Priority Queue with Customer Upgrades

Qi-Ming He,[1] Jingui Xie,[2] Xiaobo Zhao[3]

[1] *Department of Management Sciences, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1*

[2] *School of Management, University of Science and Technology of China, Hefei, People's Republic of China*

[3] *Department of Industrial Engineering, Tsinghua University, Beijing, People's Republic of China*

Received 24 December 2010; accepted 4 June 2012
DOI 10.1002/nav.21494
Published online 9 July 2012 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: This article is concerned with a general multi-class multi-server priority queueing system with customer priority upgrades. The queueing system has various applications in inventory control, call centers operations, and health care management. Through a novel design of Lyapunov functions, and using matrix-analytic methods, sufficient conditions for the queueing system to be stable or instable are obtained. Bounds on the queue length process are obtained by a sample path method, with the help of an auxiliary queueing system. © 2012 Wiley Periodicals, Inc. Naval Research Logistics 59: 362–375, 2012

Keywords: priority queue; customer upgrades; stability/ergodicity; matrix-analytic methods

1. INTRODUCTION

We consider a service system with S identical servers and N customer classes. The distinguishing feature in our model is that customers automatically progress from their current class to the next more important class after a so-called upgrading time. In the system, customers are served according to an absolute priority rule, that is, when a server frees up, the server proceeds to serve one of the customers in queue with the highest class index and, among all those, has been waiting the longest time. The service process is preemptive repeat, that is, a customer's service is preempted when all servers are busy and a new customer of a higher class index arrives; in this case, the customer's service needs to be repeated from scratch. We assume that the customer arrival processes are marked Markovian arrival processes (*MMAP*) and service times are of general phase type. The upgrade times are confined to a more specific class of phase type distributions, the so-called Coxian distributions. We derive system stability conditions and characterize the steady-state queue lengths.

Potential applications of the queueing model of interest can be found in emergency department management and perishable inventory control, where the system feature applies. In a hospital emergency department, patients are categorized into critical and noncritical groups. A patient in the critical group will be attended by a doctor, if one is available, as soon as the patient arrives. The condition of a patient in the noncritical group may deteriorate while waiting, and become critical. Then the patient has to be attended as soon as a doctor is available. For inventory control of perishable goods in a custom inspection system, perishable products such as food require immediate attention (e.g., [9]). Products may be expedited while waiting and need to be inspected at the earliest available time. In such systems, items have different service priorities. The service priority of an item may increase while waiting.

Applications of multi-class queueing models with customers' class constantly evolving over time are not limited to the above. For instance, Akan et al. [2] consider the problem of designing an efficient liver allocation system for allocating donated organs to patients waiting for transplantation, which is the only viable treatment for end-stage liver disease. They model the system as a multiclass queue, which captures the disease evolution by allowing the patients to switch between classes over time, for example, patients waiting for transplantation may get sicker/better, or may die, and use clinical databases and data from the current United Network for Organ Sharing System. In their model, both upgrades and downgrades between priorities are considered.

In the context of call centers application, customers can access the service provided by the call center either by phone or email. The incoming calls need to be answered quickly,

Correspondence to: Jingui Xie (xiej@ustc.edu.cn)

and hence have high priority. However, email requests not being responded to for a long time, will be replaced by a phone call by the customer. Then the problem is how to manage the call center with consideration of possible transfers between different types of service. Down and Lewis [11] introduce a new method of mitigating the problem of long wait times for low priority customers in a two-class queueing system, where low priority customers will be upgraded to the high priority class after they have been in queue for some time. This model is quite close to ours, except that there is an upper limit for the total number of low priority customers who can upgrade their priorities, and there are two groups of servers.

In this article, we emphasize the emergency health care application. In a hospital emergency department, every patient is triaged on presentation before its medical treatment, no matter which triage scale is used. Triage is an important concept involving prioritization using professional nursing judgment on patients' conditions [26]. The emergency department uses triage to determine the workflow within the department for the patients who need to be seen quickly, prioritizing those at high risk. This rations patient treatment efficiently when resources are insufficient for all to be treated immediately. Patients are generally classified into several priority categories. For instance, according to the Singapore Patients Acuity Category Scale [25], patients are triaged into four categories: P1, P2, P3, and P4. For P2 patients, it stated:

These patients are ill, non-ambulant, and in various forms of severe distress. …The severity of their symptoms requires very early attention. Failure to do so will like result in early deterioration of their medical status. …

We notice that there is deterioration of patient conditions in emergency departments, though there is no published data to support this. However, there are on average 41,467 (34.6%) P2 patients in the ED of Singapore General Hospital each year [25]. That implies that a certain number of patients will have medical status deterioration because of delayed medical treatment. This phenomenon will certainly affect the ED practice.

Motivated by the various applications in call centers and inventory control, especially in health care management, we introduce and analyze a preemptive repeat priority $MMAP[N]/(COX, PH)/S$ queue with priority upgrades. Our study is closely related to queueing systems with customer priorities and queueing systems with customer transfers, which have been studied extensively. Some of the existing papers focus on system stability conditions, some on the stationary analysis of the queue length(s) and waiting times, and some on customer transfer strategies [1,16,20,28,29,32–34]. However, a queueing model with both service priority and customer transfers (i.e., priority upgrades) has not been considered explicitly, except for Xie et al. [30,31].

Xie et al. [30,31] investigate a similar queueing model that has a Poisson arrival process, exponential service times, and exponential transfer times. Although the arrival process of an emergency department may be Poisson [15], the service times are almost surely not exponential. Recent empirical studies show that the arrival process of some service systems is not a Poisson process, and the service times are not exponentially distributed [4]. Thus, application of the model in Xie et al. [30,31] is limited due to its restrictive assumptions on the arrival, service, and upgrade processes.

In this article, we assume that the arrival process is a *MMAP*, the service times have phase-type distributions (*PH*), and the upgrading times have Coxian distributions. *MMAP* is a versatile process that can be used to model complicated multitype arrival processes with correlations between individual arrivals and/or with special arrival patterns. This is really helpful in modeling the arrival process in health care systems, where patients usually develop more than one disease. According to Asmussen and Koole [3], *MMAP* can adequately approximate any multitype arrival processes. Both *PH*-distributions and Coxian distributions can approximate any probability distribution defined on $[0, \infty)$. Consequently, the queueing model under investigation in this article is rather general and can be a good approximation to real systems.

For the preemptive repeat priority $MMAP[N]/(COX, PH)/S$ queue of interest, we obtain results on system stability conditions and stationary distributions of queue lengths. The results obtained in this article imply that system stability/instability depends only on the service rate of customers of the highest priority and the arrival rates of individual types of customers. That implies that system stability/instability is independent of the service rates and upgrade rates of lower priority customers. The results also imply that the correlations between the arrival processes of customers have no impact on system stability/instability. The results can be useful in the design of such queueing systems. For instance, the results indicate that, as far as system stability is concerned, there is no need to allocate excessive resource to the service of lower priority customers. Compared to that of Xie et al. [30], the queueing model considered in this article has fairly general assumptions on its arrival process, service times, and upgrade times, and captures features such as the correlations between arrivals, which do not exist in the simpler model. Therefore, the applicability of the results is extended significantly.

The mean-drift method, matrix-analytic methods, and stochastic comparison techniques are the main mathematical tools used in this article. The mean-drift method [7,8,12,13,21] has been used to classify Markov processes and queueing models. One of the key steps in using this method is the construction of the Lyapunov (test) functions. In this article, two special Lyapunov functions are introduced and they lead to the findings of simple conditions for system stability and instability. Matrix-analytic methods have been used in the analysis of complicated queueing models [24]. The methods are effective in the development of algorithms for computing performance measures. The methods have also been used in the study of stability conditions for queueing models [10, 18]. Stochastic comparison is also a widely used method in stochastic modeling [27].

The remainder of the paper is organized as follows. The queueing model of interest is introduced in Section 2. Section 3 defines a continuous time Markov chain (CTMC) associated with the queue length processes explicitly. System stability conditions are given in Section 4. Section 5 studies the queue length processes. Section 6 concludes the paper.

2. QUEUEING MODEL

The queueing model of interest consists of S identical servers serving N types of customers: type 1, type 2, …, and type N customers. Type 1, 2, …, and N customers form queue 1, 2, …, and N, respectively. Type N customers have the highest service priority, type $N-1$ the second highest service priority, …, and type 1 the lowest service priority. The priority level of a type k customer is k. The S servers are numbered as server 1, 2, …, and S.

A type k customer can upgrade to a type $k+1$ customer while it is waiting for service, for $1 \le k \le N-1$. The time for a waiting type k customer to be upgraded to a type $k+1$ customer is called the upgrade time. The clock of the upgrade time of a customer is set to zero as soon as the customer joins a queue waiting for service. A customer in service does not change its type or its priority level.

Customers are served on a preemptive repeat basis. That implies that an interrupted service is repeated. We also assume that the clock of the upgrade time of a customer whose service is interrupted is reset to zero. The service discipline is specified as follows.

a. Suppose that, when a type k customer arrives, some of the servers are idle. Then the type k customer is served immediately by one of the idle servers. Exactly which server is assigned does not affect system's stability and performance analysis.
b. Suppose that, when a type k customer arrives, all servers are busy. If the priority level of all customers in service is k or higher, then the type k customer joins queue k. The clock of the customer's upgrade time is set to zero and begins to click. If the priority level of some customers in service is lower than k, then one of the customers of the lowest priority in service is pushed out of its server and back into its queue, and the server begins to serve the type k customer immediately. The clock of the upgrade time of the customer pushed out is reset to zero and begins to click. The service of this customer will be repeated from scratch. Exactly which customer of the lowest priority in service is pushed out does not affect the system's stability analysis.
c. Suppose that, when a server completes serving a customer, at least one queue is not empty. Then the server chooses a customer from the nonempty queue of the highest priority and begins to serve it immediately. Exactly which customer is chosen from that queue does not affect the system stability and performance analysis. For mathematical convenience, a method for the server to choose a customer will be specified, after the distribution of the upgrade time is defined later in this section.
d. Suppose that, when a type k customer upgrades to a type $k+1$ customer, there are type k customers in service. Then one of the type k customers in service is pushed back into queue k. The server begins to serve the upgraded customer immediately. The clock of the upgrade time of the type k customer just pushed out is reset to zero and begins to click. That customer will repeat its service when it enters a server later.

Next, we define the arrival process, service times, and upgrade times explicitly.

The Arrival Process: The N types of customers arrive according to a MMAP (*MMAP*[N]) (Readers are referred to [3, 17, 19, 23] for more about *MMAP*[N]). The *MMAP*[N] has a matrix representation $\{D_0, D_J, J \in \Phi\}$, where Φ is a set of strings of integers defined as

$$\Phi = \{J : J = j_1 j_2 \cdots j_N, \text{ where } j_1, j_2, \ldots, j_N \ge 0, \\ J \neq 0, D_J \neq 0\}, \quad (2.1)$$

D_0 and $\{D_J, J \in \Phi\}$ are matrices of order m_a, D_0 is a matrix with negative diagonal elements and nonnegative off-diagonal elements, $\{D_J, J \in \Phi\}$ are nonnegative elements. The matrix D_J, $J \in \Phi$, is for the arrival rates of type J batches that include j_1 type 1 customers, j_2 type 2 customers, …, and j_N type N customers, conditioning on the phase of a underlying CTMC just before the arrival. Let $D = D_0 + \sum_{J\in\Phi} D_J$. We have $D\mathbf{e} = 0$, where $\mathbf{e}$ is a column vector with all elements being one. Then D is the infinitesimal generator of the underlying CTMC of the arrival process. We assume that the matrix D is irreducible, that is, the underlying CTMC is irreducible. Let $I_a(t)$ be the phase of the underlying CTMC at time t. Denote by $\boldsymbol{\theta}_a$ the nonnegative row vector satisfying $\boldsymbol{\theta}_a D = 0$ and $\boldsymbol{\theta}_a \mathbf{e} = 1$. Since D is irreducible, every element of $\boldsymbol{\theta}_a$ is positive. Then the stationary arrival rate of type k customers is given by $\lambda_k = \boldsymbol{\theta}_a \sum_{J\in\Phi} j_k D_J \mathbf{e}$, for $1 \le k \le N$.

Define $D^*(z) = D_0 + \sum_{J\in\Phi} z^{|J|} D_J$, where $|J| = j_1 + j_2 + \cdots + j_N$, which is the number of customers (regardless of their types) in the batch J. We assume that there exists $\hat{z} > 1$ such that $D^*(z)$ is a finite matrix for $0 < z < \hat{z}$. This assumption is not restrictive, since it is satisfied if the set Φ has a finite number of elements or the batch size has a discrete phase-type distribution [24].

To make it easy to understand $MMAP[N]$, we give two examples of $MMAP[N]$.

EXAMPLE 2.1: Assume that all customers arrive individually, i.e., all batch sizes are one. For this case, $\Phi = \{10\cdots0, 010\cdots0, \ldots, 0\cdots01\}$. A string $J = 0\cdots010\cdots0$, whose kth number is 1, represents a batch that has a single type k customer in it.

EXAMPLE 2.2: Assume that $N = 2$ and $\Phi = \{10, 01, 11, 22\}$. For this case, customers arrive in four forms: a single type 1 arrival ($J = 10$), a single type 2 arrival ($J = 01$), a batch with one type 1 customer and one type 2 customer ($J = 11$), and a batch with 2 type 1 customers and 2 type 2 customers ($J = 22$).

The Service Times: The service times of the type k customers have the same phase-type distribution with a *PH*-representation $(\boldsymbol{\alpha}_k, T_k)$ of order m_k, $1 \le k \le N$, where $\boldsymbol{\alpha}_k$ is a stochastic vector, i.e., $\boldsymbol{\alpha}_k$ is nonnegative and $\boldsymbol{\alpha}_k\mathbf{e} = 1$ (which implies that the service time is positive with probability one), and T_k is a *PH*-generator, i.e., T_k is invertible, diagonal elements of T_k are negative, off-diagonal elements of T_k are nonnegative, and the vector $\mathbf{T}_k^0 = -T_k\mathbf{e}$ is nonnegative. The mean service time of type k customers is $\mu_k^{-1} = -\boldsymbol{\alpha}_k T_k^{-1}\mathbf{e}$, $1 \le k \le N$. Then μ_k is the service rate of type k customers. Without loss of generality, we assume that the *PH*-representation $(\boldsymbol{\alpha}_k, T_k)$ is irreducible, which is equivalent to that the infinitesimal generator $T_k + \mathbf{T}_k^0\boldsymbol{\alpha}_k$, is irreducible. The irreducibility of the *PH*-representation is assumed to ensure that a CTMC to be defined for the queue length processes is irreducible. Let $\boldsymbol{\theta}_k$ be the nonnegative vector satisfying $\boldsymbol{\theta}_k(T_k + \mathbf{T}_k^0\boldsymbol{\alpha}_k) = 0$ and $\boldsymbol{\theta}_k\mathbf{e} = 1$. Then all elements of $\boldsymbol{\theta}_k$ are positive. In fact, it can be verified that $\boldsymbol{\theta}_k = -\mu_k\boldsymbol{\alpha}_k T_k^{-1}$. Let $I_i(t)$ be the phase of the underlying CTMC of the service undergoing in server i at time t, $1 \le i \le S$. If server i is idle, we define $I_i(t) = 0$. Note that the range of $I_i(t)$ depends on the type of the customer in service at time t. We refer to [24] for more about phase-type distributions.

The Upgrade Times: The upgrade time of a type k customer to a type $k+1$ customer has a Coxian distribution with a Coxian representation $(\boldsymbol{\beta}_k, S_k)$ of order n_k, $1 \le k \le N-1$, where $\boldsymbol{\beta}_k = (\beta_{k,1}, \beta_{k,2}, \ldots, \beta_{k,n_k})$ is a stochastic vector (i.e., $\boldsymbol{\beta}_k \ge 0$ and $\boldsymbol{\beta}_k\mathbf{e} = 1$, which implies that the upgrade time is positive with probability one) and S_k is a Coxian generator given as follows:

$$S_k = \begin{pmatrix} -s_{k,1} & s_{k,1} & & & \\ & -s_{k,2} & s_{k,2} & & \\ & & \ddots & \ddots & \\ & & & -s_{k,n_k-1} & s_{k,n_k-1} \\ & & & & -s_{k,n_k} \end{pmatrix}. \quad (2.2)$$

It is easy to see that Coxian distribution is a special case of *PH*-distribution. Without loss of generality, we assume that $\beta_{k,1} > 0$ for $1 \le k \le N-1$. Similar to the irreducibility condition made on the *PH*-representations of service times, this assumption is made to ensure that a CTMC to be defined for the queue length processes is irreducible. For the Coxian distribution, the phase of its underlying CTMC always moves up by one at a transition epoch. This property is used in the definitions of the queue length processes. For type N customers, there is no upgrade. We introduce a underlying CTMC with a single phase and no transition, i.e., $n_N = 1$, $s_{N,1} = 0$, and $\beta_{N,1} = 1$.

The distributions of upgrade times are chosen to be Coxian distributions, instead of *PH*-distributions, for simplicities of the CTMC to be defined for the queue length processes and the proof of Theorem 4.2. Since the set of Coxian distributions, similar to the set of *PH*-distributions, is dense in the set of probability distributions with nonnegative support, the assumption is not restrictive.

We assume that the arrival process, service times, and upgrade times are independent. We summarize some key technical assumptions as follows.

ASSUMPTION 2.1: The matrix D is irreducible, $D^*(z)$ is a finite matrix for $0 < z < \hat{z}$ with $\hat{z} > 1$, the *PH*-representations of all service times are *PH*-irreducible, and $\beta_{k,1} > 0$ for $1 \le k \le N$.

3. THE CTMC $\{\mathbf{X}(t), t \ge 0\}$

Based on the above definitions, we define a CTMC for queue 1, queue 2, ..., and queue N. For any type k customers waiting for service, the underlying CTMC of its upgrade time must be in a state i, $1 \le i \le n_k$. Thus, the queue k of type k customers can be decomposed into $n_k + 1$ subqueues: queue $(k, 0)$ consists of type k customers in service, queue $(k, 1)$ consists of type k customers waiting in queue k and the underlying CTMCs of their upgrade times are in state 1, ..., and queue (k, n_k) consists of type k customers waiting in queue k and the underlying CTMCs of their upgrade times are in state n_k. When a server is available to serve a type k customer, we assume that one of the type k customers whose underlying CTMC is in the highest phase is chosen for service.

Define $q_{k,i}(t)$ the number of type k customers in queue (k,i) at time t, $0 \leq i \leq n_k$, $k = 1,2,\ldots,N$. Let $\mathbf{q}_k(t) = (q_{k,0}(t), q_{k,1}(t), \ldots, q_{k,n_k}(t))$, $1 \leq k \leq N$. Because of the preemption property, we must have $q_{j,i}(t) = 0$ for $j \geq k+1$ and $1 \leq i \leq n_j$, if $q_{k,0}(t) > 0$. Define $\mathbf{q}(t) = (\mathbf{q}_1(t), \mathbf{q}_2(t), \ldots, \mathbf{q}_{N-1}(t), \mathbf{q}_N(t))$ and $\mathbf{X}(t) = (\mathbf{q}(t), I_a(t), I_1(t), \ldots, I_s(t))$. The first part of $\mathbf{X}(t)$ (i.e., $\mathbf{q}(t)$) provides information on the lengths of the N queues as well as upgrade times of customers in queues. The rest of $\mathbf{X}(t)$ (i.e., $(I_a(t), I_1(t), \ldots, I_s(t))$) provides information on the underlying phases of the arrival process and service times. It can be verified, under our assumptions, the CTMC $\{\mathbf{X}(t), t \geq 0\}$ is irreducible. Denote by Ω the state space of $\{\mathbf{X}(t), t \geq 0\}$. A typical state in Ω has the form $\mathbf{x} = (\mathbf{q}, i_a, i_1, \ldots, i_s)$ with $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{N-1}, \mathbf{q}_N)$ and $\mathbf{q}_k = (q_{k,0}, q_{k,1}, \ldots, q_{k,n_k})$, $1 \leq k \leq N$.

Denote by Q the infinitesimal generator of the CTMC $\{\mathbf{X}(t), t \geq 0\}$. First, we find Q explicitly. For that purpose, we define the following sets:

$$\begin{aligned}
&\Psi(\mathbf{q}) = \{\mathbf{x} : \mathbf{x} = (\mathbf{q}, i_a, i_1, \ldots, i_s) \in \Omega\};\\
&\mathcal{A}(J) = \left\{\boldsymbol{\delta} : \quad \boldsymbol{\delta} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{N-1}, \boldsymbol{\delta}_N) \geq 0, \right.\\
&\qquad \left. \sum_{i=1}^{n_k} \delta_{k,i} = j_k, 1 \leq k \leq N \right\}, \quad J \in \Phi;\\
&\mathcal{B}(\mathbf{q}, J) = \{\mathbf{y} : \quad \text{If batch } J \text{ arrives, the queues can}\\
&\qquad \text{changes from } \mathbf{q} \text{ to } \mathbf{y}.\}, \quad J \in \Phi. \qquad (3.1)
\end{aligned}$$

Note that $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{N-1}, \mathbf{q}_N)$, $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{N-1}, \mathbf{y}_N)$, and $\boldsymbol{\delta}_k = (0, \delta_{k,1}, \ldots, \delta_{k,n_k})$, $1 \leq k \leq N$, are all nonnegative. The set $\Psi(\mathbf{q})$ includes all the states associated with queue $\mathbf{q}$; $\mathcal{A}(J)$ consists of all the possible breakdowns of the batch J; and $\mathcal{B}(\mathbf{q}, J)$ has all the queues formed by customers from the original queue $\mathbf{q}$ and customers from the batch J.

We also define the following constant: for $J \in \Phi$ and $\boldsymbol{\delta} \in \mathcal{A}(J)$,

$$p(J, \boldsymbol{\delta}) = \prod_{k=1}^{N} \left(\frac{j_k!}{(\delta_{k,1}! \cdots \delta_{k,n_k}!)} \prod_{i=1}^{n_k} \beta_{k,i}^{\delta_{k,i}} \right). \qquad (3.2)$$

If $q_{N,0} = S$, all arriving customers join the queues. For such cases, $p(J, \boldsymbol{\delta})$ is the probability that the batch J breaks into $\boldsymbol{\delta} \in \mathcal{A}(J)$ and the queues changes from $\mathbf{q}$ to $\mathbf{y} = \mathbf{q} + \boldsymbol{\delta}$ (i.e., $y_{k,i} = q_{k,i} + \delta_{k,i}$). By the law of total probability, we have

$$1 = \sum_{\boldsymbol{\delta} \in \mathcal{A}(J)} p(J, \boldsymbol{\delta}), \qquad \text{for } \mathbf{q} \geq 0, \ J \in \Phi. \qquad (3.3)$$

Let $Q(\mathbf{q}, \mathbf{y})$ be the transition rate block in Q for the queue lengths $(\mathbf{q}_1(t), \mathbf{q}_2(t), \ldots, \mathbf{q}_{N-1}(t), \mathbf{q}_N(t))$ changing from $\mathbf{q}$ to $\mathbf{y}$, that is, the transition block from the set $\Psi(\mathbf{q})$ to $\Psi(\mathbf{y})$. The corresponding transition rates for the variables $(I_a(t), I_1(t), \ldots, I_s(t))$ are given within the matrix $Q(\mathbf{q}, \mathbf{y})$.

The status of the CTMC $\{\mathbf{X}(t), t \geq 0\}$ changes when there is an arrival, a service completion, a upgrade of customer, or a change in the underlying CTMCs of the arrival process, service times, and the upgrade times. To obtain $Q(\mathbf{q}, \mathbf{y})$, we need to consider the following five cases: (i) The queue $\mathbf{q}$ remains the same; (ii) The queue $\mathbf{q}$ changes due to an arrival; (iii) The queue $\mathbf{q}$ changes due to a service completion; (iv) The queue $\mathbf{q}$ changes due to a customer upgrade; and (v) The queue $\mathbf{q}$ changes due to an internal phase change of a upgrade time.

Case (i): The Queue q Remains the Same: For this case, only transitions in the underlying CTMCs that do not lead to arrivals, service completions, or customer upgrades can occur. Thus, we have

$$Q(\mathbf{q}, \mathbf{q}) = D_0 \otimes I + \sum_{i=1}^{S(\mathbf{q})} I \otimes T_{k_i} \otimes I - \left(\sum_{k=1}^{N-1} \sum_{i=1}^{n_k} q_{k,i} s_{k,i} \right) I, \qquad (3.4)$$

where I is the identity matrix, "$\otimes$" is for Kronecker product of matrices [14], $S(\mathbf{q}) = \sum_{k=1}^{N} q_{k,0}$ is the number of customers in service, and k_i is the type of the customer being served in server i. The orders of the identity matrices in Eq. (3.4) (and in some other equations) depend on numbers of phases of the underlying CTMCs of the arrival process and service times, which is determined by $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{N-1}, \mathbf{q}_N)$. For instance, if $q_{N,0} = S$, the expression "$I \otimes T_k \otimes I$" in Eq. (3.4) should be "$I_{m_a m_N^{i-1}} \otimes T_N \otimes I_{m_N^{S-i}}$", where I_n is an identity matrix of order n. We ignore the indices for convenience and with the understanding that the orders of the identity matrices depend on the queue $\mathbf{q}$.

Case (ii): The Queue q Changes due to an Arrival: Suppose that a type J batch arrives when the queue status is $\mathbf{q}$. Then we have, for $\mathbf{y} = \mathbf{q} + \boldsymbol{\delta} \in \mathcal{B}(\mathbf{q}, J)$,

$$Q(\mathbf{q}, \mathbf{y}) = p(J, \boldsymbol{\delta})(D_J \otimes I_{m_N^S}), \qquad \text{if } q_{N,0} = S. \qquad (3.5)$$

If $q_{N,0} < S$, the expression of $Q(\mathbf{q}, \mathbf{y})$ is quite complicated because some of the lower priority customers in service may be pushed back into their queues due to the arrival of some higher priority customers. The queues can be changed and so are the underlying CTMCs of the services. We do not give an explicit expression of $Q(\mathbf{q}, \mathbf{y})$ for this case since it is not used in the rest of the paper. Nonetheless, the following property of $Q(\mathbf{q}, \mathbf{y})$ is used in this paper: for $J \in \Phi$,

$$\sum_{\mathbf{y} \in \mathcal{B}(\mathbf{q}, J)} Q(\mathbf{q}, \mathbf{y})\mathbf{e} = (D_J \mathbf{e}) \otimes \mathbf{e}, \qquad \text{for } \mathbf{q} \geq 0. \qquad (3.6)$$

Case (iii): The Queue q Changes due to a Service Completion: For this case, we consider two situations: (1) At a service completion epoch, there are customers waiting for service; (2) At a service completion epoch, there is no customer waiting for service. Suppose that the service of a type k customer is completed and the non-empty waiting queue with the highest priority is l. Then we have, for $k \le N$,

$$Q(\mathbf{q},\ \mathbf{y}) = \begin{cases} I \otimes \left(\mathbf{T}_k^0 \boldsymbol{\alpha}_l\right) \otimes I, & \text{if } y_{l,i} = q_{l,i} - 1 \ge 0,\ q_{l,p} = 0,\ i < p \le n_l, y_{l,0} = q_{l,0} + 1, \\ & \quad q_{j,i} = 0, l < j \le N,\ 1 \le i \le n_j,\ y_{k,0} = q_{k,0} - 1, \\ & \quad \sum_{j=1}^{N} q_{j,0} = S, \text{ and for all other } (j,i),\ y_{j,i} = q_{j,i}; \\ I \otimes \mathbf{T}_k^0 \otimes I, & \text{if } y_{j,i} = q_{j,i} = 0,\ 1 \le j \le N,\ 1 \le i \le n_j, \\ & \quad y_{k,0} = q_{k,0} - 1 \ge 0,\ y_{j,0} = q_{j,0},\ j \ne k, \end{cases} \tag{3.7}$$

where the orders of the identity matrices are determined by the number of phases of the underlying CTMCs of the arrival process, the services in progress, and the service to be completed.

Case (iv): The Queue q Changes due to a Customer Upgrade:

For this case, we consider two situations: (1) Upgrade occurs to a customer whose priority level is lower than all customers in service; and (2) Upgrade occurs to a customer whose priority level equals to the lowest priority level of customers in service. For case (2), the upgraded customer pushes out a customer of the lowest service priority in service and begins its own service. Note that the priority level of a waiting customer cannot be higher than that of any customer in service due to the preemption assumption. Then we have

$$Q(\mathbf{q}, \mathbf{y}) = \begin{cases} \beta_{k+1,i} q_{k,n_k} s_{k,n_k} I, & \text{if } y_{k+1,i} = q_{k+1,i} + 1, y_{k,n_k} = q_{k,n_k} - 1 \ge 0, \\ & \quad \sum_{n=k+1}^{N} q_{n,0} = S, \text{ and for all other } (j,l),\ y_{j,l} = q_{j,l}; \\ \beta_{k,i} q_{k,n_k} s_{k,n_k} I \otimes (\mathbf{e}\boldsymbol{\alpha}_{k+1}) \otimes I, & \text{if } y_{k+1,0} = q_{k+1,0} + 1, y_{k,n_k} = q_{k,n_k} - 1 \ge 0, \\ & \quad y_{k,0} = q_{k,0} - 1 \ge 0, y_{k,i} = q_{k,i} + 1, \\ & \quad \sum_{n=k+1}^{N} q_{n,0} < \sum_{n=k}^{N} q_{n,0} = S, \\ & \quad \text{and for all other } (j,l),\ y_{j,l} = q_{j,l}. \end{cases} \tag{3.8}$$

Case (v): The Queue q Changes due to an Internal Phase Change of an Upgrade Time:

For this case, there is no customer upgrade. We have

$$Q(\mathbf{q}, \mathbf{y}) = q_{k,i} s_{k,i} I, \quad \text{if } y_{k,i+1} = q_{k,i+1} + 1,\ y_{k,i} = q_{k,i} - 1 \ge 0, \quad \text{and for all other } (j,l) : y_{j,l} = q_{j,l}. \tag{3.9}$$

For all other cases, we have $Q(\mathbf{q}, \mathbf{y}) = 0$.

4. STABILITY ANALYSIS

Define the matrix generating function

$$A^*(z) = \sum_{i=1}^{S} I_{m_a m_N^{i-1}} \otimes \left(\mathbf{T}_N^0 \boldsymbol{\alpha}_N\right) \otimes I_{m_N^{S-i}} + z \left(D_0 \otimes I_{m_N^S} + \sum_{i=1}^{S} I_{m_a m_N^{i-1}} \otimes T_N \otimes I_{m_N^{S-i}} \right) + z \sum_{J \in \Phi} z^{|J|} D_J \otimes I_{m_N^S}. \tag{4.1}$$

By the assumption on $D^*(z)$, the matrix $A^*(z)$ is well defined for $0 < z < \hat{z}$. For $z > 0$, the matrix $A^*(z)$ has nonnegative off-diagonal elements. Under Assumption 1, the matrix $A^*(z)$ is an irreducible M-matrix [14]. Denote by $\rho(z)$ the eigenvalue of $A^*(z)$ with the largest real part for $z > 0$. Then $\rho(1) = 0$, since the matrix $A^*(1)$ is an infinitesimal generator. Similar to the proof of Lemma 1.3.3 in [24], the following results can be proved.

LEMMA 4.1: Under Assumption 1, the derivative of $\rho(z)$ at $z = 1$ is given as $\rho^{(1)}(1) = \sum_{k=1}^{N} \lambda_k - S\mu_N$. In addition, we have

(4.1.1) If $\sum_{k=1}^{N} \lambda_k - S\mu_N < 0$, there exist $z^* > 1$ such that $\rho(z) < 0$ for $1 < z < z^* \leq \hat{z}$.

(4.1.2) If $\sum_{k=1}^{N} \lambda_k - S\mu_N > 0$, there exist $0 \leq z^* < 1$ such that $\rho(z) < 0$ for $z^* < z < 1$.

PROOF: For any $z > 0$ such that the matrix $A^*(z)$ is well-defined, since $A^*(z)$ is an M-matrix [14], there exist a left eigenvector $\mathbf{u}(z)$ and a right eigenvector $\mathbf{v}(z)$ satisfying $\mathbf{u}(z)A^*(z) = \rho(z)\mathbf{u}(z), A^*(z)\mathbf{v}(z) = \rho(z)\mathbf{v}(z), \mathbf{u}(z)\mathbf{e} = 1$, and $\mathbf{u}(z)\mathbf{v}(z) = 1$. Since $A^*(z)$ is irreducible, all elements of $\mathbf{u}(z)$ and $\mathbf{v}(z)$ are positive. It is easy to verify that

$$\mathbf{u}(1) = \boldsymbol{\theta}_a \otimes \left(\bigotimes_{i=1}^{S} (-\mu_N \boldsymbol{\alpha}_N T_N^{-1}) \right) \quad \text{and} \quad \mathbf{v}(1) = \mathbf{e}. \tag{4.2}$$

Note that $\boldsymbol{\theta}_N = -\mu_N \boldsymbol{\alpha}_N T_N^{-1}$. According to the proof of Lemma 1.3.3 in [24], the vectors $\mathbf{u}(z)$ and $\mathbf{v}(z)$ can be chosen to be analytic. By routine calculations, we obtain

$$\begin{aligned}
\rho^{(1)}(1) &= \mathbf{u}(1)A^{*(1)}(1)\mathbf{v}(1) \\
&= \sum_{J\in\Phi} (|J|+1)\mathbf{u}(1)(D_J \otimes I_{m_N^S})\mathbf{e} \\
&\quad + \mathbf{u}(1)\left(D_0 \otimes I_{m_N^S} + \sum_{i=1}^{S} I_{m_a m_N^{i-1}} \otimes T_N \otimes I_{m_N^{S-i}} \right)\mathbf{e} \\
&= \sum_{J\in\Phi} \left(\sum_{k=1}^{N} j_k \right) \mathbf{u}(1)(D_J \otimes I_{m_N^S})\mathbf{e} \\
&\quad + \mathbf{u}(1)\left(D_0 \otimes I_{m_N^S} + \sum_{J\in\Phi} (D_J \otimes I_{m_N^S}) \right. \\
&\qquad \left. + \sum_{i=1}^{S} I_{m_a m_N^{i-1}} \otimes T_N \otimes I_{m_N^{S-i}} \right)\mathbf{e} \\
&= \sum_{k=1}^{N} \lambda_k - S\mu_N.
\end{aligned} \tag{4.3}$$

For (4.1.1), $\rho^{(1)}(1)$ is negative. Since $\rho(z)$ is differentiable in $[0, \hat{z}]$ and $\rho(1) = 0$, we must have $\rho(z) < 0$ for $z > 1$ and z close to 1. The case (4.1.2) can be proved similarly. This completes the proof Lemma 4.1. □

We call the queueing system stable if the CTMC $\{\mathbf{X}(t), t \geq 0\}$ is ergodic (irreducible and positive recurrent). The CTMC is called non-ergodic if it is not ergodic. We call the queueing system instable if the CTMC $\{\mathbf{X}(t), t \geq 0\}$ is non-ergodic. The ergodicity of the CTMC $\{\mathbf{X}(t), t \geq 0\}$ is characterized in the following theorem.

THEOREM 4.2: Under Assumption 1, the CTMC $\{\mathbf{X}(t), t \geq 0\}$ is irreducible.

(4.2.1) The CTMC $\{\mathbf{X}(t), t \geq 0\}$ is ergodic if $\sum_{k=1}^{N} \lambda_k < S\mu_N$.

(4.2.2) The CTMC $\{\mathbf{X}(t), t \geq 0\}$ is non-ergodic if

$$\sum_{k=1}^{N} \lambda_k > S\mu_N.$$

Part (4.2.1) and part (4.2.2) indicate that the ergodicity/non-ergodicity conditions of the CTMC (or the stability/instability of the queueing system) are independent of the service rates and the upgrade rates of lower priority customers. For the special case with $\lambda_1 + \cdots + \lambda_N = S\mu_N$, our simulation results seem to suggest that the CTMC $\{\mathbf{X}(t), t \geq 0\}$ is null-recurrent. The conditions in Theorem 4.2 can be interpreted intuitively as follows. Because of the upgrades of customers, all queues, except queue N, will always be finite. Since all waiting customers of lower priority may upgrade to customers of the highest priority, for queue N to be stable, the server must have the capacity to serve all customers at the priority level N, regardless of their original priority level.

We have the following observations based on Theorem 4.2.

i. System stability is determined by the service rate of customers of the highest priority and the arrival rates of individual types of customers.
ii. The correlations between individual arrival processes, the service rates of lower priority customers, and the upgrade rates of lower priority customers have no impact on system stability.
iii. According to Asmussen and Koole [3], *MMAP*[*N*] can approximate any arrival process. It is well-known that both the set of *PH*-distributions and the set of Coxian distributions are dense in the set of all probability distributions with nonnegative support. Thus, Theorem 4.2 holds for fairly general queueing models.

In addition, we would like to remark that focusing only on the service time of the highest priority customers is crucial for stability. On the other hand, one also needs to look closely at the allocation of the resources to all customers to reduce the waiting times (delays) of lower priority customers. This is especially important for hospital applications, since it is not a good idea to allow the deterioration of patient's conditions due to prolonged waiting. Thus, the trade-off between keeping the total system queue length finite (system stability) and reducing the waiting times of lower priority customers should be taken under consideration in the design of such queueing systems.

5. QUEUE LENGTHS

To analyze the queue length processes, we first introduce an auxiliary queueing system. Then we use sample path method to compare queue lengths in the two systems, and give the bounds on the mean queue length of all lower priority queues in the original system.

Consider a queueing system that is the same as the one defined in Section 2 except that servers only serve queue N. We call this queueing system a *designated service queueing system* and our original queueing system a *priority service queueing system*. It is readily seen that, in the designated service queueing system, all customers are eventually upgraded to type N customers before being served. For the designated service queueing system, define $q^u_{k,i}(t)$ the number of type k customers in queue (k, i) at time t, $0 \le i \le n_k$, $k = 1, 2, \ldots, N$. Let $\mathbf{q}^u_k(t) = (q^u_{k,1}(t), q^u_{k,2}(t), \ldots, q^u_{k,n_k}(t))$, $1 \le k \le N$ and $\mathbf{q}^u(t) = (\mathbf{q}^u_1(t), \mathbf{q}^u_2(t), \ldots, \mathbf{q}^u_N(t))$.

To analyze the designated service queueing system, we look at the customers originally form type k $(1 \le k < N)$. The upgrade process of type k customers is independent of any other types of customers. Consequently, all lower priority customers who come to the system as a type k customer originally form an $MAP/(\text{COX}_k + \text{COX}_{k+1} + \cdots + \text{COX}_{N-1})/\infty$ queue, where $\text{COX}_k + \text{COX}_{k+1} + \cdots + \text{COX}_{N-1}$ represents the sum of the Coxian random variables $\{(\boldsymbol{\beta}_k, S_k), 1 \le k \le N - 1\}$. Recall that the stationary arrival rate of type k customers is given by $\lambda_k = \boldsymbol{\theta}_a \sum_{J \in \Phi} j_k D_J \mathbf{e}$, for $1 \le k \le N$. The mean service time of type k customers (i.e., the mean upgrade time from queue k to queue $k+1$) is $\tau_k^{-1} = -\boldsymbol{\beta}_k S_k^{-1} \mathbf{e}$, $1 \le k < N$.

LEMMA 5.1: The mean queue length of queue k $(1 \le k < N)$ in the designated service queueing system is $\sum_{i=1}^{k} \lambda_i \tau_k^{-1}$.

PROOF: For a customer originally from type k $(1 \le k < N)$, the average time in queues before join the highest priority queue is $\tau_k^{-1} + \cdots + \tau_{N-1}^{-1}$. By little's law, the average queue size formed by these customers is $\lambda_k(\tau_k^{-1} + \cdots + \tau_{N-1}^{-1})$. Similarly, the average queue size formed by type k customers in queue j $(k \le j < N)$ is $\lambda_k \tau_j^{-1}$. Note that the customers in queues come from queues $1, 2, \cdots$, and $k - 1$, and queue k itself. Then, the mean queue length of queue k is $\sum_{i=1}^{k} \lambda_i \tau_k^{-1}$. This completes the proof of Lemma 5.1. □

The designated service queueing system provides bounds on the queue length of the priority service queueing system. To obtain the bounds, we first establish some sample path relationships between $\mathbf{q}^u(t)$ and $\mathbf{q}(t)$ in Lemma 5.2. Define $|\mathbf{Y}(t)| = Y_1(t) + \cdots + Y_n(t)$, if $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_n(t))$.

LEMMA 5.2: Assume that both the priority service queueing system and the designated service queueing system are empty at time zero. For the queue length processes $\mathbf{q}(t)$ and $\mathbf{q}^u(t)$, we have, for any $1 \le k \le N$, $1 \le j \le n_k$ and $t \ge 0$,

i. $q_{k,j}(t) \le_{st} q^u_{k,j}(t) + (\sum_{i=1}^{k-1} n_i + j)S$;
ii. $q_{1,1}(t) + \cdots + q_{k,j}(t) \le_{st} q^u_{1,1}(t) + \cdots + q^u_{k,j}(t) + (\sum_{i=1}^{k-1} n_i + j)S$;
iii. $|\mathbf{q}_k(t)| \le_{st} |\mathbf{q}^u_k(t)| + S \sum_{i=1}^{k} n_i$;
iv. $|\mathbf{q}_1(t)| + \cdots + |\mathbf{q}_k(t)| \le_{st} |\mathbf{q}^u_1(t)| + \cdots + |\mathbf{q}^u_k(t)| + S \sum_{i=1}^{k} n_i$.

PROOF: It is well known that the sample path order implies stochastically order (see [27] for more about stochastic orders of random variables). We will show the inequalities for all possible sample paths in the probability space. □

For subqueue $(1, 0)$ at time t, recall that it consists of type 1 customers in service, and there are S servers in total. Hence, we have $q_{1,0}(t) \le S$.

Consider subqueue (1, 1) at time t. Customers arrived in $[0, t)$ can be categorized as follows. For all customers that have been upgraded to subqueue $(1, 2)$ without receiving service (in the priority service queueing system), they are no longer part of $q_{1,1}(t)$ (in the priority service queueing system) nor part of $q^u_{1,1}(t)$ (in the designated service queueing system). Note that upgrade times are exponentially distributed in a subqueue so that we can assume that a upgrading process is resumed after a service interruption. So, for all customers that have been upgraded to next subqueue, but was ever attended by servers in the priority service queueing system before being upgraded, they are no longer part of $q_{1,1}(t)$ nor part of $q^u_{1,1}(t)$. For all customers that have been served in the priority service queueing system, they are no longer part of $q_{1,1}(t)$ but can be part of $q^u_{1,1}(t)$ in the designated service queueing system. For all customers that have been attended by a server(s) but still in the upgrading process, they are part of $q_{1,1}(t)$ but may not be part of $q^u_{1,1}(t)$. There can be at most S such customers, since there are in total s servers and customers are served on a first-in-first-served basis. Thus, we must have $q_{1,1}(t) \le q^u_{1,1}(t) + S$.

For subqueue (1, 2) at time t, customers still in the queue can be divided into two groups: (i) customers who arrive to subqueue (1, 2) at the same epochs for both systems (from outside or other queues); (ii) customers who arrive to subqueue (1, 2) at different time epochs for the two systems (from other queues). For customers in the first group, similar to the above analysis for subqueue (1, 1), the subqueue in the priority service queueing system is at most S customers more than that in the designated service queueing system at time t. The second group includes at most S customers who are upgraded from subqueue (1, 1) and have been attended by a server in subqueue (1, 1) (but their services were never completed.) The reason is that subqueue (1, 2) must have no waiting customer before a server can attend a customer in subqueue (1, 1). These customers can be part of $q_{1,2}(t)$, but may not be part of $q^u_{1,2}(t)$. Therefore, we must have $q_{1,2}(t) \le q^u_{1,2}(t) + 2S$.

For subqueue $(1, j)$, where $1 \le j \le n_1$, among customers who arrive to subqueue $(1, j)$ at the same time epochs for both systems, at most S of them can be in the priority service queueing system, but not in the designated service queueing system. There are at most $(j-1)S$ upgraded customers who arrive to subqueue $(1, j)$ at different time epochs for the two systems—at most S from subqueue $(1, i)$ (those are customers in subqueue $(1, j)$ who have received incomplete service in subqueue $(1, i)$), $i = 1, 2, \ldots, j-1$. Therefore, we must have $q_{1,j}(t) \le q^u_{1,j}(t) + jS$.

Consider the first j $(1 \le j \le n_1)$ subqueues for type 1 customers. There can be at most s customers that have received incomplete service in subqueue $(1, j)$, are still in the priority service queueing system but not in the designated service queueing system. The reason is that all such customers have to be served in one of the higher priority subqueues $(1, j+1), (1, j+2), \ldots$, before a server can move down to serve queue j (to possibly create another such customer in queue j). Therefore, we have $q_{1,1}(t) + q_{1,2}(t) + \cdots + q_{1,j}(t) \le q^u_{1,1}(t) + q^u_{1,2}(t) + \cdots + q^u_{1,j}(t) + jS$, for $1 \le j \le n_1$ and $t \ge 0$. In addition, we have

$$q_{1,0}(t) + q_{1,1}(t) + q_{1,2}(t) + \cdots + q_{1,n_1}(t) \\ \le q^u_{1,1}(t) + q^u_{1,2}(t) + \cdots + q^u_{1,n_1}(t) + n_1 S, \quad (5.1)$$

i.e., $|\mathbf{q}_1(t)| \le |\mathbf{q}^u_1(t)| + n_1 S$. In general, for subqueue (k, j), where $1 \le k \le N$ and $1 \le j \le n_k$, among customers who arrive to subqueue (k, j) at the same time epochs for both systems, at most s of them can be in the priority service queueing system, but not in the designated service queueing system. There are at most $(\sum_{i=1}^{k-1} n_i + j - 1)S$ upgraded customers who arrive to subqueue (k, j) at different time epochs for the two systems - at most S from other queues (those are customers who have received incomplete service in the lower priority queues). Therefore, we must have $q_{k,j}(t) \le q^u_{k,j}(t) + (\sum_{i=1}^{k-1} n_i + j)S$. Similarly, we have, for any $1 \le k \le N$, $1 \le j \le n_k$ and $t \ge 0$,

$$q_{1,1}(t) + \cdots + q_{k,j}(t) \\ \le_{st} q^u_{1,1}(t) + \cdots + q^u_{k,j}(t) + \left(\sum_{i=1}^{k-1} n_i + j\right) S, \quad (5.2)$$

and $|\mathbf{q}_k(t)| \le_{st} |\mathbf{q}^u_k(t)| + S \sum_{i=1}^{k} n_i$. This completes the proof of Lemma 5.2.

Lemmas 5.1 and 5.2 lead to the following results on the mean queue lengths.

COROLLARY 5.3: For the priority service system, an upper bounds on the mean queue length of queue k $(1 \le k \le N-1)$ is given by $\sum_{i=1}^{k} (\lambda_i \tau_k^{-1} + n_i S)$, and an upper bound on the average number of customers in all lower priority queue is given by $\sum_{k=1}^{N-1} (\sum_{i=1}^{k} \lambda_i \tau_k^{-1} + n_k S)$.

According to the definition of the designated service queueing system, the upper bounds on the mean queue length can serve as an approximation if (i) $\lambda_N/(S\mu_N)$ is close to 1 (Note: the system must be stable, though.) or (ii) the service rates of lower priority customers are small as compared to their transfer rates. For case (i), all servers have to serve customers of the highest priority most of the time. Then most of the lower priority customers are transferred to queue N for service. Consequently, the priority service queue behaves similar to the designated service queue. For case (ii), since the service rates are small, most of the lower priority customers are transferred to queue N for service. Again, the two queueing systems behave similarly.

6. CONCLUSION AND DISCUSSION

In this article, we have obtained conditions for system stability/instability for the multi-class queue with customer upgrades defined in Section 2. The results indicate that the service processes and the upgrade processes of lower priority customers have no impact on the stability/instability of the queueing system of interest, which implies that, in the design of such queueing systems, resource should be allocated to serving customers of the highest priority as much as possible for the consideration of system stability. However, such allocation may result in long waiting time for lower priority customers, and hence more customer upgrades. Therefore, how to allocate limited resource to, on one hand, ensure a stable queue, and, on the other hand, keep all queues short, is an interesting issue for future research. In addition, the system stability issue has not been solved for the boundary case where $S\mu_N = \lambda_1 + \lambda_2 + \cdots + \lambda_N$, which is difficult to solve by using the current methodology and is an interesting problem for future research.

We have also obtained bounds on the queue lengths. Further analysis of the stationary distribution of the queue lengths is an interesting issue. Since the state space of the CTMC $\{\mathbf{X}(t), t \geq 0\}$ is multiple-dimensional, it is not straightforward to find the steady state probability distribution of the queue lengths. Asymptotic methods such as the one studied in [22] might be useful. In addition, it is intuitive to see that the stability results might be obtained for similar queueing systems with a preemptive resume, non-preemptive repeat, or non-preemptive resume service discipline. The proofs might be similar but much more tedious.

APPENDIX

THEOREM A.1 (Theorem 1.18 in [5]): Given an irreducible Q-matrix $Q = (q_{i,j})$ with state space S. Suppose that there exist a compact function $f(j)$ and constants $K \geq 0$ and $\eta > 0$ such that

$$\sum_{j\in S} q_{i,j}(f(j) - f(i)) \leq K - \eta f(i), \quad i \in S.$$

Then the Markov chain is positive recurrent and hence has a unique stationary distribution.

THEOREM A.2 (Theorem 1 in [6]): Consider an irreducible regular CTMC with state space S and Q-matrix $Q = (q_{ij})$. Let $f : S \to (-\infty, \infty)$ be a function satisfying

$$\sup_{i\in S} \sum_{j\in S} q_{ij}(f(i) - f(j))^+ < \infty.$$

Then, the CTMC is not ergodic if anyone of the following holds:

a. The function $f(\cdot)$ is not constant and $\sum_{j\in S} q_{ij} f(j) \geq 0, \forall i \in S$.
b. $\sum_{j\in S} q_{ij} f(j) \geq 0, \forall i \in S$. and $\sum_{j\in S} q_{ij} f(j) > 0$ for some state i.
c. There exists a non-empty subset B of S satisfying $\sum_{j\in S} q_{ij} f(j) \geq 0, \forall i \in B$, $B \neq S$, and $f(i) > \sup_{j\notin B} f(j)$ for some $i \in B$.

PROOF OF THEOREM 4.2: In the proof, Lemma 4.1 and the mean-drift method ([8]) are utilized. Theorem A.1 and Theorem A.2 are applied.

PROOF OF (4.2.1) OF THEOREM 4.2: The proof consists of three steps: (1) Construction of a Lyapunov function (test function); (2) Calculation of the mean-drift; and (3) Application of Theorem 1.18 in [5].

First, we construct a Lyapunov function. Since $\sum_{k=1}^{N} \lambda_k - S\mu_N < 0$, by (4.1.1) of Lemma 4.1, there exist $z^* > 1$ such that $\rho(z) < 0$ for $1 < z < z^* \leq \hat{z}$. Choose constant $a_{N,0}$ such that $1 < a_{N,0} < z^*$. Then we must have $\rho(a_{N,0}) < 0$. Choose $\mathbf{v} = \mathbf{v}(a_{N,0})$, the right eigenvector corresponding to $\rho(a_{N,0})$ defined in the proof of Lemma 4.1. Since the matrix $A^*(a_{N,0})$ is an irreducible M-matrix, all elements of the column vector $\mathbf{v}$ are positive. Choose constants $\{a_{1,0}, a_{1,1}, \ldots, a_{N-1,0}, \ldots, a_{N-1.n_{N-1}}, a_{N,1}\}$ satisfying the following conditions:

1. $1 < a_{N,1} = a_{N,0} < a_{N-1,n_{N-1}} < \cdots < a_{N-1,1} < a_{N-1,0} < \cdots < a_{1,n_1} < \cdots < a_{1,1} < a_{1,0} < z^*$;
2. $$\frac{1}{a_{N,0}}\rho(a_{N,0})\mathbf{v} + \sum_{J\in\Phi}\sum_{\boldsymbol{\delta}\in\mathcal{A}(J)} \left(\mathbf{a}^{\boldsymbol{\delta}} - a_{N,0}^{|J|}\right)p(J,\boldsymbol{\delta})(D_J \otimes I)\mathbf{v} \leq -\varepsilon\mathbf{v} < 0, \tag{A1}$$

where ε is a positive constant, vector $\mathbf{a} = (a_{1,0}, a_{1,1}, \ldots, a_{N-1,0}, \ldots, a_{N-1.n_{N-1}}, a_{N,0}, a_{N,1})$, and

$$\mathbf{a}^{\boldsymbol{\delta}} = \prod_{k=1}^{N}\left(\prod_{i=0}^{n_k} a_{k,i}^{\delta_{k,i}}\right). \tag{A2}$$

The existence of $\{a_{1,0}, a_{1,1}, \ldots, a_{N-1,0}, \ldots, a_{N-1.n_{N-1}}\}$ for Eq. (A1) is guaranteed by the facts that all elements of $\mathbf{v}$ are positive and $\rho(a_{N,0}) < 0$. Also note $\mathbf{a}^{\boldsymbol{\delta}} \geq a_{N,0}^{|J|}$ for $\boldsymbol{\delta} \in \mathcal{A}(J)$. Now, we are ready to introduce the following Lyapunov function. For $\mathbf{q} = (\mathbf{q}_1, \ldots, \mathbf{q}_{N-1}, \mathbf{q}_N)$, define

$$\mathbf{f}(\mathbf{q}) = \begin{cases} \prod_{k=1}^{N}\left(\prod_{i=0}^{n_k} a_{k,i}^{q_{k,i}}\right)\mathbf{v} = \mathbf{a}^{\mathbf{q}}\mathbf{v}, & \text{if } q_{N,0} = S; \\ \prod_{k=1}^{N}\left(\prod_{i=0}^{n_k} a_{k,i}^{q_{k,i}}\right)\mathbf{e} = \mathbf{a}^{\mathbf{q}}\mathbf{e}, & \text{if } q_{N,0} < S. \end{cases} \tag{A3}$$

Note that the size of $\mathbf{e}$ equals the number of state in the set $\Psi(\mathbf{q})$. It is easy to show that, for any $d > 0$, the set $\{\mathbf{q} : \mathbf{f}(\mathbf{q}) < d, \mathbf{q} \geq 0\}$ is bounded and has a finite number of elements. Thus, the function $\mathbf{f}(\mathbf{q})$ is a compact function (see [5]).

Next, we calculate the mean-drifts at states in the set $\Psi(\mathbf{q})$ for a given $\mathbf{q}$, which is defined in vector form as $\sum_{\mathbf{y}\geq 0} Q(\mathbf{q},\mathbf{y})\mathbf{f}(\mathbf{y})$ for a given $\mathbf{q}$. To calculate the mean drifts, we need to consider the following three cases: (i) $q_{N,0} = S$ and $q_{N,1} \geq 1$; (ii) $q_{N,0} = S$ and $q_{N,1} = 0$; and (iii) $q_{N,0} < S$ and $\sum_{j=1}^{N} q_{j,0} = S$. In the following calculations, we consider the five types of transitions given in Eqs. (3.4)–(3.9). For convenience, we omit the orders of the identity matrices.

Case (i): If $q_{N,0} = S$ and $q_{N,1} \geq 1$, it is easy to see that (1) All arriving customers have to join the queues; (2) If a server completes a service, the server begins to serve another type N customer immediately. We have the following calculations

$$\begin{aligned}
&\sum_{\mathbf{y}\geq 0} Q(\mathbf{q},\mathbf{y})\mathbf{f}(\mathbf{y}) \\
&= \sum_{J\in\Phi}\sum_{\boldsymbol{\delta}\in\mathcal{A}(J)} p(J,\boldsymbol{\delta})(D_J \otimes I)\mathbf{a}^{\mathbf{q}+\boldsymbol{\delta}}\mathbf{v} + \sum_{i=1}^{S}\left(I \otimes \mathbf{T}_N^0\boldsymbol{\alpha}_N \otimes I\right)\frac{1}{a_{N,0}}\mathbf{a}^{\mathbf{q}}\mathbf{v} \\
&\quad + \sum_{k=1}^{N-1}\sum_{i=1}^{n_k-1} s_{k,i}q_{k,i}\frac{a_{k,i+1}}{a_{k,i}}\mathbf{a}^{\mathbf{q}}\mathbf{v} + \sum_{k=1}^{N-1} q_{k,n_k}s_{k,n_k}\sum_{i=1}^{n_{k+1}}\beta_{k+1,i}\frac{a_{k+1,i}}{a_{k,n_k}}\mathbf{a}^{\mathbf{q}}\mathbf{v} \\
&\quad + \left(D_0 \otimes I + \sum_{i=1}^{S}(I \otimes T_N \otimes I) - \left(\sum_{k=1}^{N-1}\sum_{i=1}^{n_k} s_{k,i}q_{k,i}\right)I\right)\mathbf{a}^{\mathbf{q}}\mathbf{v} \\
&= \mathbf{a}^{\mathbf{q}}\left\{D_0 \otimes I + \sum_{J\in\Phi}\sum_{\boldsymbol{\delta}\in\mathcal{A}(J)} p(J,\boldsymbol{\delta})(D_J \otimes I)\mathbf{a}^{\boldsymbol{\delta}}\right. \\
&\quad + \sum_{i=1}^{S} I \otimes\left(\frac{1}{a_{N,0}}\mathbf{T}_N^0\boldsymbol{\alpha}_N + T_N\right)\otimes I + \sum_{k=1}^{N-1}\sum_{i=1}^{n_k-1} s_{k,i}q_{k,i}\left(\frac{a_{k,i+1}}{a_{k,i}} - 1\right)I \\
&\quad \left. + \sum_{k=1}^{N-1} q_{k,n_k}s_{k,n_k}\sum_{i=1}^{n_{k+1}}\beta_{k+1,i}\left(\frac{a_{k+1,i}}{a_{k,n_k}} - 1\right)I\right\}\mathbf{v} \\
&\leq \mathbf{a}^{\mathbf{q}}\left\{D_0 \otimes I + \sum_{J\in\Phi}\sum_{\boldsymbol{\delta}\in\mathcal{A}(J)} p(J,\boldsymbol{\delta})(D_J \otimes I)\mathbf{a}^{\boldsymbol{\delta}}\right. \\
&\qquad \left. + \sum_{i=1}^{S} I \otimes\left(\frac{1}{a_{N,0}}\mathbf{T}_N^0\boldsymbol{\alpha}_N + T_N\right)\otimes I\right\}\mathbf{v}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{a}^{\mathbf{q}} \left\{ \frac{1}{a_{N,0}} A^*(a_{N,0}) + \sum_{J\in\Phi} \sum_{\boldsymbol{\delta}\in\mathcal{A}(J)} p(J,\boldsymbol{\delta})(D_J \otimes I)(\mathbf{a}^{\boldsymbol{\delta}} - a_{N,0}^{|J|}) \right\} \mathbf{v} \\
&= \mathbf{a}^{\mathbf{q}} \left\{ \frac{1}{a_{N,0}} \rho(a_{N,0})\mathbf{v} + \sum_{J\in\Phi} \sum_{\boldsymbol{\delta}\in\mathcal{A}(J)} p(J,\boldsymbol{\delta})(D_J \otimes I)\big(\mathbf{a}^{\boldsymbol{\delta}} - a_{N,0}^{|J|}\big)\mathbf{v} \right\} \\
&\le -\varepsilon \mathbf{a}^{\mathbf{q}}\, \mathbf{v}. \qquad \text{(A4)}
\end{aligned}
$$

The first inequality in Eq. (A4) is due to $a_{k,i+1} < a_{k,i}$ and $a_{k+1,i} < a_{k,n_k}$. The second inequality in Eq. (A4) follows from part 2) of Eq. (A1). Also note $a_{N,0} = a_{N,1}$.

Case (ii): If $q_{N,0} = S$ and $q_{N,1} = 0$, suppose the non-empty queue with the highest priority is queue n ($<N$) with $q_{n,j} > 0$ and $q_{n,i} = 0$ for $0 < j < i \le n_n$. Then we have

$$
\begin{aligned}
&\sum_{\mathbf{y}\ge 0} Q(\mathbf{q},\mathbf{y})\mathbf{f}(\mathbf{y}) \\
&= \sum_{J\in\Phi} \sum_{\boldsymbol{\delta}\in\mathcal{A}(J)} p(J,\boldsymbol{\delta})(D_J \otimes I)\mathbf{a}^{\mathbf{q}+\boldsymbol{\delta}}\mathbf{v} + \sum_{i=1}^{S} \big(I \otimes \big(\mathbf{T}_N^0 \boldsymbol{\alpha}_n\big) \otimes I\big) \frac{a_{n,0}}{a_{n,j} a_{N,0}} \mathbf{a}^{\mathbf{q}}\mathbf{e} \\
&\quad + \sum_{k=1}^{N-1} \sum_{i=1}^{n_k-1} s_{k,i} q_{k,i} \frac{a_{k,i+1}}{a_{k,i}} \mathbf{a}^{\mathbf{q}}\mathbf{v} + \sum_{k=1}^{N-1} q_{k,n_k} s_{k,n_k} \sum_{i=1}^{n_{k+1}} \beta_{k+1,i} \frac{a_{k+1,i}}{a_{k,n_k}} \mathbf{a}^{\mathbf{q}}\mathbf{v} \\
&\quad + \left(D_0 \otimes I + \sum_{i=1}^{S} (I \otimes T_N \otimes I) - \left(\sum_{k=1}^{N-1} \sum_{i=1}^{n_k} s_{k,i} q_{k,i} \right) I \right) \mathbf{a}^{\mathbf{q}}\mathbf{v} \\
&= \left\{ D_0 \otimes I + \sum_{J\in\Phi} \sum_{\boldsymbol{\delta}\in\mathcal{A}(J)} p(J,\boldsymbol{\delta})(D_J \otimes I)\mathbf{a}^{\boldsymbol{\delta}} + \sum_{i=1}^{S} I \otimes T_N \otimes I \right\} \mathbf{a}^{\mathbf{q}}\mathbf{v} \\
&\quad + \left\{ \sum_{i=1}^{S} I \otimes \left(\frac{a_{n,0}}{a_{n,j} a_{N,0}} \mathbf{T}_N^0 \boldsymbol{\alpha}_n \right) \otimes I \right\} \mathbf{a}^{\mathbf{q}}\mathbf{e} \\
&\quad + \left\{ \sum_{k=1}^{N-1} \sum_{i=1}^{n_k-1} s_{k,i} q_{k,i} \left(\frac{a_{k,i+1}}{a_{k,i}} - 1 \right) \right. \\
&\quad \left. + \sum_{k=1}^{N-1} q_{k,n_k} s_{k,n_k} \sum_{i=1}^{n_{k+1}} \beta_{k+1,i} \left(\frac{a_{k+1,i}}{a_{k,n_k}} - 1 \right) \right\} \mathbf{a}^{\mathbf{q}}\mathbf{v} \\
&\le \mathbf{a}^{\mathbf{q}} \left\{ \phi_N \mathbf{e} - \left(\sum_{k=1}^{N-1} \sum_{i=1}^{n_k} q_{k,i} \right) \xi_N \mathbf{v} \right\}, \qquad \text{(A5)}
\end{aligned}
$$

where

$$
\begin{aligned}
\phi_N &= \left| \left(D_0 \otimes I + \sum_{J\in\Phi} \sum_{\boldsymbol{\delta}\in\mathcal{A}(J)} p(J,\boldsymbol{\delta})(D_J \otimes I)\mathbf{a}^{\boldsymbol{\delta}} + \sum_{i=1}^{S} I \otimes T_N \otimes I \right) \mathbf{v} \right|_{\max} \\
&\quad + \max_{\substack{1\le k\le N-1 \\ 1\le j\le n_k}} \left\{ \left| \left(\sum_{i=1}^{S} I \otimes \left(\frac{a_{k,0}}{a_{k,j} a_{N,0}} \mathbf{T}_N^0 \boldsymbol{\alpha}_k \right) \otimes I \right) \mathbf{e} \right|_{\max} \right\}; \qquad \text{(A6)}
\end{aligned}
$$

$$
\begin{aligned}
\xi_N = \min \Bigg\{ &\min_{\substack{1\le k\le N-1 \\ 1\le i\le n_k-1}} \left\{ s_{k,i} \left(1 - \frac{a_{k,i+1}}{a_{k,i}} \right) \right\}, \\
&\min_{1\le k\le N-1} \left\{ s_{k,n_k} \sum_{i=1}^{n_{k+1}} \beta_{k+1,i} \left(1 - \frac{a_{k+1,i}}{a_{k,n_k}} \right) \right\} \Bigg\} > 0. \qquad \text{(A7)}
\end{aligned}
$$

Note that, in Eqs. (A5) and (A6), the summation $\sum_{i=1}^{S} \bullet$ is over the S servers. For different i, the identity matrices in the summation have different size, which is not explicitly shown.

In Eq. (A6), the notation $|\mathbf{x}|_{\max}$ represents the absolute value of the element in the vector $\mathbf{x}$ that has the greatest absolute value. The finiteness of ϕ_N is guaranteed by our assumption on the finiteness of $D^*(z)$ for $1 < z \le z^* < \hat{z}$ and $a_{k,i} < z^*$ for all possible (k, i). Also note that $\mathbf{a}^{\boldsymbol{\delta}} \le a_{1,0}^{|J|}$ for $\boldsymbol{\delta} \in \mathcal{A}(J)$. Detailed calculations are given as follows:

$$
\begin{aligned}
&\left| \left(D_0 \otimes I + \sum_{J\in\Phi} \sum_{\boldsymbol{\delta}\in\mathcal{A}(J)} p(J,\boldsymbol{\delta})(D_J \otimes I)\mathbf{a}^{\boldsymbol{\delta}} + \sum_{i=1}^{S} I \otimes T_N \otimes I \right) \mathbf{v} \right|_{\max} \\
&\le |\mathbf{v}|_{\max} \left| \left(D_0 \otimes I + \sum_{J\in\Phi} \sum_{\boldsymbol{\delta}\in\mathcal{A}(J)} p(J,\boldsymbol{\delta})(D_J \otimes I)\mathbf{a}^{\boldsymbol{\delta}} \right) \mathbf{e} \right|_{\max} \\
&\quad + \left| \left(\sum_{i=1}^{S} I \otimes T_N \otimes I \right) \mathbf{v} \right|_{\max} \\
&\le |\mathbf{v}|_{\max} \left| \left(D_0 \otimes I + \sum_{J\in\Phi} (D_J \otimes I)|a_{1,0}|^{|J|} \right) \mathbf{e} \right|_{\max} \\
&\quad + \left| \left(\sum_{i=1}^{S} I \otimes T_N \otimes I \right) \mathbf{v} \right|_{\max} \\
&\le |\mathbf{v}|_{\max} \big| \big(D^*(a_{1,0}) \otimes I\big)\, \mathbf{e} \big|_{\max} + \left| \left(\sum_{i=1}^{S} I \otimes T_N \otimes I \right) \mathbf{v} \right|_{\max} < \infty, \qquad \text{(A8)}
\end{aligned}
$$

and, since $\boldsymbol{\alpha}_k \mathbf{e} = 1$,

$$
\begin{aligned}
&\max_{\substack{1\le n\le N-1 \\ 1\le j\le n_n}} \left\{ \left| \sum_{i=1}^{S} \left(I \otimes \left(\frac{a_{n,0}}{a_{n,j} a_{N,0}} \mathbf{T}_N^0 \boldsymbol{\alpha}_k \right) \otimes I \right) \mathbf{e} \right|_{\max} \right\} \\
&\qquad \le \frac{a_{1,0}}{a_{N,0}^2} \left| \sum_{i=1}^{S} \mathbf{e} \otimes \mathbf{T}_N^0 \otimes \mathbf{e} \right|_{\max} < \infty. \qquad \text{(A9)}
\end{aligned}
$$

Positivity of ξ_N is guaranteed by $a_{k,i+1} < a_{k,i}$ and $a_{k+1,i} < a_{k,n_k}$. If $q_{N,0} = S$ and all queues are empty, then term $I \otimes (\mathbf{T}_N^0 \boldsymbol{\alpha}_n) \otimes I$ in Eq. (A5) becomes $I \otimes \mathbf{T}_N^0 \otimes I$. Since $(I \otimes \mathbf{T}_N^0 \otimes I)\mathbf{e} = \mathbf{e} \otimes \mathbf{T}_N^0 \otimes \mathbf{e} = (I \otimes (\mathbf{T}_N^0 \boldsymbol{\alpha}_n) \otimes I)\mathbf{e}$, Eq. (A9) is still valid. Consequently, Eq. (A5) is valid for this case.

Case (iii): If $q_{N,0} < S$ and $\sum_{k=1}^{N} q_{k,0} = S$, there exists j ($1 \le j \le N-1$) such that $\sum_{k=j+1}^{N} q_{k,0} < S$, $\sum_{k=j}^{N} q_{k,0} = S$. For this case, if the queue n is not empty ($n \le j$), which is the nonempty queue of the highest priority, and $q_{n,l} > 0$ and $q_{n,i} = 0$ for $0 < l < i \le n_n$, we have

$$
\begin{aligned}
&\sum_{\mathbf{y}\ge 0} Q(\mathbf{q},\mathbf{y})\mathbf{f}(\mathbf{y}) \\
&= \sum_{J\in\Phi} \sum_{\mathbf{y}\in\mathcal{B}(\mathbf{q},J)} Q(\mathbf{q},\mathbf{y})\mathbf{f}(\mathbf{y}) + \sum_{i=1}^{S} \big(I \otimes \big(\mathbf{T}_{k_i}^0 \boldsymbol{\alpha}_n\big) \otimes I\big) \frac{a_{n,0}}{a_{k_i,0} a_{n,l}} \mathbf{a}^{\mathbf{q}}\mathbf{e} \\
&\quad + \sum_{k=1}^{j} \sum_{i=1}^{n_k-1} s_{k,i} q_{k,i} \frac{a_{k,i+1}}{a_{k,i}} \mathbf{a}^{\mathbf{q}}\mathbf{e} + \sum_{k=1}^{j-1} q_{k,n_k} s_{k,n_k} \sum_{i=1}^{n_{k+1}} \beta_{k+1,i} \frac{a_{k+1,i}}{a_{k,n_k}} \mathbf{a}^{\mathbf{q}}\mathbf{e} \\
&\quad + \sum_{i=1}^{n_j} \beta_{j,i} s_{j,n_j} q_{j,n_j} (I \otimes (\mathbf{e}\boldsymbol{\alpha}_{j+1}) \otimes I) \frac{a_{j,i} a_{j+1,0}}{a_{j,n_j} a_{j,0}} \mathbf{a}^{\mathbf{q}}\mathbf{e} \\
&\quad + \left(D_0 \otimes I + \sum_{i=1}^{S} (I \otimes T_{k_i} \otimes I) - \left(\sum_{k=1}^{N-1} \sum_{i=1}^{n_k} s_{k,i} q_{k,i} \right) I \right) \mathbf{a}^{\mathbf{q}}\mathbf{e} \\
&\le \left\{ \phi_j \mathbf{e} - \left(\sum_{k=1}^{N-1} \sum_{i=1}^{n_k} q_{k,i} \right) \xi_j \mathbf{e} \right\} \mathbf{a}^{\mathbf{q}}\mathbf{e}, \qquad \text{(A10)}
\end{aligned}
$$

where k_i is the type of customer being served by server i, and

$$\phi_j = \max_{\mathbf{q}\geq 0}\left\{\left|(D_0 \otimes I)\mathbf{e} + \sum_{J\in\Phi}\sum_{\mathbf{y}\in\mathcal{B}(\mathbf{q},J)} Q(\mathbf{q},\mathbf{y})\frac{\mathbf{f}(\mathbf{y})}{\mathbf{a}^{\mathbf{q}}}\right|_{\max}\right\} + \max_{\mathbf{q}\geq 0}\left\{\left|\sum_{i=1}^{S}\frac{a_{n,0}}{a_{k_i,0}a_{n,l}}\left(I\otimes\left(T_{k_i}+\mathbf{T}^0_{k_i}\boldsymbol{\alpha}_n\right)\otimes I\right)\mathbf{e}\right|_{\max}\right\}; \quad \text{(A11)}$$

$$\xi_j = \min\left\{\min_{\substack{1\leq k\leq N-1\\ 1\leq i\leq n_k-1}}\left\{s_{k,i}\left(1-\frac{a_{k,i+1}}{a_{k,i}}\right)\right\}, \min_{1\leq k\leq j-1}\left\{s_{k,n_k}\sum_{i=1}^{n_{k+1}}\beta_{k+1,i}\left(1-\frac{a_{k+1,i}}{a_{k,n_k}}\right)\right\}, s_{j,n_j}\sum_{i=1}^{n_{j+1}}\beta_{j,i}\left(1-\frac{a_{j,i}a_{j+1,0}}{a_{j,n_j}a_{j,0}}\right)\right\} > 0. \quad \text{(A12)}$$

Note that in Eq. (A11), ϕ_j is a function of j since n depends on j. Again, the finiteness of ϕ_j is guaranteed by our assumption on the finiteness of $D^*(z)$ for $1 < z < z^* < \hat{z}$ and $1 < a_{k,i} < z^*$ for all possible (k,i). Details are given as follows:

$$\begin{aligned}&\max_{\mathbf{q}\geq 0}\left\{\left|(D_0 \otimes I)\mathbf{e} + \sum_{J\in\Phi}\sum_{\mathbf{y}\in\mathcal{B}(\mathbf{q},J)} Q(\mathbf{q},\mathbf{y})\frac{\mathbf{f}(\mathbf{y})}{\mathbf{a}^{\mathbf{q}}}\right|_{\max}\right\}\\ &\leq \max_{\mathbf{q}\geq 0}\left\{\max\{1,|\mathbf{v}|_{\max}\}\left|(D_0\otimes I)\mathbf{e} + \sum_{J\in\Phi}(a_{1,0})^{|J|}\sum_{\mathbf{y}\in\mathcal{B}(\mathbf{q},J)}Q(\mathbf{q},\mathbf{y})\mathbf{e}\right|_{\max}\right\}\\ &\leq \max_{\mathbf{q}\geq 0}\{\max\{1,|\mathbf{v}|_{\max}\}|(D^*(a_{1,0})\otimes I)\mathbf{e}|_{\max}\} < \infty, \end{aligned} \quad \text{(A13)}$$

and

$$\begin{aligned}&\max_{\mathbf{q}\geq 0}\left\{\left|\sum_{i=1}^{S}\frac{a_{n,0}}{a_{k_i,0}a_{n,l}}\left(I\otimes\left(T_{k_i}+\mathbf{T}^0_{k_i}\boldsymbol{\alpha}_n\right)\otimes I\right)\mathbf{e}\right|_{\max}\right\}\\ &\leq \frac{a_{1,0}}{a^2_{N,0}}\max_{\mathbf{q}\geq 0}\left\{\left|\sum_{i=1}^{S}\left(I\otimes\left(T_{k_i}+\mathbf{T}^0_{k_i}\boldsymbol{\alpha}_n\right)\otimes I\right)\mathbf{e}\right|_{\max}\right\} < \infty. \end{aligned} \quad \text{(A14)}$$

Positivity of ξ_j is guaranteed by $a_{k,i+1} < a_{k,i}$, $a_{k+1,i} < a_{k,n_k}$, $a_{j,i} < a_{j,0}$, and $a_{j+1,0} < a_{j,n_j}$. If $q_{N,0} < S$ and all queues are empty, for the same reason given after Eq. (A9), Eq. (A10) is still valid. Define

$$\begin{aligned}\phi^* &= \max_{1\leq j\leq N}\{\phi_j\} < \infty;\\ \xi^* &= \min_{1\leq k\leq N}\{\xi_k\}|\mathbf{v}|_{\min} > 0,\end{aligned} \quad \text{(A15)}$$

where $|\mathbf{v}|_{\min}$ is the absolution value of the element in the vector $\mathbf{v}$ that has the smallest absolute value. Then we have, if $q_{N,1} = 0$ and $\sum_{k=1}^{N} q_{k,0} = S$,

$$\sum_{\mathbf{y}\geq 0} Q(\mathbf{q},\mathbf{y})\mathbf{f}(\mathbf{y}) \leq \left\{\phi^* - \left(\sum_{k=1}^{N-1}\sum_{i=1}^{n_k} q_{k,i}\right)\xi^*\right\}\mathbf{a}^{\mathbf{q}}\mathbf{e}. \quad \text{(A16)}$$

Note that the size of the vector $\mathbf{e}$ depends on the given vector $\mathbf{q}$.

Last, we apply Theorem 1.18 in [5] to show that the Markov chain is ergodic. According to Theorem 1.18 in [5] (see Theorem A.1), we need to show that

$$\sum_{\mathbf{y}\geq 0} Q(\mathbf{q},\mathbf{y})\mathbf{f}(\mathbf{y}) + \eta\mathbf{f}(\mathbf{q}) \leq K\mathbf{e} \quad \text{(A17)}$$

holds for all possible $\mathbf{q}$ for some constants $K \geq 0$ and $\eta > 0$. For that purpose, we choose $0 < \eta < \varepsilon$ [ε is defined in Eq. (A1)]. For the chosen η, by the last expression in Eq. (A16), there exists a finite number $q^* > 0$ such that, for any $\mathbf{q}$ satisfying $q_{N,1} = 0$ and $\sum_{k=1}^{N-1}\sum_{i=1}^{n_k} q_{k,i} > q^*$, we have

$$\begin{aligned}&\sum_{\mathbf{y}\geq 0} Q(\mathbf{q},\mathbf{y})\mathbf{f}(\mathbf{y}) + \eta\mathbf{f}(\mathbf{q})\\ &\leq \left\{\phi^* + \eta\max\{1,|\mathbf{v}|_{\max}\} - \left(\sum_{k=1}^{N-1}\sum_{i=1}^{n_k} q_{k,i}\right)\xi^*\right\}\mathbf{a}^{\mathbf{q}}\mathbf{e} \leq 0. \end{aligned} \quad \text{(A18)}$$

Define

$$K = \max\left\{0, \max_{\{\mathbf{q}:\ \mathbf{q}\geq 0,\ q_{N,1}=0,\ \sum_{k=1}^{N}\sum_{i=0}^{n_k} q_{k,i}\leq q^*+S\}} \times\left\{\left|\sum_{\mathbf{y}\geq 0} Q(\mathbf{q},\mathbf{y})\mathbf{f}(\mathbf{y}) + \eta\mathbf{f}(\mathbf{q})\right|_{\max}\right\}\right\}. \quad \text{(A19)}$$

Since the set $\{\mathbf{q}:\ \mathbf{q}\geq 0,\ q_{N,1} = 0,\ \sum_{k=1}^{N}\sum_{i=0}^{n_k} q_{k,i} \leq q^* + S\}$ has a finite number of elements, it is easy to see that K is nonnegative and finite, i.e., $0 \leq K < \infty$. Note that $\sum_{k=0}^{N} q_{k,0} \leq S$.

For the chosen η and K, by their definitions, Eq. (A17) holds if $q_{N,1} = 0$ and $\sum_{k=1}^{N}\sum_{i=0}^{n_k} q_{k,i} \leq q^* + S$. By Eq. (A18), Eq. (A17) holds if $q_{N,1} = 0$ and $\sum_{j=1}^{N}\sum_{i=0}^{n_j} q_{j,i} > q^* + S$. If $q_{N,1} \geq 1$, Eq. (A4) leads to

$$\sum_{\mathbf{y}\geq 0} Q(\mathbf{q},\mathbf{y})\mathbf{f}(\mathbf{y}) + \eta\mathbf{f}(\mathbf{q}) \leq (-\varepsilon+\eta)\mathbf{a}^{\mathbf{q}}\mathbf{v} \leq 0. \quad \text{(A20)}$$

Consequently, we have shown that Eq. (A17) holds for all possible $\mathbf{q}$. Therefore, by Theorem 1.18 in [5], the CTMC $\{\mathbf{X}(t), t \geq 0\}$ is ergodic. This completes the proof of (4.2.1) of Theorem 4.2.

PROOF OF (4.2.2) OF THEOREM 4.2: To prove (4.2.2), instead of Theorem 1.18 in [5], part c) of Theorem 1 in [6] is utilized (see Theorem A.2). The steps of the proof are similar to that of (4.2.1).

By (4.1.2) of Lemma 4.1, we choose positive constant $a_{N,0}$ such that $0 < z^* < a_{N,0} < 1$ and $\rho(a_{N,0}) < 0$. Again, we choose $\mathbf{v} = \mathbf{v}(a_{N,0})$, the right eigenvector corresponding to $\rho(a_{N,0})$ defined in the proof of Lemma 4.1. All elements of v are positive. Choose constants $\{a_{1,0}, a_{1,1}, \ldots, a_{N-1,0}, \ldots, a_{N-1.n_{N-1}}, a_{N,1}\}$ satisfying the following conditions:

1) $z^* < a_{N,1} = a_{N,0} < a_{N-1,n_{N-1}} < \cdots < a_{N-1,1} < a_{N-1,0} < \cdots < a_{1,n_1} < \cdots < a_{1,1} < a_{1,0} < 1$;

2) $$\frac{1}{a_{N,0}}\rho(a_{N,0})\mathbf{v} + \sum_{J\in\Phi}\sum_{\boldsymbol{\delta}\in\mathcal{A}(J)}\left(\mathbf{a}^{\boldsymbol{\delta}} - a_{N,0}^{|J|}\right)p(J,\boldsymbol{\delta})(D_J\otimes I)\mathbf{v} \leq -\varepsilon\mathbf{v} < 0, \quad \text{(A21)}$$

where ε is a positive constant. The Lyapunov function for this case is defined as: for $\mathbf{q}$,

$$\mathbf{f}(\mathbf{q}) = \begin{cases} -\prod_{k=1}^{N}\left(\prod_{i=0}^{n_k} a_{k,i}^{q_{k,i}}\right)\mathbf{v} = -\mathbf{a}^{\mathbf{q}}\mathbf{v}, & \text{if } q_{N,0} = S;\\ -\prod_{k=1}^{N}\left(\prod_{i=0}^{n_k} a_{k,i}^{q_{k,i}}\right)\mathbf{e} = -\mathbf{a}^{\mathbf{q}}\mathbf{e}, & \text{if } q_{N,0} < S. \end{cases} \quad \text{(A22)}$$

For our problem, part (c) of Theorem 1 in [6] can be restated as follows: the CTMC $\{\mathbf{X}(t), t \geq 0\}$ is non-ergodic if

$$\begin{aligned} &1) \quad \sup_{\mathbf{q}\geq 0}\left\{\max\left\{\sum_{\mathbf{y}\geq 0} Q(\mathbf{q},\mathbf{y})(\mathbf{f}(\mathbf{q})-\mathbf{f}(\mathbf{y}))^{+}\right\}\right\} < \infty;\\ &2) \quad \sum_{\mathbf{y}\geq 0} Q(\mathbf{q},\mathbf{y})\mathbf{f}(\mathbf{y}) \geq 0, \quad \text{if } \mathbf{q} \in B;\\ &3) \quad \exists \mathbf{q} \in B, \max\{\mathbf{f}(\mathbf{q})\} > \sup_{\mathbf{y}\notin B,\mathbf{y}\geq 0}\{\max\{\mathbf{f}(\mathbf{y})\}\}, \end{aligned} \tag{A23}$$

where $(\mathbf{x})^{+} = (\max\{0, x_i\})$, B is a subset of the set $\{\mathbf{q}\text{: } \mathbf{q} \geq 0\}$, and $\max\{\mathbf{w}\}$ is the value of the largest element in the vector $\mathbf{w}$.

First note that, for any constant $0 < x < 1$, we have $nx^n \to 0$ if $n \to \infty$. Condition (1) in Eq. (A23) holds since $0 < a_{k,i} < 1$ for all possible (k,i) and $0 \leq (\mathbf{f}(\mathbf{q}) - \mathbf{f}(\mathbf{y}))^{+} \leq \max\{1, |\mathbf{v}|_{\max}\}\mathbf{e}$ for $\mathbf{q} \geq 0$ and $\mathbf{y} \geq 0$. To verify conditions (2) and (3) in Eq. (A23), we consider two cases: $q_{N,1} \geq 1$ and $q_{N,1} = 0$. If $q_{N,1} \geq 1$, similar to Eq. (A4), we obtain

$$\sum_{\mathbf{y}\geq 0} Q(\mathbf{q},\mathbf{y})\mathbf{f}(\mathbf{y}) \geq \varepsilon \mathbf{a}^{\mathbf{q}}\mathbf{v} \geq 0. \tag{A24}$$

If $q_{N,1} = 0$ and $\sum_{k=1}^{N} q_{k,0} = S$, similar to the proof of equation (A16), we obtain

$$\sum_{\mathbf{y}\geq 0} Q(\mathbf{q},\mathbf{y})\mathbf{f}(\mathbf{y}) \geq -\left\{\phi^{*} - \left(\sum_{k=1}^{N-1}\sum_{i=1}^{n_k} q_{k,i}\right)\xi^{*}\right\}\mathbf{a}^{\mathbf{q}}\mathbf{e}, \tag{A25}$$

for some $0 \leq \phi^{*} < \infty$ and $\xi^{*} > 0$. Equation (A25) implies that there exists $q^{*} > 0$ such that $\sum_{\mathbf{y}\geq 0} Q(\mathbf{q},\mathbf{y})\mathbf{f}(\mathbf{y}) \geq 0$ if $\sum_{j=1}^{N}\sum_{i=0}^{n_j} q_{j,i} > q^{*} + S$. For this case, the finiteness of ϕ^{*} is guaranteed by $a_{k,i} < 1$ for all possible (k,i). Define

$$\begin{aligned} B = &\{\mathbf{q} : q_{N,1} \geq 1, \mathbf{q} \geq 0\}\\ &\cup \left\{\mathbf{q} : q_{N,1} = 0 \text{ and } \sum_{j=1}^{N}\sum_{i=0}^{n_j} q_{j,i} > q^{*} + S, \mathbf{q} \geq 0\right\}. \end{aligned} \tag{A26}$$

Then condition (2) in Eq. (A23) holds for this set B. It is easy to see $\sup_{\mathbf{y}\in B}\{\max\{\mathbf{f}(\mathbf{y})\}\} = 0$. Since there are only a finite number of possible $\mathbf{y} \geq 0$ not in the set B, it is easy to see $\sup_{\mathbf{y}\notin B,\mathbf{y}\geq 0}\{\max\{\mathbf{f}(\mathbf{y})\}\} < 0$. Then Condition (3) in Eq. (A23) holds for the set B. Therefore, the CTMC $\{X(t), t \geq 0\}$ is non-ergodic if $\lambda_1 + \cdots + \lambda_N > S\mu_N$. This completes the proof of (4.2.2) of Theorem 4.2.

ACKNOWLEDGMENTS

The authors thank the editor-in-chief, the associate editor, and two referees for their valuable comments and suggestions. The research was partially supported by NSF of China under grant 71121061 and 71031005.

REFERENCES

[1] I.J.B.F. Adan, J. Wessels, and W.H.M. Zijm, Analysis of the asymmetric shortest queue problem with threshold jockeying, Stochastic Models 7 (1991), 615–628.

[2] M. Akan, O. Alagoz, B. Ata, F.S. Erenay, and A. Said, A broader view of designing the liver allocation system, Oper Res, (in press).

[3] S. Asmussen and G. Koole, Marked point processes as limits of Markovian arrival streams, J Appl Probab 30 (1993), 365–372.

[4] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, Statistical analysis of a telephone call center: A queueing-science perspective, J Am Stat Assoc 100 (2005), 36–50.

[5] M.F. Chen, On three classical problems for Markov chains with continuous time parameters, J Appl Probab 28 (1991), 305–320.

[6] B.D. Choi and B. Kim, Non-ergodicity criteria for denumerable continuous time Markov processes, Oper Res Lett 32 (2004), 574–580.

[7] K.L. Chung, Markov chains with stationary transition probabilities, 2nd ed., Springer-Verlag, Berlin, 1967.

[8] J.W. Cohen, The single server queue, North-Holland, Amsterdam, 1982.

[9] B. Deniz, I. Karaesmen, and A.A. Sheller-Wolf, Managing perishables with substitution: Inventory issuance and replenishment heuristics, Manufacturing Service Oper Manage 12 (2010), 319–329.

[10] J. Diamond and A.S. Alfa, Matrix analytic methods for $M/PH/1$ retrial queues, Stochastic Models 11 (1995), 447–470.

[11] D.G. Down and M.E. Lewis, The N-Network model with upgrades, Probab Eng Inf Sci 24 (2010), 171–200.

[12] G. Fayolle, V.A. Malyshev, and M.V. Menshikov, Topics in the constructive theory of countable Markov chains, Cambridge University Press, Cambridge, UK, 1995.

[13] F.G. Foster, On stochastic matrices associated with certain queueing processes, Ann Math Stat 24 (1953), 355–360.

[14] F.R. Gantmacher, Matrix theory, Volume Volume I, Chelsea Publishing Company, New York, 1960.

[15] L.V. Green, J. Soares, J.F. Giglio, and R.A. Green, Using queueing theory to increase the effectiveness of emergency department provider staffing, Acad Emergency Med 13 (2006), 61–68.

[16] B. Hajek, Optimal control of two interacting service stations, IEEE Trans Automat Control AC-29 (1984), 491–499.

[17] Q.M. He and M.F. Neuts, Markov chains with marked transitions, Stochastic Processes Appl 74 (1998), 37–52.

[18] Q.M. He, H. Li, and Q. Zhao, Ergodicity of the $BMAP/PH/s/s + K$ retrial queue with PH-retrial times, Queueing Syst 35 (2000), 323–347.

[19] Q.M. He, The versatility of $MMAP[K]$ and the $MMAP[K]/G[K]/1$ queue, Queueing Syst 38 (2001), 397–418.

[20] S.A. Lippman, Applying a new device in the optimization of exponential queuing systems, Oper Res 23 (1975), 687–710.

[21] S.P. Meyn and R. Tweedie, Markov chains and stochastic stability, Springer Verlag, London, 1993.

[22] M. Miyazawa and Y.Q. Zhao, The stationary tail asymptotics in the $GI/G/1$ type queue with countably many background states, Adv Appl Probab 36 (2004), 1231–1251.

[23] M.F. Neuts, A versatile Markovian point process, J Appl Probab 16 (1979), 764–779.

[24] M.F. Neuts, Matrix-geometric solutions in stochastic models: An algorithmic approach, The Johns Hopkins University Press, Baltimore, 1981.

[25] M.E.H. Ong, K.K. Ho, T.P. Tan, S.W. Koh, Z. Almuthar, J. Overton, and S.H. Lim, Using demand analysis and system status management for predicting ED attendances and rostering, Am J Emergency Med 27 (2009), 16–22.

[26] M. Standing, Clinical judgment and decision making in nursing—Nine modes of practice in a revised cognitive continuum, J Adv Nurs 62 (2008), 124–134.

[27] D. Stoyan and D.J. Daley, Comparison methods for queues and other stochastic models, Wiely, New York, 1983.

[28] H. Takagi, Queueing systems, Vacation and priority systems, Vol. 1, Elsevier, Amsterdam, 1991.

[29] W. Whitt, Deciding which queue to join: Some counterexamples, Oper Res 34 (1986), 55–62.

[30] J.G. Xie, Q.M. He, and X.B. Zhao, Stability of a priority queueing system with customer transfers, Oper Res Lett 36 (2008), 705–709.

[31] J. Xie, Q.M. He, and X. Zhao, On the stationary distribution of queue lengths in a multi-class priority queueing system with customer transfers, Queueing Syst 62 (2009), 255–277.

[32] S.H. Xu and H. Chen, On the asymptote of the optimal routing policy for two service stations, IEEE Trans Automatic Control 38 (1990), 187–189.

[33] S.H. Xu and Y.Q. Zhao, Dynamic routing and jockeying controls in a two-station queueing system, Adv Appl Probab 28 (1996), 1201–1226.

[34] Y. Zhao and W.K. Grassmann, Queueing analysis of a jockeying model, Oper Res 43 (1995), 520–529.