Performance Analysis of a Discrete-Time Queue with Working Breakdowns and Searching for the Optimum Service Rate in Working Breakdown Period

Shaojun LAN
School of Mathematics and Statistics, Sichuan University of Science and Engineering, Zigong 643000, China; School of Mathematics and Software Science, Sichuan Normal University, Chengdu 610066, China
E-mail: sjlan89@163.com

Yinghui TANG
School of Fundamental Education, Sichuan Normal University, Chengdu 610066, China
E-mail: tangyh@uestc.edu.cn

Abstract This paper deals with a discrete-time Geo/Geo/1 queueing system with working breakdowns in which customers arrive at the system in variable input rates according to the states of the server. The server may be subject to breakdowns at random when it is in operation. As soon as the server fails, a repair process immediately begins. During the repair period, the defective server still provides service for the waiting customers at a lower service rate rather than completely stopping service. We analyze the stability condition for the considered system. Using the probability generating function technique, we obtain the probability generating function of the steady-state queue size distribution. Also, various important performance measures are derived explicitly. Furthermore, some numerical results are provided to carry out the sensitivity analysis so as to illustrate the effect of different parameters on the system performance measures. Finally, an operating cost function is formulated to model a computer system and the parabolic method is employed to numerically find the optimum service rate in working breakdown period.

Keywords discrete-time queue; working breakdowns; different arrival rates; performance measures; optimum service rate

1 Introduction

Since the digital computer and communication systems, such as broadband integrated services digital network (BISDN), time division multiple access (TDMA) and asynchronous transfer mode (ATM), operate on a discrete-time basis where the events (arrival of packets and their forward transmissions) can only take place at regularly spaced epochs, discrete-time queues are more suitable than their continuous-time equivalents for characterizing the behaviors of data communication and computer networks. In recent years, there has been an increasing interest in
investigating discrete-time queueing systems. The detailed analysis in the field of discrete-time queueing theory can be found in the monographs by Hunter[1], Bruneel and Kim[2], Takagi[3], Woodward[4] and Alfa[5].

In many real-life service systems such as computer and communication networks, flexible manufacturing systems, transportation systems and production systems, we often meet the situation that the service devices may fail more or less frequently when rendering service to jobs. The breakdowns of service facilities will result in a temporarily unavailable period of systems and therefore the performances of the systems will be heavily affected. In this context, the research of unreliable queueing system is well worth doing from the viewpoint of queueing and reliability theory. In the recent past, remarkable contributions regarding the unreliable queueing systems in discrete time regime have been made by many authors. Using supplementary variable method, Atencia and Moreno[6] studied a discrete-time Geo/G/1 retrial queue in which the server is unreliable. Tang, et al.[7] considered a batch arrival discrete-time queueing system with repairable server and multiple adaptive delayed vacations by total probability decomposition law. Wang and Zhang[8] investigated a Geo/G/1 retrial queue with negative customers and unreliable server. More discussion on discrete-time unreliable queues with various features can be found in Lin and Ke[9], Wang, et al.[10], Gao and Liu[11], Wang[12], Atencia[13], and references therein.

In the aforementioned papers, it is generally assumed that the failed server completely stops service during the breakdown period and the present customers in the system have to wait until the broken server is repaired. However, in many day-to-day congestion problems, we frequently encounter the situation where the defective server can still render service to the waiting customers at a slower rate instead of completely stopping service during the breakdown period. For instance, in flexible manufacturing systems, a product is processed by a system consisting of many parallel facilities. When some of these facilities are subject to failures, the system does not stop working and can still operates. But the production efficiency of the whole system will be degraded because of the breakdowns of some facilities. The failed facilities can be repaired so that the system resumes to normal working level. Another real-life example is the presence of a virus in computer system. When a computer is subject to the invasion of a virus, the CPU of the computer will not stop running completely and is still able to operate at a lower speed. Meanwhile, the antivirus software begins to repair the system until the virus is cleared and the system recovers to its normal state.

The above phenomena are referred to as working breakdown which is first proposed by Kalidass and Kasturi[14]. Kalidass and Kasturi[14] analyzed the M/M/1-type queue with working breakdowns and obtained various queueing characteristics in steady state. Later, several researchers addressed the related topic concerning working breakdowns. Li, et al.[15] presented an economic analysis of customer behaviors in Markovian queues with partial breakdowns. Kim and Lee[16] considered an M/G/1 queueing model with disasters and working breakdowns. Liu and Song[17] studied a batch arrival MX/M/1 queue with working breakdowns. Recently, Yang and Wu[18] investigated the transient behavior of a finite capacity M/M/1/N queueing system with working breakdowns and multiple vacations. Liou[19] discussed an M/M/1 queue with working breakdowns and impatient customers. It should be pointed out that the concept of
working breakdowns differs from the concept of working vacations which is first introduced by Servi and Finn\cite{20}. A working breakdown can occur at arbitrary time epoch at which the server is busy at normal service rate, no matter how many customers are waiting in the system. But a working vacation occurs only when the system becomes empty. The introduction of working breakdowns not only allows the defective server to deal with emergencies occurring during the repair period but also reduces the queue length of system, the system cost and the system congestion. Therefore, the incorporation of working breakdown service in unreliable queueing systems is more reasonable and significant.

In the queueing literature, a considerable number of queueing systems have been analyzed under the assumption that the arrival rates of jobs are fixed. However, in practice, the arrival rates of jobs may be changed along with the server's status. In most cases, if the server is available, the arrival rate is relatively high while if the server cannot provide normal service for customers (e.g., the server breaks down or leaves for vacation), some arriving customers may not join the queue and abandon the system without service, which clearly leads to the decrease of the arrival rate. In real-world situations, the kind of queueing systems in which the arrival rates of customers depend on the server's state can be found in hospitals, in banks, in telecommunication systems, and so forth. A few authors have investigated some queueing models with variable arrival rate, see, e.g., Hur and Paik\cite{21}, Sun, et al.\cite{22}, Wei, et al.\cite{23}, and Luo, et al.\cite{24}.

Although it is observed that some continuous-time queues with working breakdowns have been considered in the recent past, their discrete-time counterparts seem to receive very little attention in the literature. In addition, from the practical point of view, many common things in our life, especially the computer systems, are modeled by discrete-time behavior. Thus, inspired by the natural and reasonable applications of discrete-time queues, in the present study, we will develop an analytic model that allows us to consider a discrete-time Geo/Geo/1 queue with working breakdowns and variable input rates. To the best of our knowledge, there is no research work on the proposed model. We analyze the sufficient and necessary condition for the concerned system to be stable. Applying the probability generating function technique, we obtain the probability generating function for the number of customers in the system as well as various performance measures. Some numerical experiments are implemented to examine the effect of different system parameters on the characteristics of our queueing system. Additionally, we develop a long-run expected operating cost function per unit time and use the parabolic method to find the optimum value of service rate in the working breakdown period.

The remaining part of this paper is structured as follows. In Section 2, the mathematical description of the queueing model is given. Section 3 is devoted to discussing the stability condition of the our model. The probability generating functions of the queue size under different system states are derived in Section 4. Various performance characteristics of the queueing system are obtained in Section 5. In Section 6, some numerical results for sensitivity analysis of the crucial performance measures with respect to different system parameters are provided. Then, in Section 7, we establish a cost structure for a computer system and consider a cost minimization problem through parabolic method. At last, some conclusions and extensions are given in Section 8.
2 Model Formulation

Consider a discrete-time Geo/Geo/1 queueing system with working breakdown and different arrival rates. In discrete-time queues, all the queueing activities (arrivals, departures, breakdowns and repairs) are nonnegative integer-valued random variables. The time axis is divided into equal time intervals and is marked with $0, 1, \cdots, n, \cdots$. All the queueing events only happen at boundary epochs of time slots in discrete-time setting. In view of this fact, different events may take place simultaneously within a slot. Therefore, it is necessary to establish the order of these events in case of simultaneity. Generally speaking, there are two types of discrete-time models, namely, the early arrival system (EAS) and the late arrival system (LAS). And the late arrival system (LAS) can further be subdivided into late arrival system with delayed access (LAS-DA) and late arrival system with immediate access (LAS-IA). More discussion regarding these concepts can be referred to Hunter\cite{1}. In the present research, we consider the early arrival system (EAS), that is, the arrivals take place in time interval $(n, n^+)$, $n = 0, 1, \cdots$ and the departures occur in time interval $(n^-, n)$, $n = 1, 2, \cdots$. To make it clearer, the various time epochs at which queueing events occur are displayed in Figure 1. Throughout this paper, for any real number $x \in [0, 1]$ we denote $\bar{x} = 1 - x$. The detailed mathematical model is described as follows.

![Figure 1 Various time epochs in an early arrival system (EAS)](image)

The arrival rates of customers depend on the states of the server. That is, customers arrive at system in the normal state and in the working breakdown state according to Bernoulli process with parameters $\lambda_1$ and $\lambda_2$, respectively. From the practical point of view, it is assumed that $\lambda_2 < \lambda_1$. Customers are served based on the order of their arrivals, i.e., first-come first-served (FCFS) discipline. The service times in the normal state are geometrically distributed with parameter $\mu_1$. The server may be subject to working breakdowns only when the server is operating in the normal busy period. The time interval until the occurrence of working breakdown follows a geometric distribution with parameter $\alpha$. As soon as the working breakdown occurs, a repair process begins. The repair times follow geometric distribution with parameter $\beta$. During the working breakdown period, the server still renders service to the waiting customers and the service times are governed by a geometric distribution with parameter $\mu_2$ ($< \mu_1$). It is further supposed that when the repair is completed, the server functions as well as before breakdowns and immediately resumes to normal working level. The inter-arrival times, the service times during the normal period, the service times during the working breakdown period, the inter-breakdown times and the repair times are independent of each other.
3 The Stability Condition of the System

Let \( N(m) \) be the number of customers present in the system at time \( m^+ \) and \( J(m) \) be the state of the server at time \( m^+ \) with

\[
J(m) = \begin{cases} 
0, & \text{the server is in working breakdown period at time } m^+, \\
1, & \text{the server is in normal busy period at time } m^+.
\end{cases}
\]

Then, the two-dimensional stochastic process

\[
\{X_m, m \in \mathbb{N}\} = \{(N(m), J(m)), m \in \mathbb{N}\}
\]

is a discrete time Markov process with the state space \( \Omega = \{(i, j) : i \geq 0, j = 0, 1\} \). Let

\[
\xi_1 = (\bar{\lambda}_2 \bar{\mu}_2 + \lambda_2 \mu_2) \beta, \quad \xi_2 = (\bar{\lambda}_1 \bar{\mu}_1 + \lambda_1 \mu_1) \bar{\alpha}, \quad \xi_3 = (\bar{\lambda}_2 \bar{\mu}_2 + \lambda_2 \mu_2) \beta, \quad \text{and} \quad \xi_4 = (\bar{\lambda}_1 \bar{\mu}_1 + \lambda_1 \mu_1) \alpha.
\]

The state transition diagram of the system is shown in Figure 2.

![State transition diagram of the system](image)

Figure 2 State transition diagram of the system

Using the lexicographical order for the states, we can express the one-step transition probabilities of the Markov process \( \{X_m, m \in \mathbb{N}\} \) in the following matrix form, which is called transition probability matrix of the quasi-birth-and-death (QBD) process under consideration

\[
P = \begin{pmatrix}
A_{00} & A_{01} & B_{10} & A_1 & A_0 \\
B_{10} & A_1 & A_0 \\
A_2 & A_1 & A_0 \\
\vdots & \ddots & \ddots & \ddots & \ddots
\end{pmatrix},
\]

where

\[
A_{00} = \begin{pmatrix}
\bar{\lambda}_2 \bar{\beta} & \bar{\lambda}_2 \bar{\beta} \\
0 & \bar{\lambda}_1
\end{pmatrix}, \quad A_{01} = \begin{pmatrix}
\lambda_2 \bar{\beta} & \lambda_2 \beta \\
\lambda_1 \bar{\alpha} & \lambda_1 \bar{\alpha}
\end{pmatrix}, \quad B_{10} = \begin{pmatrix}
\bar{\lambda}_2 \mu_2 \bar{\beta} & \bar{\lambda}_2 \mu_2 \beta \\
\bar{\lambda}_1 \mu_1 \bar{\alpha} & \bar{\lambda}_1 \mu_1 \alpha
\end{pmatrix},
\]

\[
A_0 = \begin{pmatrix}
\lambda_2 \bar{\beta} & \lambda_2 \beta \\
\lambda_1 \bar{\alpha} & \lambda_1 \bar{\alpha}
\end{pmatrix}, \quad A_1 = \begin{pmatrix}
(\bar{\lambda}_2 \bar{\mu}_2 + \lambda_2 \mu_2) \bar{\beta} & (\bar{\lambda}_2 \bar{\mu}_2 + \lambda_2 \mu_2) \beta \\
(\bar{\lambda}_1 \bar{\mu}_1 + \lambda_1 \mu_1) \alpha & (\bar{\lambda}_1 \bar{\mu}_1 + \lambda_1 \mu_1) \bar{\alpha}
\end{pmatrix}.
\]
\[ A_2 = \begin{pmatrix} \lambda_2 \mu_2 \beta & \lambda_2 \mu_2 \beta \\ \lambda_1 \mu_1 \alpha & \lambda_1 \mu_1 \alpha \end{pmatrix}. \]

**Theorem 1** The QBD process \( \{X_m, m \in \mathbb{N}\} \) is ergodic (irreducible and positive recurrent) if and only if \( \rho \equiv \frac{\beta \lambda_1 + \alpha \lambda_2}{\beta \mu_1 + \alpha \mu_2} < 1 \).

**Proof** It is easy to see from transition probability matrix \( P \) that the QBD process \( \{X_m, m \in \mathbb{N}\} \) is irreducible and aperiodic. In order to prove the ergodicity condition, we only need to prove that Markov process \( \{X_m, m \in \mathbb{N}\} \) is positive recurrent if and only if \( \frac{\beta \lambda_1 + \alpha \lambda_2}{\beta \mu_1 + \alpha \mu_2} < 1 \). To this end, we define matrix \( A \) as

\[
A = A_0 + A_1 + A_2 = \begin{pmatrix} \beta & \beta \\ \alpha & \alpha \end{pmatrix}.
\]

Obviously, matrix \( A \) is irreducible. Thus, from Neuts\(^{[25]}\), the Markov process \( \{X_m, m \in \mathbb{N}\} \) is positive recurrent if and only if \( \theta A_0 e < \theta A_2 e \), where \( e \) denotes a column vector of ones, and \( \theta \) is the stationary probability vector of matrix \( A \), i.e., \( \theta A = \theta \) and \( \theta e = 1 \). After some algebraic manipulation, we have \( \beta \lambda_1 + \alpha \lambda_2 < \beta \mu_1 + \alpha \mu_2 \), i.e., the QBD process \( \{X_m, m \in \mathbb{N}\} \) is positive recurrent if and only if \( \frac{\beta \lambda_1 + \alpha \lambda_2}{\beta \mu_1 + \alpha \mu_2} < 1 \).

### 4 System Queue Size Distribution

In this section, we discuss the stationary probability distributions for \( \{X_m, m \in \mathbb{N}\} \). Let \( \pi = (\pi_0, \pi_1, \pi_2, \ldots) \) be the steady-state probability vector of \( P \), i.e., \( \pi P = \pi \) and \( \pi e = 1 \), where

\[
\pi_i = (\pi_{i,0}, \pi_{i,1}), \quad \pi_{i,j} = \lim_{m \to \infty} P \{N(m) = i, J(m) = j\}, \quad i \geq 0, \ j = 0, 1.
\]

Then the equilibrium equations for the stationary distributions of the system under study are given by

\[
\begin{align*}
\pi_{0,0} &= \lambda_0 \beta \pi_{0,0} + \lambda_2 \mu_2 \beta \pi_{1,0} + \lambda_1 \mu_1 \alpha \pi_{1,1}, \\
\pi_{i,0} &= (\lambda_2 \mu_2 + \lambda_2 \mu_2) \beta \pi_{i,0} + (\lambda_1 \mu_1 + \lambda_1 \mu_1) \alpha \pi_{i,1} + \lambda_2 \mu_2 \beta \pi_{i+1,0} \\
&\quad + \lambda_1 \mu_1 \alpha \pi_{i+1,1} + (1 - \delta_{i,1}) \lambda_2 \mu_2 \beta \pi_{i-1,0} + (1 - \delta_{i,1}) \lambda_1 \mu_1 \alpha \pi_{i-1,1} \\
&\quad + \delta_{i,1} \lambda_1 \alpha \pi_{0,0} + \delta_{i,1} \lambda_2 \beta \pi_{0,0}, \quad i \geq 1, \\
\pi_{0,1} &= \lambda_1 \pi_{0,1} + \lambda_2 \beta \pi_{0,0} + \lambda_2 \mu_2 \beta \pi_{1,0} + \lambda_1 \mu_1 \alpha \pi_{1,1}, \\
\pi_{i,1} &= (\lambda_2 \mu_2 + \lambda_2 \mu_2) \beta \pi_{i,0} + (\lambda_1 \mu_1 + \lambda_1 \mu_1) \alpha \pi_{i,1} + \lambda_2 \mu_2 \beta \pi_{i+1,0} \\
&\quad + \lambda_1 \mu_1 \alpha \pi_{i+1,1} + (1 - \delta_{i,1}) \lambda_2 \mu_2 \beta \pi_{i-1,0} + (1 - \delta_{i,1}) \lambda_1 \mu_1 \alpha \pi_{i-1,1} \\
&\quad + \delta_{i,1} \lambda_1 \alpha \pi_{0,0} + \delta_{i,1} \lambda_2 \beta \pi_{0,0}, \quad i \geq 1,
\end{align*}
\]

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
\[ \sum_{i=0}^{\infty} \pi_{i,0} + \sum_{i=0}^{\infty} \pi_{i,1} = 1. \]

In order to derive the solutions of (1)~(4), we introduce the following partial probability generating functions
\[ \varphi_0(z) = \sum_{i=0}^{\infty} z^i \pi_{i,0}, \quad \varphi_1(z) = \sum_{i=0}^{\infty} z^i \pi_{i,1}. \]

Multiplying Equations (1) and (2) by appropriate powers of \( z \) and summing over \( i \) from 0 to \( \infty \), it follows that
\begin{align*}
\varphi_0(z) &= \left( \hat{\lambda}_2 \hat{\mu}_2 + \lambda_2 \mu_2 + \lambda_2 \bar{\mu}_2 z + \frac{\lambda_2 \mu_2}{z} \right) \beta [\varphi_0(z) - \pi_{0,0}] \\
&\quad + \left( \hat{\lambda}_1 \bar{\mu}_1 + \lambda_1 \mu_1 + \lambda_1 \bar{\mu}_1 z + \frac{\lambda_1 \mu_1}{z} \right) \alpha [\varphi_1(z) - \pi_{0,1}] \\
&\quad + \lambda_1 \alpha z \pi_{0,1} + (\hat{\lambda}_2 + \lambda_2) \beta \pi_{0,0}.
\end{align*}

Multiplying Equations (3) and (4) by appropriate powers of \( z \) and summing over \( i \) from 0 to \( \infty \), we have
\begin{align*}
\varphi_1(z) &= \left( \hat{\lambda}_2 \hat{\mu}_2 + \lambda_2 \mu_2 + \lambda_2 \bar{\mu}_2 z + \frac{\lambda_2 \mu_2}{z} \right) \beta [\varphi_0(z) - \pi_{0,0}] \\
&\quad + \left( \hat{\lambda}_1 \bar{\mu}_1 + \lambda_1 \mu_1 + \lambda_1 \bar{\mu}_1 z + \frac{\lambda_1 \mu_1}{z} \right) \alpha [\varphi_1(z) - \pi_{0,1}] \\
&\quad + (\lambda_1 + \lambda_1 \bar{\alpha} z) \pi_{0,1} + (\hat{\lambda}_2 + \lambda_2) \beta \pi_{0,0}.
\end{align*}

Solving (5) and (6) for \( \varphi_0(z) \) and \( \varphi_1(z) \) yields
\begin{align*}
\varphi_0(z) &= \frac{\mu_2 f(z) \pi_{0,0}}{\Psi(z)} + \lambda_1 \alpha z \left( \bar{\lambda}_1 \mu_1 z - \bar{\lambda}_1 \mu_1 - z \right) \pi_{0,1}, \quad (7) \\
\varphi_1(z) &= \frac{-\mu_2 (\hat{\lambda}_2 + \lambda_2) \beta z \pi_{0,0} + t(z) \pi_{0,1}}{\Psi(z)}, \quad (8)
\end{align*}

where
\begin{align*}
f(z) &= (\hat{\lambda}_2 + \lambda_2 z) \left[ (\beta - \alpha) (z - 1) (\bar{\lambda}_1 \mu_1 - \bar{\lambda}_1 \mu_1) - \alpha z \right], \\
t(z) &= \left[ (1 - z) (\hat{\lambda}_2 \mu_2 - \lambda_2 \bar{\mu}_2 z) \beta - \beta z \right] \mu_1 (\bar{\lambda}_1 + \lambda_1 z) \\
&\quad + \alpha (\hat{\lambda}_2 \mu_2 - \lambda_2 \bar{\mu}_2 z) (z (\lambda_1 z - 1) + (z - 1) (\bar{\lambda}_1 \mu_1 - \bar{\lambda}_1 \mu_1)), \\
\Psi(z) &= (1 - z) (\bar{\lambda}_1 \mu_1 - \bar{\lambda}_1 \mu_1) (\hat{\lambda}_2 \mu_2 - \lambda_2 \bar{\mu}_2) (1 + \alpha + \beta) \\
&\quad - \alpha z (\hat{\lambda}_2 \mu_2 - \lambda_2 \bar{\mu}_2) - \beta z (\bar{\lambda}_1 \mu_1 - \bar{\lambda}_1 \mu_1).
\end{align*}

From the model assumptions \( \lambda_2 < \lambda_1, \mu_2 < \mu_1, \) and the stability condition \( \rho < 1 \), we can obtain \( \frac{\lambda_2 \mu_2}{\lambda_1 \mu_1} > 1 \), which gives \( \frac{\lambda_2 \mu_2}{\lambda_1 \mu_1} > 1 \). By some direct calculations, we have that
\[ \Psi(0) = \bar{\lambda}_1 \hat{\lambda}_2 \mu_1 \mu_2 (1 + \alpha + \beta) > 0, \]
The various queue size distributions for the different system states are established. Now some

Theorem 2 If \( \rho = \frac{\beta \lambda_1 + \alpha \lambda_2}{\beta \mu_1 + \alpha \mu_2} < 1 \), the stationary distribution of the Markov process \( \{X_m, m \in \mathbb{N}\} \) has the following generating function

\[
\varphi(z) = \varphi_0(z) + \varphi_1(z),
\]

where the expressions of \( \varphi_0(z) \) and \( \varphi_1(z) \) are given by (11) and (12), respectively.

5 Performance Measures

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

1) The probability that the server is in working breakdown period is given by

\[
p_0 = \lim_{z \to 1} \varphi_0(z) = \frac{\lambda_1 \alpha [\alpha h(z_0) + f(z_0)(\mu_1 + \lambda_1)]}{(\alpha + \beta) \lambda_1 \alpha h(z_0) + [\alpha \lambda_1 (\mu_2 - \lambda_2 + \lambda_1) + \mu_1 (\beta + \alpha \lambda_1)] f(z_0)}.
\]
2) The probability that the server is in normal busy period is given by
\[
p_1 = \lim_{z \to 1} \varphi_1(z) = \frac{\lambda_1 \alpha \beta h(z_0) + [\alpha \lambda_1 (\mu_2 - \lambda_2) + \beta \mu_1] f(z_0)}{(\alpha + \beta) \lambda_1 \alpha h(z_0) + [\alpha \lambda_1 (\mu_2 - \lambda_2 + \lambda_1) + \mu_1 (\beta + \alpha \lambda_1)] f(z_0)}.
\]

3) The probability that the server is free in working breakdown period is given by
\[
\pi_{0,0} = \frac{\lambda_1 \alpha h(z_0)}{\mu_2 \{\alpha + \beta \lambda_1 \alpha h(z_0) + [\alpha \lambda_1 (\mu_2 - \lambda_2 + \lambda_1) + \mu_1 (\beta + \alpha \lambda_1)] f(z_0)\}}.
\]

4) The probability that the server is free in normal busy period is given by
\[
\pi_{0,1} = \frac{f(z_0) [\alpha (\mu_2 - \lambda_2) + \beta (\mu_1 - \lambda_1)]}{(\alpha + \beta) \lambda_1 \alpha h(z_0) + [\alpha \lambda_1 (\mu_2 - \lambda_2 + \lambda_1) + \mu_1 (\beta + \alpha \lambda_1)] f(z_0)}.
\]

5) The probability that the system is empty is given by
\[
\pi_{0,0} + \pi_{0,1} = \frac{(\mu_2 f(z_0) + \lambda_1 \alpha h(z_0)) [\alpha (\mu_2 - \lambda_2) + \beta (\mu_1 - \lambda_1)]}{\mu_2 \{\alpha + \beta \lambda_1 \alpha h(z_0) + [\alpha \lambda_1 (\mu_2 - \lambda_2 + \lambda_1) + \mu_1 (\beta + \alpha \lambda_1)] f(z_0)\}}.
\]

6) The expected number of the customers in the system when the server is in working breakdown period, denoted by \( E[L_0] \), is given by
\[
E[L_0] = \frac{d}{dz} \varphi_0(z) |_{z=1}
= \frac{\lambda_1 \alpha}{\Delta} \left\{ [\alpha (1 + \lambda_2) + (\lambda_1 - \mu_1) (\beta - \bar{\alpha})] h(z_0) + [2 \bar{\lambda}_1 + \mu_1 (1 + \lambda_1)] f(z_0) \right.
+ \frac{[\alpha h(z_0) + (\mu_1 + \bar{\lambda}_1) f(z_0)] \Psi'(1)}{\alpha (\mu_2 - \lambda_2) + \beta (\mu_1 - \lambda_1)} \Bigg\},
\]
where
\[
\Delta = (\alpha + \beta) \lambda_1 \alpha h(z_0) + [\alpha \lambda_1 (\mu_2 - \lambda_2 + \lambda_1) + \mu_1 (\beta + \alpha \lambda_1)] f(z_0),
\]
\[
\Psi'(1) = (\mu_1 - \lambda_1) (\mu_2 - \lambda_2) (\beta - \bar{\alpha}) - \alpha \mu_2 (1 + \lambda_2) - \beta \mu_1 (1 + \lambda_1) + 2 (\alpha \lambda_2 + \beta \lambda_1).
\]

7) The expected number of the customers in the system when the server is in normal busy period, denoted by \( E[L_1] \), is given by
\[
E[L_1] = \frac{d}{dz} \varphi_1(z) |_{z=1}
= \frac{1}{\Delta} \left\{ \alpha \beta \lambda_1 (1 + \lambda_2) h(z_0) + t'(1) f(z_0)
+ \frac{[\alpha \beta \lambda_1 h(z_0) + (\beta \mu_1 + \alpha \lambda_1 (\mu_2 - \lambda_2)) f(z_0)] \Psi'(1)}{\alpha (\mu_2 - \lambda_2) + \beta (\mu_1 - \lambda_1)} \Bigg\},
\]
where
\[
t'(1) = \alpha \left[ \bar{\lambda}_1 \lambda_2 \bar{\mu}_2 - (\bar{\mu}_1 - \lambda_1) (\mu_2 - \lambda_2) \right] + \beta \mu_1 (\lambda_2 - \mu_2) - \beta \mu_1 (1 + \lambda_1).
\]
8) The expected number of the customers in the system, denoted by $E[L]$, is given by

$$E[L] = \frac{d}{dz} \phi(z) |_{z=1} = E[L_0] + E[L_1].$$

9) According to the well-known Little’s formula, the average sojourn time of an arbitrary customer in the system, denoted by $E[S]$, is given by

$$E[S] = \frac{E[L]}{\lambda_e},$$

where $\lambda_e = \lambda_2 p_0 + \lambda_1 p_1$ is the effective average arrival rate.

6 Numerical Experience and Sensitivity Analysis

In this section, some numerical examples are presented to qualitatively describe the behavior of the queueing system under consideration. To help system designer or manager better understand the system behavior, we examine the effect of the system parameters on some crucial performance measures of our model. All the numerical results are obtained by developing program in Matlab software. Of course, the values of the parameters are chosen so as to satisfy the stability condition $\rho < 1$.

Figure 3 depicts the impact of $\lambda_2$ on $\pi_{0,0}$ for various values of $\beta$. As intuitively expected, $\pi_{0,0}$, the probability that the server is free in working breakdown period, decreases with the increase of $\lambda_2$ for any $\beta$. Further, for a fixed arrival rate $\lambda_2$, it can be also observed that $\pi_{0,0}$ decreases with increasing values of $\beta$. This is due to the fact that as the repair rate $\beta$ increases, the defective system can be repaired in a shorter time and therefore the system enters normal state. The effect of $\lambda_1$ on $\pi_{0,1}$ for different values of $\alpha$ is shown in Figure 4. It is seen that $\pi_{0,1}$ decreases as $\lambda_1$ increases for fixed $\alpha$. Also, there is also a decreasing trend in $\pi_{0,1}$ with the growth of $\alpha$. This is because the larger the failure rate $\alpha$ is, the shorter the time interval of the system being normal state is, which is accordance with our expectation.

![Figure 3](image_url)

**Figure 3** The effect of $\lambda_2$ on $\pi_{0,0}$ for different values of $\beta$ ($\lambda_1 = 0.4, \mu_1 = 0.8, \mu_2 = 0.45, \alpha = 0.4$)
Figure 4 The effect of $\lambda_1$ on $\pi_{0,1}$ for different values of $\alpha$ ($\lambda_2 = 0.35$, $\mu_1 = 0.8$, $\mu_2 = 0.45$, $\beta = 0.25$)

Figure 5 The effect of $\lambda_2$ on $p_0$ for different values of $\beta$ ($\lambda_1 = 0.4$, $\mu_1 = 0.8$, $\mu_2 = 0.45$, $\alpha = 0.4$)

Figure 6 The effect of $\lambda_1$ on $p_1$ for different values of $\alpha$ ($\lambda_2 = 0.35$, $\mu_1 = 0.8$, $\mu_2 = 0.45$, $\beta = 0.25$)

In Figure 5, the probability $p_0$ of the server being in working breakdown period is plotted against $\lambda_2$ with various values of $\beta$. One can see from Figure 5 that $p_0$ increases with the increment of $\lambda_2$. However, it decreases with increase of $\beta$. The reason is that the larger the repair rate $\beta$ is, the greater the probability of the system switching to the normal working level is, which results in the decrease of $p_0$. This situation is coincident with the practical situation. Figure 6 examines the influence of $\lambda_1$ on $p_1$ for different values of $\alpha$. It is observed that $p_1$ decreases as a function of $\lambda_1$. This could be due to the fact that the increasing arrival rate $\lambda_1$ will
lead to the excessive accumulation of customers in the system, and the system operates under an overload condition, which makes the system be prone to failures and therefore the probability that the system is in normal state decreases. This observation suggests that system designers must pay attention to controlling arrival rate to avoid the congestion situation. Further, as $\alpha$ increases, $p_1$ also shows a decreasing trend, which matches with our intuition. In particular, when $\alpha = 0$, it is noted that the value of $p_1$ is always equal to 1. This is because for $\alpha = 0$, our queueing model reduces to the classical discrete-time Geo/Geo/1 queue without breakdowns, and the system is undoubtedly in normal working state with probability 1.

The effects of $\mu_2$ on $E[L]$ and $E[S]$ are plotted in Figures 7 and 8, respectively. As is expected, both $E[L]$ and $E[S]$ decrease with the increasing values of the service rate $\mu_2$. The similar trend is shown with the increase of $\beta$. This is because the mean repair time is becoming shorter with the increasing repair rate $\beta$, and therefore the waiting customers have a greater chance to be served by normal service rate, which can reduce the sojourn time of customers and the system queue length. Moreover, one can see that when $\mu_2$ approaches to $\mu_1 = 0.8$, the values of $E[L]$ and $E[S]$ become fixed numbers for any $\beta$. The reason is that when $\mu_2 = \mu_1$, i.e., the service rate during working breakdown period equals that during normal busy period, our model becomes the corresponding queue without breakdowns, no matter how long the repair times are.

![Figure 7](image1.png)

**Figure 7** The effect of $\mu_2$ on $E[L]$ for different values of $\beta$ ($\lambda_1 = 0.5$, $\lambda_2 = 0.25$, $\mu_1 = 0.8$, $\alpha = 0.25$)

![Figure 8](image2.png)

**Figure 8** The effect of $\mu_2$ on $E[S]$ for different values of $\beta$ ($\lambda_1 = 0.5$, $\lambda_2 = 0.25$, $\mu_1 = 0.8$, $\alpha = 0.25$)
Figures 9 and 10 describe the behaviors of $E[L]$ and $E[S]$ with the changes of $\mu_2$ and $\alpha$, respectively. Obviously, $E[L]$ and $E[S]$ also show a decreasing trend with the increment of $\mu_2$. But the effects get reversed when $\alpha$ varies from 0 to 0.4. This is due to the fact that as $\alpha$ increases, the system is more likely to break down and the service rate will decrease, which in turn increases the sojourn time of customers and the system queue size. Additionally, it should be noted that for $\alpha = 0$, $E[L]$ and $E[S]$ are insensitive to the change of $\mu_2$. The reason is that when $\alpha = 0$, our system becomes the classical discrete-time Geo/Geo/1 queue without failures. Similar to the case in Figures 7 and 8, when $\mu_2 = \mu_1$, $E[L]$ and $E[S]$ are not related to the failure rate $\alpha$ and achieve fixed values.

![Figure 9](image1.png)

**Figure 9** The effect of $\mu_2$ on $E[L]$ for different values of $\alpha$ ($\lambda_1 = 0.3, \lambda_2 = 0.15, \mu_1 = 0.8, \beta = 0.15$)

![Figure 10](image2.png)

**Figure 10** The effect of $\mu_2$ on $E[S]$ for different values of $\alpha$ ($\lambda_1 = 0.3, \lambda_2 = 0.15, \mu_1 = 0.8, \beta = 0.15$)

The above numerical analysis not only demonstrates the validity of our analytical results, but also can provide insight to the concerned system designers and decision makers so as to reduce the congestion problem encountered in computer systems and communication network.

7 The Optimum Service Rate in a Working Breakdown Period

In practice, the operating cost of system is closely related to the system benefit. Therefore, from the perspective of economic profit, system designers or system managers are interested in minimizing operating cost of unit time. Thus, in order to demonstrate the applicability of
the results given in the previous discussion, in this section, we consider a practical problem concerning computer system. Data packets arrive at the computer system according to a Bernoulli process. The arrival rates of data packets depend on the states of the computer system. When the computer system is operating in the normal state, the arrival rate of data packets is \(\lambda_1\) and the processing time (service time) for each data packet is geometrically distributed with parameter \(\mu_1\). The computer system may be subject to the invasion of a virus during the normal operation period. It is assumed that the time interval until the presence of virus follows a geometric distribution with parameter \(\alpha\). After the computer system is invaded by a virus, the CPU of the computer system will not stop running completely and is still able to work at a lower speed. Under such circumstances, the arrival rate of data packets is \(\lambda_2\) and the processing time for each data packet is governed by a geometric distribution with parameter \(\mu_2\). Meanwhile, the antivirus software begins to repair the system until the virus is cleared and the system recovers to its normal working state. The repair times follow a geometric distribution with parameter \(\beta\).

Thus, this computer system can be modeled by the discrete-time Geo/Geo/1 queueing system with working breakdown and different arrival rates investigated in this paper. In order to realize precise control and make the system profitable, the system designers or managers consider the following cost elements.

- \(C_h\) \equiv \text{unit time cost of every customer present in the system};
- \(C_{p_0}\) \equiv \text{unit time cost for keeping the server in working breakdown period};
- \(C_{p_1}\) \equiv \text{unit time cost for keeping the server in the normal busy period};
- \(C_{\mu_1}\) \equiv \text{fixed service cost per unit time during the normal busy period};
- \(C_{\mu_2}\) \equiv \text{fixed service cost per unit time during working breakdown period};
- \(C_{\beta}\) \equiv \text{fixed repair cost per unit time for broken server}.

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

\[
TC(\mu_2) = C_h E[L] + C_{p_0} p_0 + C_{p_1} p_1 + C_{\mu_1} \mu_1 + C_{\mu_2} \mu_2 + C_{\beta} \beta,
\]

where \(p_0, p_1,\) and \(E[L]\) are given by 1), 2), and 8) of Section 5, respectively.

In the above cost function, the service rate \(\mu_2\) in working breakdown period is a decision variable. Our objective is to determine the optimum value of service rate in working breakdown period, say \(\mu_2^*\), so as to minimize the expected operating cost function per unit time. The optimum service rate \(\mu_2^*\) can provide some insight for system designers and decision makers so as to help them model real time system.

One may note that it would be a hard task to solve the cost minimization problem (15) by using analytic method because \(TC(\mu_2)\) is highly non-linear and complex. Here, we use the parabolic method to find the optimum value \(\mu_2^*\). The details about the parabolic method can be referred to Ronald[26]. According to the polynomial approximation theory, the unique optimum of the quadratic function agreeing with the objective function \(g(x)\) at 3-point pattern \(\{x^{(l)}, x^{(m)}, x^{(r)}\}\) occurs at
\[
x^{(q)} = \frac{g(x^{(l)})(x^{(m)})^2 - (x^{(r)})^2 + g(x^{(m)})(x^{(r)})^2 - (x^{(l)})^2 + g(x^{(l)})(x^{(l)})^2 - (x^{(m)})^2}{2[g(x^{(l)})(x^{(m)} - x^{(r)}) + g(x^{(m)})(x^{(r)} - x^{(l)}) + g(x^{(l)})(x^{(l)} - x^{(m)})]}. \quad (16)
\]

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 steps 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 \(|x^{(q)} - x^{(m)}| \leq \varepsilon\), stop and report approximate optimum solution \(x^{(m)}\).

**Step 3 (Quadratic fit)** Compute a quadratic fit optimum \(x^{(q)}\) according to the formulas \((15)\) and \((16)\). If \(x^{(q)} \leq x^{(m)}\), go to Step 4. If \(x^{(q)} > x^{(m)}\), go to Step 5.

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

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

In the following numerical example, we apply the procedure of the parabolic method to search for the optimum service rate in working breakdown period \(\mu_2\). Assume that the values of the system parameters and the cost elements are taken as \(\lambda_1 = 0.37, \lambda_2 = 0.26, \mu_1 = 0.8, \beta = 0.38, \alpha = 0.4, C_b = 225, C_{p_0} = 50, C_{p_1} = 220, C_{\mu_1} = 800, C_{\mu_2} = 600\), and \(C_{\beta} = 100\). The effect of service rate during working breakdown period \(\mu_2\) on the system operating cost is illustrated in Figure 11. From the information of Figure 11, we observe that there is an optimum value of \(\mu_2\) to minimize the system cost and we choose the initial 3-point pattern \(\mu_2^{(l)} = 0.2, \mu_2^{(m)} = 0.25\) and \(\mu_2^{(r)} = 0.3\). Applying the parabolic method as mentioned above with the stopping tolerance \(\varepsilon = 10^{-6}\), after five iterations, one can see from Table 1 that the minimum expected operating cost per unit time converges to the solution \(\mu_2^* = 0.264572\) with value 1204.618908.

![Figure 11](image_url) The effect of \(\mu_2\) on the expected operating cost per unit time
8 Concluding Remarks

In this paper, we have carried out an analysis of discrete-time Geo/Geo/1 queue with working breakdowns and variable input rates. We derived the sufficient and necessary condition for the considered system to be stable. By employing probability generating function technique, the probability generating function for the number of customers in the system was found. Various important performance measures such as the probabilities that the server is in working breakdown period, the server is in normal busy period and the server is free, and average queue length were obtained in explicit form. Furthermore, some numerical results were presented to discuss the effect of some key parameters on the characteristics of the model. Finally, we applied the parabolic method to search for the optimum service rate in working breakdown period under a given cost structure. This queueing system can be found in many practical situations such as telecommunication systems, flexible manufacturing system, and machine replace problem. Our queueing model has the capability of dealing with emergencies which may occur during the repair period.

| Table 1 | The parabolic method in searching for the optimum solution |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                | No. of iterations |                |                |                |                |
|                | 0                | 1               | 2               | 3               | 4               | 5               |
| $\mu_2^{(l)}$  | 0.200000         | 0.250000        | 0.250000        | 0.250000        | 0.250000        | 0.250000        |
| $\mu_2^{(m)}$  | 0.250000         | 0.266730        | 0.265146        | 0.264618        | 0.264582        | 0.264573        |
| $\mu_2^{(r)}$  | 0.300000         | 0.300000        | 0.266730        | 0.265146        | 0.264618        | 0.264573        |
| $TC\left(\mu_2^{(l)}\right)$ | 1211.959366        | 1204.945793      | 1204.945793      | 1204.945793      | 1204.945793      | 1204.945793      |
| $TC\left(\mu_2^{(m)}\right)$ | 1204.945793        | 1204.619397      | 1204.619397      | 1204.618908      | 1204.618908      | 1204.618908      |
| $TC\left(\mu_2^{(r)}\right)$ | 1206.335718        | 1206.335718      | 1204.619397      | 1204.618908      | 1204.618908      | 1204.618908      |
| $\mu_2^{*}$    | 0.266730          | 0.265146         | 0.264618         | 0.264582         | 0.264573         | 0.264572         |
| $TC\left(\mu_2^{*}\right)$ | 1204.625792        | 1204.619397      | 1204.618908      | 1204.618908      | 1204.618908      | 1204.618908      |
| Tolerance      | 0.016730          | 0.001585         | 5.272443×10^{-4} | 3.585264×10^{-5} | 9.906920×10^{-6} | 8.069918×10^{-7} |

For further research, one can extend this model by incorporating more complex scenarios like Markovian arrival process (MAP) of customers, general service times, impatient customers, and so on.

References