mathematics

MDPI

Article

Analysis of a Pre-Emptive Two-Priority Queuing System with Impatient Customers and Heterogeneous Servers

Mingang Yin [1], Ming Yan [1,*], Yu Guo [2] and Minghe Liu [3]

[1] School of Mechanical Engineering, Shenyang University of Technology, Shenyang 110870, China; yinmingang@163.com
[2] School of Mechanical Engineering & Automation, Northeastern University, Shenyang 110819, China; guoyu@mail.neu.edu.cn
[3] School of Mechanical Engineering, Shenyang Jianzhu University, Shenyang 110168, China; mhliu@sjzu.edu.cn
* Correspondence: yanming7802@163.com

Abstract: This paper presents a queuing system model that incorporates multiple priorities, multiple abandonments, and heterogeneous servers. Waiting for service easily leads to impatient behaviors. The impact of two kinds of impatient behaviors, balking and reneging, on queueing system performance is examined. The problem is formulated as continuous-time Markov chains. It also introduces a special state called the non-sojourn state to record the number of customers who abandon the system. The state transition rate matrix is transformed into a block tridiagonal matrix by appropriately setting the state numbers. A novel indicator called interstate transition frequency is proposed, which aids in distinguishing state transitions during the system evaluation process. Based on the interstate transition frequency, a set of indicators is derived to offer additional analytical perspectives for the queuing system. Finally, the proposed model is applied to an automobile repair shop to validate its effectiveness in practical scenarios.

Keywords: interstate transition frequency; Markov processes; queueing analysis; systems modeling

MSC: 60K30

check for updates

Citation: Yin, M.; Yan, M.; Guo, Y.; Liu, M. Analysis of a Pre-Emptive Two-Priority Queuing System with Impatient Customers and Heterogeneous Servers. *Mathematics* **2023**, *11*, 3878. https://doi.org/10.3390/math11183878

Academic Editors: Steve Drekic and Andrea Scozzari

Received: 22 July 2023
Revised: 30 August 2023
Accepted: 31 August 2023
Published: 11 September 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The increasing integration and intelligentization of industrial modernization have led to a gradual rise in the proportion of equipment maintenance and labor costs within the total cost. Therefore, controlling this cost has become a new focal point in the field of equipment maintenance and management. Queuing theory and Markov processes have laid a solid foundation for operating cost analysis and have attracted widespread attention. However, in practice, traditional models are difficult to widely use in actual production processes due to their lack of flexibility. Therefore, conducting research on a new queuing model that considers priority, abandonment, and heterogeneous servers is of practical significance.

The existing research on priority and abandonment in queuing systems can be summarized as follows. Jaiswal [1] introduced the concept of priority scheduling with the resumption of service of a low-priority customer. Kleinrock [2] proposed a time-dependent priority queue and derived the expressions for the average waiting times for different priorities. Priorities can be classified into two categories [3]: pre-emptive priority [4–6], in which a higher-priority customer entering the system will cause any ongoing work to be interrupted, and the maintainer will immediately serve the higher-priority customer, and non-pre-emptive priority [7], in which a customer with priority entering the system will be placed at the front of the queue automatically.

The concept of abandonment was initially proposed by Halfin [8]. Aguir [9] categorized impatient actions into three situations: balking, reneging, and retrial. In terms of

balking rules, Height [10] and Garnett [11] determined whether malfunctioning equipment entered the system with queue length and system load. If the number of waiting customers or the workload falls below a threshold, the customer will enter the system; otherwise, they will leave. In the case of reneging, there is an upper boundary of waiting time generated when a customer enters the system. If the waiting time exceeds the threshold, the customer will leave due to a lack of patience. Wei [12] defined the waiting time as a constant, Wang [13] considered individual differences and modeled the waiting time as an exponential distribution, and Ward [14] extended the distribution of the waiting time to accommodate arbitrary distributions. Comprehensive reviews of retrials were provided by Falin [15] and Artalejo [16], wherein customers have the option to re-enter the system and join the queue after leaving.

Most studies on queuing systems traditionally separate the prioritization and abandonment mechanisms. However, as queuing theory advanced, the integration of these two aspects has become increasingly prevalent. For instance, research has emerged on service differentiation problems in multiclass queuing systems that consider class-dependent services and abandonment distributions. Takagi [17] derived explicit expressions using Laplace-Stieltjes transforms for an M/M/m pre-emptive-loss last-come-first-serve (LCFS) system with impatient customers, as well as an M/M/m pre-emptive-resume LCFS queue with patient customers. Additionally, Jouini [18] examined scenarios involving multiple servers and non-pre-emptive customers.

The quantity and types of personnel are critical factors in research. Brandwajn [19] proposed a mathematical approach to compute the steady-state queue length distribution in first-come-first-serve (FCFS) queues with homogeneous server numbers. Bouchentouf [20] conducted an analysis of a heterogeneous two-server queuing system. Liu [21] investigated a service differentiation problem in a multiclass queuing system with class-dependent services and abandonment distributions. Heterogeneous servers are an important feature of multiserver queuing systems. This feature allows for the implementation of a scheduling method that enables customers to receive different levels of service quality. In this paper, this feature is achieved through the classification of service personnel. By classifying the personnel, wage management can be carried out based on their working intensity. Additionally, this approach can also serve as a means to motivate personnel and encourage them to improve their skill levels.

In terms of model solution, Bhulai [22] noted that the majority of studies analyzing multiple priority queues relied on Markov processes, with some improved algorithms applied to compute the sojourn probability of the system state. For example, Matthieu [23] estimated the system state sojourn probability by reducing the state dimension through customer type aggregation and server reduction. The model solution for multiple priorities with abandonment was derived from that of the multiple priority model, with the main focus being the calculation of the state sojourn probability. The state sojourn probability represents the relative proportion of the system's time spent in each state over a period. However, one drawback of the aforementioned methods based on state sojourn probability is that they often overlook shorter sojourn states, which leads to incomplete construction of model performance indicators and a disconnect with reality. For instance, the number of reneging cannot be directly determined through the state sojourn probability.

Summarizing leads to the conclusion that the structures of multiple priority models, multiple abandonment models, and multiple server models are relatively simplistic and lack the necessary flexibility and accuracy to handle complex problems in practical scenarios. Limited studies have successfully integrated all three aspects, whereas the development of the indicator system remains at a relatively primitive stage. This paper presents a comprehensive model that incorporates multiple priorities, multiple abandonments, and heterogeneous servers. An indicator system based on the frequency of interstate transitions is proposed as a solution to address these limitations. Moreover, a more robust theoretical foundation is established to guide service providers in optimizing the system configuration with profit as the objective function.

The remainder of this paper is organized as follows: In Section 1, the problem statements are discussed to address the fundamental characteristics of the studied queuing systems, such as multiple priorities, multiple abandonments, and heterogeneity servers. Section 2 focuses on the analysis of the state transition rate matrix elements, taking into consideration the allocation scheme of maintainers and the various arrival scenarios of customers. In Section 3, the steady-state solution for the system is provided, including state sojourn probability and interstate transition frequency. Section 4 presents the methodology for calculating the system indicators, which encompass critical interstate transition frequency and system operation cost. An illustrative example is further presented in Section 5 to evaluate the performance of the proposed model, followed by a conclusion in Section 6.

2. Model Introduction

In order to better reflect real-world processes, the model introduces a repair shop as an example. The maintainers are classified as senior technicians and general operators. The customers are divided into two levels: special customers with pre-emptive priority, requiring one senior technician and multiple general operators for service, and ordinary customers serviced by one maintainer. The service rate can be improved by increasing the number of general operators assigned to a customer. Let c_n and c_g represent the number of general operators and senior technicians, respectively. k represents the number of general operators required for a single special customer. N and G denote the capacity of the system for ordinary and special customers, respectively. $n(t)$ and $g(t)$ represent the numbers of ordinary and special customers in the system at time t, respectively. Two kinds of impatient behaviors, i.e., balking and reneging, are considered. Additionally, $v(t)$ represents the number of ordinary customers who leave the system via reneging at time t. Due to the property of the Markov process, two customers cannot renege the system simultaneously. When a customer leaves, the value of $v(t)$ will be reset to zero. Therefore, the maximum value of $v(t)$ is 1. The system state can be represented as $(n(t), g(t), v(t))$ and, subsequently, abbreviated as (n, g, v). The model can be represented as an $M/M/c_1,c_2/N$ queuing system. Ordinary customers have two ways to leave the system: one is to leave after completing the service, and the other is to renege due to a long waiting time without patience. In order to distinguish between these two ways of leaving, this study adds an abandonment state. As other waiting customers continue to queue when an ordinary customer reneges the system, it does not occupy the runtime of the system. Therefore, the time the system sojourns in the abandonment state approaches zero. Some of the specific assumptions are as follows.

(1) Ordinary and special customers arrive at the system according to Poisson processes at rates λ_1 and λ_2, respectively;
(2) The total number of maintainers in the system is $c_n + c_g$. It is assumed that each ordinary customer is serviced by a single maintainer, which can be either a senior technician or a general operator. The service time for an ordinary customer follows an exponential distribution, with parameter μ_{n1} for a senior technician or μ_{g1} for a general operator;
(3) Each special customer is serviced by one senior technician and k general operators simultaneously. If the idle maintainers are insufficient, the system will allocate the maintenance resources from ordinary customers being serviced to create a team of $k + 1$ maintainers to handle the special service immediately. The service time for a special customer follows an exponential distribution with parameter μ_2. In order to ensure priority, there are, at most, $\min(c_g, \lfloor c_n/k \rfloor)$ special customers in the system simultaneously;
(4) If all the maintainers are occupied, the ordinary customers entering the system have to wait. An ordinary customer will leave the system (renege) if their waiting time exceeds a certain patience limit, which follows an exponential distribution with parameter ρ. Ordinary customers will not re-enter the system after leaving. When $v = 0$, it means that no customer is reneging the system at time t, but it does not imply that no customer is waiting. $v = 1$ indicates that one ordinary customer is reneging the

system, and once they have left, the system state immediately returns to $v = 0$. In order to represent the instantaneous transitions between these states, the duration of the abandonment state is modeled by an exponential distribution with a parameter of μ_3, where $\mu_3 \to \infty$;

(5) Regarding balking, the definition of this type of abandonment differs from the traditional definition [10,13]. It can be referred to as negative balking, or balking for short, in this research. If an ordinary customer is informed that all maintainers are occupied and that the repair shop compensates for waiting time, it is reasonable that the ordinary customer may choose to wait in the parking lot. When the capacity of ordinary customers reaches the upper limit N, the repair shop will no longer provide parking space, whereas ordinary customers can still attempt to reach the repair shop, but they will leave immediately upon arrival. The time for a newly arrived customer to leave the system follows an exponential distribution with parameter μ_4, where $\mu_4 \to \infty$;

(6) The arrival interval and service time of the two levels of customers are mutually independent. For ordinary service, a first-come-first-serve (FCFS) rule is followed. Additionally, a general operator will be given priority for an ordinary service to improve maintenance ability. When there is an idle general operator, a senior technician will not be assigned to an ordinary service;

(7) In order to fulfill the requirement of timely service for special customers, it is assumed that $N \geq c_n + c_g$, $c_n \geq kG$, $c_g \geq G$.

The abandonment states referred to in assumptions (4) and (5), which are called non-sojourn states, are added to simulate the state of abandonment and to keep track of the number of customers who abandon the system. A waiting ordinary customer takes time to decide whether to renege, but the actual departure is usually instantaneous once the decision is made. Therefore, the duration in the abandonment states tends to be zero, and it cannot be directly described by the state sojourn probability. Thus, the concept of interstate transition frequency is introduced to characterize the process of transitions from these states.

These assumptions define the system as a three-dimensional Markov process. The state space is denoted by

$$\boldsymbol{E} = \{(n, g, v) | n = 0, 1, 2, \ldots, N + 1;\ g = 0, 1, \ldots, G;\ v = 0, 1\} \tag{1}$$

When referring to the idea of partitioning, the system states are divided based on the number of ordinary customers n. When n = 0, (0, 0, 0) is referred to as state 1, (0, 0, 1) as state 2, (0, 1, 0) as state 3, and so on, up to (0, G, 1), which is denoted as state 2(G + 1). When the number of ordinary customers reaches n, (n, 0, 0) is denoted as state 2(G + 1)n + 1, and (n, G, 1) is referred to as state 2(G + 1)(n + 1). When $n = N$ due to the specific transition mechanism, (N, 0, 0) is labeled as state 2(G + 1)N + 1, (N + 1, 0, 0) as state 2(G + 1)N + 2, (N, 1, 0) as state 2 (G + 1)N + 3, and so on, up to (N + 1, G, 0), which is denoted as state 2(G + 1)(N + 1). Let $\boldsymbol{S}$ be a set that contains all the states. $\boldsymbol{S}$ is defined as {1, 2,. . ., GT}, where GT equals 2(G + 1)(N + 1). It should be noted that the states in both sets ($\boldsymbol{S}$ and $\boldsymbol{E}$) are identical, except for differences in their expression format. Some states have mathematical expressions that are not practically feasible. For example, (0, 0, 1) indicates that there are no customers in the system, yet an ordinary customer is leaving. It is not feasible for the system to be in such a state, resulting in a transition rate of 0 for that state. The state is introduced to simplify the process of building the model.

3. System Analysis

When there are g special customers in the system, we need to allocate $w_{gg} = g$ senior technicians and $w_{ng} = kg$ general operators to service them when considering their priority. As a result, for ordinary customers, the number of senior technicians available is $l_g = c_g - w_{gg} = c_g - g$,

and the number of general operators available is $l_n = c_n - w_{ng} = c_n - kg$. Consequently, within the range of n and g, w_{gn} and w_{nn} represent the number of senior technicians and general operators, respectively, who are engaged in ordinary services. In order to simplify the expression, we omit the independent variables g and n unless explicitly stated.

$$\begin{cases} w_{nn} = n, w_{gn} = 0 & \text{if } l_n \geq n \\ w_{nn} = l_n, w_{gn} = (n - l_n) & \text{if } l_n < n \text{ and } l_n + l_g \geq n \\ w_{nn} = l_n, w_{gn} = l_g & \text{if } l_n + l_g < n \end{cases} \tag{2}$$

where the first case occurs when there are enough general operators. The second case arises when the number of general operators is insufficient, resulting in senior technicians assisting with ordinary services; the third case indicates a shortage of workforce, where there remain ordinary customers waiting for service.

The ordinary service rate from state $(n, g, 0)$ to $(n - 1, g, 0)$ is

$$\widetilde{\mu} = w_{nn}\mu_{n1} + w_{gn}\mu_{g1} \tag{3}$$

When considering pre-emptive priority, there is no special customer in the system waiting for service. We define the number of ordinary customers waiting for service as y:

$$y = n - w_{nn} - w_{gn} \tag{4}$$

Therefore, the reneging rate $\widetilde{\rho}$ of the system is

$$\widetilde{\rho} = y\rho \tag{5}$$

Let us consider $(n_3, G, 0)$ in Figure 1 as an illustrative example to explain the process of state transition, where $n_1 = c_n - kG$, $n_2 = c_g - G$ and $n_3 = n_1 + n_2$ with all maintainers currently in operation. Once a special service is completed, a group of maintainers comprising one senior technician and k general operators is released. As a result, the system transitions to $(n_3, G - 1, 0)$. This transition releases $k + 1$ maintainers. If the number of general operators released exceeds the number of senior technicians currently servicing ordinary customers, i.e., $k \geq n_2$, the general operators will take over all the senior technicians' service for ordinary customers, resulting in a service rate of $n_3 \cdot \mu_{n1}$ for ordinary customers. If $k < n_2$, only k senior technicians can be replaced, leading to $n_1 + k$ general operators and $n_2 - k$ senior technicians performing ordinary services. The service rate for ordinary customers, in this case, is $(n_1 + k)\mu_{n1} + (n_2 - k)\mu_{g1}$. The spatial states are sorted according to a predefined order, and the state transition mechanism of the model is depicted in Figure 1.

Figure 1 illustrates three special states, which are depicted by dashed boxes: $(n_1 + 1, G, 0)$ indicates the state where the first senior technician provides ordinary service; $(n_3, G, 0)$ signifies that all maintainers are currently working, and $(n_3 + 1, G, 0)$ denotes the state where the first ordinary customer reneges due to impatience. The order of these states is determined by the set $\boldsymbol{S}$.

A state sequence comprises states that share the same value of n and v. There are three sequences that require special attention. The first sequence is the fully loaded sequence of general operators, where the number of ordinary customers in the system is c_n and all the general operators are working. The second sequence is the fully loaded sequence of personnel. In this case, the number of ordinary customers reaches $c_n + c_g$, and all the maintainers are working. The third sequence is the balking sequence, in which the subsequent ordinary customers will promptly leave the system upon entering once the number of ordinary customers reaches N.

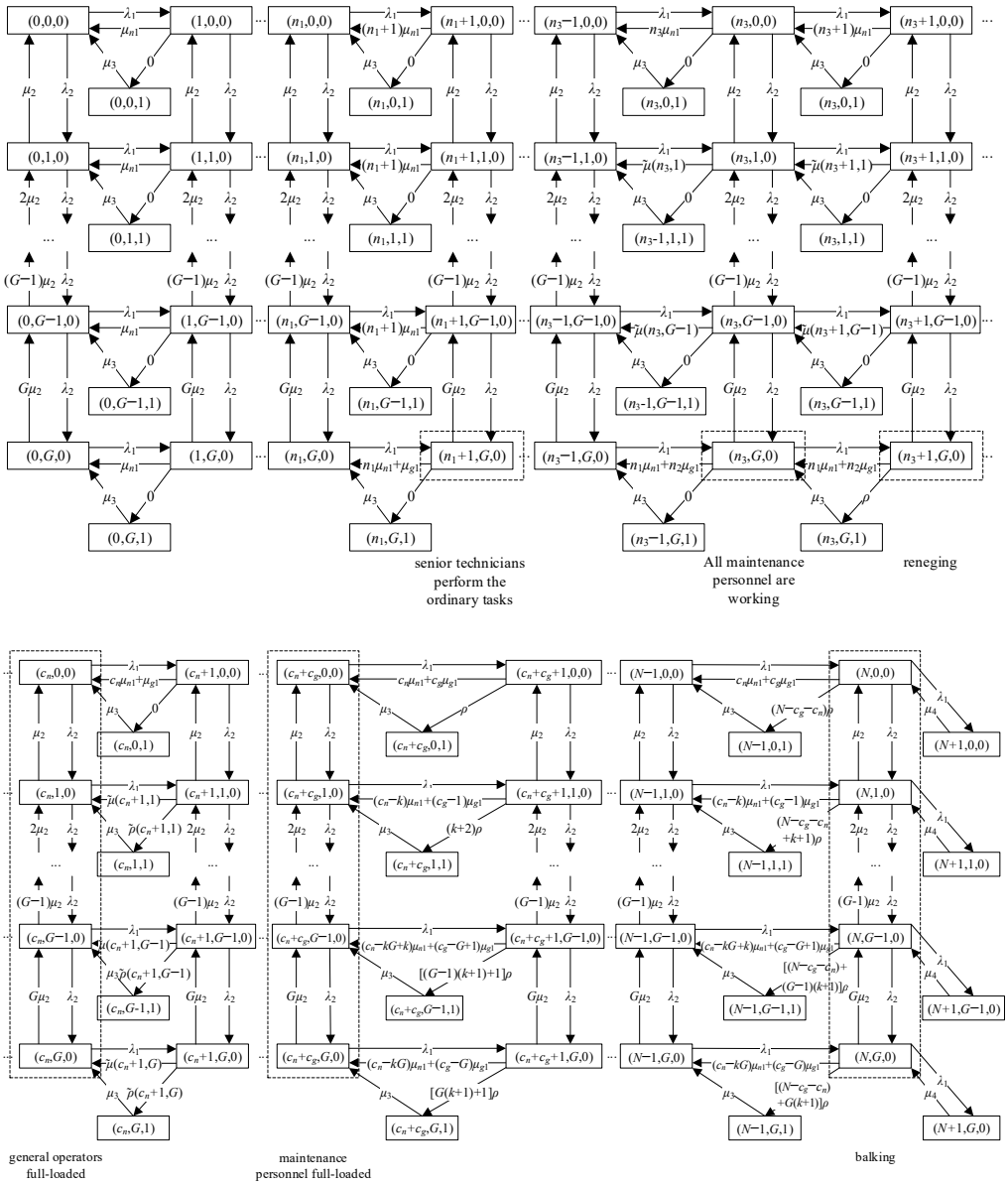


Figure 1. Mechanism of the state transition of the model.

The state transition process related to $(n, g, 0)$ is shown in Figure 2 for the values of n ranging from 1 to $N-1$ and g ranging from 1 to $G-1$. $\widetilde{\mu}$ and $\widetilde{\rho}$ are calculated using Equation (3) and Equation (5), respectively.

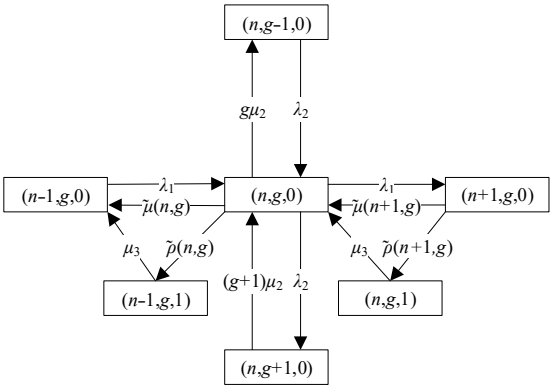


Figure 2. State transition process relating to $(n, g, 0)$.

The state transition rate matrix of the system can be represented by

$$\begin{aligned} \boldsymbol{Q} &= \{q_{i,j}\}_{GT\times GT} \\ &= \begin{pmatrix} \boldsymbol{A}_0 & \boldsymbol{C}_0 & & & & & \\ \boldsymbol{B}_1 & \boldsymbol{A}_1 & \boldsymbol{C}_1 & & & & \\ & \boldsymbol{B}_2 & \boldsymbol{A}_2 & \boldsymbol{C}_2 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \boldsymbol{B}_n & \boldsymbol{A}_n & \boldsymbol{C}_n & \\ & & & & \ddots & \ddots & \ddots \\ & & & & & \boldsymbol{B}_N & \boldsymbol{A}_N \end{pmatrix} \end{aligned} \tag{6}$$

where $\boldsymbol{A}_n$, $\boldsymbol{B}_n$, $\boldsymbol{C}_n$ are the submatrices of $\boldsymbol{Q}$. $q_{i,j}$ denotes the element in the ith row and the jth column of the transition rate matrix $\boldsymbol{Q}$.

$\boldsymbol{B}_n$ is the transition rate matrix that represents the reduction of ordinary customers.

$$\boldsymbol{B}_n = \left\{B_{i,j}^{(n)}\right\}_{4(G+1)\times(G+1)},\ n = 1,2,\cdots,N \tag{7}$$

where

$$\begin{cases} B_{2g+1,2g+1}^{(n)} = \widetilde{\mu}(n,g) \\ B_{2g+1,2g+2}^{(n)} = \widetilde{\rho}(n,g) \end{cases},\ g = 0,1,2,\cdots,G \tag{8}$$

$\boldsymbol{C_n}$ is the transition rate matrix that models the arrival of ordinary customers.

$$\boldsymbol{C}_n = \left\{C_{i,j}^{(n)}\right\}_{4(G+1)\times(G+1)},\ n = 0,1,2,\cdots,N-1 \tag{9}$$

where

$$C_{2g+1,2g+1}^{(n)} = \lambda_1,\ g = 0,1,2,\cdots,G \tag{10}$$

For values of n ranging from 0 to $N-1$, $\boldsymbol{A_n}$ is the transition rate matrix between two adjacent state sequences with the same value of n. When $n = N$, $\boldsymbol{A_N}$ denotes the transition rate matrix between $n = N, N+1$, while $v = 0$.

$$\boldsymbol{A}_n = \left\{A_{i,j}^{(n)}\right\}_{4(G+1)\times(G+1)},\ n = 0,1,2,\cdots,N \tag{11}$$

where $i = 2(G+1)(n-1)g$.

When $n = 0, 1, 2, \ldots, N-1$,

$$\begin{cases} A_{2g+2,2g+1}^{(n)} = \mu_3 & g = 0,1,\cdots,G \\ A_{2g+3,2g+1}^{(n)} = (g+1)\mu_2 & g = 0,1,\cdots,G-1 \\ A_{2g+1,2g+3}^{(n)} = \lambda_2 & g = 0,1,\cdots,G-1 \\ A_{g,g}^{(n)} = q_{i,i} & g = 0,1,\cdots,2(G+1) \end{cases} \tag{12}$$

When $n = N$,

$$\begin{cases} A_{2g+2,2g+1}^{(n)} = \mu_4 & g = 0,1,\cdots,G \\ A_{2g+3,2g+1}^{(n)} = (i+1)\mu_2 & g = 0,1,\cdots,G-1 \\ A_{2g+1,2g+3}^{(n)} = \lambda_2 & g = 0,1,\cdots,G-1 \\ A_{2g+1,2g+2}^{(n)} = \lambda_1 & g = 0,1,\cdots,G \\ A_{g,g}^{(n)} = q_{i,i} & g = 0,1,\cdots,2(G+1) \end{cases} \tag{13}$$

Note that the nondiagonal elements in $\boldsymbol{A}_n$ correspond to the transitions among the $2(G+1)$ states, whereas the diagonal element $A_{i,i}^{(n)}$ represents the Markov process element.

$$A_{i,i}^{(n)} = q_{i,i} = -\sum_{j \in \boldsymbol{S}, j \neq i} q_{i,j} \tag{14}$$

4. Steady-State Solution

The steady-state solution of the system comprises the steady-state sojourn probability and the steady-state interstate transition frequency.

4.1. Steady-State Sojourn Probability

Let $\boldsymbol{P}(t) = [P_1(t), P_2(t), \ldots, P_{GT}(t)]^{\mathrm{T}}$ represent the probability distribution of the system in each state at time t.

$$\begin{cases} \boldsymbol{P}'(t) = \boldsymbol{Q} \cdot \boldsymbol{P}(t) \\ \boldsymbol{P}(0) = \left(1, \underbrace{0, 0, \cdots, 0}_{GT-1}\right)^{\mathrm{T}} \end{cases} \tag{15}$$

Applying Laplace transformation to both sides yields the following equation:

$$\boldsymbol{P}^*(s) = (s\boldsymbol{I} - \boldsymbol{Q})^{-1}\boldsymbol{P}(0), s > 0 \tag{16}$$

where $\boldsymbol{I}$ is an identity matrix. According to Tauberian Theorem,

$$\boldsymbol{P} = \lim_{t \to \infty} \boldsymbol{P}(t) = \lim_{s \to 0} s\boldsymbol{P}^*(s) \tag{17}$$

where $\boldsymbol{P} = [P_1, P_2, \ldots, P_{GT}]^{\mathrm{T}}$ represents the probabilities of each state in each steady state. These probabilities reflect the sojourn time in each state over a given period.

4.2. Steady-State Interstate Transition Frequency

This paper introduces a novel indicator called the interstate transition frequency, which includes information about both the starting state and entering state. It distinguishes the traditional transition frequency, which only considers the entering state. Let $M_{l,z}^{(i)}(t + \Delta t)$ denote the number of transitions from state l to state z within the time interval Δt, giving the initial state i. When $l \neq z$,

$$M_{l,z}^{(i)}(t + \Delta t) = \begin{cases} \left[\sum\limits_{j=1, j \neq z}^{GT} \left(P_{l,j}(\Delta t) \cdot M_{l,z}^{(j)}(t)\right) \right. \\ \left. + P_{l,z}(\Delta t)\left(1 + M_{l,z}^{(z)}(t)\right) + o(\Delta t)\right], \ i = l \\ \sum\limits_{j=1}^{GT} P_{i,j}(\Delta t) \cdot M_{l,z}^{(j)}(t) + o(\Delta t) \ , \ i \neq l \end{cases} \tag{18}$$

where 1 in the first expression indicates a single transition from state l to z within Δt, denoted as $M_{l,z}^{(l)}(\Delta t) = 1$. $P_{i,j}(\Delta t)$ represents the probability of the transition from state i to j. This probability can be calculated based on the properties of the exponential distribution and can be expressed as

$$P_{i,j}(\Delta t) = \begin{cases} q_{i,j}\Delta t & j \neq i \\ 1 - \sum\limits_{j=1, j \neq i}^{GT} q_{i,j}\Delta t & j = i \end{cases} \tag{19}$$

By combining Equation (14) with Equation (19), we simplify the expression as follows:

$$P_{i,j}(\Delta t) = \begin{cases} q_{i,j}\Delta t & j \neq i \\ 1 + q_{i,i}\Delta t & j = i \end{cases} \tag{20}$$

By substituting Equation (20) into Equation (18), one can obtain the following derivation:

$$\frac{M_{l,z}^{(i)}(t+\Delta t) - M_{l,z}^{(i)}(t)}{\Delta t} = \begin{cases} \sum_{j=1}^{GT} q_{l,j} \cdot M_{l,z}^{(j)}(t) + q_{l,z} + \frac{o(\Delta t)}{\Delta t} & i = l \\ \sum_{j=1}^{GT} q_{i,j} \cdot M_{l,z}^{(j)}(t) + \frac{o(\Delta t)}{\Delta t} & i \neq l \end{cases} \tag{21}$$

When taking the limit as Δt approaches 0, the right-hand side of Equation (21) converges to a finite value, implying the left-hand side also converges. Therefore, we can conclude that $M_{l,z}^{(i)}(t)$ is differentiable based on this deduction. Furthermore, it is evident from the definition of $M_{l,z}^{(i)}(t)$ that $M_{l,z}^{(i)}(0)$ is equal to 0, where $i \in \boldsymbol{S}$. The matrix representation of Equation (21) is as follows:

$$\begin{cases} \boldsymbol{M}'_{l,z}(t) = \boldsymbol{Q}\boldsymbol{M}_{l,z}(t) + \boldsymbol{W}_{l,z} \\ \boldsymbol{M}_{l,z}(0) = \boldsymbol{0} \end{cases} \tag{22}$$

where $\boldsymbol{M}_{l,z}(t) = [M_{l,z}^{(1)}(t), M_{l,z}^{(2)}(t), \ldots, M_{l,z}^{(GT)}(t)]^{\mathrm{T}}$; $\boldsymbol{W}_{l,z} = [W_{l,z}^{(1)}, W_{l,z}^{(2)}, \ldots, W_{l,z}^{(GT)}]^{\mathrm{T}}$. It should be noted that all elements in $\boldsymbol{W}_{l,z}$, except for $W_{l,z}^{(l)} = q_{l,z}$, are equal to 0.

Let $m_{l,z}^{(i)}(t)$ represents the derivative of $M_{l,z}^{(i)}(t)$.

$$m_{l,z}^{(i)}(t) = \mathrm{d}M_{l,z}^{(i)}(t) \,/\mathrm{d}t \quad , i \in S \tag{23}$$

where $m_{l,z}^{(i)}(t)$ represent the transient frequency from state l to z at time t given the initial state i. Apply the Laplace transform to Equation (23).

$$\begin{aligned} m_{l,z}^{(i)*}(s) &= \int_0^{\infty} e^{-st} m_{l,z}^{(i)}(t)\mathrm{d}t \\ &= \int_0^{\infty} e^{-st} \mathrm{d}M_{l,z}^{(i)}(t), \quad s > 0 \end{aligned} \tag{24}$$

One can derive

$$s\boldsymbol{m}_{l,z}^{*}(s) = \boldsymbol{Q}\boldsymbol{m}_{l,z}^{*}(s) + \boldsymbol{W}_{l,z} \tag{25}$$

where $\boldsymbol{m}_{l,z}^{*}(s) = \left[m_{l,z}^{(1)*}(s), m_{l,z}^{(2)*}(s), \ldots, m_{l,z}^{(GT)*}(s)\right]^{\mathrm{T}}$; that is,

$$\boldsymbol{m}_{l,z}^{*}(s) = (s\boldsymbol{I} - \boldsymbol{Q})^{-1}\boldsymbol{W}_{l,z} \tag{26}$$

By applying the Laplace inversion method, we can obtain $\boldsymbol{m}_{l,z}(t)$. According to the Tauberian Theorem, the steady-state interstate transition frequency $\boldsymbol{m}_{l,z}$ can be determined as

$$\boldsymbol{m}_{l,z} = \lim_{t\to\infty} \boldsymbol{m}_{l,z}(t) = \lim_{s\to 0} s\boldsymbol{m}_{l,z}^{*}(s) \tag{27}$$

According to the steady-state properties, $\boldsymbol{m}_{l,z}$ is independent of the initial state, resulting in the elements in the column vector $\boldsymbol{m}_{l,z}$ being the same. The subsequent sections exclusively focus on steady-state analysis. For simplicity, we express $m_{l,z}^{(i)}$ without explicitly stating (i), and this can be given by

$$m_{l,z} = m_{l,z}^{(i)}, i \in \boldsymbol{S} \tag{28}$$

According to the properties of the Markov process, two state transitions cannot occur within Δt. Hence, in the case $l = z$, $m_{z,z} = 0$.

5. System Operation Indicators

By building upon the models of Kim [24] and Vasiliadis [25], we propose the following performance indicators.

5.1. Critical Interstate Transition Frequency

The transition frequencies between critical states play a crucial role in system operation indicators, primarily encompassing the service frequency and the abandonment frequency.

(1) Frequency of ordinary service L_n

$$L_n = \sum_{g=0}^{G} \sum_{n=0}^{N-1} m_{2(G+1)(n+1)+2g+1,2(G+1)n+2g+1} \tag{29}$$

(2) Frequency of special service L_g

$$L_g = \sum_{g=0}^{G-1} \sum_{n=0}^{N} m_{2(G+1)n+2g+3,2(G+1)n+2g+1} \tag{30}$$

From Equations (29) and (30), it is observed that the entering states are identical. When the starting state is indistinguishable, it becomes impossible to determine the specific level of the service that has been completed. Therefore, interstate transition frequency, which quantifies the frequency of transitions between states, becomes more meaningful than the traditional transition frequency.

(3) Reneging frequency L_r

When $n > l_n + l_g$, indicating a lack of workforce, the ordinary customers who cannot be serviced in time will renege. In such instances, when a customer reneges, the system state transitions from $(n, g, 0)$ to $(n - 1, g, 1)$ and instantaneously from $(n - 1, g, 1)$ to $(n - 1, g, 0)$ due to $\mu_3 \to \infty$. By calculating the frequency of the transition from $(n - 1, g, 1)$ to $(n - 1, g, 0)$, one can obtain L_r:

$$L_r = \sum_{g=0}^{G} \sum_{n=l_n+l_g+1}^{N} m_{2(G+1)n+2g+1,2(G+1)(n-1)+2g+2} \tag{31}$$

Since $\mu_3 \to \infty$, the system's duration in $(n - 1, g, 1)$ approaches 0, which indicates that the probability of being in that state at any given time tends to zero. Consequently, employing the sojourn probability to count the number of reneging is not feasible. However, Equation (31) reveals that the frequency of reneging can be effectively addressed by interstate transition frequency. Therefore, interstate transition frequency offers greater flexibility in calculating system indicators compared to state sojourn probability. This advantage is also reflected in the formulation of the balking frequency.

(4) Balking frequency L_b

When the number of ordinary customers reaches N, and an additional ordinary customer arrives, the system state transitions from $(N, g, 0)$ to $(N + 1, g, 0)$ and then immediately reverts to $(N, g, 0)$ due to assumption 4). L_b can be determined by calculating the transition frequency from $(N + 1, g, 0)$ to $(N, g, 0)$.

$$L_b = \sum_{g=0}^{G} m_{2(G+1)N+2g+2,2(G+1)N+2g+1} \tag{32}$$

5.2. Waiting Time Analysis

The waiting time T_d represents the cumulative waiting time of all customers throughout the system's operation interval T_s. In order to accurately reflect this, it is important to

consider the sojourn time in each state as well as the number of waiting customers in the corresponding state. Additionally, capturing the reasons for state transitions during the system operation allows T_d to be categorized into four distinct parts.

$$T_d = T_{d1} + T_{d2} + T_{d3} + T_{d4} \tag{33}$$

(1) T_{d1} represents the waiting time caused by ordinary services. During this time, the system transitions from $(n, g, 0)$ to $(n-1, g, 0)$ is

$$T_{d1} = T_s \sum_{g=0}^{G} \sum_{n=l_n+l_g+1}^{N} y P_{i,j} \tag{34}$$

where $i = 2(G+1)n + 2g + 1$, $j = 2(G+1)(n-1) + 2g + 1$.

(2) T_{d2} represents the waiting time caused by special services. During this time, the system transitions from $(n, g, 0)$ to $(n, g-1, 0)$ is

$$T_{d2} = T_s \sum_{g=0}^{G-1} \sum_{n=l_n+l_g+1}^{N} y P_{i,j} \tag{35}$$

where $i = 2(G+1)n + 2g + 3$, $j = 2(G+1)n + 2g + 1$.

(3) T_{d3} represents the waiting time caused by reneging. The transition occurs from $(n, g, 0)$ to $(n-1, g, 1)$, then transits to $(n-1, g, 0)$. However, as $\mu_3 \to \infty$, the state transitions to $(n-1, g, 0)$ instantaneously, making the transition time negligible.

$$T_{d3} = T_s \sum_{g=0}^{G} \sum_{n=l_n+l_g+1}^{N} y P_{i,j} \tag{36}$$

where $i = 2(G+1)n + 2g + 1$, $j = 2(G+1)(n-1) + 2g + 2$.

(4) T_{d4} represents the waiting time due to balking. The system transitions from $(N, g, 0)$ to $(N+1, g, 0)$, and as $\mu_4 \to \infty$, it returns to $(N, g, 0)$ immediately.

$$T_{d4} = T_s \sum_{g=0}^{G} \left(N - l_n - l_g\right) P_{i,j} \tag{37}$$

where $i = 2(G+1)N + 2g + 1$, $j = 2(G+1)N + 2g + 2$.

The meaning of Equations (33)–(37) is as follows: $T_s \cdot P_{i,j}$ represents the sojourn time at state i before transitioning to state j during the system's runtime. This time is then multiplied by the number of waiting customers y to obtain the waiting time at each state. Finally, the waiting time for the corresponding state is summed, considering the different reasons for the transitions.

5.3. Working Time Analysis

It is known to all that a maintainer should not be continuously occupied during runtime. Hence, it is reasonable to establish a correlation between wages and working intensity. In order to calculate the working intensity P_{lw} for two-level maintainers, it is crucial to compute the corresponding working time T_l individually. T_l is influenced by both the steady-state sojourn probability and the number of maintainers at the same level.

$$T_l = \frac{T_{lg} + T_{ln}}{c_l} = \frac{T_s}{c_l} \sum_{n=1}^{N} \sum_{g=0}^{G} \left(w_{lg} + w_{ln}\right) P_i \tag{38}$$

where T_l denotes the average working time for level l, $l = g, n$, representing senior technicians and general operators, respectively. T_{lg} and T_{ln} represent the total working time of level l maintainers servicing special and ordinary customers, respectively. $i = (2G+1)n + 2g + 1$.

Accordingly, the actual working intensity, denoted as P_{lw}, can be expressed as

$$P_{lw} = \frac{T_l}{T_S} \tag{39}$$

5.4. Operation Profit Analysis

Referring to the utilization of Markov process in networks [26] and queuing theory in hospitals [27], a comprehensive profit analysis is performed.

The profit of the system can be calculated based on the system parameters k, c_n, c_g, and N. By taking profit as the objective function, the system parameters can be optimized. The details of operation profit are as follows:

(1) The income of the system per unit of time is divided into two parts: the income from ordinary customers (L_nC_n), which is derived from the frequency of ordinary customers and the income per ordinary customer, and the income from special customers (L_gC_g), which is calculated by multiplying the frequency of special customer by the income per special customer;
(2) The cost of abandonment per unit of time is composed of two parts. One is the number of reneging customers per unit of time and the cost per reneging customer, denoted as L_rC_r. The other is calculated by the number of balking customers per unit of time and the cost per balking customer, denoted as L_bC_b;
(3) The waiting cost, denoted as T_dC_d, is calculated by multiplying the waiting time by the cost per unit of time of waiting;
(4) The wages of maintenance personnel are determined through a floating mechanism according to P_{gw} and P_{nw}. The personnel cost, denoted as C_w, can be expressed as

$$C_w = \frac{T_s}{T_w} \sum_{l=n,g} S_l(P_{lw}) \tag{40}$$

where $S_l(P_{lw})$ denotes the weekly wages of class l, and T_w is the weekly business time.

It should be noted that this paper studies the system configuration before the actual operation. It assumes that service rates for different levels of maintainers are constant, ensuring that there are no intentional service delays due to wages.

(5) The site cost is calculated as $(N + G)C_N$. N and G are the capacities of ordinary customers and special customers, respectively. C_N is the cost of a single space.

Therefore, the total profit of the system C can be calculated as:

$$C = T_s\left(L_nC_n + L_gC_g - L_aC_a - L_eC_e\right) - T_dC_d - C_w - (N+G)C_N \tag{41}$$

6. Numerical Example

An automobile repair shop is considered here as an illustrative example. The repair shop primarily offers two types of services: vehicle maintenance and repair. The staff includes c_n general operators and c_g senior technicians, and each maintenance work is assigned to a single maintainer. When a car requires a repair, the customer is granted pre-emptive priority. The repair progress is promptly handled by k general operators and one senior technician, with greater values of k equating to increased efficiency. Generally, k is chosen from the set {1, 2, 3} due to spatial limitations. In cases of insufficient manpower, some maintainers may temporarily cease ongoing maintenance to prioritize repair services. Maintenance services typically favor general operators. In order to ensure repair service priority and consider working site constraints, the repair shop can only simultaneously accommodate G vehicles for repair and N vehicles for maintenance. For the sake of controlling the system scale while maintaining generality, we set $G = 2$.

The arrival interval and service time of vehicles in the system follow exponential distributions. The average arrival interval for maintenance vehicles is 40 min, whereas it is 200 min for repair vehicles. A senior technician requires an average maintenance time of 200 min per vehicle, whereas a general operator needs 240 min per vehicle. The average repair times per vehicle for k = 1, 2, and 3 are 360, 330, and 300 min, respectively. The

maintenance vehicles waiting in the system will renege with an exponential distribution of 120 min per vehicle. The average duration of both abandonment and barking is 1 s.

The repair shop earns USD 1000 per vehicle for maintenance work and USD 5000 per vehicle for repair work. The waiting cost is USD 10 per minute. The shop incurs a loss of USD 1000 per vehicle when a customer reneges and USD 2500 per vehicle when balking occurs. The site cost for each working space is USD 2500 per month. The weekly wages of personnel vary based on the actual work intensity. The weekly business time, denoted as T_w, is calculated as $T_w = 5 \times 8 \times 60 = 2400$ (min). The specific allocation scheme of wages can be found in Table 1.

Table 1. Details of wages according to fluctuating mechanism $S_l(p_{lw})$.

Senior Technicians		General Operators	
Weekly Wages C_g (USD/Person)	**Working Intensity P_{gw}**	**Weekly Wages C_n (USD/Person)**	**Working Intensity P_{nw}**
1500	[0, 40%)	1000	[0, 60%)
2000	[40%, 60%)	1250	[60%, 70%)
2500	[60%, 100%)	1500	[70%, 100%)

Then, the system parameters k, c_n, c_g, and N can be determined by optimizing the profit of the shop.

6.1. Comparison of Interstate Transition Frequency between Theory and Simulation

In order to verify the theoretical calculation of interstate transition frequency, the MCMC (Markov chain Monte Carlo) method is employed. The system with parameters $k = 2$, $c_n = 6$, $c_g = 3$, and $N = 10$ was selected, and we compared several frequencies of interstate transition that were more likely to occur. The results are demonstrated in Figure 3.

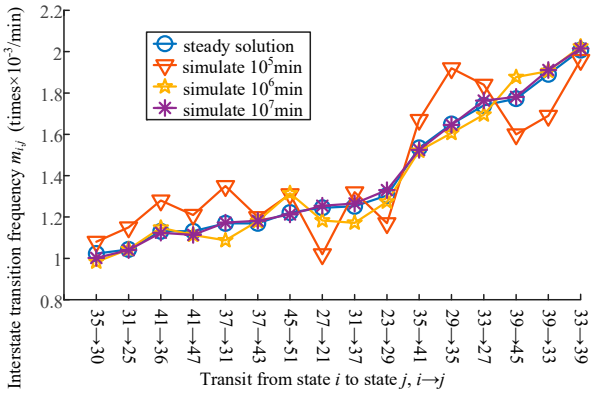


Figure 3. Comparison of the simulation and theoretical results.

As can be seen from Figure 3, it is evident that the simulation results gradually approach the theoretical results as the simulation runtime increases. At a runtime of 10^6 min (8.7 years), the simulation results are closely aligned with the theoretical results. Moreover, as the system reaches 10^7 min, the two curves nearly coincide, indicating that the system has reached a steady state. Thus, the proposed method is demonstrated to be valid.

6.2. Analysis of Optimization

In order to calculate the system's profit, we consider a month-long period consisting of 20 days and 8 h per day, resulting in $T_s = 20 \times 8 \times 60 = 9600$ (min). The maximum system profit is then determined for various combinations of k and N, and the results are presented in Figure 4.

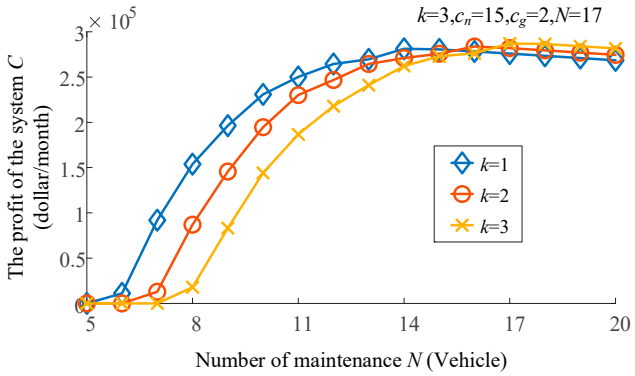


Figure 4. The maximum profit with different k and N.

The results displayed in Figure 4 indicate that the maximum profit initially increases and then decreases as ordinary customer capacity N increases. An obvious gradual trend is observed when $N \geq 14$. Specifically, when $k = 3$, $c_n = 15$, $c_g = 2$, and $N = 17$, the system achieves a maximum profit of USD 287,000. The proportion of income and cost in each part is shown in Figure 5.

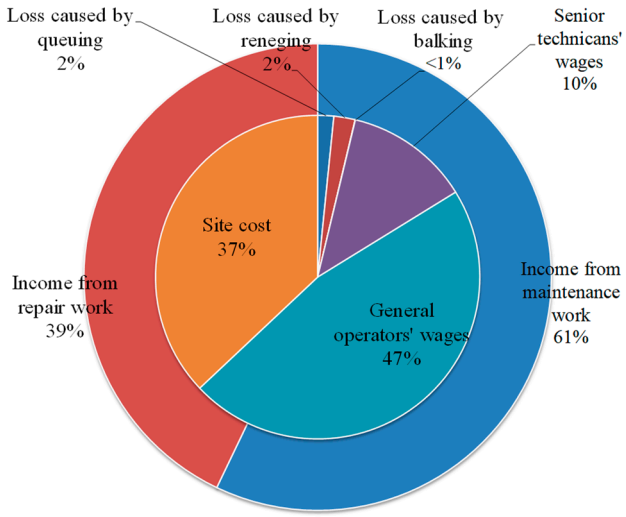


Figure 5. The proportion of income and cost in each part of the system.

Figure 5 depicts the income and cost of the system, which are represented by the outer and inner circles, respectively. The total income is USD 415,000, whereas the total cost amounts to USD 128,000.

The repair shop handles an average of 35.63 vehicle repairs and 237.27 vehicle maintenances each month. The majority of the income, 61%, is generated from maintenance services.

The system experiences a relatively high proportion of costs from the personnel and site, constituting 57% and 37%, respectively. Among the personnel expenses, ordinary operators have a working intensity of 58.08%, resulting in weekly wages of USD 2000 per person, according to Table 1. Similarly, the senior technicians have a working intensity of 46.55%, and their weekly wages are also set at USD 2000 per person.

It is noteworthy that the losses attributed to waiting and reneging account for 2% each, and the loss due to balking is less than 1%. These numbers indicate that customers receive satisfactory service under the current configuration. The total waiting time for maintenance vehicles is 205.8 min, resulting in an average waiting time of less than 1 min. Furthermore, the monthly number of reneging vehicles is 2.71, and the number of balking vehicles is 0.01, both of which can be considered marginal.

7. Conclusions

This study focuses on the analysis of a queuing system model that integrates multiple priorities, multiple abandonments, and heterogeneous servers. The model was solved and optimized by employing the combination of the Markov process with block matrix theory. The findings of this study are summarized as follows.

(1) A novel system indicator called interstate transition frequency is introduced. In the context of CTMC, interstate transition frequency is proposed as the new computational target, replacing the sojourn probability. This substitution is based on the ability of interstate transition frequency to decompose the sojourn probability. Consequently, the transitions with the same entering state can be distinguished during the evolution process. In order to accurately simulate the abandonment process, a non-sojourn state is introduced in the CTMC framework, and this process can only be described by using the new indicator;

(2) Based on interstate transition frequency, several system operation indicators are obtained, including service frequency, reneging frequency, balking frequency, total waiting time, and system profit. These indicators significantly enhance the evaluation methodology for queuing systems with multiple priority, multiple abandonments, and heterogeneous servers;

(3) The proposed queuing model was implemented in an automobile repair shop. The accuracy of interstate transition frequency was validated. The shop configuration was optimized, with a focus on maximizing system profit. Based on the optimization results, a succinct financial report is provided.

Author Contributions: Methodology, M.Y. (Ming Yan); Formal analysis, M.L.; Resources, M.Y. (Mingang Yin); Writing—review & editing, Y.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Foundation of China [grant number 52205451].

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jaiswal, N. Preemptive resume priority queue. *Oper. Res.* **1961**, *9*, 732–742. [CrossRef]
2. Kleinrock, L.; Finkelstein, R.P. Time dependent priority queues. *Oper. Res.* **1967**, *15*, 104–116. [CrossRef]
3. Chang, W. Queuing with Nonpreemptive and Preemptive-Resume Priorities. *Oper. Res.* **1965**, *13*, 1020–1022. [CrossRef]
4. Wu, J.B.; Liu, Z.M.; Peng, Y. A discrete-time Geo/G/1 retrial queue with preemptive resume and collisions. *Appl. Math. Model.* **2011**, *35*, 837–847. [CrossRef]
5. Lin, J.S.; Huang, C.Y.; Fang, C.C. Analysis and assessment of software reliability modeling with preemptive priority queueing policy. *J. Syst. Softw.* **2022**, *187*, 111249. [CrossRef]
6. Wang, Y.L.; Fan, Y.M.; Zhu, T.Z.; Cao, P.; Xie, J.G. An approximation method for 2-chain flexible queues with preemptive priority. *Int. J. Prod. Res.* **2019**, *57*, 5935–5950. [CrossRef]
7. Bai, X.J.; Jin, S.F. Performance analysis of an energy-saving strategy in cloud data centers based on a MMAP K/MK/N−1+N−2 non-preemptive priority queue. *Futur. Gener. Comput. Syst.* **2022**, *136*, 205–220. [CrossRef]
8. Halfin, S.; Whitt, W. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **1981**, *29*, 567–588. [CrossRef]
9. Aguir, S.; Karaesmen, F.; Aksin, O.Z.; Chauvet, F. The impact of retrials on call center performance. *OR Spectr.* **2004**, *26*, 353–376. [CrossRef]
10. Haight, F.A. Queueing with Balking. II. *Biometrika* **1960**, *47*, 285–296. [CrossRef]
11. Garnett, O.; Mandelbaum, A.; Reiman, M. Designing a Call Center with Impatient Customers. *Manuf. Serv. Oper. Manag.* **2002**, *4*, 208–227. [CrossRef]
12. We, X.; Tayfur, A. An approximation for multi-server queues with deterministic reneging times. *Ann. Oper. Res.* **2009**, *172*, 143–151. [CrossRef]
13. Wang, K.H.; Chang, Y.C. Cost analysis of a finite M/M/R queueing system with balking, reneging, and server breakdowns. *Math. Methods Oper. Res.* **2002**, *56*, 169–180. [CrossRef]

14. Ward, A.R.; Glynn, P.W. A diffusion approximation for a GI/GI/1 queue with balking or reneging. *Queueing Syst.* **2005**, *50*, 371–400. [CrossRef]
15. Falin, G. A survey of retrial queues. *Queueing Syst.* **1990**, *7*, 127–167. [CrossRef]
16. Artalejo, J.R.; Gómez-Corral, A. *Retrial Queueing Systems*; Springer: Berlin/Heidelberg, Germany, 1999. [CrossRef]
17. Takagi, H. Times until service completion and abandonment in an m/m/m preemptive-resume lcfs queue with impatient customers. *J. Ind. Manag. Optim.* **2018**, *14*, 1701–1726. [CrossRef]
18. Jouini, O.; Roubos, A. On multiple priority multi-server queues with impatience. *J. Oper. Res. Soc.* **2014**, *65*, 616–632. [CrossRef]
19. Brandwajn, A.; Begin, T. First-come-first-served queues with multiple servers and customer classes. *Perform. Eval.* **2019**, *130*, 51–63. [CrossRef]
20. Bouchentouf, A.A.; Messabihi, A. Heterogeneous two-server queueing system with reverse balking and reneging. *Opsearch* **2018**, *55*, 251–267. [CrossRef]
21. Liu, Y.N.; Sun, X.; Hovey, K. Scheduling to Differentiate Service in a Multiclass Service System. *Oper. Res.* **2022**, *70*, 527–544. [CrossRef]
22. Bhulai, S.; Blok, H.; Spieksma, F.M. K competing queues with customer abandonment: Optimality of a generalised cμ-rule by the Smoothed Rate Truncation method. *Ann. Oper. Res.* **2022**, *317*, 387–416. [CrossRef]
23. Matthieu, H.; Harten, A.; Sleptchenko, A. Approximations for Markovian multi-class queues with preemptive priorities. *Oper. Res. Lett.* **2004**, *32*, 273–282. [CrossRef]
24. Kim, B.; Kim, J.; Bueker, O. Non-preemptive priority M/M/m queue with servers' vacations. *Comput. Ind. Eng.* **2021**, *160*, 107390. [CrossRef]
25. Vasiliadis, G. Transient Analysis of the M/M/k/N/N Queue using a Continuous Time Homogeneous Markov System with Finite State Size Capacity. *Commun. Stat. Theory Methods* **2014**, *43*, 1548–1562. [CrossRef]
26. Sheikh, S.M.; Wolhuter, R.; Engelbrecht, H.A. A model for analyzing the performance of wireless multi-hop networks using a contention-based CSMA/CA strategy. *KSII Trans. Internet Inf. Syst.* **2017**, *11*, 2499–2522. [CrossRef]
27. Li, N.; Kong, N.; Li, Q.; Jiang, Z. Evaluation of reverse referral partnership in a tiered hospital system—A queuing-based approach. *Int. J. Prod. Res.* **2017**, *55*, 5647–5663. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.