RAIRO-Oper. Res. 56 (2022) 395–414
https://doi.org/10.1051/ro/2022008

RAIRO Operations Research
www.rairo-ro.org

PERFORMANCE ANALYSIS OF AN $M/G/1$ QUEUE WITH BI-LEVEL RANDOMIZED (p, N_1, N_2)-POLICY

Xinyu Kuang[1], Yinghui Tang[1,*], Miaomiao Yu[1] and Wenqing Wu[2]

Abstract. This paper proposes an $M/G/1$ queueing model with bi-level randomized (p, N_1, N_2)-policy. That is, after all of the customers in the system are served, the server is closed down immediately. 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. Using the total probability decomposition technique and the Laplace transform, we study the transient queue length distribution and obtain the expressions of the Laplace transform of the transient queue-length distribution with respect to time t. Then, employing L'Hospital's rule and some algebraic operations, the explicit recursive formulas of the steady-state queue-length distribution, which can be used to accurately evaluate the probabilities of queue length, are presented. Moreover, some other important queuing performance indices, such as the explicit expressions of its probability generating function of the steady-state queue-length distribution, the expected queue size and so on, are derived. Meanwhile, we investigate the system capacity optimization design by the steady-state queue-length distribution. Finally, an operating cost function is constructed, and by numerical calculation, we find the minimum of the long-run average cost rate and the optimal bi-level threshold policy (N_1^*, N_2^*) that satisfies the average waiting time constraints.

Mathematics Subject Classification. 60K25, 68M20, 90B22.

Received June 2, 2021. Accepted January 11, 2022.

1. Introduction

As we know, the most extensive research regarding queueing models is the optimal design and control of the queue. The main objective of investigating controllable queueing systems is to economize the running cost and improve operational efficiency. In general, the issue of controlling the service includes the N-policy introduced by Yadin and Naor [26], the T-policy proposed by Heyman [1], and the D-policy presented by Balachandran [4]. Since their seminal works, considerable efforts have been devoted to study these types of controllable queueing models, such as Teghem [20], Tian and Zhang [21], Tang and Tang [18]. Considering the customer's sensitivity to time delays and increasing the flexibility of the service system, some queueing systems with the joint control policy have been presented. For example, Lee and Seo [13] considered the $M/G/1$ queueing system with the

Keywords. $M/G/1$ queue, (p, N_1, N_2)-policy, queue-length distribution, capacity optimization design, cost optimization.

[1] School of Mathematical Sciences, Sichuan Normal University, Chengdu 610068, Sichuan, P.R. China.

[2] School of Science, Southwest University of Science and Technology, Mianyang 621000, Sichuan, P.R. China.

*Corresponding author: tangyh@sicnu.edu.cn

dyadic $\text{Min}(N, D)$-policy combined with the N-policy and the D-policy, in which the server resumes its service if either N customers accumulate in the system or the sum of the service times of the waiting customers exceeds D, whichever occurs first. And Lee *et al.* [15] further extended the model to the $MAP/G/1$ system with the $\text{Min}(N, D)$-policy. Tang *et al.* [19] considered the $M/G/1$ queueing system with the $\text{Min}(N, V)$-policy combined with the N-policy and server multiple vacations. To prolong the lifetime of wireless sensor nodes, Jiang *et al.* [7] proposed a novel design strategy for mitigating the average power consumption of sensor nodes using the $M/G/1$ queueing model with $\text{Min}(N, T)$-policy. Lan and Tang [9] studied the optimal control strategy for a discrete-time $Geo/G/1$ queue in which the system operates under the control of multiple server vacations and $\text{Min}(N, V)$-policy. Li and Liu [16] studied an $M/G/1$ queue operating in a multi-phase random environment with $\text{Min}(N, V)$-policy. Using the supplementary variable technique, the distribution of the stationary system size at arbitrary epoch and the sojourn time distribution are derived. Lee and Park [12] considered a bi-level threshold control policy with set-up time. That is, the server starts the system immediately if there are $m(m \geq 1)$ customers accumulate in the system. When the system star up is complete, the server begins service at once if the number of waiting customers is no less than another given positive integer threshold $N(N \geq m)$. Otherwise, the server stays on standby until the number of waiting customers reaches N. Lee *et al.* [14] extended the model constructed by Lee and Park [12] to the batch arrival systems with/without server's vacations. In discrete-time case, Luo *et al.* [17] discussed the $Geo/G/1$ queue with (r, N)-policy and different input rates. Some other research on queueing systems with joint control policy can be found in [3, 8, 11, 24].

In the aforementioned papers about the N-policy queue, it is generally assumed that the idle server must resume service as soon as N customers accumulate in the system. However, in a certain situation (*e.g.*, the extra work that he/she is engaged in cannot be interrupted immediately), even if the number of customers has exactly accumulated to N, the dormant server may also not provide service for customers. Motivated by such a queuing phenomenon, Feinberg and Kim [2] first investigated an $M/G/1$ queueing system with (p, N)-policy, in which the (p, N)-policy means that when the number of customers in the system becomes N, the server is turned on with probability $p(0 \leq p \leq 1)$ or still left off with probability $1 - p$. The server begins serving immediately if there are more than N customers in the system. Later, the (p, N)-policy in the queueing system has a series of in-depth research [5, 6, 10, 22, 23, 25]. Wang and Huang [22] analyzed an $M/G/1$ queue with (p, N)-policy and unreliable server, and the maximum entropy approach was employed to develop the approximate formulae for the queue-length distribution. Wang and Ke [23] first introduced (p, N)-policy into the discrete-time $Geo/G/1$ queueing system. Jia and Chen [6] studied the $Geo/G/1$ queue model with (p, N)-policy, set-up time, multiple vacation and disasters. Jain and Kaur [5] considered an $M^x/G/1$ queueing model and studied the performance of phase service queue along with realistic features of the unreliable server, vacation, balking and feedback.

In many real-world production systems, the precise setting of the threshold for starting service is one of the most critical factors for the cost-effective operation of the systems. If the set threshold for starting service is too small, it will generate a large amount of switching costs when the system frequently switches its state for a long time. If the set threshold for starting service is too large, it will increase the waiting time of customers, which will degrade the customers satisfaction and lead to the loss of customers. Based on the above situation, this paper proposes a new $M/G/1$ queueing model with the bi-level randomized (p, N_1, N_2)-policy. That is, 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. First of all, employing the total probability decomposition and the Laplace transform, we study the transient queue-length distribution of the system starting from any initial state, and obtain the expressions of the Laplace transform of the transient queue-length distribution. Secondly, applying L'Hospital's rule and using some algebraic operations, the recursive formulas of the steady-state queue-length distribution and the average queue size are obtained. Thirdly, we use the recursive formulas to discuss the system capacity optimization design. Finally, we establish the cost structure model and obtain the explicit expression of the long-run expected cost rate of the system by the renewal reward theorem. Moreover, a numerical example is used to discuss the constrained optimization problem under the limit of the average waiting time. It is worthwhile to point out that the analysis technique used in this paper is different from some traditional analysis techniques (*e.g.*, embedded Markov chain and supplementary variable techniques). Using the method proposed in our paper, we can investigate the transient queue size distribution at any epoch t of the system that starts from any initial state. Based on the transient results, we can easily obtain the recursive formula of the steady-state queue length distribution that is numerically tractable. Compared with our method, the traditional approaches such as the supplementary variable technique and the embedded Markov chain method can only obtain the probability generating function of the stationary queue size instead of the queue-length distribution. Moreover, to analyze the evolution of the queueing process by using the supplementary variable technique, the difference-differential equations governing the $M/G/1$ queueing model can be established under the assumption that the service time follows an arbitrary distribution with a probability density function.

The remainder of this paper is originated as follows. In the next section, we formulate the considered mathematical model and give some definitions and lemmas. In Section 3, some queueing characteristics are analyzed. Some special cases are given in Section 4. The system capacity optimization design is discussed in Section 5. In Section 6, applying the renewal reward theorem, we obtain the explicit expression of the long-run expected cost rate under a given cost model. Then, we construct a constrained optimization problem under the limit of the average waiting time and discuss the optimal bi-level threshold policy that minimizes the expected cost function by a numerical example. Finally, Section 7 provides the conclusions.

2. Model formulation and some preliminaries

In this paper, we investigate the $M/G/1$ queueing model with bi-level randomized (p, N_1, N_2)-policy by making the following assumptions.

Assumption 2.1. *The inter-arrival times $\tau_n, n = 1, 2, \cdots$, are independent identically distributed (i.i.d.) random variables each with the exponential distribution $F(t) = 1 - e^{-\lambda t}$, $\lambda > 0, t \geq 0$. The customers are served one by one. The service times $\chi_n, n = 1, 2, \cdots$, are also i.i.d. random variables each with the distribution $G(t), t \geq 0$, and the average service time is $\frac{1}{\mu}, 0 < \mu < \infty$. The inter-arrival time τ and service time χ are all independent of each other.*

Assumption 2.2. *The bi-level randomized (p, N_1, N_2)-policy: Whenever the system is empty, the server keeps dormant in the system. If the number of customers reaches $N_1(\geq 1)$ in the system, the deactivated server is turned on with probability $p(0 \leq p \leq 1)$ or is 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.*

Assumption 2.3. *The server will not adopt the control policy if the system is empty at initial time $t = 0$, then the server will stay in the system until the next customer arrives and starts its service immediately. If there are $j(\geq 1)$ customers in the system at initial time $t = 0$, the server begins serving the customers immediately.*

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

Definition 2.4. System idle period is a period of time during which the system is continuously idle (no customers). Obviously, the system idle period is the remaining time of an arrival interval. Let $\hat{\tau}_j$ represent the length of the jth system idle period. Thus, $\hat{\tau}_j (j = 1, 2, \cdots)$ are independent mutually and satisfy the same exponential distribution $F(t) = 1 - e^{-\lambda t}, t \geq 0$.

Definition 2.5. System busy period is the time interval that starts at the instant at which the first customer arriving at the idle system and ends at the instant when the system becomes empty again.

Definition 2.6. Server idle period is the time interval commences when the system is completely empty and finishes when server begins to serve the waiting customers.

Definition 2.7. Server busy period is the time interval from the server begins service until the system becomes empty again. Let b denote the server busy period that begins with only one customer, and let $B(t) = P\{b \le t\}, b(s) = \int_0^\infty e^{-st}\mathrm{d}B(t)$. Then, similar to the discussing in Tang *et al.* [18], we have the following lemma.

Lemma 2.8. *For $\Re(s) > 0$, $b(s)$ is the root with the smallest absolute value of the equation $z = g(s + \lambda - \lambda z)$ within $|z| < 1$, and*

$$B(t) = \sum_{k=1}^{\infty} \int_0^t e^{-\lambda x} \frac{(\lambda x)^{k-1}}{k!} \mathrm{d}G^{(k)}(x),$$

$$\lim_{t\to\infty} B(t) = \begin{cases} 1, & \rho \le 1, \\ \omega < 1, & \rho > 1, \end{cases} \quad E(b) = \begin{cases} \frac{\rho}{\lambda(1-\rho)}, & \rho < 1, \\ \infty, & \rho \ge 1, \end{cases}$$

where $g(s) = \int_0^\infty e^{-st}\mathrm{d}G(t)$, $\omega(0 < \omega < 1)$ is the root of the equation $z = g(\lambda - \lambda z)$. The traffic intensity of system is given by $\rho = \frac{\lambda}{\mu}$. $\Re(s)$ denotes the real part of the complex variable s. Furthermore, $G^{(k)}(t)$ is the k-fold convolution of $G^{(k)}(t)$, i.e., $G^{(k)}(t) = \int_0^t G^{(k-1)}(t - x)\mathrm{d}G(x), k \ge 1$ and $G^{(0)}(t) = 1$.

Let $b^{(i)}$ be the generalized server busy period initiated with i customers. Because the arrival process is a Poisson process, the probability distribution function of $b^{(i)}$ is given by $P\{b^{(i)} \le t\} = B^{(i)}(t)$, $t \ge 0, i \ge 1$.

We now define the joint distribution of the queue-length and the server busy period by

$$Q_j(t) = P\{b > t \ge 0; N(t) = j\}.$$

So that $Q_j(t)$ denotes the transient probability of there being j customers at epoch t during the time interval $(0, b]$. It implies that the state $N(t) = j$ in $Q_j(t)$ depends on the initial state $N(0) = 1$, $(0, t] \subset (0, b]$ and the epoch $t = 0$ is the beginning of the server period b. We have boundary condition $Q_1(0) = 1, Q_j(0) = 0, j > 1$. Similar to the discussing in Tang *et al.* [18], we have the following lemma.

Lemma 2.9. *Let $q_j^*(s) = \int_0^\infty e^{-st}Q_j(t)\mathrm{d}t$ be the Laplace transform of $Q_j(t)$,then for $\Re(s) > 0$ and $j \ge 1$, we have*

$$q_1^*(s) = \frac{b(s)[1 - g(s+\lambda)]}{(s+\lambda)g(s+\lambda)},$$

$$q_j^*(s) = \frac{b(s)}{g(s+\lambda)} \int_0^\infty e^{-st}\bar{G}(t)\frac{(\lambda t)^{j-1}}{(j-1)!}e^{-\lambda t}\mathrm{d}t + \frac{1}{g(s+\lambda)} \sum_{k=1}^{j-1} \frac{q_{j-k}^*(s)}{b^k(s)} \left\{ b(s) - \sum_{i=0}^{k} \int_0^\infty e^{-(s+\lambda)t} \frac{[\lambda b(s)t]^i}{i!} \mathrm{d}G(t) \right\}, j > 1,$$

where $b(s)$ is defined by Lemma 2.8, $\sum_{k=i}^{j} = 0$ if $j < i$.

Proof. See Appendix A. □

3. Transient and steady-state distribution of system queue size

In this section, using the renewal theory, the law of total probability decomposition and the Laplace transform, we first investigate the transient distribution of the queue-length and derive the expressions of the Laplace transform of the transient queue-length distribution with respect to time t. Then, employing L'Hospital's rule, we obtain the recursive formulas of the steady-state queue-length distribution. Moreover, some other important queueing performance indices are derived by some algebraic manipulations.

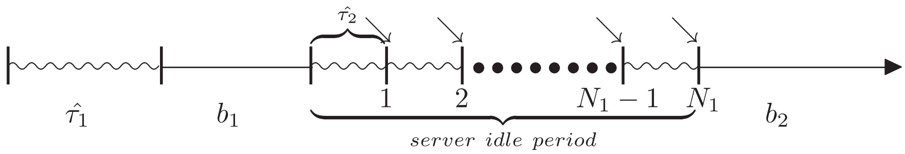

FIGURE 1. When N_1 customers arrive after the system becomes empty, the server is activated for service with probability p.

Let

$$p_{ij}(t) = P\{N(t) = j|N(0) = i\}$$

denote the conditional probability that there are j customers at time point t under initial state $N(0) = i(i = 0, 1, 2, \cdots)$, and

$$p^*_{ij}(s) = \int_0^\infty e^{-st} p_{ij}(t)\mathrm{d}t, \qquad i \geq 0, j \geq 0.$$

Theorem 3.1. *For $\Re(s) > 0$ and $i \geq 1$, we have*

$$p^*_{00}(s) = \frac{1 - f(s)}{s}\left[1 + \frac{f(s)b(s)}{\Delta\ (s)}\right], \tag{3.1}$$

$$p^*_{i0}(s) = \frac{1 - f(s)}{s} \cdot \frac{b^i(s)}{\Delta\ (s)}, \tag{3.2}$$

where $\Delta(s) = 1 - pf^{N_1}(s)b^{N_1}(s) - (1 - p)f^{N_2}(s)b^{N_2}(s)$, and $b(s)$ is defined by Lemma 2.8.

Proof. Let $l_n = \sum_{i=1}^n \tau_i, n \geq 1$, and $l_0 = 0$. It is noted that $p_{00}(t)$ indicates that there are no customers in the system at time point t under initial state $N(0) = 0$, that is, the time point t is located in the system idle period. Since the beginning and ending epochs of the server busy period are renewal points, using the total probability decomposition, it gets

$$\begin{aligned} p_{00}(t) &= P\{\hat{\tau}_1 > t\} + P\{\hat{\tau}_1 + b \leq t < \hat{\tau}_1 + b + \hat{\tau}_2\} + P\{\hat{\tau}_1 + b + \hat{\tau}_2 \leq t, N(t) = 0\} \\ &= \bar{F}(t) + \int_0^t \bar{F}(t - x)\mathrm{d}[F(x) * B(x)] + P\{\hat{\tau}_1 + b + \hat{\tau}_2 \leq t, N(t) = 0\}. \end{aligned} \tag{3.3}$$

The third term of equation (3.3) can be decomposed into the following two cases:

(i) When N_1 customers arrive after the system becomes empty, the server is activated for service with probability $p(0 \leq p \leq 1)$ (see Fig. 1).
(ii) When N_1 customers arrive after the system becomes empty, the server is not activated for service with probability $1 - p$, and then the server is turned on by the arrival of the N_2th customer in the system (see Fig. 2).

So, the third term of equation (3.3) is given by

$$\begin{aligned} P\{\hat{\tau}_1 + b + \hat{\tau}_2 \leq t, N(t) = 0\} &= pP\{\hat{\tau}_1 + b + \hat{\tau}_2 + l_{N_1-1} \leq t, N(t) = 0\} \\ &\quad + (1 - p)P\{\hat{\tau}_1 + b + \hat{\tau}_2 + l_{N_2-1} \leq t, N(t) = 0\} \\ &= p\int_0^t \int_0^{t-x} p_{N_1 0}(t - x - y)\mathrm{d}F^{(N_1)}(y)\mathrm{d}[F(x) * B(x)] \end{aligned}$$

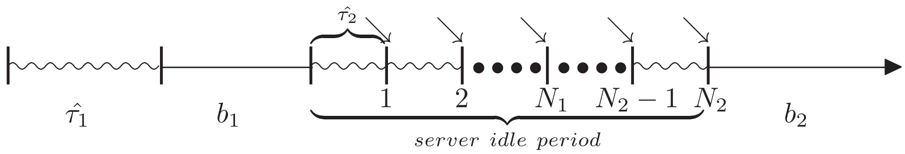

FIGURE 2. The server is turned on by the arrival of the N_2th customer.

$$+ (1-p)\int_0^t \int_0^{t-x} p_{N_2 0}(t-x-y)\mathrm{d}F^{(N_2)}(y)\mathrm{d}[F(x) * B(x)]. \tag{3.4}$$

Substituting equation (3.4) into equation (3.3) and applying the Laplace transform to equation (3.3), $p_{00}^*(s)$ is given by

$$p_{00}^*(s) = \frac{[1-f(s)][1+b(s)f(s)]}{s} + pb(s)f^{N_1+1}(s)p_{N_1 0}^*(s) + (1-p)b(s)f^{N_2+1}(s)p_{N_2 0}^*(s). \tag{3.5}$$

For $i > 0$, similar to the discussion of $p_{00}(t)$, we have

$$\begin{aligned} p_{i0}(t) &= P\left\{b^{(i)} \le t < b^{(i)} + \hat{\tau}_1\right\} + P\left\{b^{(i)} + \hat{\tau}_1 \le t, N(t) = 0\right\} \\ &= \int_0^t \bar{F}(t-x)\mathrm{d}B^{(i)}(x) + p\int_0^t \int_0^{t-x} p_{N_1 0}(t-x-y)\mathrm{d}F^{(N_1)}(y)\mathrm{d}B^{(i)}(x) \\ &\quad + (1-p)\int_0^t \int_0^{t-x} p_{N_2 0}(t-x-y)\mathrm{d}F^{(N_2)}(y)\mathrm{d}B^{(i)}(x). \end{aligned} \tag{3.6}$$

By taking the Laplace transform of equation (3.6), we obtain

$$p_{i0}^*(s) = \frac{[1-f(s)]b^i(s)}{s} + pb^i(s)f^{N_1}(s)p_{N_1 0}^*(s) + (1-p)b^i(s)f^{N_2}(s)p_{N_2 0}^*(s). \tag{3.7}$$

From equations (3.5) and (3.7), the relationship between $p_{00}^*(s)$ and $p_{i0}^*(s)$ can be obtained as follows

$$p_{i0}^*(s) = \frac{b^i(s)}{b(s)f(s)}\left\{p_{00}^*(s) - \frac{1-f(s)}{s}\right\}, \qquad i \ge 1. \tag{3.8}$$

Substituting equation (3.8) back to equation (3.5) and solving equation (3.5) gives equation (3.1). Substituting equation (3.1) into equation (3.7) leads to equation (3.2). □

Theorem 3.2. *For $\Re(s) > 0$ and $i \ge 1$, we have*

(1) *If $j = 1, 2, \cdots, N_1 - 1$, then*

$$p_{0j}^*(s) = q_j^*(s)f(s) + \frac{[1-f(s)]f^{j+1}(s)b(s) + sf(s)\delta_j(s)}{s\Delta(s)}, \tag{3.9}$$

$$p_{ij}^*(s) = \sum_{k=1}^{i} q_{j-i+k}^*(s)b^{k-1}(s) + \frac{[1-f(s)]f^j(s)b^i(s) + sb^{i-1}(s)\delta_j(s)}{s\Delta(s)}. \tag{3.10}$$

(2) *If $j = N_1, \cdots, N_2 - 1$, then*

$$p_{0j}^*(s) = q_j^*(s)f(s) + \frac{(1-p)[1-f(s)]f^{j+1}(s)b(s) + sf(s)\delta_j(s)}{s\Delta(s)}, \tag{3.11}$$

$$p_{ij}^*(s) = \sum_{k=1}^{i} q_{j-i+k}^*(s)b^{k-1}(s) + \frac{(1-p)[1-f(s)]f^j(s)b^i(s) + sb^{i-1}(s)\delta_j(s)}{s\Delta(s)}. \tag{3.12}$$

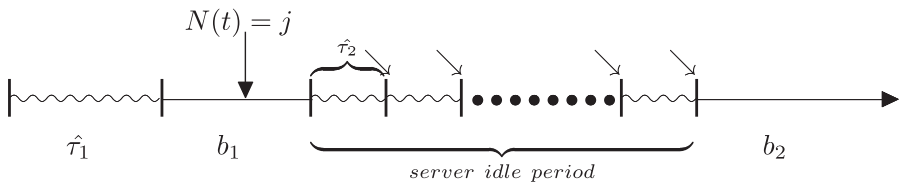

FIGURE 3. The time point t is located in the server idle period with j customers.

(3) *If* $j = N_2, N_2+1, N_2+2, \cdots,$ *then*

$$p_{0j}^*(s) = q_j^*(s)f(s) + \frac{f(s)\delta_j(s)}{\Delta(s)}, \tag{3.13}$$

$$p_{ij}^*(s) = \sum_{k=1}^{i} q_{j-i+k}^*(s)b^{k-1}(s) + \frac{b^{i-1}(s)\delta_j(s)}{\Delta(s)}, \tag{3.14}$$

where $\Delta(s)$ *is given in Theorem 3.1, and* $\delta_j(s) = pf^{N_1}(s)\sum_{k=1}^{N_1} q_{j-N_1+k}^*(s)b^k(s) + (1-p)f^{N_2}(s)\sum_{k=1}^{N_2} q_{j-N_2+k}^*(s)b^k(s)$.

Proof. (1) When $j = 1, 2, \cdots, N_1 - 1$, it is noted that $p_{0j}(t)$ indicates that there are j customers in the system at time point t under initial state $N(0) = 0$. Therefore, the event $\{N_t = j\}$ must satisfy one of the following conditions:

(i) The time point t is located in the server busy period and there are j customers in the system at time point t.

(ii) The time point t is located in the server idle period and there are j customers in the system at time point t.

Thus

$$\begin{aligned} p_{0j}(t) = {} & P\{\hat{\tau}_1 \le t < \hat{\tau}_1 + b_1, N(t) = j\} + P\{\hat{\tau}_1 + b_1 + \hat{\tau}_2 + l_{j-1} \le t < \hat{\tau}_1 + b_1 + \hat{\tau}_2 + l_j\} \\ & + pP\{\hat{\tau}_1 + b_1 + \hat{\tau}_2 + l_{N_1-1} < t, N(t) = j\} + (1-p)P\{\hat{\tau}_1 + b_1 + \hat{\tau}_2 + l_{N_2-1} < t, N(t) = j\}. \end{aligned} \tag{3.15}$$

The first term of equation (3.15) indicates that the time point t is located in the first server busy period with j customers (see Fig. 3).

So, it yields

$$P\{\hat{\tau}_1 \le t < \hat{\tau}_1 + b_1, N(t) = j\} = \int_0^t Q_j(t-x)\mathrm{d}F(x). \tag{3.16}$$

The second term of equation (3.15) indicates that the time point t is located in the server idle period with j customers (see Fig. 4).

So that

$$P\{\hat{\tau}_1 + b_1 + \hat{\tau}_2 + l_{j-1} \le t < \hat{\tau}_1 + b_1 + \hat{\tau}_2 + l_j\} = \int_0^t \bar{F}(t-x)\mathrm{d}\Big[F^{(j+1)}(x) * B(x)\Big]. \tag{3.17}$$

The third term of equation (3.15) indicates that the server is activated for service with probability p when N_1 customers arrive after the system becomes empty, and the time point t is located in after the start of the second server busy period with j customers (see Fig. 5).

It yields

$$pP\{\hat{\tau}_1 + b_1 + \hat{\tau}_2 + l_{N_1-1} < t, N(t) = j\} = p\int_0^t \int_0^{t-x} p_{N_1 j}(t-x-y)\mathrm{d}F^{(N_1)}(y)\mathrm{d}[F(x) * B(x)]. \tag{3.18}$$

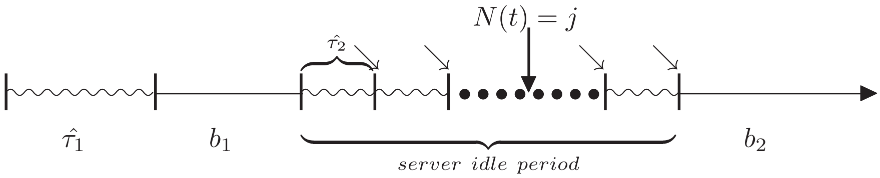

FIGURE 4. The time point t is located in the first server busy period with j customers.

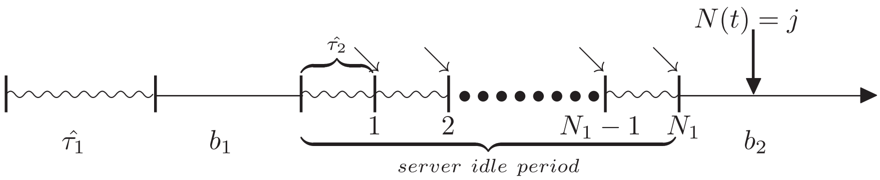

FIGURE 5. The server is activated for service with probability p when N_1 customers arrive, and the time point t is located in after the start of the second server busy period with j customers.

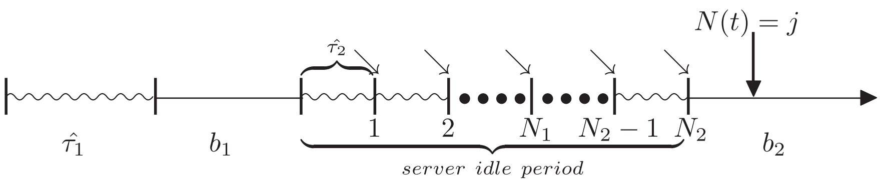

FIGURE 6. The server is not activated for service with probability $1-p$ when N_1 customers arrive, and the time point t is located in after the start of the second server busy period with j customers.

The fourth term of equation (3.15) indicates that the server is not activated for service with probability $1-p$ when N_1 customers arrive after the system becomes empty, but the server is turned on by the arrival of the N_2th customer, and the time point t is located in after the start of the second server busy period with j customers (see Fig. 6).
Therefore

$$(1-p)P\{\hat{\tau}_1+b_1+\hat{\tau}_2+l_{N_2-1}<t, N(t)=j\}=(1-p)\int_0^t\int_0^{t-x} p_{N_2j}(t-x-y)\mathrm{d}F^{(N_2)}(y)\mathrm{d}[F(x)*B(x)]. \tag{3.19}$$

Substituting equations (3.16)–(3.19) into equation (3.15), and applying the Laplace transform to equation (3.15), $p_{0j}^*(s)$ is given by

$$p_{0j}^*(s)=q_j^*(s)f(s)+\frac{[1-f(s)]f^{j+1}(s)b(s)}{s}+pb(s)f^{N_1+1}(s)p_{N_1j}^*(s)+(1-p)b(s)f^{N_2+1}(s)p_{N_2j}^*(s). \tag{3.20}$$

For $i>0$, similar to the discussion of $p_{0j}(t)$, we have

$$\begin{aligned}p_{ij}(t)=P\Big\{t\le b^{(i)}, N(t)=j\Big\}+P\Big\{b^{(i)}+\hat{\tau}_1+l_{j-1}\le t<b^{(i)}+\hat{\tau}_1+l_j\Big\}\\+pP\Big\{b^{(i)}+l_{N_1}<t, N(t)=j\Big\}+(1-p)P\Big\{b^{(i)}+l_{N_2}<t, N(t)=j\Big\}\end{aligned}$$

$$= \sum_{k=1}^{i} \int_0^t Q_{j-i+k}(t-x)\mathrm{d}B^{(k-1)}(x) + \int_0^t \bar{F}(t-x)\mathrm{d}\Big[F^{(j)}(x) * B^{(i)}(x)\Big]$$
$$+ p \int_0^t \int_0^{t-x} p_{N_1 j}(t-x-y)\mathrm{d}F^{(N_1)}(y)\mathrm{d}B^{(i)}(x)$$
$$+ (1-p) \int_0^t \int_0^{t-x} p_{N_2 j}(t-x-y)\mathrm{d}F^{(N_2)}(y)\mathrm{d}B^{(i)}(x). \tag{3.21}$$

By taking the Laplace transform of equation (3.21), we obtain

$$p_{ij}^*(s) = \sum_{k=1}^{i} q_{j-i+k}^*(s) b^{k-1}(s) + \frac{[1-f(s)]f^j(s)b^i(s)}{s} + pb^i(s)f^{N_1+1}(s)p_{N_1 j}^*(s) + (1-p)b^i(s)f^{N_2+1}(s)p_{N_2 j}^*(s). \tag{3.22}$$

From equations (3.20) and (3.22), the relationship between $p_{0j}^*(s)$ and $p_{ij}^*(s)$ can be obtained as

$$p_{ij}^*(s) = \sum_{k=1}^{i} q_{j-i+k}^*(s) b^{k-1}(s) + \frac{b^{i-1}(s)}{f(s)}\big\{p_{0j}^*(s) - q_j^*(s)f(s)\big\}, \qquad i \geq 1. \tag{3.23}$$

Substituting equation (3.23) back to equation (3.20) and solving equation (3.20) gives equation (3.9). Substituting equation (3.9) into equation (3.22) leads to equation (3.10).

(2) For $j = N_1, \cdots, N_2 - 1$, we have

$$p_{0j}(t) = P\{\hat{\tau}_1 \leq t < \hat{\tau}_1 + b_1, N(t) = j\} + (1-p)P\{\hat{\tau}_1 + b_1 + \hat{\tau}_2 + l_{N_1-1} + l_{j-N_1} \leq t < \hat{\tau}_1 + b_1 + \hat{\tau}_2 + l_{N_1-1}$$
$$+ l_{j-N_1+1}\} + pP\{\hat{\tau}_1 + b_1 + l_{N_1} < t, N(t) = j\} + (1-p)P\{\hat{\tau}_1 + b_1 + l_{N_2} < t, N(t) = j\}$$
$$= \int_0^t Q_j(t-x)\mathrm{d}F(x) + (1-p)\int_0^t \int_0^{t-x} \bar{F}(t-x-y)\mathrm{d}F^{(j-N_1)}(y)\mathrm{d}\Big[F^{(N_1+1)}(x) * B(x)\Big]$$
$$+ p \int_0^t \int_0^{t-x} p_{N_1 j}(t-x-y)\mathrm{d}F^{(N_1)}(y)\mathrm{d}[F(x) * B(x)]$$
$$+ (1-p) \int_0^t \int_0^{t-x} p_{N_2 j}(t-x-y)\mathrm{d}F^{(N_2)}(y)\mathrm{d}[F(x) * B(x)] \tag{3.24}$$

$$p_{ij}(t) = P\Big\{t \leq b^{(i)}, N(t) = j\Big\} + (1-p)P\Big\{b^{(i)} + \hat{\tau}_1 + l_{N_1-1} + l_{j-N_1} \leq t < b^{(i)} + \hat{\tau}_1 + l_{N_1-1} + l_{j-N_1+1}\Big\}$$
$$+ pP\Big\{b^{(i)} + l_{N_1} < t, N(t) = j\Big\} + (1-p)P\Big\{b^{(i)} + l_{N_2} < t, N(t) = j\Big\}$$
$$= \sum_{k=1}^{i} \int_0^t Q_{j-i+k}(t-x)\mathrm{d}B^{(k-1)}(x) + (1-p)\int_0^t \int_0^{t-x} \bar{F}(t-x-y)\mathrm{d}F^{(j-N_1)}(y)\mathrm{d}\Big[F^{(N_1)}(x) * B^{(i)}(x)\Big]$$
$$+ p \int_0^t \int_0^{t-x} p_{N_1 j}(t-x-y)\mathrm{d}F^{(N_1)}(y)\mathrm{d}B^{(i)}(x)$$
$$+ (1-p) \int_0^t \int_0^{t-x} p_{N_2 j}(t-x-y)\mathrm{d}F^{(N_2)}(y)\mathrm{d}B^{(i)}(x). \tag{3.25}$$

Taking the Laplace transform of equations (3.24) and (3.25), we have

$$p_{0j}^*(s) = q_j^*(s)f(s) + \frac{(1-p)[1-f(s)]f^{j+1}(s)b(s)}{s} + pb(s)f^{N_1+1}(s)p_{N_1 j}^*(s)$$
$$+ (1-p)b(s)f^{N_2+1}(s)p_{N_2 j}^*(s), \tag{3.26}$$

$$p_{ij}^*(s) = \sum_{k=1}^{i} q_{j-i+k}^*(s) b^{k-1}(s) + \frac{(1-p)[1-f(s)]f^j(s)b^i(s)}{s} + pb^i(s)f^{N_1+1}(s)p_{N_1 j}^*(s)$$
$$+ (1-p)b^i(s)f^{N_2+1}(s)p_{N_2 j}^*(s). \tag{3.27}$$

From equations (3.26) and (3.27), we obtain the relationship between $p_{0j}^*(s)$ and $p_{ij}^*(s)$ as follows

$$p_{ij}^*(s) = \sum_{k=1}^{i} q_{j-i+k}^*(s) b^{k-1}(s) + \frac{b^{i-1}(s)}{f(s)} \{p_{0j}^*(s) - q_j^*(s)f(s)\}, \qquad i \geq 1. \tag{3.28}$$

Substituting equation (3.28) back to equation (3.26) and solving equation (3.26) gives equation (3.11). Substituting equation (3.11) into equation (3.27) leads to equation (3.12).

(3) For $j = N_2, N_2+1, N_2+2, \cdots$, we have

$$\begin{aligned} p_{0j}(t) &= P\{\hat{\tau}_1 \leq t < \hat{\tau}_1 + b_1, N(t) = j\} + pP\{\hat{\tau}_1 + b_1 + l_{N_1} < t, N(t) = j\} \\ &\quad + (1-p)P\{\hat{\tau}_1 + b_1 + l_{N_2} < t, N(t) = j\} \\ &= \int_0^t Q_j(t-x)\mathrm{d}F(x) + p\int_0^t \int_0^{t-x} p_{N_1 j}(t-x-y)\mathrm{d}F^{(N_1)}(y)\mathrm{d}[F(x)*B(x)] \\ &\quad + (1-p)\int_0^t \int_0^{t-x} p_{N_2 j}(t-x-y)\mathrm{d}F^{(N_2)}(y)\mathrm{d}[F(x)*B(x)] \end{aligned} \tag{3.29}$$

$$\begin{aligned} p_{ij}(t) &= P\Big\{t \leq b^{(i)}, N(t) = j\Big\} + pP\Big\{b^{(i)} + l_{N_1} < t, N(t) = j\Big\} + (1-p)P\Big\{b^{(i)} + l_{N_2} < t, N(t) = j\Big\} \\ &= \sum_{k=1}^{i} \int_0^t Q_{j-i+k}(t-x)\mathrm{d}B^{(k-1)}(x) + p\int_0^t \int_0^{t-x} p_{N_1 j}(t-x-y)\mathrm{d}F^{(N_1)}(y)\mathrm{d}B^{(i)}(x) \\ &\quad + (1-p)\int_0^t \int_0^{t-x} p_{N_2 j}(t-x-y)\mathrm{d}F^{(N_2)}(y)\mathrm{d}B^{(i)}(x). \end{aligned} \tag{3.30}$$

By taking the Laplace transform of equations (3.29) and (3.30), we have

$$p_{0j}^*(s) = q_j^*(s)f(s) + pb(s)f^{N_1+1}(s)p_{N_1 j}^*(s) + (1-p)b(s)f^{N_2+1}(s)p_{N_2 j}^*(s), \tag{3.31}$$

$$p_{ij}^*(s) = \sum_{k=1}^{i} q_{j-i+k}^*(s) b^{k-1}(s) + pb^i(s)f^{N_1+1}(s)p_{N_1 j}^*(s) + (1-p)b^i(s)f^{N_2+1}(s)p_{N_2 j}^*(s). \tag{3.32}$$

From equations (3.31) and (3.32), we obtain the relationship between $p_{0j}^*(s)$ and $p_{ij}^*(s)$ as follows

$$p_{ij}^*(s) = \sum_{k=1}^{i} q_{j-i+k}^*(s) b^{k-1}(s) + \frac{b^{i-1}(s)}{f(s)} \{p_{0j}^*(s) - q_j^*(s)f(s)\}, \qquad i \geq 1. \tag{3.33}$$

Substituting equation (3.33) back to equation (3.31) and solving equation (3.31) gives equation (3.13). Substituting equation (3.13) into equation (3.32) leads to equation (3.14).

□

On the basis of the expressions of the Laplace transform of the transient queue-length distribution obtained in Theorems 3.1 and 3.2 above, employing L'Hospital's rule, we obtain the explicit recursive formulas of the steady-state queue-length distribution. Moreover, some other important queueing performance indices, such as the explicit expressions of its probability generating function of the steady-state queue-length distribution, the expected queue size, the stochastic decomposition structure of the steady-state queue size and so on, are derived by some algebraic manipulations.

Theorem 3.3. *Let* $p_j = \lim_{t\to\infty} p\{N(t) = j\}, j = 0, 1, 2\cdots,$ *we have*

(1) *When* $\rho = \frac{\lambda}{\mu} \geq 1$, $p_j = 0, j = 0, 1, 2\cdots$. *Therefore,* $\{p_j, j \geq 0\}$ *do not constitute a probability distribution.*
(2) *When* $\rho < 1$, *the recursive expressions of* $\{p_j, j = 0, 1, 2, \cdots\}$ *are listed as follows*

$$p_0 = \frac{(1-\rho)}{pN_1 + (1-p)N_2}, \tag{3.34}$$
$$p_j = \frac{(1-\rho)(1+\lambda\delta_j)}{pN_1 + (1-p)N_2}, \qquad j = 1, 2, \cdots, N_1 - 1, \tag{3.35}$$
$$p_j = \frac{(1-\rho)(1-p+\lambda\delta_j)}{pN_1 + (1-p)N_2}, \qquad j = N_1, N_1+1, \cdots, N_2 - 1, \tag{3.36}$$
$$p_j = \frac{(1-\rho)\lambda\delta_j}{pN_1 + (1-p)N_2}, \qquad j \geq N_2, \tag{3.37}$$

and $\{p_j, j = 0, 1, 2, \cdots\}$ *forms a probability distribution, where* $\delta_j = p\sum_{k=1}^{N_1} q_{j-N_1+k} + (1-p)\sum_{k=1}^{N_2} q_{j-N_2+k}, j \geq 1$, $q_j = \frac{1}{g(\lambda)}\int_0^\infty e^{-\lambda t}\frac{(\lambda t)^{j-1}}{(j-1)!}\bar{G}(t)\mathrm{d}t + \frac{1}{g(\lambda)}\sum_{k=1}^{j-1} q_{j-k}\Big\{1 - \sum_{i=0}^{k}\int_0^\infty e^{-\lambda t}\frac{(\lambda t)^i}{i!}\mathrm{d}G(t)\Big\}, j \geq 1.$

Proof. It is noted that $p_j = \sum_{i=0}^\infty p\{N(0) = i\}\cdot \lim\limits_{t\to\infty} p_{ij}(t)$, and

$$\begin{aligned}\lim_{t\to\infty} p_{ij}(t) &= \lim_{t\to\infty}\int_0^x \mathrm{d}p_{ij}(x) + p_{ij}(0)\\ &= \lim_{t\to\infty}\lim_{s\to 0^+}\int_0^x e^{-sx}\mathrm{d}p_{ij}(x) + p_{ij}(0)\\ &= \lim_{s\to 0^+}\int_0^\infty e^{-sx}\mathrm{d}p_{ij}(x) + p_{ij}(0) = \lim_{s\to 0^+} sp^*_{ij}(s).\end{aligned} \tag{3.38}$$

So we only need to calculate $\lim\limits_{s\to 0^+} sp^*_{ij}(s)$.

(1) When $\rho = \frac{\lambda}{\mu} > 1$, we know $\lim\limits_{s\to 0^+} b(s) = \omega(0 < \omega < 1)$ from Lemma 2.8. So we have

$$\lim_{s\to 0^+}\Delta(s) = 1 - p\omega^{N_1} - (1-p)\omega^{N_2} \neq 0.$$

Therefore, by combining the expressions of Theorem 3.1 and 3.2, we obtain $\lim\limits_{s\to 0^+} sp^*_{ij}(s) = 0$ $(i, j = 0, 1, 2, \cdots)$. So that $p_j = 0, j = 0, 1, 2\cdots$.
(2) When $\rho = \frac{\lambda}{\mu} = 1$, we know $\lim\limits_{s\to 0^+} b(s) = 1$ and $E[b] = \infty$. From Lemma 2.8, we get

$$\lim_{s\to 0^+}\Delta(s) = 1 - p - (1-p) = 0, \tag{3.39}$$

and

$$\lim_{s\to 0^+}\Delta'(s) = pN_1E[b] + \frac{p}{\lambda}N_1 + pN_2E[b] + \frac{p}{\lambda}N_2 = \infty. \tag{3.40}$$

Using L'Hospital's rule leads to $\lim\limits_{s\to 0^+} sp^*_{ij}(s) = 0$ $(i, j = 0, 1, 2, \cdots)$. Therefore, for $\rho = \frac{\lambda}{\mu} = 1$, we obtain $p_j = 0, j = 0, 1, 2\cdots$.

(3) When $\rho = \frac{\lambda}{\mu} < 1$, we know $\lim\limits_{s\to 0^+} b(s) = 1$ and $E[b] = \frac{\rho}{\lambda(1-\rho)}$ from Lemma 2.8, so we get

$$\begin{aligned}\lim_{s\to 0^+} \Delta'(s) &= pN_1\left[E[b] + \frac{1}{\lambda}\right] + (1-p)N_2\left[E[b] + \frac{1}{\lambda}\right] \\ &= \frac{pN_1 + (1-p)N_2}{\lambda(1-\rho)}.\end{aligned} \tag{3.41}$$

Applying L'Hospital's rule again, the recursive expressions of $\{p_j, j = 0, 1, 2, \cdots\}$ can be obtained by a direct calculation.

When $\rho = \frac{\lambda}{\mu} < 1$, since

$$\sum_{j=0}^{\infty} p_j = \frac{(1-\rho)}{pN_1 + (1-p)N_2}\left[pN_1 + (1-p)N_2 + \lambda\sum_{j=1}^{\infty}\delta_j\right]. \tag{3.42}$$

By a calculation, we have that the following formulas

$$\begin{aligned}\lambda\sum_{j=1}^{\infty}\delta_j &= \lambda\left[p\sum_{j=1}^{\infty}\sum_{k=1}^{N_1} q_{j-N_1+k} + (1-p)\sum_{j=1}^{\infty}\sum_{k=1}^{N_2} q_{j-N_2+k}\right] \\ &= \lambda[pN_1E[b] + (1-p)N_2E[b]]\end{aligned} \tag{3.43}$$

and

$$\sum_{j=1}^{\infty} q_j = \frac{1}{\mu-\lambda} = E[b] \tag{3.44}$$

hold. Substituting equations (3.43) and (3.44) into equation (3.42) lead to $\sum_{j=0}^{\infty} p_j = 1$. That is, for $\rho = \frac{\lambda}{\mu} < 1$, the equilibrium distribution $\{p_j, j = 0, 1, 2, \cdots\}$ at epoch t exists and forms a probability distribution. □

Theorem 3.4. *When $\rho < 1$, let $P(z)$ denote the probability generating function of $\{p_j, j = 0, 1, 2, \cdots\}$, then*

$$P(z) = \frac{(1-\rho)(1-z)g(\lambda(1-z))}{g(\lambda(1-z)) - z}\cdot\frac{\left(1-z^{N_2}\right) + p\left(z^{N_2} - z^{N_1}\right)}{(1-z)[pN_1 + (1-p)N_2]}, |z| < 1. \tag{3.45}$$

And the average queue-length, denoted by $\overline{L}$, is presented by

$$\overline{L} = \rho + \frac{\lambda^2 E\left[\chi^2\right]}{2(1-\rho)} + \frac{pN_1(N_1-1) + (1-p)N_2(N_2-1)}{2[pN_1 + (1-p)N_2]}, \tag{3.46}$$

where $g(\lambda(1-z)) = \int_0^{\infty} e^{-\lambda(1-z)t}\mathrm{d}G(t)$.

Proof. According to $P(z) = \sum_{j=0}^{\infty} p_j z^j$ and the expression of p_j given in Theorem 3.3, we get

$$P(z) = \frac{(1-\rho)}{pN_1 + (1-p)N_2}\left[\frac{1-z^{N_1}}{1-z} + (1-p)\frac{z^{N_1} - z^{N_2}}{1-z} + \lambda\sum_{j=1}^{\infty}\delta_j z^j\right]. \tag{3.47}$$

By a calculation, we can obtain

$$\sum_{j=1}^{\infty}\delta_j z^j = p\sum_{j=1}^{\infty} z^j\sum_{k=1}^{N_1} q_{j-N_1+k} + (1-p)\sum_{j=1}^{\infty} z^j\sum_{k=1}^{N_2} q_{j-N_2+k}$$

$$= p\left(\sum_{j=1}^{\infty} z^j q_j\right) \cdot \frac{1-z^{N_1}}{1-z} + (1-p)\left(\sum_{j=1}^{\infty} z^j q_j\right) \cdot \frac{1-z^{N_2}}{1-z}$$

$$= \left(\sum_{j=1}^{\infty} z^j q_j\right) \cdot \left[p\frac{1-z^{N_1}}{1-z} + (1-p)\frac{1-z^{N_2}}{1-z}\right] \tag{3.48}$$

$$\sum_{j=1}^{\infty} z^j q_j = \frac{z[1-g(\lambda(1-z))]}{\lambda[g(\lambda(1-z))-z]}. \tag{3.49}$$

Thus, substituting equations (3.48) and (3.49) into equation (3.47) to get equation (3.45), and then the equation (3.46) can be derived by using $\overline{L} = \frac{\mathrm{d}}{\mathrm{d}z}[P(z)]\big|_{z=1}$. □

Corollary 3.5 (Stochastic decomposition structure of queue size)**.** *For $\rho = \frac{\lambda}{\mu} < 1$, the steady-state queue size L of the queueing system studied in this paper can be decomposed into the sum of two independent parts: one is the steady-state queue size of the classic $M/G/1$ queueing system considered by Tang et al. [18], and the other is that the additional queue size L_d caused by the bi-level randomized (p, N_1, N_2)-policy. Furthermore, the additional queue size L_d has the following discrete distribution*

$$P\{L_\mathrm{d} = m\} = \frac{1}{pN_1 + (1-p)N_2}, \qquad m = 0, 1, 2, \cdots, N_1 - 1,$$

$$P\{L_\mathrm{d} = m\} = \frac{1\text{-}p}{pN_1 + (1-p)N_2}, \qquad m = N_1, N_1 + 1, \cdots, N_2 - 1.$$

Corollary 3.6. *For $\rho = \frac{\lambda}{\mu} < 1$, the average waiting time of customer, denoted by $\bar{W}_q$, is given by*

$$\bar{W}_q = \frac{\lambda E\left[\chi^2\right]}{2(1-\rho)} + \frac{pN_1(N_1-1) + (1-p)N_2(N_2-1)}{2\lambda[pN_1 + (1-p)N_2]}. \tag{3.50}$$

Proof. In the queuing system studied in this paper, customers are served by first-come-first-served rule. Therefore, the Little's law holds. Then, employing the Little's law, the equation (3.50) can be derived by $\bar{W}_q = \frac{\bar{L}}{\lambda} - \frac{1}{\mu}$. □

4. Some special cases

Special cases 4.1. If the service time χ of each customer obeys the exponential distribution $G(t) = 1 - e^{-\mu t}$, then some more concise expressions of the queuing performance indices of the system studied in this paper are as follows:

(1) The expressions of $\{p_j, j = 0, 1, 2, \cdots\}$ are given by

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

$$p_j = \frac{(1-\rho)+\delta_j}{pN_1 + (1-p)N_2}, \qquad j = 1, 2, \cdots, N_1 - 1,$$

$$p_j = \frac{(1-\rho)(1-p)+\delta_j}{pN_1 + (1-p)N_2}, \qquad j = N_1, N_1 + 1, \cdots, N_2 - 1,$$

$$p_j = \frac{\delta_j}{pN_1 + (1-p)N_2}, \qquad j \geq N_2,$$

where $\delta_j = p\left(1-\rho^{N_1}\right)\rho^{j-N_1+1} + (1-p)\left(1-\rho^{N_2}\right)\rho^{j-N_2+1}, j \geq 1.$

(2) The probability generating function of $\{p_j, j = 0, 1, 2, \cdots\}$

$$P(z) = \frac{(1-\rho)(1-z)\rho}{\rho - \lambda^2 z(1+\rho-z)} \cdot \frac{\left(1-z^{N_2}\right) + p\left(z^{N_2} - z^{N_1}\right)}{(1-z)[pN_1 + (1-p)N_2]}, |z| < 1.$$

And the average queue size is

$$\overline{L} = \frac{\rho}{(1-\rho)} + \frac{pN_1(N_1-1) + (1-p)N_2(N_2-1)}{2[pN_1 + (1-p)N_2]}\cdot$$

Special cases 4.2. When $p = 1$ or $p = 0$ or $N_1 = N_2$=N, the $M/G/1$ queueing system studied in this paper is equivalent to the $M/G/1$ queueing system with the classic N-policy studied in the references [21,26].

Special cases 4.3. When $N_2 = N_1 + 1$, the $M/G/1$ queueing system studied in this paper is equivalent to the $M/G/1$ queueing system with (p, N)-policy studied by Feinberg and Kim [2]. The probability generating function of steady-state queue-length distribution is

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

and the average queue-length is

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

Special cases 4.4. When $p = 1$ and $N_1 = 1$, or $p = 0$ and $N_2 = 1$, or $N_1 = N_2 = 1$, the $M/G/1$ queueing system studied in this paper becomes the classic $M/G/1$ queuing system studied by Tang *et al.* [18].

5. The system capacity optimization design

In practice, the capacity of the system will directly affect the cost and benefit of the system. If the capacity of the system is too large, it will lead to an increase in construction cost and operation cost. If the capacity of the system is too small, it will lead to the loss of customers. Therefore, it is very important to design a reasonable system capacity. In most occasions, the decision-makers design the buffer space by employing the mean steady-state queue size. In fact, the great irrationality of the modus operandi can be seen from the following example.

Example 5.1. When the service time χ of customer is the exponential distribution $G(t) = 1 - e^{-\mu t}$, we select $\lambda = 0.6, \mu = 2.8, p = 0.6, N_1 = 5, N_2 = 12$, and use *MATLAB* program to calculate the values of $\{p_j, j = 0, 1, 2, \cdots\}$ and the mean queue size $\bar{L}$ by Special case 4.1. See Table 1.

From the numerical results in Table 1, we know that the value of p_j is close to 0 when j exceeds a certain value. Therefore, the system capacity does not need to be designed to be infinite at all. After some calculations, we can get

$$P\left\{L > \overline{L}\right\} = 1 - \sum_{j=0}^{\bar{L}} p_j = 1 - \sum_{j=0}^{4} p_j = 0.3941 \tag{5.1}$$

$$P\left\{L > \overline{L} + 1\right\} = 1 - \sum_{j=0}^{\bar{L}+1} p_j = 1 - \sum_{j=0}^{5} p_j = 0.3264. \tag{5.2}$$

TABLE 1. When $\lambda = 0.6, \mu = 2.8, p = 0.6, N_1 = 5, N_2 = 12$, numerical results of the steady-state queue-length distribution.

p_0	p_1	p_2	p_3	p_4	p_5	p_6
0.1007	0.1223	0.1269	0.1279	0.1281	0.0677	0.0548
p_7	p_8	p_9	p_{10}	p_{11}	p_{12}	p_{13}
0.0520	0.0514	0.0513	0.0512	0.0510	0.0110	0.0024
p_{14}	p_{15}	p_{16}	p_{17}	p_{18}	p_{19}	$\overline{L}$
0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	4.4266

That is to say, if the system capacity is designed by the average queue- length $\overline{L}$, the probability of loss of customer will reach 39.41% due to no waiting space available for new arrivals. Even if we increase one unit on the mean queue size to design the system capacity, the probability of loss as presented in (5.2) is also up to 32.64%. Therefore, designing the system capacity only by using the mean queue size is very unsuitable.

Thus, in an attempt to reduce the loss probability of arrivals and increase system's profit, the decision-makers can design the system capacity according to the following case. Let M denote the system capacity to be determined. If it requires that the loss probability of new arrivals should be no more than 0.0001, that is,

$$P\{L > M\} = \sum_{j=M+1}^{\infty} p_j = 1 - \sum_{j=0}^{M} p_j \leq 0.0001.$$

Using the numerical results in Table 1, we can obtain $M \geq 15$, namely, the system capacity is designed to be at least $M = 15$. Form the discussing above, it can be seen that the steady-state queue-length distribution $\{p_j, j = 0, 1, 2, \cdots\}$ may play an important role in designing the system capacity, which also reflects the potential application value of Theorem 3.3.

6. Control strategy of the system under the constraint of customers waiting time

In this section, we first consider a cost structure that consists of linear waiting cost with rate h and fixed startup cost R for each busy period. Using the renewal reward theorem for a cycle period that is defined as the finite interval between two consecutive server busy period ending instants and the average queue length $\bar{L}$, we can obtain the long-run expected cost rate given by

$$F = h \cdot \overline{L} + \frac{R}{E[C]},$$

where C denotes a busy cycle, it consists of a server idle period and a server busy period, denoted by I and B, respectively.

Since long wait times cause customer frustration and system revenue losses, making it a lose-lose situation for all concerned. Thus, our aim here is to determine the optimal joint policy (N_1^*, N_2^*) such that the expected cost rate function F is minimized under the premise that the average waiting time of customers does not exceed the predetermined threshold $\bar{W}_{q_0}$. Hence, the optimization problem can be stated as

$$\begin{cases} \min F = h \cdot \bar{L} + \frac{R}{E[C]}, \\ \text{s.t.} \qquad \bar{W}_q \leq \bar{W}_{q_0} \end{cases} \tag{6.1}$$

where $\bar{W}_q$ is given in Corollary 3.6.

Next, we will find the expression of the expected busy cycle $E[C]$ in the objective function. Let V denote the number of customers in the system at the beginning of a server busy period. Clearly, for $\rho < 1$, the average number of customers in the system is given

$$E[V] = pN_1 + (1-p)N_2.$$

Thus, it follows that

$$E[B] = E[b]\cdot E[V] = \frac{\rho E[V]}{\lambda(1-\rho)}, \qquad \rho < 1.$$

Since the number of customers in the system at the beginning of a busy period is equal to the number of customers who arrived during the server idle period, and the arrival process is a Poisson process with rate λ, then the expected length of the server idle period is given by

$$E[I] = \frac{E[V]}{\lambda} = \frac{pN_1 + (1-p)N_2}{\lambda}. \tag{6.2}$$

Therefore, we can obtain the expected length of the busy cycle as below

$$E[C] = E[I] + E[B] = \frac{E[V]}{\lambda(1-\rho)}, \qquad \rho < 1. \tag{6.3}$$

Substituting equations (3.46), (3.50) and (6.3) into equation (6.1), the constrained optimization problem can be rewritten as

$$\begin{cases} \min F = h\cdot\left\{\rho + \frac{\lambda^2 E\left[\chi^2\right]}{2(1-\rho)} + \frac{pN_1(N_1-1)+(1-p)N_2(N_2-1)}{2[pN_1+(1-p)N_2]}\right\} + \frac{R(1-\rho)}{pN_1+(1-p)N_2}, \\ \text{s.t.} \qquad \bar{W}_q = \frac{\lambda E\left[\chi^2\right]}{2(1-\rho)} + \frac{pN_1(N_1-1)+(1-p)N_2(N_2-1)}{2\lambda[pN_1+(1-p)N_2]} \le \bar{W}_{q_0}. \end{cases} \tag{6.4}$$

From equation (6.4) we can observe that F and $\bar{W}_q$ are extremely complex and non-linear with respect to the decision variables N_1 and N_2, which poses a hard task to achieve the analytic results for the optimum values ${N_1}^*$ and ${N_2}^*$. Thus, to solve the optimization problem, we will utilize the direct search method to find the optimum values of N_1 and N_2. The following numerical experiments are performed on a PC having $Corei7$ processor using $MATLAB$ software package in Windows 10 environment, and the numerical results are reported in tables up to six decimal places because of lack of space.

Example 6.1. Assume that customers arrive at the service facility according to a Poisson process with a mean rate λ, and the service time follows an exponential distribution having a mean service rate μ. Under these assumptions, equation (6.4) can be reduced to the following minimization problem:

$$\begin{cases} \min F = h\cdot\left\{\frac{\rho}{1-\rho} + \frac{pN_1(N_1-1)+(1-p)N_2(N_2-1)}{2[pN_1+(1-p)N_2]}\right\} + \frac{R(1-\rho)}{pN_1+(1-p)N_2}, \\ \text{s.t.} \qquad \bar{W}_q = \frac{\rho}{\mu(1-\rho)} + \frac{pN_1(N_1-1)+(1-p)N_2(N_2-1)}{2\lambda[pN_1+(1-p)N_2]} \le \bar{W}_{q_0}. \end{cases} \tag{6.5}$$

Furthermore, we choose $R = 1500, h = 15, p = 0.7, \lambda = 0.2, \mu = 2$ and vary the values of N_1 and N_2, the corresponding expected cost rate and the average waiting time for different values of N_1 and N_2 are detailed reported in Table 2. From Table 2, we may reveal that

(1) If the upper bound of the mean waiting time is set to be 5, namely $\bar{W}_{q_0} = 5$, we find that $F = 119.9359$ is the minimum of the long-run average cost rate, and $({N_1}^*, {N_2}^*) = (2, 4)$ is the optimal bi-level threshold policy that satisfies constraints of the average waiting time $\bar{W}_{q_0} = 5$.
(2) If the upper bound of the mean waiting time is set to be 10, namely $\bar{W}_{q_0} = 10$, we may observe that $F = 88.7319$ is the minimum of the long-run average cost rate, and $({N_1}^*, {N_2}^*) = (4, 6)$ is the optimal bi-level threshold policy that satisfies constraints of the average waiting time $\bar{W}_{q_0} = 10$.
(3) If the upper bound of the mean waiting time is set to be 12, namely $\bar{W}_{q_0} = 12$, we may observe that $F = 88.7319$ is the minimum of the long-run average cost rate, and $({N_1}^*, {N_2}^*) = (5, 6)$ is the optimal bi-level threshold policy that satisfies constraints of the average waiting time $\bar{W}_{q_0} = 12$.

TABLE 2. The changes of F and $\bar{W}_q$ with different values of N_1 and N_2.

	N_1						
N_2	1	2	3	4	5	6	7
1	$F = 271.6667$ $\bar{W}_q = 0.0556$	–	–	–	–	–	–
2	$F = 212.8205$ $\bar{W}_q = 1.2094$	$F = 144.166$ $\bar{W}_q = 2.5556$	–	–	–	–	–
3	$F = 178.8542$ $\bar{W}_q = 2.8681$	$F = 129.4928$ $\bar{W}_q = 3.5338$	$F = 106.6667$ $\bar{W}_q = 5.0556$	–	–	–	–
4	$F = 157.9825$ $\bar{W}_q = 4.7924$	$\boldsymbol{F = 119.9359}$ $\boldsymbol{\bar{W}_q = 4.8632}$	$F = 101.2121$ $\bar{W}_q = 5.9646$	$F = 91.6667$ $\bar{W}_q = 7.5556$	–	–	–
5	$F = 144.8485$ $\bar{W}_q = 6.8737$	$F = 113.9080$ $\bar{W}_q = 6.4349$	$F = 97.9167$ $\bar{W}_q = 7.1389$	$F = 89.5736$ $\bar{W}_q = 8.4276$	$F = 85.6667$ $\bar{W}_q = 10.0556$	–	–
6	$F = 136.6667$ $\bar{W}_q = 9.0556$	$F = 110.4167$ $\bar{W}_q = 8.1806$	$F = 96.2821$ $\bar{W}_q = 8.5171$	$\boldsymbol{F = 88.7319}$ $\boldsymbol{\bar{W}_q = 9.5121}$	$\boldsymbol{F = 85.1572}$ $\boldsymbol{\bar{W}_q = 10.9046}$	$F = 84.1667$ $\bar{W}_q = 12.5556$	–
7	$F = 131.8452$ $\bar{W}_q = 11.3056$	$F = 108.8095$ $\bar{W}_q = 10.0556$	$F = 95.9524$ $\bar{W}_q = 10.0556$	$F = 88.9116$ $\bar{W}_q = 10.7698$	$F = 85.5060$ $\bar{W}_q = 11.9306$	$F = 84.5238$ $\bar{W}_q = 13.3889$	$F = 85.2381$ $\bar{W}_q = 15.0556$
8	$F = 129.4086$ $\bar{W}_q = 13.6039$	$F = 108.6404$ $\bar{W}_q = 12.0292$	$F = 96.6667$ $\bar{W}_q = 11.7222$	$F = 89.9359$ $\bar{W}_q = 12.1709$	$F = 86.5819$ $\bar{W}_q = 13.1064$	$F = 85.5303$ $\bar{W}_q = 14.3737$	$F = 86.1187$ $\bar{W}_q = 15.8775$
9	$F = 128.7255$ $\bar{W}_q = 15.9379$	$F = 109.5935$ $\bar{W}_q = 14.0799$	$F = 98.2292$ $\bar{W}_q = 13.4931$	$F = 91.6667$ $\bar{W}_q = 13.6919$	$F = 88.2796$ $\bar{W}_q = 14.4104$	$F = 87.1014$ $\bar{W}_q = 15.4903$	$F = 87.5219$ $\bar{W}_q = 16.8319$
10	$F = 129.3694$ $\bar{W}_q = 18.2988$	$F = 111.4394$ $\bar{W}_q = 16.1919$	$F = 100.4902$ $\bar{W}_q = 15.3497$	$F = 93.9943$ $\bar{W}_q = 15.3142$	$F = 90.5128$ $\bar{W}_q = 15.8248$	$F = 89.1667$ $\bar{W}_q = 16.7222$	$F = 89.3882$ $\bar{W}_q = 17.9037$
11	$F = 131.0417$ $\bar{W}_q = 20.6806$	$F = 114.0071$ $\bar{W}_q = 18.3534$	$F = 103.3333$ $\bar{W}_q = 17.2778$	$F = 96.8306$ $\bar{W}_q = 17.0228$	$F = 93.2108$ $\bar{W}_q = 17.3350$	$F = 91.6667$ $\bar{W}_q = 18.0556$	$F = 91.6667$ $\bar{W}_q = 19.0799$
12	$F = 133.5271$ $\bar{W}_q = 23.0788$	$F = 117.1667$ $\bar{W}_q = 20.5556$	$F = 106.6667$ $\bar{W}_q = 19.2661$	$F = 100.1042$ $\bar{W}_q = 18.8056$	$F = 96.3146$ $\bar{W}_q = 18.9288$	$F = 94.5513$ $\bar{W}_q = 19.4786$	$F = 94.3137$ $\bar{W}_q = 20.3497$

7. CONCLUSIONS

The present investigation explored a new $M/G/1$ queueing model with bi-level randomized (p, N_1, N_2)-policy. The queueing model proposed in this paper is not only an extension of mathematical meaning, but also makes the structure of the queueing model richer and more practical, which makes the operation and control of the system more flexible. Employing the total probability decomposition and the Laplace transform, some important queuing performance indices were obtained. Then, the probability generating function of steady-state queue-length and the expression of average queue-length were obtained by the obtained results in some special cases. Furthermore, a numerical example was used to discussed the optimal design of system capacity, which showed the importance of obtaining the specific expression of the stable length distribution in the design of system capacity. Finally, on the basis of the establishment of the cost structure model, we constructed a constrained optimization problem under the limit of the average waiting time and discussed the optimal double threshold policy that minimizes the expected cost function under the limit of the average waiting time by a numerical example. The analysis of this paper will provide a potentially practical application for system managers in related application areas such as telecommunication systems, flexible manufacturing systems, and so forth.

APPENDIX A.

Proof of Lemma 2.9. Consider the first customer served in the server busy period b. Let χ denote the service time of this customer and the number of customers arriving during χ. We regard these customers who arrive during χ as primary customers, and the others who arrive after the primary customers as secondary customers. Furthermore, let $A_1, A_2, \cdots, A_v$ denote primary customers. Since the order of server for arriving customers is irrelevant to the length of server busy period, we consider the following service order: Primary customers are served in order of $A_1, A_2, \cdots, A_v$. After serving each primary customer, however, the server will serve any secondary customers that are present until there are no secondary customers present. So we have that $b = \chi + L_1 + \cdots + L_v$, where $L_v(i = 1, 2, \cdots, v)$ denotes the time interval from the epoch when the server begins to serve ith primary customer until the next time point when the service of the $(i+1)$th primary customer begins. Hence, $L_1, L_2, \cdots, L_v$ are independent of each other with the same distribution as b, and also independent of χ and v. Let $L_1 + L_2 + \cdots + L_v = 0$ only if $v = 0$. Since the inter-arrivals are generated by Poisson process, and the ending epoch of every server busy period is a renewal epoch, we have

$$\begin{aligned} Q_j(t) &= P\{\chi + L_1 + \cdots + L_v > t \geq 0; N(t) = j\} \\ &= P\{\chi > t \geq 0; N(t) = j\} + \sum_{i=0}^{\infty} \int_0^t \frac{(\lambda x)^i}{i!} \mathrm{e}^{-\lambda x} P\{L_1 + \cdots + L_i > t - x; N(t-x) = j\} \mathrm{d}G(x) \\ &= \bar{G}(t) \frac{(\lambda t)^{j-1}}{(j-1)!} \mathrm{e}^{-\lambda t} + \sum_{i=1}^{\infty} \int_0^t \frac{(\lambda x)^i}{i!} \mathrm{e}^{-\lambda x} P\{L_1 + \cdots + L_i > t - x; N(t-x) = j\} \mathrm{d}G(x), \quad j \geq 1. \end{aligned} \tag{A.1}$$

It is important to note that each L_i follows the same stochastic laws as the server busy period b, $i = 1, 2, \cdots$. According to the service order of primary customer mentioned above, there are $j - (i - k)$ primary customers waiting for service. That is, at the time point $t - x$, the number of primary customers waiting for service in the system is equal to $j - (i - k)$. Hence,

$$\begin{aligned} P\{L_1 + \cdots + L_i > t - x; N(t-x) = j\} &= \sum_{k=1}^{i} P\{L_1 + \cdots + L_k > t - x; L_1 + \cdots + L_{k-1} \leq t - x; \\ &\quad N(t-x) = j - (i-k)\} \\ &= \sum_{k=1}^{i} Q_{j-(i-k)}(t) * B^{(k-1)}(t), \qquad j \geq 1, \end{aligned} \tag{A.2}$$

where $Q_j(t) = 0$ if $j \leq 0$, "$*$" denotes the convolution. Substituting equation (A.2) into equation (A.1), it gets

$$Q_j(t) = \bar{G}(t) \frac{(\lambda t)^{j-1}}{(j-1)!} \mathrm{e}^{-\lambda t} + \sum_{i=1}^{\infty} \int_0^t \frac{(\lambda x)^i}{i!} \mathrm{e}^{-\lambda x} \sum_{k=1}^{i} Q_{j-(i-k)}(t-x) * B^{(k-1)}(t-x) \mathrm{d}G(x), \qquad j \geq 1. \tag{A.3}$$

The Laplace transform of equation (A.3) is given by

$$\begin{aligned} q_j^*(s) &= \int_0^{\infty} e^{-st} \bar{G}(t) \frac{(\lambda t)^{j-1}}{(j-1)!} \mathrm{e}^{-\lambda t} \mathrm{d}t + \sum_{i=1}^{\infty} \sum_{k=1}^{i} q_{j-(i-k)}^*(s) b^{k-1}(s) \int_0^t \frac{(\lambda t)^i}{i!} \mathrm{e}^{-(s+\lambda)x} \mathrm{d}G(x) \\ &= \int_0^{\infty} e^{-st} \bar{G}(t) \frac{(\lambda t)^{j-1}}{(j-1)!} \mathrm{e}^{-\lambda t} \mathrm{d}t + \sum_{k=1}^{\infty} b^{k-1}(s) \sum_{i=k}^{j+k-1} q_{j-(i-k)}^*(s) \int_0^t \frac{(\lambda t)^i}{i!} \mathrm{e}^{-(s+\lambda)x} \mathrm{d}G(x) \\ &= \int_0^{\infty} e^{-st} \bar{G}(t) \frac{(\lambda t)^{j-1}}{(j-1)!} \mathrm{e}^{-\lambda t} \mathrm{d}t + \frac{q_j^*(s)}{b(s)} \{g(s + \lambda - \lambda b(s)) - g(s + \lambda)\} \\ &\quad + \frac{q_{j-1}^*(s)}{b^2(s)} \left\{ g(s + \lambda - \lambda b(s)) - \sum_{i=0}^{1} \int_0^{\infty} e^{-(s+\lambda)t} \frac{[\lambda b(s) t]^i}{i!} \mathrm{d}G(t) \right\} + \cdots \\ &\quad + \frac{q_1^*(s)}{b^j(s)} \left\{ g(s + \lambda - \lambda b(s)) - \sum_{i=0}^{j-1} \int_0^{\infty} e^{-(s+\lambda)t} \frac{[\lambda b(s) t]^i}{i!} \mathrm{d}G(t) \right\}, \qquad j \geq 1. \end{aligned} \tag{A.4}$$

Noting $b(s) = g(s + \lambda - \lambda b(s))$, it yields the conclusion given by Lemma 2.9. □

Acknowledgements. This research is supported by the National Natural Science Foundation of China (No.71571127) and the National Natural Science Youth Foundation of China (No.71301111).

References

[1] K.R. Balachandran, Control policies for a single server system. *Manag. Sci.* **19** (1973) 1013–1018.

[2] E.A. Feinberg and D.J. Kim, Bicriterion optimization of an $M/G/1$ queue with a removable server. *Probab. Eng. Inf. Sci.* **10** (1996) 57–73.

[3] J.X. Gu, Y.Y. Wei, Y.H. Tang and M.M. Yu, Queue size distribution of $Geo/G/1$ queue under the $Min(N, D)$-policy. *J. Syst. Sci. Complexity* **29** (2016) 752–771.

[4] D.P. Heyman, N-policy for the $M/G/1$ queue. *Manag. Sci.* **23** (1977) 775–778.

[5] M. Jain and S. Kaur, (p, N)-policy for unreliable server bulk queue with Bernoulli feedback. *Int. J. Appl. Comput. Math.* **6** (2020) 170–198.

[6] S.F. Jia and Y.H. Chen, The Geo/G/1 queue model with (p, N)-policy set-up time, multiple vacation and disasters. *Int. J. Sci. Eng. Technol.* **2** (2013) 991–995.

[7] F.C. Jiang, D.C. Huang, C.T. Yang, C.H. Lin and K.H. Wang, Design strategy for optimizing power consumption of sensor node with $Min(N, T)$-policy $M/G/1$ queuing models. *Int. J. Commun. Syst.* **25** (2012) 652–671.

[8] J.C. Ke, Bi-level control for batch arrival queues with an early startup and unreliable server. *Appl. Math. Model.* **28** (2004) 469–485.

[9] S.J. Lan and Y.H. Tang, The structure of departure process and optimal control strategy N^* for $Geo/G/1$ discrete-time queue with multiple server vacations and $Min(N, V)$-policy. *J. Syst. Sci. Complexity* **30** (2017) 1382–1402.

[10] S.J. Lan and Y.H. Tang, Performance and reliability analysis of a repairable discrete-time $Geo/G/1$ queue with Bernoulli feedback and randomized policy. *Appl. Stoch Model. Bus.* **33** (2017) 522–543.

[11] S.J. Lan and Y.H. Tang, Analysis of D-policy discrete-time $Geo/G/1$ queue with second J-optional service and unreliable server. *RAIRO-Oper. Res.* **51** (2017) 101–122.

[12] H.W. Lee and J.G. Park, Optimal strategy in N-policy production system with early set-up. *J. Oper. Res. Soc.* **48** (1997) 306–313.

[13] H.W. Lee and W.J. Seo, The performance of the $M/G/1$ queue under the dyadic $Min(N, D)$-policy and its cost optimization. *Perform. Eval.* **65** (2008) 742–758.

[14] H.W. Lee, N.I. Park and J. Jeon, Queue length analysis of batch arrival queues under bi-level threshold control with early set-up. *Int. J. Syst. Sci.* **34** (2003) 195–204.

[15] H.W. Lee, W.J. Seo, S.W. Lee and J. Jeon, Analysis of the $MAP/G/1$ queue under the $Min(N, D)$-policy. *Stoch. Models* **26** (2010) 98–123.

[16] J. Li and L.W. Liu, On an $M/G/1$ queue in random environment with $Min(N, V)$-policy. *RAIRO-Oper. Res.* **52** (2018) 61–77.

[17] C.Y. Luo, Y.H. Tang and K.Z. Yu, Optimal (r, N)-policy for discrete-time $Geo/G/1$ queue with different input rate and setup time. *Appl. Stoch. Model. Bus.* **31** (2015) 405–423.

[18] Y.H. Tang and X.W. Tang, Queuing Theory-Foundations and Analysis Techniques. Science Press, Beijing (2006) (in Chinese).

[19] Y.H. Tang, W.Q. Wu, Y.P. Liu and X.Y. Liu, The queue length distribution of $M/G/1$ queuing system with $Min(N, V)$-policy based on multiple server vacations. *Syst. Eng-Theor. Practice* **34** (2014) 1533–1546 (in Chinese).

[20] J. Teghem, Control of the service process in a queuing system. *Eur. J. Oper. Res.* **23** (1986) 141–158.

[21] N.S. Tian, Stochastic Service System with Vacations. Peking University Press, Beijing (2001) (in Chinese).

[22] K.H. Wang and K.B. Huang, A maximum entropy approach for the (p, N)-policy $M/G/1$ queue with a removable and unreliable server. *Appl. Math. Model.* **33** (2009) 2024–2034.

[23] T.Y. Wang and J.C. Ke, The randomized threshold for the discrete-time $Geo/G/1$ queue. *Appl. Math. Model.* **33** (2009) 3178–3185.

[24] Y.Y. Wei, Y.H. Tang and M.M. Yu, Recursive solution of queue length distribution for $Geo/G/1$ queue with delayed $Min(N, D)$-Policy. *J. Syst. Sci. Inf.* **4** (2020) 1478–9906.

[25] D.Y. Yang and J.C. Ke, Cost optimization of a repairable $M/G/1$ queue with a randomized policy and single vacation. *Appl. Math. Model.* **38** (2014) 5113–5125.

[26] M. Yadin and P. Naor, Queuing systems with a removable service station. *Oper. Res. Q.* **14** (1963) 393–405.