

Analysis and Constraint Cost Optimization for Discrete-time Queue with Bi-Level Randomized (p, N_1, N_2) -Policy

Xinyu KUANG*

School of Mathematics and Statistics, Sichuan University of Science and Engineering, Zigong 643000, China; Key Laboratory of Higher Education of Sichuan Province for Enterprise Informationalization and Internet of Things, Zigong 643000, China

E-mail: xyukuang@163.com

Yinghui TANG

School of Mathematical Sciences, Sichuan Normal University, Chengdu 610068, China

E-mail: tangyh@sicnu.edu.cn

Shaojun LAN

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

E-mail: sjlan89@163.com

Abstract This paper proposes a new discrete-time $Geo/G/1$ queueing model under the control of bi-level randomized (p, N_1, N_2) -policy. That is, the server is closed down immediately when the system is empty. If $N_1 (\geq 1)$ customers are accumulated in the queue, the server is activated for service with probability p ($0 \leq p \leq 1$) or still left off with probability $(1 - p)$. When the number of customers in the system becomes $N_2 (\geq N_1)$, the server begins serving the waiting customers until the system becomes empty again. For the model, firstly, we obtain the transient solution of the queue size distribution and the explicit recursive formulas of the stationary queue length distribution by employing the total probability decomposition technique. Then, the expressions of its probability generating function of the steady-state queue size and the expected steady-state queue size are presented. Additionally, numerical examples are conducted to discuss the effect of the system parameters on some performance indices. Furthermore, the steady-state distribution of queue length at epochs n^- , n and outside observer's observation epoch are explored, respectively. Finally, we establish a cost function to investigate the cost optimization problem under the constraint of the average waiting time. And the presented model provides a less expected cost as compared to the traditional N -policy.

Received June 12, 2024, accepted March 12, 2025

Supported by the National Natural Science Foundation of China (71571127); the Opening Fund of Key Laboratory of Higher Education of Sichuan Province for Enterprise Informationalization and Internet of Things (2023WZJ02)

*Corresponding author

Keywords discrete-time $Geo/G/1$ queue; (p, N_1, N_2) -policy; the total probability decomposition technique; queue length distribution; cost optimization

1 Introduction

During the last few decades, there has been increasing interest in discrete-time queueing models and operating policies, particularly in production/inventory systems and computer science and communication systems. Therefore, since Meisling^[1] studied a single-server discrete-time queueing model, the analysis of discrete-time queues has seen substantial development in recent years^[2–14]. For instance, Zhang and Tian^[2] considered a discrete-time $Geo/G/1$ queue with multiple adaptive vacations and applied the stochastic decomposition property of the steady-state queue size to obtain the probability generating function of the steady-state queue length. Yu et al.^[6] presented a finite buffer size discrete-time multiple working vacation queue with changeover time, employing the supplementary variable and embedded Markov chain techniques. Lan and Tang^[10] proposed 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. Fiems and Turck^[11] considered a discrete-time single server queueing system, where arrivals stem from a multi-type Galton-Watson branching process with migration.

To improve the system's operating economic benefit and reduce the large setup cost, it is necessary to introduce a threshold policy in practical queueing systems to control the frequency of startup, where N -policy is a typical threshold policy. Until now, an enormous number of works on N -policy have been published^[15–24]. For example, Luo et al.^[17] considered a discrete-time queue $Geo/G/1$ with N -policy and LAS-DA (late arrival system with delayed access) discipline, by using renewal process theory and probability decomposition techniques. Unlike Luo et al.^[17], Lim^[19] investigated a discrete-time single server N -policy $GI/Geo/1$ queueing system by taking into account the stochastic nature of customer arrival times and using an embedded Markov chain and a trial solution approach. Liu et al.^[22] studied a discrete-time $Geo/G/1$ queue controlled by the combination of the N and D policies (called ND -policy). In this system, when there are N waiting customers or the service time backlog of all waiting customers exceeds a given threshold D , whichever emerges first, the idle server immediately resumes its service. Wang^[24] examined a new ND policy in the discrete-time $Geo/G/1$ queue, unlike Liu et al.^[22] under this ND policy, the idle server restarts its service when the N and D policies are simultaneously satisfied.

It is widely assumed in the articles above about the N -policy queue that the idle server must start providing service as soon as N customers enter the system. However, in some circumstances (for example, the extra work that he/she is engaged in cannot be interrupted immediately), even though the total number of customers has increased to N , the idle server may still be unable to serve customers. Therefore, based on this actual situation, in discrete-time queues, Wang and Ke^[25] first introduced (p, N) -policy into the discrete-time $Geo/G/1$ queueing system. Xu et

al.^[26] discussed a *Geo/Geo/1* queueing model with multiple vacations, negative customers, and the (p, N) -policy set-up time. Later, based on Xu et al.^[26], Jia and Chen^[27] further studied the *Geo/G/1* queueing model with (p, N) -policy, set-up time, multiple vacations and disasters. Lan and Tang^[28] investigated a discrete-time *Geo/G/1* repairable queueing system with Bernoulli feedback and randomized (p, N) -policy. Recently, Kuang et al.^[29] proposed the concept of bi-level randomized (p, N_1, N_2) -policy in a continuous-time *M/G/1* queueing model. The bi-level randomized (p, N_1, N_2) -policy means that whenever the system becomes empty, the server is closed down immediately when the number of customers arriving in the system reaches a given low threshold value $N_1 (\geq 1)$, the server is activated for work with probability $p (0 \leq p \leq 1)$ or still left off with probability $(1 - p)$. If the number of customers arriving in the system reaches a given high threshold value $N_2 (\geq N_1)$, the server begins serving the waiting customers until the system becomes empty again. The advantage of this policy is that the high and low thresholds N_1 and N_2 are used to limit the opening behavior of the system, which not only improves the system's energy efficiency but also enhances the flexibility of server operations. Building on the work of Kuang et al.^[29], Kumar^[30] examined the randomness of the customer arrival process and studied the *GI/M/1* queueing model with a bi-level randomized (p, N_1, N_2) -policy using the supplementary variable method and difference equations. These literature studies are all conducted in a continuous-time environment. However, analytic solutions for a discrete-time queue under this policy remain unexplored, which serves as the motivation for this study.

In this work, we introduce the bi-level randomized (p, N_1, N_2) -policy into a discrete-time *Geo/G/1* queue. This model can be used in the information processing system, in manufacturing systems, computer and telecommunication systems, etc. For example, in an information processing system, workers may close the information channel after processing incoming data and allocate storage space for newly received information while the system is idle. To prevent information congestion and frequent system restarts, the system operates as follows: when the number of received messages reaches the lower threshold $N_1 (\geq 1)$, the information channel is opened with probability p to transmit data or remains idle. If the queue length reaches the upper threshold $N_2 (\geq N_1)$, the system must activate the information channel to prevent data overflow.

First, we discuss the transient probability distribution of the queue size at arbitrary time epoch n^+ starting from an arbitrary initial state. Then, the recursive expressions of the steady-state distribution for the queue size at arbitrary time epoch n^+ are derived using L'Hospital rule. Furthermore, the steady-state distributions of the queue size at epochs n^- , n and outside observer's observation epoch are explored, respectively. Also, the important relationships between the steady-state queue length distributions at different epochs n^- , n , n^+ and outside observer's observation epoch are derived. To analyze system performance, numerical examples are presented to analyze the impact of system parameters on key performance measures, such as the probability of an empty system and the mean queue size $\bar{L}^+$. Finally, we construct an operating cost function to discuss the cost optimization problem under the constraint of the

average waiting time.

The following is a summary of the paper's main contributions.

1) To the authors' knowledge, that is the first study in a discrete-time environment the bi-level randomized (p, N_1, N_2) -policy.

2) The analysis method employed in this paper differs from some traditional analysis methods (e.g., embedded Markov chain and supplementary variable techniques). Using the approach described in our paper, we can study the transient queue size distribution at any epoch n^+ of the system, starting from any initial state. Based on the transient results, we can easily obtain the recursive formula of the steady-state queue length distribution that can be used to calculate accurately the probabilities of queue length.

3) We explore the behavior of the system as it relates to changes in the parameter values using numerical examples. In addition, we establish a cost function, study the cost optimization problem under the average waiting time constraint to obtain the optimum threshold values (N_1^*, N_2^*) , and the model provides a lower expected cost compared to the traditional N -policy.

The rest of this work is structured as follows. Section 2 presents the assumptions of the model considered in this paper. Some significant system performance measures are discussed in Section 3. Section 4 investigates the sensitivity of system parameters on several performance characteristics. In Section 5, we construct the cost function to numerically find the optimal values of N_1 and N_2 for minimizing the system cost. Finally, the conclusions are offered in Section 6.

2 Model Formulation and Some Preliminaries

In a discrete-time queue, all the queueing activities (arrivals, services) are nonnegative integer-valued random variables. The time axis is divided into equal time intervals and is marked with $0, 1, \dots, n, \dots$. 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. In the present research, we consider the queueing system is a late arrival system with delayed access (LAS-DA)^[31], that is, the arrival takes place within (n^-, n) , $n = 0, 1, 2, \dots$ and the departure only occurs within (n, n^+) , even if a customer arrives at the system in the interval (n^-, n) and the service station is empty at this moment, its service will also be postponed to the interval (n, n^+) . To make it clear, the various time epochs at which queueing activities occur are displayed in Figure 1.

This paper considers a discrete-time $Geo/G/1$ queueing model with the bi-level randomized (p, N_1, N_2) -policy. The detailed assumptions of the model are given as follows.

1) The inter-arrival times τ_k , $k = 1, 2, \dots$ are independent and identically distributed (i.i.d.) discrete-time random variables with the same geometrical distribution $P\{\tau_k = j\} = \lambda(1 - \lambda)^{j-1}$, $j = 1, 2, \dots$, $0 < \lambda < 1$. That is to say, a customer arrives with probability λ , and no customer arrives with probability $(1 - \lambda)$ in every time interval (n^-, n) , $n = 0, 1, 2, \dots$.

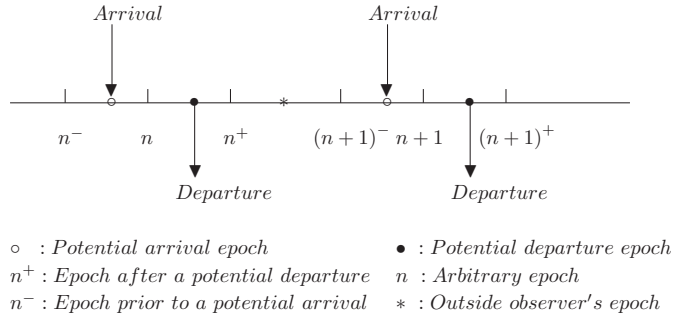


Figure 1 Various time epochs in a late arrival system with delayed access (LAS-DA)

2) There is only one server in the system with infinite capacity, and customers are served on a first-come, first-served basis. The service times $\{\chi_k, k = 1, 2, \dots\}$ are i.i.d. discrete-time random variables and arbitrarily distributed with the probability mass function (p.m.f.) $P\{\chi_k = j\} = g_j$, $j = 1, 2, \dots$, and the probability generating function (p.g.f.) $G(z) = \sum_{j=1}^{\infty} g_j z^j$, $|z| < 1$. Both the mean service time μ and the second moment are finite.

3) The queueing system operates under the bi-level randomized (p, N_1, N_2) -policy, that is, whenever the system is empty, the server keeps dormant in the system. If the number of customers reaches N_1 (≥ 1) in the system, the deactivated server is turned on with probability p ($0 \leq p \leq 1$) or still left off with complementary probability $(1 - p)$. If the number of customers reaches N_2 ($\geq N_1$) in the system, the service starts to serve customers immediately. Furthermore, once the server is activated, it will keep rendering service until the system becomes empty (see Figure 2).

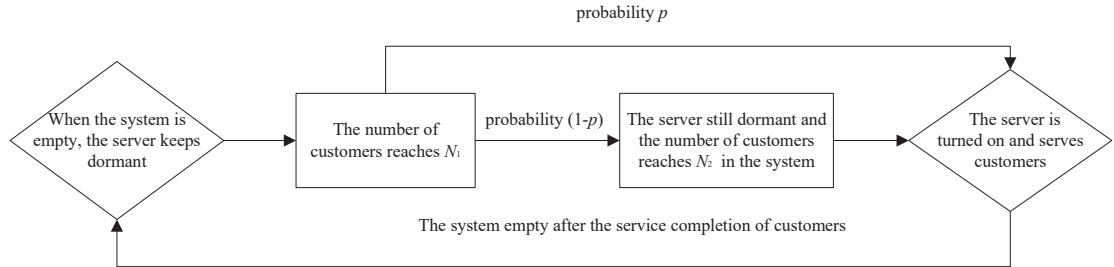


Figure 2 Diagrammatic representation of the bi-level randomized (p, N_1, N_2) -policy

In addition, we assume that the inter-arrival time τ and the service time χ are mutually independent of each other. It is also supposed that the server stays standby and waits for the first arrival if there is no customer at time $n^+ = 0^+$ (this hypothesis makes our model more realistic in real-world situations).

For later discussions, we first present some definitions and lemmas as follows.

Definition 1 (System idle period) The time interval from the system becomes empty until the first customer arrives. Let $\hat{\tau}_k$ ($k = 1, 2, \dots$) be the k th system idle period. Thus, $\hat{\tau}_k$ ($k = 1, 2, \dots$) are independent mutually and satisfy the same geometric distribution as τ (i.e.,

$P\{\hat{\tau}_k = j\} = \lambda(1 - \lambda)^{j-1}$, $j \geq 1$, $0 < \lambda < 1$) because of the Markov property of geometric distribution.

Definition 2 (Server's idle period) The time period from the moment that the system just becomes empty until the moment when server begins to serve the waiting customers.

Definition 3 (Server's busy period) The time interval from the server begins service until the system becomes empty again.

Definition 4 (busy cycle) It is the time interval from the moment that the server's idle period just begins until the moment that the next server's busy period ends.

Let b denote the length of the server's busy period initiated with only one customer. Then, we have the following lemma.

Lemma 1 For $|z| < 1$, let $B(z) = \sum_{j=1}^{\infty} P\{b = j\}z^j$ be the p.g.f. of b , then $B(z)$ is the unique solution of equation $u = G(z(\bar{\lambda} + \lambda u))$ within $|u| < 1$, and the mean value is given by

$$E[b] = \begin{cases} \frac{\rho}{\lambda(1-\rho)}, & \rho < 1, \\ \infty, & \rho \geq 1, \end{cases}$$

where $\bar{\lambda} = 1 - \lambda$, and $\rho = \lambda\mu$ represents the traffic intensity of the model under consideration.

Proof See Appendix A. ■

Let $b^{<i>}$ denote the length of the server's busy period starting from i customers, because of the "memoryless" property of geometric distribution, then $\sum_{j=i}^{\infty} P\{b^{<i>} = j\}z^j = B^i(z)$, $i = 1, 2, \dots$, $|z| < 1$.

Let $Q_j(n^+) = P\{b > n^+ \geq 0; N(n^+) = j\}$ denote the transient probability of there being j customers at epoch n^+ in the busy period b , we have the boundary condition $Q_1(0^+) = 1$, $Q_j(0^+) = 0$, $j = 2, 3, \dots$. We have the following lemma.

Lemma 2 For $|z| < 1$, let $q_j^+(z) = \sum_{n=0}^{\infty} Q_j(n^+)z^n$ be the z -transform of $Q_j(n^+)$, the recursive formulas of $q_j^+(z)$, $j \geq 1$, can be expressed as

$$q_j^+(z) = \frac{1}{G(z\bar{\lambda})} \left\{ B(z) \sum_{n=j-i}^{\infty} z^n \sum_{k=n+1}^{\infty} g_k \binom{n}{j-1} \lambda^{j-1} (\bar{\lambda})^{n-j+1} + \sum_{i=1}^{j-1} \frac{q_{j-i}^+(z)}{B^i(z)} \left[B(z) - G(z\bar{\lambda}) - \sum_{m=1}^i \sum_{k=m}^{\infty} g_k z^k \binom{k}{m} (\lambda B(z))^m (\bar{\lambda})^{k-m} \right] \right\}, \quad j \geq 1,$$

where $B(z)$ is defined by Lemma 2.1, $G(z\bar{\lambda}) = \sum_{j=1}^{\infty} g_j (z\bar{\lambda})^j$, $\sum_{k=i}^j = 0$ if $j < i$, $\binom{k}{m} = \frac{k!}{m!(k-m)!}$, $k \geq m \geq 1$, $\binom{k}{0} = 1$, and $\binom{k}{m} = 0$ if $k < m$.

Proof The proof is similar to that of Theorem 1 in [18]. ■

3 Analysis for Performance Measures

3.1 Analysis of Queue Length

In this subsection, by the total probability decomposition law, the renewal theory and the z -transformation, we first derive the z -transformation of the transient queue length distribution at any epoch n^+ . Then, on the basis of the transient analysis, the explicit formulas for the steady-state queue size at various time epochs are obtained.

3.1.1 The Transient Distribution of the Queue Size at Any Epoch n^+

Let $p_{ij}(n^+) = P\{N(n^+) = j | N(0^+) = i\}$ be the condition probability that there are j customers at epoch n^+ under arbitrary initial state $N(0^+) = i$ ($i = 0, 1, \dots$). The PGF of $p_{ij}^+(z) = \sum_{n=0}^{\infty} p_{ij}(n^+)z^n$, $i, j = 0, 1, \dots$. The expressions of $p_{ij}^+(z)$ will be derived in the following Theorems 1–2.

Theorem 1 For $|z| < 1$, $i \geq 1$, we have

$$p_{00}^+(z) = \frac{1}{1 - z\bar{\lambda}} \left\{ 1 + \frac{f(z)B(z)}{\Delta(z)} \right\}, \quad (1)$$

$$p_{i0}^+(z) = \frac{1}{1 - z\bar{\lambda}} \cdot \frac{B^i(z)}{\Delta(z)}, \quad (2)$$

where $B(z)$ is defined by Lemma 2.1, $f(z) = \frac{z\lambda}{1-z\bar{\lambda}}$, $\Delta(z) = 1 - pf^{N_1}(z)B^{N_1}(z) - (1-p)f^{N_2}(z)B^{N_2}(z)$.

Proof See Appendix B. ■

Theorem 2 For $|z| < 1$, $i \geq 1$:

1) If $j = 1, 2, \dots, N_1 - 1$, then

$$p_{0j}^+(z) = q_j^+(z)f(z) + \frac{f(z)}{\Delta(z)} \left[\frac{f^{j+1}(z)B(z)}{1 - z\bar{\lambda}} + \delta_j(z) \right], \quad (3)$$

$$p_{ij}^+(z) = \sum_{k=1}^i q_{j-i+k}^+(z)B^{k-1}(z) + \frac{B^{i-1}(z)}{\Delta(z)} \left[\frac{f^j(z)B(z)}{1 - z\bar{\lambda}} + \delta_j(z) \right]; \quad (4)$$

2) If $j = N_1, \dots, N_2 - 1$, then

$$p_{0j}^+(z) = q_j^+(z)f(z) + \frac{f(z)}{\Delta(z)} \left[\frac{(1-p)f^{j+1}(z)B(z)}{1 - z\bar{\lambda}} + \delta_j(z) \right], \quad (5)$$

$$p_{ij}^+(z) = \sum_{k=1}^i q_{j-i+k}^+(z)B^{k-1}(z) + \frac{B^{i-1}(z)}{\Delta(z)} \left[\frac{(1-p)f^j(z)B(z)}{1 - z\bar{\lambda}} + \delta_j(z) \right]; \quad (6)$$

3) If $j = N_2, N_2 + 1, N_2 + 2, \dots$, then

$$p_{0j}^+(z) = q_j^+(z)f(z) + \frac{f(z)\delta_j(z)}{\Delta(z)}, \quad (7)$$

$$p_{ij}^+(z) = \sum_{k=1}^i q_{j-i+k}^+(z)B^{k-1}(z) + \frac{B^{i-1}(z)\delta_j(z)}{\Delta(z)}; \quad (8)$$

where $B(z)$ is defined by Lemma 1, $q_j^+(z)$ is defined by Lemma 2, $\Delta(z)$ and $f(z)$ are defined by Theorem 1, $\delta_j(z) = pf^{N_1}(z) \sum_{k=1}^{N_1} q_{j-N_1+k}^+(z) B^k(z) + (1-p)f^{N_2}(z) \sum_{k=1}^{N_2} q_{j-N_2+k}^+(z) B^k(z)$, $j \geq 1$.

Proof See Appendix C. ■

3.1.2 The Recursive Formulas of the Steady-State Distribution for Queue Length at Any Epoch n^+

On the basis of the transient (Theorems 1~2) analysis above, the recursive expressions of the steady-state distribution for queue size at epoch n^+ are explicitly derived by using L'Hospital rule. Moreover, the stochastic decomposition structure of the stationary queue size and the distribution of the additional queue size at epoch n^+ are presented.

Theorem 3 Let $p_j^+ = \lim_{n \rightarrow \infty} P\{N(n^+) = j\}$, $j = 0, 1, 2, \dots$, we have

1) When $\rho \geq 1$, $p_j^+ = 0$, $j = 0, 1, 2, \dots$.

2) When $\rho < 1$, the recursive expressions of $\{p_j^+, j = 0, 1, 2, \dots\}$ for the queue size at epoch n^+ are listed as follows

$$p_0^+ = \frac{(1-\rho)}{pN_1 + (1-p)N_2}, \quad (9)$$

$$p_j^+ = \frac{(1-\rho)(1+\lambda\delta_j)}{pN_1 + (1-p)N_2}, \quad j = 1, 2, \dots, N_1 - 1, \quad (10)$$

$$p_j^+ = \frac{(1-\rho)(1-p+\lambda\delta_j)}{pN_1 + (1-p)N_2}, \quad j = N_1, N_1 + 1, \dots, N_2 - 1, \quad (11)$$

$$p_j^+ = \frac{(1-\rho)\lambda\delta_j}{pN_1 + (1-p)N_2}, \quad j \geq N_2, \quad (12)$$

and the equilibrium distribution $\{p_j^+, j = 0, 1, 2, \dots\}$ forms a probability distribution, where $\delta_j = p \sum_{k=1}^{N_1} q_{j-N_1+k} + (1-p) \sum_{k=1}^{N_2} q_{j-N_2+k}$, $j \geq 1$, $q_j = \frac{1}{G(\bar{p})} \sum_{n=j-i}^{\infty} \sum_{k=n+1}^{\infty} g_k \binom{n}{j-1} p^{j-1} \bar{p}^{n-j+1} + \sum_{i=1}^{j-1} q_{j-i} \left[1 - G(\bar{p}) - \sum_{m=1}^i \sum_{k=m}^{\infty} g_k \binom{k}{m} p^m \bar{p}^{k-m} \right]$, $j \geq 1$,

Proof See Appendix D. ■

Theorem 4 When $\rho < 1$, let $P^+(z)$ denote the probability generating function of $\{p_j^+, j = 0, 1, 2, \dots\}$, then

$$P^+(z) = \frac{(1-\rho)(1-z)G(z\lambda + \bar{\lambda})}{G(z\lambda + \bar{\lambda}) - z} \cdot \frac{(1-z^{N_2}) + p(z^{N_2} - z^{N_1})}{(1-z)[pN_1 + (1-p)N_2]}, \quad |z| < 1, \quad (13)$$

and the mean queue size, denoted by $\bar{L}^+$, is presented by

$$\bar{L}^+ = \rho + \frac{\lambda^2 E[\chi(\chi-1)]}{2(1-\rho)} + \frac{pN_1(N_1-1) + (1-p)N_2(N_2-1)}{2[pN_1 + (1-p)N_2]}, \quad (14)$$

where $E[\chi(\chi-1)] = \sum_{j=2}^{\infty} j(j-1)P(\chi=j)$.

Proof See Appendix E. ■

Special Case 1 When $p = 1$ and $N_1 = 1$ or $p = 0$ and $N_2 = 1$ or $N_1 = N_2 = 1$, the $Geo/G/1$ queueing system studied in this paper becomes the classic $Geo/G/1$ queueing system studied by Luo et al.^[3]. Then

$$P^+(z) = \frac{(1-\rho)(1-z)G(z\lambda + \bar{\lambda})}{G(z\lambda + \bar{\lambda}) - z}, \quad |z| < 1,$$

$$\bar{L}^+ = \rho + \frac{\lambda^2 E[\chi(\chi - 1)]}{2(1-\rho)}, \quad \rho < 1,$$

which agree with Luo et al.^[3].

Special Case 2 When $p = 1$ or $p = 0$ or $N_1 = N_2$, the $Geo/G/1$ queueing system studied in this paper is equivalent to the $Geo/G/1$ queueing system with N -policy studied by Luo et al.^[17]. Then

$$P^+(z) = \frac{(1-\rho)(1-z)G(z\lambda + \bar{\lambda})}{G(z\lambda + \bar{\lambda}) - z} \cdot \frac{1 - z^{N_1}}{N_1(1-z)}, \quad |z| < 1,$$

$$\bar{L}^+ = \rho + \frac{\lambda^2 E[\chi(\chi - 1)]}{2(1-\rho)} + \frac{N_1 - 1}{2}, \quad \rho < 1,$$

which agree with Luo et al.^[17].

Special Case 3 When $N_2 = N_1 + 1$, the $Geo/G/1$ queueing system studied in this paper is equivalent to the $Geo/G/1$ queueing system with $\langle p, N \rangle$ -policy studied by Wang and Ke^[25]. Then

$$P^+(z) = \frac{(1-\rho)(1-\lambda + \lambda z)}{1 - \lambda + \lambda z - \rho z} \cdot \frac{1 - pz^{N_1} - (1-p)z^{N_1+1}}{(1-z)(1-p + N_1)}, \quad |z| < 1,$$

$$\bar{L}^+ = \rho + \frac{\lambda^2 E[\chi(\chi - 1)]}{2(1-\rho)} + \frac{pN_1(N_1 - 1) + (1-p)N_1(N_1 + 1)}{2(1-p + N_1)}, \quad \rho < 1,$$

which agree with Wang and Ke^[25].

Special Case 4 When the service time χ of each customer obeys the geometric distribution with parameter η ($0 < \eta < 1$), then

$$P^+(z) = \frac{(1-\rho)(1-z)(1-\lambda + \lambda z)}{1 - \lambda + \lambda z - z\rho} \cdot \frac{(1 - z^{N_2}) + p(z^{N_2} - z^{N_1})}{(1-z)[pN_1 + (1-p)N_2]}, \quad |z| < 1,$$

$$\bar{L}^+ = \rho + \frac{\rho^2(1-\eta)}{1-\rho} + \frac{pN_1(N_1 - 1) + (1-p)N_2(N_2 - 1)}{2[pN_1 + (1-p)N_2]}, \quad \rho = \frac{\lambda}{\eta} < 1.$$

3.1.3 The Steady-State Distributions for Queue Length at Epochs n^- , n and Outside Observer's Observation Epoch

Let $p_{ij}(n^-) = PN(n^-) = j|N(0^-) = i\}$ be the condition probability that there are j customers at epoch n^- under arbitrary initial state $N(0^-) = i$ ($i = 0, 1, \dots$). And let $p_j^- = \lim_{n \rightarrow \infty} P\{N(n^-) = j\}$ denote the equilibrium distribution that there are j customers at epoch n^- . Similarly, $p_{ij}(n) = N(n) = j|N(0) = ii\}$ is the condition probability that there are j customers at epoch n under arbitrary initial state $N(0) = i$. $p_j = \lim_{n \rightarrow \infty} P\{N(n) = j\}$ is the

equilibrium distribution that there are j customers at epoch n . Let $P^-(z)$ denote the probability generating function of $\{p_j^-, j = 0, 1, 2, \dots\}$. From the model assumptions in Section 2, we have $P\{N(0^-) = i\} = P\{N(0^+) = i\} = P\{N(0) = i\}$.

Therefore, for $i \geq 0, j \geq 0, n \geq 1$, we get

$$\begin{aligned} p_{ij}(n^-) &= N(n^-) = j | N(0^-) = i = i\}, \\ N((n-1)^+) &= j | N(0^+) = i = i\} = p_{ij}((n-1)^+). \end{aligned} \quad (15)$$

Theorem 5 For $\rho < 1$, we have

$$p_j^- = p_j^+, j = 0, 1, \dots, P^-(z) = P^+(z), \quad (16)$$

where p_j^+ ($j = 0, 1, 2, \dots$) is determined by Theorem 3, $P^+(z)$ is determined by Theorem 4.

The results in Theorem 5 indicate that the steady state distribution of queue length at epoch n^+ is the same as the one at epoch n^- .

Since $p_0 = (1-p)p_0^-, p_j = pp_{j-1}^- = (1-p)p_j^-, j \geq 1$, we have the following Theorem 6.

Theorem 6 For $\rho < 1$, we have

1)

$$p_0 = (1-p)p_0^-, p_j = pp_{j-1}^- = (1-p)p_j^-, j = 1, 2, \dots; \quad (17)$$

2) Let $P(z)$ denote the probability generating function of $\{p_j, j = 0, 1, 2, \dots\}$, then

$$P(z) = (1 - \lambda + \lambda z)P^+(z); \quad (18)$$

3) The mean of steady-state queue size, denoted by $\bar{L}$, is presented by

$$\bar{L} = \lambda + \bar{L}^+; \quad (19)$$

where p_j^+ ($j = 0, 1, 2, \dots$) is determined by Theorem 3, $P^+(z)$ and $\bar{L}^+$ are determined by Theorem 4.

Remark 1 From Theorems 3, 5 and 6, we can obtain the following relations $p_j^- = p_j^+ \neq p_j, j = 0, 1, \dots$. The relations indicate the important difference between discrete-time queueing system and continuous-time queueing system.

Remark 2 We study the steady-state queue size distribution at outside observer's epoch, designated by $p_j^{\text{out}}, j = 0, 1, \dots$. As can be seen from Figure 1, outside observer's epoch falls in the middle of interval $(n^+, (n+1)^-)$. Since there is neither arrival nor service completion between time epochs n^+ and $(n+1)^-$, the system state at the outside observer's epoch is the same as that at time epochs n^+ and $(n+1)^-$. So we know $p_j^{\text{out}} = p_j^- = p_j^+, j = 0, 1, \dots$.

Theorem 7 When $\rho < 1$, the average waiting time of customer, denoted by $\bar{W}_q$, is given by

$$\bar{W}_q = \frac{\lambda E[\chi(\chi-1)]}{2(1-\rho)} + \frac{pN_1(N_1-1) + (1-p)N_2(N_2-1)}{2\lambda[pN_1 + (1-p)N_2]}. \quad (20)$$

Proof In the queueing system examined in this study, customers are served on a first-come, first-served basis. Therefore, Little's law is valid. In addition to this, from Remark 2, $p_j^{\text{out}} = p_j^+$; hence, applying $\overline{W}_q = \frac{\overline{L}^+}{\lambda} - \mu$ leads to Equation (20). ■

3.2 Some Other Important Performance Measures

1) Distribution of customers at the beginning of the server's busy period

Let Q_b be the number of customers present in the system at the beginning of the server's busy period, and $E[Q_b]$ be the expected value of Q_b . Thus, on the basis of the model assumption, the distribution of Q_b is given by

$$P\{Q_b = N_1\} = p, \quad P\{Q_b = N_2\} = 1 - p.$$

Obviously, $E[Q_b]$ can be calculated as

$$E[Q_b] = pN_1 + (1 - p)N_2. \quad (21)$$

2) Expected value of the server's busy period

Let B denote the length of the server's busy period initiated with Q_b customers. With the help of Lemma 1, the expected value of the server busy period B is given by

$$E[B] = E[b]E[Q_b] = \frac{\rho[pN_1 + (1 - p)N_2]}{\lambda(1 - \rho)}, \quad \rho < 1. \quad (22)$$

3) Expected value of the server's idle period

Because the inter-arrival times follow independent and identical geometric distribution with parameter λ , the expected value of the server's idle period, denoted by $E[I]$, is given by

$$E[I] = \frac{E[Q_b]}{\lambda} = \frac{pN_1 + (1 - p)N_2}{\lambda}. \quad (23)$$

4) Expected value of the busy cycle

We note that the length of a busy cycle, denoted by S , consists of a server's idle period and a server's busy period. Therefore, under the stability condition, we can obtain the expected value of the busy cycle $E[S]$ as follows

$$E[S] = E[I] + E[B] = \frac{pN_1 + (1 - p)N_2}{\lambda(1 - \rho)}, \quad \rho < 1. \quad (24)$$

4 Sensitivity Analysis of Operating Characteristics

In this section, some numerical results are presented to illustrate the effect of various system parameters on the performance measures of p_0^+ and $\overline{L}^+$. All the computations are done by developing program in MATLAB software and it is assumed that the service time is governed by a geometric distribution with parameter η .

Figures 3 and 4 are depicted to explore the impact of p on p_0^+ and $\overline{L}^+$ for different values of λ . We set default parameters for Figures 3 and 4 as $\eta = 0.6$, $N_1 = 6$, $N_2 = 11$. It can be observed from Figure 3 that for fixed λ , p_0^+ increases with the increase of p . The reason is

that the greater the probability p increases (i.e., the more likely the service is to start serving customers when the number of customers in the system is N_1), the fewer customers waiting in the system, which undoubtedly leads to the increase of p_0^+ . On the other hand, as λ increases from 0.05 to 0.4, p_0^+ shows a decreasing trend. The reason is that as λ increases, the number of customers in the system gradually increases, which undoubtedly leads to the decrease of p_0^+ . From Figure 4, one can see that the mean system size $\bar{L}^+$ decreases with the increase of p while increases as λ increases, which is in accordance with our expectation. In fact, when λ increases, more and more customers accumulate in the system, which certainly causes the increase of the mean queue length of system $\bar{L}^+$.

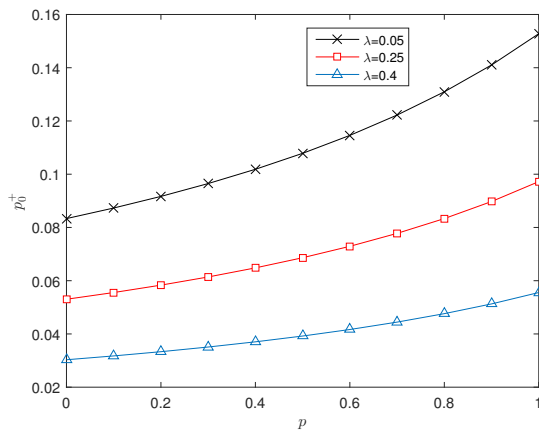


Figure 3 The effect of p on p_0^+ for different values of λ

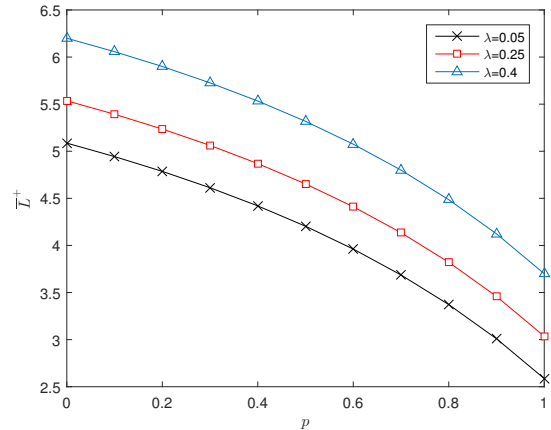


Figure 4 The effect of p on $\bar{L}^+$ for different values of λ

Figures 5 and 6 describe the influence of p on p_0^+ and $\bar{L}^+$ for different values of η . We set default parameters for Figures 5 and 6 as $\lambda = 0.25$, $N_1 = 6$, $N_2 = 11$. Figure 5 shows that p_0^+ shows an increasing trend as η increases for a fixed p . In fact, as the value of η 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 p and η on $\bar{L}^+$ is plotted in Figure 6.

5 Cost Optimization Under the Constraint of the Average Waiting Time

In practice, operating costs and benefits are closely tied. Thus, system designers or system managers are interested in reducing the running cost of the system. In order to show the applicability of the findings from the previous discussion, a cost optimization analysis is carried out from the economic view. Let us construct the cost structure model as follows:

$C_h \equiv$ cost per unit time for each customer present in the system (this cost originates from the customer's waiting time and service time).

$C_r \equiv$ fixed setup cost per unit time per busy cycle (the transition between busy and idle periods is what causes this cost).

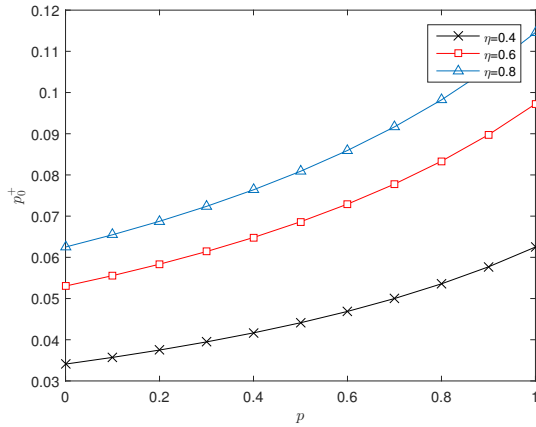


Figure 5 The effect of p on p_0^+ for different values of η

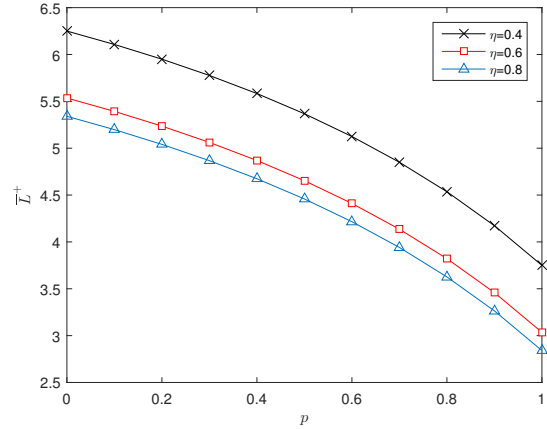


Figure 6 The effect of p on $\bar{L}^+$ for different values of η

Let $F(N_1, N_2)$ be the long-run expected cost per unit time of the system. Applying the renewal reward theorem, it gets

$$\begin{aligned} F(N_1, N_2) &= C_h \bar{L}^+ + \frac{C_r}{E[S]} \\ &= C_h \left\{ \rho + \frac{\lambda^2 E[\chi(\chi-1)]}{2(1-\rho)} + \frac{pN_1(N_1-1) + (1-p)N_2(N_2-1)}{2[pN_1 + (1-p)N_2]} \right\} + \\ &\quad \frac{C_r \lambda (1-\rho)}{pN_1 + (1-p)N_2}, \end{aligned} \quad (25)$$

where $\bar{L}^+$ is determined by Theorem 4 and $E[S]$ is determined by Equation (24) as mentioned in Section 3.

In previous system optimization problems with control policies, scholars seldom considered the effect of customer waiting time length on the optimal control policy. The (p, N_1, N_2) -policy can reduce the frequent start-up costs of the system, but it increases queue length and the time customers spend in the system, which can lead to system congestion and even customer churn. Therefore, based on the cost function $F(N_1, N_2)$ above, we construct the cost optimization problem under the constraint of the average waiting time as follows:

$$\begin{cases} \min F(N_1, N_2), \\ \text{s.t. } \bar{W}_q \leq \bar{W}_{q_0}, \end{cases} \quad (26)$$

where $\bar{W}_q$ is given in Theorem 7, and $\bar{W}_{q_0}$ is a predetermined non-negative threshold value.

It is clear from Equation (26) that $F(N_1, N_2)$ and $\bar{W}_q$ are quite complicated and non-linear in relation to the choice variables N_1 and N_2 . As a result, it is very difficult to obtain the analytical solutions for the ideal values of N_1 and N_2 . Next, we will use the direct search method to solve the optimization problem by a numerical experiment.

Example 1 Since the PH distribution is widely used in science and engineering, we suppose that the service time χ of each customer follows PH distribution with the representation

(γ, K) of order m , where K is a random square matrix of order m . Vector $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_m)$ can be interpreted as the initial probability vector among the transient state, where $\gamma e_m = 1$ and e_m is a column vector of dimension m with all elements equal to one. Therefore, Equation (26) can be reduced to the following minimization problem:

$$\begin{cases} \min F(N_1, N_2) = C_h \cdot \left\{ \lambda \gamma (I_m - K)^{-1} e_m + \frac{\lambda^2 \gamma K (I_m - K)^{-2} e_m}{1 - \lambda \gamma (I_m - K)^{-1} e_m} \right. \\ \quad \left. + \frac{p N_1 (N_1 - 1) + (1 - p) N_2 (N_2 - 1)}{2[p N_1 + (1 - p) N_2]} \right\} + \frac{C_r \lambda (1 - \lambda \gamma (I_m - K)^{-1} e_m)}{p N_1 + (1 - p) N_2}, \\ \text{s.t. } \bar{W}_q = \frac{\lambda \gamma K (I_m - K)^{-2} e_m}{1 - \lambda \gamma (I_m - K)^{-1} e_m} + \frac{p N_1 (N_1 - 1) + (1 - p) N_2 (N_2 - 1)}{2\lambda[p N_1 + (1 - p) N_2]} \leq \bar{W}_{q_0}. \end{cases} \quad (27)$$

where I_m represents the identity matrix of order m and $I_m - K$ is a nonsingular matrix.

We select $\lambda = 0.2$, $p = 0.8$, $C_h = 6$, $C_r = 350$, $m = 3$, $\gamma = \begin{pmatrix} 0.35 \\ 0.35 \\ 0.3 \end{pmatrix}^T$, $K = \begin{bmatrix} 0.2 & 0.3 & 0.1 \\ 0.2 & 0.3 & 0 \\ 0.05 & 0.1 & 0.4 \end{bmatrix}$, and put them into Equation (27). By using *MATLAB* software, we get the computational results of the expected cost per unit time and the expected waiting time $\bar{W}_q$ under different values of N_1 and N_2 . The numerical results are presented in Table 1.

Table 1 $F(N_1, N_2)$ and $\bar{W}_q$ vary with different values of N_1 and N_2

(N_1, N_2)	Under $\bar{W}_q \leq \bar{W}_{q_0} = 7$		Under $\bar{W}_q \leq \bar{W}_{q_0} = 5$	
	$F(N_1, N_2)$	$\bar{W}_q$	$F(N_1, N_2)$	$\bar{W}_q$
(1, 1)	43.8931	0.1608	43.8931	0.1608
(1, 2)	38.3625	0.9941	38.3625	0.9941
(1, 3)	35.2693	2.3036	35.2693	2.3036
(2, 2)	27.3014	2.6608	27.3014	2.6608
(2, 3)	26.3385	3.3426	26.3385	3.3426
(1, 4)	33.6993	3.9108	33.6993	3.9108
(2, 4)	26.0361	4.3274	26.0361	4.3274
(3, 3)	23.7708	5.1608	-	-
(1, 5)	33.1449	5.7163	-	-
(3, 4)	23.7045	5.7858	-	-
(2, 5)	26.2418	5.9454	-	-
(3, 5)	23.9989	6.6314	-	-
(2, 6)	26.8466	6.9465	-	-

For the same set of parameters, Table 2 represents the expected cost per unit time $F(N)$ and the expected waiting time $\bar{W}_q$ under $N_1 = N_2 = N$.

From Tables 1 and 2, we can find

1) If the upper bound of the mean waiting time is set to be 7, namely $\bar{W}_{q_0} = 7$, we find that $F(3, 4) = 23.7045$ is the minimum of the long-run average cost rate, and $(N_1^*, N_2^*) = (3, 4)$ is the optimal bi-level threshold policy that satisfies the average waiting time $\bar{W}_{q_0} = 7$ constraints. However, for $\bar{W}_{q_0} = 7$, we have $F(N) = 23.7708$ at $N = 3$. So, for $\bar{W}_{q_0} = 7$, we get $F(N_1^*, N_2^*) < F(N)$.

Table 2 $F(N)$ and $\bar{W}_q$ vary with different values of N

N	Under $\bar{W}_q \leq \bar{W}_{q_0} = 7$		Under $\bar{W}_q \leq \bar{W}_{q_0} = 5$	
	$F(N)$	$\bar{W}_q$	$F(N)$	$\bar{W}_q$
1	43.8931	0.1608	43.8931	0.1608
2	27.3014	2.6608	27.3014	2.6608
3	23.7708	5.1608	—	—

2) If the upper bound of the mean waiting time is set to be 5, namely $\bar{W}_{q_0} = 5$, we may observe that $F(2, 4) = 26.0361$ is the minimum of the long-run average cost rate, and $(N_1^*, N_2^*) = (2, 5)$ is the optimal bi-level threshold policy that satisfies the average waiting time $\bar{W}_{q_0} = 5$ constraints. However, for $\bar{W}_{q_0} = 5$, we have $F(N) = 27.3014$ at $N = 2$. So, for $\bar{W}_{q_0} = 5$, we get $F(N_1^*, N_2^*) < F(N)$.

From the above, it can be seen that the minimum of the long-run average cost rate obtained by the bi-level randomized policy is smaller than that obtained by the traditional N -policy, subject to a given $\bar{W}_{q_0}$. Therefore, the bi-level randomized policy is more advantageous compared to the traditional N -policy.

6 Conclusion

In this paper we carried out a new discrete-time $Geo/G/1$ queue with bi-level randomized (p, N_1, N_2) -policy. Firstly, employing the renewal process theory, the total probability decomposition technique and z -transform, we analyzed the transient queue length distribution of the system at any time n^+ that starts from any initial state, and derived the expressions of the z -transform of the transient queue length distribution with respect to time n^+ . As we know that it is very difficult and complex to analyze the transient behavior of the queue size at any time n^+ . Then, based on the transient analysis, the recursive expressions of the steady-state distribution for queue size at any time n^+ are obtained. Moreover, we also explored the steady-state distribution of queue length at epochs n^- , n and outside observer's observation epoch. And we find the important relationships between the equilibrium queue length distributions at different time epochs n^- , n and n^+ , which leads to the introduction of an expression for the key metric average waiting time. Next, some numerical examples were provided to study the effect of some key parameters on the operating characteristics of the system. These numerical analyses not only demonstrate the validity of our analytical results, but also can provide significant insight to the concerned system designers or decision makers. Finally, under the constraint of the average waiting time, numerical experiment was provided to determine the optimal bi-level control policy for economizing the system cost. And a comparison with the N -policy concludes that the bi-level randomized (p, N_1, N_2) -policy gives more efficient results than the N -policy queue.

Our analysis can help policy-makers set a reasonable threshold for cost optimization in many applications, such as information processing system. The system studied in this paper can be

extended to more complex queueing models, such as impatient customers, server vacations, and feedback phenomenon.

Acknowledgements The authors gratefully acknowledge the Editor and anonymous referees for their insightful comments and helpful suggestions that led to a marked improvement of the article.

References

- [1] Meisling T. Discrete-time queueing theory. *Operations Research*, 1958, 6(1): 96–105.
- [2] Zhang Z G, Tian N S. Discrete time $Geo/G/1$ queue with multiple adaptive vacations. *Queueing Systems*, 2001, 38: 419–429.
- [3] Luo C Y, Tang Y H, Liu R B. Transient solution for queue-length distribution of $Geometry/G/1$ queueing model. *Applied Mathematics-A Journal of Chinese Universities*, 2007, 22: 95–100.
- [4] Samanta S K, Chaudhry M L, Gupta U C. Discrete-time queues with single and multiple vacations. *Mathematical and Computer Modelling*, 2007, 45: 93–180.
- [5] Tang Y H, Yun X, Huang, S J. Discrete-time $Geo^X/G/1$ queue with unreliable server and multiple adaptive delayed vacations. *Journal of Computational and Applied Mathematics*, 2008, 220(2): 439–455.
- [6] Yu M M, Tang Y H, Fu Y H, et al. $GI/Geom/1/N/MWV$ queue with changeover time and searching for the optimum service rate in working vacation period. *Journal of Computational and Applied Mathematics*, 2011, 235(8): 2170–2184.
- [7] Luo C Y, Xiang K L, Yu M M, et al. Recursive solution of queue length distribution for $Geo/G/1$ queue with single server vacation and variable input rate. *Computers & Mathematics with Applications*, 2011, 61(9): 2401–2411.
- [8] Wang T Y, Ke J C, Chang F M. On the discrete-time $Geo/G/1$ queue with randomized vacations and at most J vacations. *Applied Mathematical Modelling*, 2011, 35(5): 2297–2308.
- [9] Gao S, Wang J J. On a discrete-time $GI^X/GEO/1/N-G$ queue with randomized working vacations and at most J vacations. *Journal of Industrial and Management Optimization*, 2015, 11(3): 779–806.
- [10] Lan S J, Tang Y H. An unreliable discrete-time retrial queue with probabilistic preemptive priority, balking customers and replacements of repair times. *AIMS Mathematics*, 2020, 5(5): 4322–4344.
- [11] Fiems D, Turck K D. Analysis of discrete-time queues with branching arrivals. *Mathematics*, 2023, 11(4): 2227–7390.
- [12] Atencia I, María Ángeles Galán-García M A, Aguilera-Venegas G, et al. A non markovian retrial queueing system. *Journal of Computational and Applied Mathematics*, 2023, 431: 115277.
- [13] Deng L, Shu Z, Chen T. Event-Triggered Stochastic Model Predictive Control for Constrained Queueing Networks. *IEEE Transactions on Network Science and Engineering*, 2023, 11(2): 1744–1755.
- [14] El-Taha M. On BASTA for discrete-time queues. *Operations Research Letters*, 2024, 57: 107206.
- [15] Moreno P. A discrete-time single-server queue with a modified N -policy. *International Journal of Systems Science*, 2007, 38(6): 483–492.

- [16] Moreno P. Analysis of a $Geo/G/1$ queueing system with a generalized N -policy and setup-closedown times. *Quality Technology & Quantitative Management*, 2008, 5(2): 111–128.
- [17] Luo C Y, Tang Y H, Li W, et al. The recursive solution of queue length for $Geo/G/1$ queue with N -policy. *Journal of Systems Science and Complexity*, 2012, 25(2): 293–302.
- [18] Wang T Y. Random N -policy $Geo/G/1$ queue with startup and closedown times. *Journal of Applied Mathematics*, 2012, (2012): 1–19.
- [19] Lim D E, Lee D H, Yang W S, et al. Analysis of the $GI/Geo/1$ queue with N -policy. *Applied Mathematical Modelling*, 2013, 37(7): 4643–4652.
- [20] Ma Z Y, Wang P C, Yue W Y. Performance analysis and optimization of a pseudo-fault $Geo/Geo/1$ repairable queueing system with N -policy setup time and multiple working vacations. *Journal of Industrial and Management Optimization*, 2017, 13: 1467–1481.
- [21] Lan S J, Tang Y H. An N -policy discrete-time $Geo/G/1$ queue with modified multiple server vacations and Bernoulli feedback. *RAIRO-Operations Research*, 2019, 53(2): 367–387.
- [22] Liu R B, Alfa A S, Yu M M. Analysis of an ND -policy $Geo/G/1$ queue and its application to wireless sensor networks. *Operational Research*, 2019, 19: 449–477.
- [23] Sikdar K. Modelling and analysis of a discrete-time $GI^X/Geom/1/K$ queue with N threshold policy. *International Journal of Mathematics in Operational Research*, 2020, 17(3): 301–317.
- [24] Wang R X. Queue size distribution on a new ND policy $Geo/G/1$ queue and its computation designs. *Discrete Dynamics in Nature and Society*, 2021, 1: 1–16.
- [25] Wang T Y, Ke J C. The randomized threshold for the discrete-time $Geo/G/1$ queue. *Applied Mathematical Modelling*, 2009, 33(7): 3178–3185.
- [26] Xu Z R, Zhu Y J, Sun G Y, et al. The $Geo/Geo/1$ queue model with multiple vacations, negative customers, and the (p, N) policy set-up time. *Chinese Journal of Engineering Mathematics*, 2011, 28(6): 794–802.
- [27] Jia S F, Chen Y H. The $Geo/G/1$ queue model with (p, N) policy set-up time, multiple vacation and disasters. *International Journal of Scientific Engineering & Technology*, 2013, 2(10): 991–995.
- [28] Lan S J, Tang Y H. Performance and reliability analysis of a repairable discrete-time $Geo/G/1$ queue with Bernoulli feedback and randomized policy. *Applied Stochastic Models in Business and Industry*, 2017, 33(5): 522–543.
- [29] Kuang X Y, Tang Y H, Yu M M, et al. Performance analysis of an $M/G/1$ queue with bi-level randomized (p, N_1, N_2) -policy. *RAIRO-Operations Research*, 2022, 56: 395–414.
- [30] Kumar N. Analysis of a manufacturing queueing system using bi-level randomized policy. *Top*, 2024, 17: 1–28.
- [31] Hunter J J. *Mathematical techniques of applied probability*, Vol. II, Discrete time models: Techniques and applications. New York: Academic Press, 1983.
- [32] Bruneel H, Kim B G. *Discrete-time models for communication systems including ATM*. Boston: Kluwer Academic, 1993.

A Proof of Lemma 1

Consider the first customer served in the server busy period b . Let χ denote the service time of this customer, and ν denote the number of arriving customers during χ . We call these customers who arrive during χ are the “descendants of the first generation”. Since the length of the busy period depends on the number of customers served during the busy period and is independent of the order of service. Therefore, for ease of calculation, it is assumed that customers are served on a first-come-first-served basis. According to the rule of last-come-first-served, the first customer who leaves will serve the “descendants of the first generation” in the last customer to arrive, and then serve all the descendants of that customer until there are no descendants of that customer in the system before serving another customer in the “descendants of the first generation”. In this way we have

$$b = \chi + b_1 + b_2 + \cdots + b_\nu,$$

where $b_j (j = 1, 2, \cdots, \nu)$ denotes the time period from when the server starts serving the j th master client until all his descendants have been served. once the server starts to serve, the operation of the system depends entirely on the number of initial customers, the arrival process and the service process, and has nothing to do with the cumulative number of customers at the beginning of the busy period. Additionally, since the arrival process of customers is a Poisson process, the start time of the server busy period is the point of regeneration of the system. So that $b_1, \cdots, b_\nu$ are identical and independently distributed with b , and are independent of χ and ν , which are also independent of each other. Therefore, similar to the rest proof of the busy period initiated with one customer in reference^[32] the proof of Lemma 1 can be completed.

B Proof of Theorem 1

Let $l_k = \sum_{i=1}^k \tau_i$, and $l_0 = 0$. For $p_{00}(n^+)$, it is noted that $p_{00}(n^+)$ indicates that there is no customer in the queue at epoch n^+ under initial state $N(0^+) = 0$, i.e., epoch n^+ is located in the system idle period. Because the arrival time interval of customers obeys the geometric distribution (with “memoryless”), according to the hypothesis of the model, employing renewal theory and the total probability decomposition law, we obtain

$$\begin{aligned} p_{00}(n^+) &= P\{0 \leq n^+ < \hat{\tau}_1\} + P\{\hat{\tau}_1 + b \leq n^+ < \hat{\tau}_1 + b + \hat{\tau}_2\} + P\{\hat{\tau}_1 + b + \hat{\tau}_2 \leq n^+, N(n^+) = 0\} \\ &= \bar{\lambda}^n + \sum_{t=2}^n P\{\hat{\tau}_1 + b = t\} \bar{\lambda}^{n-t} + P\{\hat{\tau}_1 + b + \hat{\tau}_2 \leq n^+, N(n^+) = 0\}. \end{aligned} \quad (B1)$$

The third term of Equation (B1) can be decomposed into the following two cases:

i) When N_1 customers arrive after the system becomes empty, the server is activated for services with probability $p(0 \leq p \leq 1)$, and the epoch n^+ is located after the server’s idle period with no customer (see Figure 7).

ii) When N_1 customers arrive after the system becomes empty, the server is not activated for services with probability $(1 - p)$, then the server is turned on by the arrival of the N_2 th customer

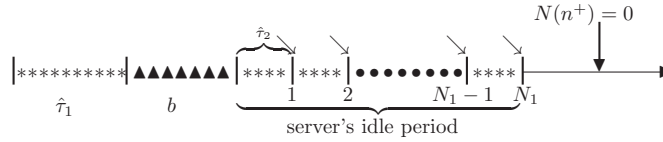


Figure 7 There are N_1 customers in system during the server's idle period and the epoch n^+ is located after the server's idle period with no customer

in the system, and the epoch n^+ is located after the server's idle period with no customer (see Figure 8).

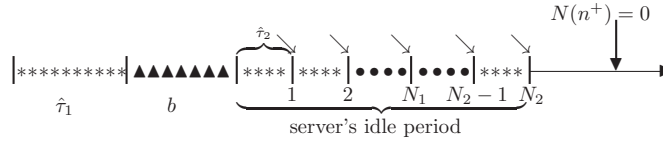


Figure 8 There are N_2 customers in system during the server's idle period and the epoch n^+ is located after the server's idle period with no customer

So, the third term of Equation (B1) is

$$\begin{aligned}
 & P\{\hat{\tau}_1 + b + \hat{\tau}_2 \leq n^+, N(n^+) = 0\} \\
 &= pP\{\hat{\tau}_1 + b + \hat{\tau}_2 + l_{N_1-1} \leq n^+, N(n^+) = 0\} + \\
 & \quad (1-p)P\{\hat{\tau}_1 + b + \hat{\tau}_2 + l_{N_2-1} \leq n^+, N(n^+) = 0\} \\
 &= p \sum_{t=2}^n P\{\hat{\tau}_1 + b = t\} \sum_{k=N_1}^{n-t} P\{\hat{\tau}_2 + l_{N_1-1} = k\} p_{N_1 0}((n-t-k)^+) + \\
 & \quad (1-p) \sum_{t=2}^n P\{\hat{\tau}_1 + b = t\} \sum_{k=N_2}^{n-t} P\{\hat{\tau}_2 + l_{N_2-1} = k\} p_{N_2 0}((n-t-k)^+). \quad (B2)
 \end{aligned}$$

Substituting Equation (B2) into Equation (B1), multiplying Equation (B1) by z^n and summing over n , it finally leads to

$$p_{00}^+(z) = \frac{1 + f(z)B(z)}{1 - z\bar{\lambda}} + pB(z)f^{N_1+1}(z)p_{N_1 0}^+(z) + (1-p)B(z)f^{N_2+1}(z)p_{N_2 0}^+(z). \quad (B3)$$

For $i \geq 1$, similar to the analysis of $p_{00}(n^+)$, we can get

$$\begin{aligned}
 p_{i0}(n^+) &= P\{b^{<i>} \leq n^+ < b^{<i>} + \hat{\tau}_1\} + P\{b^{<i>} + \hat{\tau}_1 \leq n^+, N(n^+) = 0\} \\
 &= \sum_{t=2}^n P\{b^{<i>} = t\} \bar{\lambda}^{n-t} + pP\{b^{<i>} + \hat{\tau}_1 + l_{N_1-1} \leq n^+, N(n^+) = 0\} + \\
 & \quad (1-p)P\{b^{<i>} + \hat{\tau}_1 + l_{N_2-1} \leq n^+, N(n^+) = 0\} \\
 &= \sum_{t=2}^n P\{b^{<i>} = t\} \bar{p}^{n-t} + \\
 & \quad p \sum_{t=i}^n P\{b^{<i>} = t\} \sum_{k=N_1}^{n-t} P\{\hat{\tau}_1 + l_{N_1-1} = k\} p_{N_1 0}((n-t-k)^+) +
 \end{aligned}$$

$$(1-p) \sum_{t=i}^n P\{b^{<i>} = t\} \sum_{k=N_2}^{n-t} P\{\hat{\tau}_1 + l_{N_2-1} = k\} p_{N_2 0}((n-t-k)^+). \quad (B4)$$

Multiplying Equation (B4) by z^n and summing over n , it gives

$$p_{i0}^+(z) = \frac{B^i(z)}{1-z\bar{\lambda}} + pB^i(z)f^{N_1}(z)p_{N_1 0}^+(z) + (1-p)B^i(z)f^{N_2}(z)p_{N_2 0}^+(z). \quad (B5)$$

From Equations (B3) and (B5), we obtain the relationship between $p_{00}^+(z)$ and $p_{i0}^+(z)$ as follows

$$p_{i0}^+(z) = \frac{B^{i-1}(z)}{f(z)} \left\{ p_{00}^+(z) - \frac{1}{1-z\bar{\lambda}} \right\}, \quad i \geq 1. \quad (B6)$$

Taking Equation (B6) into Equation (B3) gives Equation (1). Furthermore, we can get Equation (2) by substituting Equation (1) into Equation (B6).

C Proof of Theorem 2

1) For $j = 1, 2, \dots, N_1 - 1$, there are j customers in the system at epoch n^+ , one of the following conditions must be satisfied:

- i) The epoch n^+ is located in the server's busy period with j customers,
- ii) The epoch n^+ is located in the server's idle period with j customers.

So similar to the decomposition of Theorem 1, we get

$$\begin{aligned} p_{0j}(n^+) &= P\{\hat{\tau}_1 \leq n^+ < \hat{\tau}_1 + b\} + P\{\hat{\tau}_1 + b + \hat{\tau}_2 \leq n^+, N(n^+) = j\} \\ &= \sum_{t=1}^n P\{\hat{\tau}_1 = t\} Q_j((n-t)^+) + P\{\hat{\tau}_1 + b + \hat{\tau}_2 + l_{j-1} \leq n^+ < \hat{\tau}_1 + b + \hat{\tau}_2 + l_j\} + \\ &\quad pP\{\hat{\tau}_1 + b + \hat{\tau}_2 + l_{N_1-1} \leq n^+, N(n^+) = j\} + \\ &\quad (1-p)P\{\hat{\tau}_1 + b + \hat{\tau}_2 + l_{N_2-1} \leq n^+, N(n^+) = j\} \\ &= \sum_{t=1}^n P\{\hat{\tau}_1 = t\} Q_j((n-t)^+) + \\ &\quad \sum_{t=2}^n P\{\hat{\tau}_1 + b = t\} \sum_{k=j}^{n-t} P\{\hat{\tau}_2 + l_{j-1} = k\} P\{\tau_j > (n-t-k)^+\} + \\ &\quad p \sum_{t=2}^n P\{\hat{\tau}_1 + b = t\} \sum_{k=N_1}^{n-t} P\{\hat{\tau}_2 + l_{N_1-1} = k\} p_{N_1 j}((n-t-k)^+) + \\ &\quad (1-p) \sum_{t=2}^n P\{\hat{\tau}_1 + b = t\} \sum_{k=N_2}^{n-t} P\{\hat{\tau}_2 + l_{N_2-1} = k\} p_{N_2 j}((n-t-k)^+). \end{aligned} \quad (C7)$$

Analogously, for $i \geq 1$, we obtain

$$p_{ij}(n^+) = \sum_{k=1}^i \sum_{t=k-1}^n P\{b^{<k-1>} = t\} Q_{j-i-k}((n-t)^+) +$$

$$\begin{aligned}
& \sum_{t=i}^n P\{b^{<i>} = t\} \sum_{k=j}^{n-t} P\{\hat{\tau}_2 + l_{j-1} = k\} P\{\tau_j > (n-t-k)^+\} + \\
& p \sum_{t=i}^n P\{b^{<i>} = t\} \sum_{k=N_1}^{n-t} P\{\hat{\tau}_2 + l_{N_1-1} = k\} p_{N_1j}((n-t-k)^+) + \\
& (1-p) \sum_{t=i}^n P\{b^{<i>} = t\} \sum_{k=N_2}^{n-t} P\{\hat{\tau}_2 + l_{N_2-1} = k\} p_{N_2j}((n-t-k)^+). \quad (C8)
\end{aligned}$$

Taking z -transformation on Equations (C7) and (C8), $p_{0j}^+(z)$ and $p_{ij}^+(z)$ are given as follows

$$\begin{aligned}
p_{0j}^+(z) &= f(z)q_j^+(z) + \frac{f^{j+1}(z)B(z)}{1-z\bar{\lambda}} + pB(z)f^{N_1+1}(z)p_{N_1j}^+(z) + \\
& (1-p)B(z)f^{N_2+1}(z)p_{N_2j}^+(z), \quad (C9)
\end{aligned}$$

$$\begin{aligned}
p_{ij}^+(z) &= \sum_{k=1}^i B^{k-1}(z)q_{j-i+k}^+(z) + \frac{f^j(z)B^i(z)}{1-z\bar{\lambda}} + pB(z)f^{N_1}(z)p_{N_1j}^+(z) + \\
& (1-p)B^i(z)f^{N_2}(z)p_{N_2j}^+(z). \quad (C10)
\end{aligned}$$

From Equations (C9) and (C10), the relationship between $p_{0j}^+(z)$ and $p_{ij}^+(z)$ are given as follows

$$p_{ij}^+(z) = \sum_{l=1}^i q_{j-i+l}^+(z)B^{l-1}(z) + \frac{B^{i-1}(z)}{f(z)}[p_{0j}^+(z) - f(z)q_j^+(z)], \quad i \geq 1. \quad (C11)$$

Taking Equation (C11) into Equation (C9) gives Equation (3). Furthermore, we can get Equation (4) by substituting Equation (3) into Equation (C11).

2) For $j = N_1, N_1 + 1, \dots, N_2 - 1$, there are j customers in the system at epoch n^+ , one of the following conditions must be satisfied:

- i) The epoch n^+ is located in the server's busy period with j customers,
- ii) The epoch n^+ is located in the server's idle period with j customers.

So we get

$$\begin{aligned}
p_{0j}(n^+) &= P\{\hat{\tau}_1 \leq n^+ < \hat{\tau}_1 + b\} + P\{\hat{\tau}_1 + b + \hat{\tau}_2 \leq n^+, N(n^+) = j\} \\
&= \sum_{t=1}^n P\{\hat{\tau}_1 = t\} Q_j((n-t)^+) + \\
& (1-p)P\{\hat{\tau}_1 + b + \hat{\tau}_2 + l_{N_1-1} + l_{j-N_1} \leq n^+ < \hat{\tau}_1 + b + \hat{\tau}_2 + l_{N_1-1} + l_{j-N_1+1}\} + \\
& pP\{\hat{\tau}_1 + b + \hat{\tau}_2 + l_{N_1-1} \leq n^+, N(n^+) = j\} + \\
& (1-p)P\{\hat{\tau}_1 + b + \hat{\tau}_2 + l_{N_2-1} \leq n^+, N(n^+) = j\} \\
&= \sum_{t=1}^n P\{\hat{\tau}_1 = t\} Q_j((n-t)^+) + \\
& (1-p) \sum_{t=N_1+2}^n P\{\hat{\tau}_1 + b + \hat{\tau}_2 + l_{N_1-1} = t\} \sum_{k=j-N_1}^{n-t} P\{l_{j-N_1} = k\} P\{\tau_j > (n-t-k)^+\} +
\end{aligned}$$

$$\begin{aligned}
 & p \sum_{t=2}^n P\{\hat{\tau}_1 + b = t\} \sum_{k=N_1}^{n-t} P\{\hat{\tau}_2 + l_{N_1-1} = k\} p_{N_1j}((n-t-k)^+) + \\
 & (1-p) \sum_{t=2}^n P\{\hat{\tau}_1 + b = t\} \sum_{k=N_2}^{n-t} P\{\hat{\tau}_2 + l_{N_2-1} = k\} p_{N_2j}((n-t-k)^+), \quad (C12)
 \end{aligned}$$

$$\begin{aligned}
 p_{ij}(n^+) &= \sum_{k=1}^i \sum_{t=k-1}^n P\{b^{<k-1>} = t\} Q_{j-i-k}((n-t)^+) + \\
 & (1-p) \sum_{t=i+N_1}^n P\{b^{<i>} + \hat{\tau}_2 + l_{N_1-1} = t\} \sum_{k=j}^{n-t} P\{\hat{\tau}_2 + l_{j-1} = k\} P\{\tau_j > (n-t-k)^+\} + \\
 & p \sum_{t=i}^n P\{b^{<i>} = t\} \sum_{k=N_1}^{n-t} P\{\hat{\tau}_2 + l_{N_1-1} = k\} p_{N_1j}((n-t-k)^+) + \\
 & (1-p) \sum_{t=i}^n P\{b^{<i>} = t\} \sum_{k=N_2}^{n-t} P\{\hat{\tau}_2 + l_{N_2-1} = k\} p_{N_2j}((n-t-k)^+), i \geq 1. \quad (C13)
 \end{aligned}$$

Taking the z -transformation on Equations (C12) and (C13), we get

$$\begin{aligned}
 p_{0j}^+(z) &= f(z) q_j^+(z) + \frac{(1-p)f^{j+1}(z)B(z)}{1-z\bar{\lambda}} + pB(z)f^{N_1+1}(z)p_{N_1j}^+(z) + \\
 & (1-p)B(z)f^{N_2+1}(z)p_{N_2j}^+(z), \quad (C14)
 \end{aligned}$$

$$\begin{aligned}
 p_{ij}^+(z) &= \sum_{k=1}^i B^{k-1}(z) q_{j-i+k}^+(z) + \frac{(1-p)f^j(z)B^i(z)}{1-z\bar{\lambda}} + pB(z)f^{N_1}(z)p_{N_1j}^+(z) + \\
 & (1-p)B^i(z)f^{N_2}(z)p_{N_2j}^+(z). \quad (C15)
 \end{aligned}$$

Solving Equations (C14) and (C15) leads to the expressions (5) and (6) of $p_{0j}^+(z)$ and $p_{ij}^+(z)$.

(3) For $j \geq N_2$, there are j customers in the system at the epoch n^+ if and only if the epoch n^+ is located in the server's busy period with j customers. We have

$$\begin{aligned}
 p_{0j}(n^+) &= P\{\hat{\tau}_1 \leq n^+ < \hat{\tau}_1 + b\} + P\{\hat{\tau}_1 + b + \hat{\tau}_2 \leq n^+, N(n^+) = j\} \\
 &= \sum_{t=1}^n P\{\hat{\tau}_1 = t\} Q_j((n-t)^+) + pP\{\hat{\tau}_1 + b + \hat{\tau}_2 + l_{N_1-1} \leq n^+, N(n^+) = j\} + \\
 & (1-p)P\{\hat{\tau}_1 + b + \hat{\tau}_2 + l_{N_2-1} \leq n^+, N(n^+) = j\} \\
 &= \sum_{t=1}^n P\{\hat{\tau}_1 = t\} Q_j((n-t)^+) + \\
 & p \sum_{t=2}^n P\{\hat{\tau}_1 + b = t\} \sum_{k=N_1}^{n-t} P\{\hat{\tau}_2 + l_{N_1-1} = k\} p_{N_1j}((n-t-k)^+) + \\
 & (1-p) \sum_{t=2}^n P\{\hat{\tau}_1 + b = t\} \sum_{k=N_2}^{n-t} P\{\hat{\tau}_2 + l_{N_2-1} = k\} p_{N_2j}((n-t-k)^+), \quad (C16)
 \end{aligned}$$

$$\begin{aligned}
p_{ij}(n^+) &= \sum_{k=1}^i \sum_{t=k-1}^n P\{b^{<k-1>} = t\} Q_{j-i-k}((n-t)^+) + \\
&\quad (1-p) \sum_{t=i+N_1}^n P\{b^{<i>} + \hat{\tau}_2 + l_{N_1-1} = t\} \sum_{k=j}^{n-t} P\{\hat{\tau}_2 + l_{j-1} = k\} P\{\tau_j > (n-t-k)^+\} + \\
&\quad p \sum_{t=i}^n P\{b^{<i>} = t\} \sum_{k=N_1}^{n-t} P\{\hat{\tau}_2 + l_{N_1-1} = k\} p_{N_1j}((n-t-k)^+) + \\
&\quad (1-p) \sum_{t=i}^n P\{b^{<i>} = t\} \sum_{k=N_2}^{n-t} P\{\hat{\tau}_2 + l_{N_2-1} = k\} p_{N_2j}((n-t-k)^+), i \geq 1. \quad (C17)
\end{aligned}$$

Taking the z -transformation on Equations (C16) and (C17), we get

$$p_{0j}^+(z) = f(z)q_j^+(z) + pB(z)f^{N_1+1}(z)p_{N_1j}^+(z) + (1-p)B(z)f^{N_2+1}(z)p_{N_2j}^+(z), \quad (C18)$$

$$p_{ij}^+(z) = \sum_{k=1}^i B^{k-1}(z)q_{j-i+k}^+(z) + pB(z)f^{N_1}(z)p_{N_1j}^+(z) + (1-p)B^i(z)f^{N_2}(z)p_{N_2j}^+(z), i \geq 1. \quad (C19)$$

Solving Equations (C18) and (C19) leads to the expressions (7) and (8) of $p_{0j}^+(z)$ and $p_{ij}^+(z)$.

D Proof of Theorem 3

In discrete-time situation we have

$$\begin{aligned}
p_j^+ &= \lim_{n \rightarrow \infty} P\{N(n^+) = j\} = \lim_{n \rightarrow \infty} \sum_{i=0}^{\infty} P\{N(n^+) = j; N(0^+) = i\} \\
&= \lim_{n \rightarrow \infty} \sum_{i=0}^{\infty} p_{ij}(n^+) \cdot P\{N(0^+) = i\} = \sum_{i=0}^{\infty} P\{N(0^+) = i\} \lim_{n \rightarrow \infty} p_{ij}(n^+).
\end{aligned}$$

And

$$\lim_{n \rightarrow \infty} p_{ij}(n^+) = \lim_{z \rightarrow 1^-} (1-z)p_{ij}^+(z),$$

so we only need to calculate $\lim_{z \rightarrow 1^-} (1-z)p_{ij}^+(z)$. Combining the expressions of Theorem 1, we give

$$p_0^+ = \lim_{z \rightarrow 1^-} (1-z)p_{i0}^+(z) = \lim_{z \rightarrow 1^-} \frac{1-z}{1-z\bar{\lambda}} \cdot \frac{B^i(z)}{\Delta(z)}. \quad (D20)$$

Part 1 according to Lemma 2, due to $\lim_{z \rightarrow 1^-} B(z) = 1$ and $E[b] = \infty$, we have

$$\lim_{z \rightarrow 1^-} \Delta(z) = 1 - p - (1-p) = 0, \quad (D21)$$

and

$$\lim_{z \rightarrow 1^-} \Delta'(z) = pN_1E[b] + \frac{p}{\lambda}N_1 + pN_2E[b] + \frac{p}{\lambda}N_2 = \infty. \quad (D22)$$

So, using L'Hospital rule gives

$$p_0^+ = \lim_{z \rightarrow 1^-} \frac{1-z}{1-z\bar{\lambda}} \left\{ 1 + \frac{f(z)B(z)}{\Delta(z)} \right\} = \lim_{z \rightarrow 1^-} \frac{1-z}{1-z\bar{\lambda}} \cdot \frac{B^i(z)}{\Delta(z)} = 0.$$

Similarly, $p_j^+ = 0$, $j = 0, 1, 2, \dots$ can be obtained by means of Theorem 1.

Part 2 when $\rho < 1$, we know $E[b] = \frac{\rho}{\lambda(1-\rho)}$ from Lemma 2, so we get

$$\lim_{z \rightarrow 1^-} \Delta'(z) = pN_1 \left[E[b] + \frac{1}{\lambda} \right] + (1-p)N_2 \left[E[b] + \frac{1}{\lambda} \right] = \frac{pN_1 + (1-p)N_2}{\lambda(1-\rho)}. \quad (D23)$$

Applying L'Hospital rule again, we get

$$\begin{aligned} p_0^+ &= \lim_{z \rightarrow 1^-} \frac{1-z}{1-z\bar{\lambda}} \left\{ 1 + \frac{f(z)B(z)}{\Delta(z)} \right\} = \lim_{z \rightarrow 1^-} \frac{1-z}{1-z\bar{\lambda}} \cdot \frac{B^i(z)}{\Delta(z)} = \frac{(1-\rho)}{pN_1 + (1-p)N_2}, \\ p_j^+ &= \lim_{z \rightarrow 1^-} (1-z) \left[q_j^+(z)f(z) + \frac{f(z)\delta_j(z)}{\Delta(z)} \right] \\ &= \lim_{z \rightarrow 1^-} (1-z) \left[\sum_{k=1}^i q_{j-i+k}^+(z)B^{k-1}(z) + \frac{B^{i-1}(z)\delta_j(z)}{\Delta(z)} \right] \\ &= \frac{(1-\rho)(1+\lambda\delta_j)}{pN_1 + (1-p)N_2}, \quad j = 1, 2, \dots, N_1 - 1, \\ p_j^+ &= \lim_{z \rightarrow 1^-} (1-z) \left\{ q_j^+(z)f(z) + \frac{f(z)}{\Delta(z)} \left[\frac{(1-p)f^{j+1}(z)B(z)}{1-z\bar{\lambda}} + \delta_j(z) \right] \right\} \\ &= \lim_{z \rightarrow 1^-} (1-z) \left\{ \sum_{k=1}^i q_{j-i+k}^+(z)B^{k-1}(z) + \frac{B^{i-1}(z)}{\Delta(z)} \left[\frac{(1-p)f^j(z)B(z)}{1-z\bar{\lambda}} + \delta_j(z) \right] \right\} \\ &= \frac{(1-\rho)(1-p+\lambda\delta_j)}{pN_1 + (1-p)N_2}, \quad j = N_1, N_1 + 1, \dots, N_2 - 1, \\ p_j^+ &= \lim_{z \rightarrow 1^-} (1-z) \left[q_j^+(z)f(z) + \frac{f(z)\delta_j(z)}{\Delta(z)} \right] \\ &= \lim_{z \rightarrow 1^-} (1-z) \left[\sum_{k=1}^i q_{j-i+k}^+(z)B^{k-1}(z) + \frac{B^{i-1}(z)\delta_j(z)}{\Delta(z)} \right] \\ &= \frac{(1-\rho)\lambda\delta_j}{pN_1 + (1-p)N_2}, \quad j \geq N_2. \end{aligned}$$

For $\rho < 1$, we know

$$\sum_{j=0}^{\infty} p_j^+ = \frac{(1-\rho)}{pN_1 + (1-p)N_2} \left[pN_1 + (1-p)N_2 + \lambda \sum_{j=1}^{\infty} \delta_j \right]. \quad (D24)$$

By a calculation, we have that the following formulas

$$\sum_{j=1}^{\infty} q_j = \frac{\rho}{\lambda(1-\rho)} = E[b], \quad (D25)$$

and

$$\lambda \sum_{j=1}^{\infty} \delta_j = \lambda \left[p \sum_{j=1}^{\infty} \sum_{k=1}^{N_1} q_{j-N_1+k} + (1-p) \sum_{j=1}^{\infty} \sum_{k=1}^{N_2} q_{j-N_2+k} \right]$$

$$= \lambda[pN_1E[b] + (1-p)N_2E[b]]. \quad (\text{D26})$$

By substituting Equations (D25) and (D26) into Equation (D24) leads to $\sum_{j=0}^{\infty} p_j = 1$. That is, for $\rho = \lambda\mu < 1$, we can see that $\{p_j^+, j = 0, 1, 2, \dots\}$ is a probability distribution.

E Proof of Theorem 4

According to $P^+(z) = \sum_{j=0}^{\infty} z^j p_j^+$, we can get

$$P^+(z) = \frac{(1-\rho)}{pN_1 + (1-p)N_2} \left[\frac{1-z^{N_1}}{1-z} + (1-p) \frac{z^{N_1}-z^{N_2}}{1-z} + \lambda \sum_{j=1}^{\infty} \delta_j z^j \right]. \quad (\text{E27})$$

By a calculation, we can get

$$\begin{aligned} \sum_{j=1}^{\infty} \delta_j z^j &= p \sum_{j=1}^{\infty} z^j \sum_{k=1}^{N_1} q_{j-N_1+k} + (1-p) \sum_{j=1}^{\infty} z^j \sum_{k=1}^{N_2} q_{j-N_2+k} \\ &= p \left(\sum_{j=1}^{\infty} q_j z^j \right) \cdot \frac{1-z^{N_1}}{1-z} + (1-p) \left(\sum_{j=1}^{\infty} z^j q_j \right) \cdot \frac{1-z^{N_2}}{1-z} \\ &= \left(\sum_{j=1}^{\infty} q_j z^j \right) \cdot \left[p \frac{1-z^{N_1}}{1-z} + (1-p) \frac{1-z^{N_2}}{1-z} \right], \end{aligned} \quad (\text{E28})$$

$$\sum_{j=1}^{\infty} q_j z^j = \frac{z[1-G(z\lambda + \bar{\lambda})]}{\lambda[G(z\lambda + \bar{\lambda}) - z]}. \quad (\text{E29})$$

Thus, substituting Equations (E28) and (E29) into Equation (E27) to get Equation (13), and then Equation (14) can be derived by using $\bar{L}^+ = \frac{d}{dz}[P^+(z)]|_{z=1}$.