classified bibliography of research on retrial queues: Progress in 1990–1999*, Top, **7** (1999), 187–211.
4. J. R. Artalejo, *Accessible bibliography on retrial queues: Progress in 2000–2009*, Math. Comput. Modell., **51** (2010), 1071–1081.
5. J. Kim, B. Kim, *A survey of retrial queueing systems*, Ann. Oper. Res., **247** (2016), 3–36.
6. G. I. Falin, J. G. C. Templeton, *Retrial Queues*, London: Chapman & Hall, 1997.
7. J. R. Artalejo, A. Gómez-Corral, *Retrial Queueing Systems: A Computational Approach*, Berlin: Springer, 2008.
8. T. Meisling, *Discrete-time queueing theory*, Oper. Res., **6** (1956), 96–105.
9. J. J. Hunter, *Mathematical Techniques of Applied Probability, Vol. 2, Discrete Time Models: Techniques and Applications*, New York: Academic Press, 1983.
10. H. Bruneel, B. G. Kim, *Discrete-Time Models for Communication Systems Including ATM*, Boston: Kluwer Academic Publishers, 1993.
11. M. E. Woodward, *Communication and Computer Networks: Modelling with Discrete-Time Queues*, California: IEEE Computer Society Press, 1994.

12. T. Yang, H. Li, *On the steady-state queue size distribution of the discrete-time $Geo/G/1$ queue with repeated customers*, Queueing Syst., **21** (1995), 199–215.

13. I. Atencia, P. Moreno, *A discrete-time $Geo/G/1$ retrial queue with general retrial times*, Queueing Syst., **48** (2004), 5–21.

14. J. R. Artalejo, I. Atencia, P. Moreno, *A discrete-time $Geo^{[X]}/G/1$ retrial queue with control of admission*, Appl. Math. Model., **29** (2005), 1100–1120.

15. A. K. Aboul-Hassan, S. I. Rabia, F. A. Taboly, *Performance evaluation of a discrete-time $Geo^{[X]}/G/1$ retrial queue with general retrial times*, Comput. Math. Appl., **58** (2009), 548–557.

16. I. Atencia, I. Fortes, S. Nishimura, et al. *A discrete-time retrial queueing system with recurrent customers*, Comput. Oper. Res., **37** (2010), 1167–1173.

17. J. Wang, *Discrete-time $Geo/G/1$ retrial queues with general retrial time and Bernoulli vacation*, J. Syst. Sci. Complex., **25** (2012), 504–513.

18. I. Atencia-Mc.Killop, J. L. Galán-García, G. Aguilera-Venegas, et al. *A $Geo^{[X]}/G^{[X]}/1$ retrial queueing system with removal work and total renewal discipline*, Appl. Math. Comput., **319** (2018), 245–253.

19. S. Gao, X. Wang, *Analysis of a single server retrial queue with server vacation and two waiting buffers based on ATM networks*, Math. Probl. Eng., **2019** (2019), 1–14.

20. I. Atencia, P. Moreno, *A discrete-time $Geo/G/1$ retrial queue with server breakdowns*, Asia Pac. J. Oper. Res., **23** (2006), 247–271.

21. I. Atencia, P. Moreno, *A discrete-time $Geo/G/1$ retrial queue with the server subject to starting failures*, Ann. Oper. Res., **141** (2006), 85–107.

22. J. Wang, Q. Zhao, *A discrete-time $Geo/G/1$ retrial queue with starting failures and second optional service*, Comput. Math. Appl., **53** (2007), 115–127.

23. I. Atencia, I. Fortes, S. Sánchez, *A discrete-time retrial queueing system with starting failures, Bernoulli feedback and general retrial times*, Comput. Ind. Eng., **57** (2009), 1291–1299.

24. J. Wang, P. Zhang, *A discrete-time retrial queue with negative customers and unreliable server*, Comput. Ind. Eng., **56** (2009), 1216–1222.

25. S. Gao, Z. Liu, *A repairable $Geo^X/G/1$ retrial queue with Bernoulli feedback and impatient customers*, Acta Math. Appl. Sin. E., **30** (2014), 205–222.

26. I. Atencia, *A discrete-time queueing system with server breakdowns and changes in the repair times*, Ann. Oper. Res., **235** (2015), 37–49.

27. I. Atencia, *A $Geo/G/1$ retrial queueing system with priority services*, Eur. J. Oper. Res., **256** (2017), 178–186.

28. S. Lan, Y. Tang, *Performance analysis of a discrete-time $Geo/G/1$ retrial queue with non-preemptive priority, working vacations and vacation interruption*, J. Ind. Manag. Optim., **15** (2019), 1421–1446.

29. H. Li, T. Yang, *$Geo/G/1$ discrete time retrial queue with Bernoulli schedule*, Eur. J. Oper. Res., **111** (1998), 629–649.

30. B. D. Choi, J. W. Kim , *Discrete-time* $Geo_1, Geo_2/G/1$ *retrial queueing systems with two types of calls*, Comput. Math. Appl., **33** (1997), 79–88.

31. J. Wu, Z. Liu, Y. Peng, *A discrete-time* $Geo/G/1$ *retrial queue with preemptive resume and collisions*, Appl. Math. Model., **35** (2011), 837–847.

32. M. Jain, A. Bhagat, C. Shekhar, *Double orbit finite retrial queues with priority customers and service interruptions*, Appl. Math. Comput., **253** (2015), 324–344.

33. S. Upadhyaya, *Performance prediction of a discrete-time batch arrival retrial queue with Bernoulli feedback*, Appl. Math. Comput., **283** (2016), 108–119.

34. F. A. Haight, *Queueing with balking*, Biometrika, **44** (1957), 360–369.

35. A. K. Aboul-Hassan, S. I. Rabia, A. Kadry, *Analytical study of a discrete-time retrial queue with balking customers and early arrival scheme*, Alex. Eng. J., **44** (2005), 911–917.

36. A. K. Aboul-Hassan, S. I. Rabia, F. A. Taboly, *A discrete time* $Geo/G/1$ *retrial queue with general retrial times and balking customers*, J. Korean Stat. Soc., **37** (2008), 335–348.

37. M. Lozano, P. Moreno, *A discrete time single-server queue with balking: economic applications*, Appl. Econ., **40** (2008), 735–748.

38. V. Laxmi, V. Goswami, K. Jyothsna, *Analysis of discrete-time single server queue with balking and multiple working vacations*, Qual. Technol. Quant. Manag., **10** (2013), 443–456.

39. A. G. Pakes, *Some conditions for ergodicity and recurrence of Markov chains*, Oper. Res., **17** (1969), 1058–1061.

40. R. L. Rardin, *Optimization in Operations Research*, New Jersey: Prentice Hall, 1998.

AIMS Press