Queueing Syst (2009) 62: 255–277
DOI 10.1007/s11134-009-9130-0

On the stationary distribution of queue lengths in a multi-class priority queueing system with customer transfers

Jingui Xie · Qi-Ming He · Xiaobo Zhao

Received: 1 September 2008 / Revised: 1 July 2009 / Published online: 30 July 2009

Abstract This paper deals with a multi-class priority queueing system with customer transfers that occur only from lower priority queues to higher priority queues. Conditions for the queueing system to be stable/unstable are obtained. An auxiliary queueing system is introduced, for which an explicit product-form solution is found for the stationary distribution of queue lengths. Sample path relationships between the queue lengths in the original queueing system and the auxiliary queueing system are obtained, which lead to bounds on the stationary distribution of the queue lengths in the original queueing system. Using matrix-analytic methods, it is shown that the tail asymptotics of the stationary distribution is exact geometric, if the queue with the highest priority is overloaded.

Keywords Priority queueing system · Tail asymptotics · Matrix-analytic methods · Sample path relationship

Mathematics Subject Classification (2000) 60K25 · 90B22

1 Introduction

This paper studies a priority queueing system with multi-types of customers subject to transferring from lower priority queues to higher priority queues. The queueing

J. Xie · X. Zhao (✉)
Department of Industrial Engineering, Tsinghua University, Beijing 100084,
People's Republic of China
e-mail: xbzhao@tsinghua.edu.cn

J. Xie
e-mail: xiejingui@tsinghua.org.cn

Q.-M. He
Department of Industrial Engineering, Dalhousie University, Halifax, NS B3J 2X4, Canada

system of interest is an extension of the one studied in Xie et al. [32]. Xie et al. [32] find simple stability/instability conditions for the queueing system, which are also valid for the extended queueing system. In this paper, the stationary distribution of queue lengths is analyzed.

The study of the queueing system is motivated by potential applications in the design of emergency departments in healthcare systems, custom inspection systems, and systems alike. For such systems, customers (patients or products) can be categorized into groups with different service priorities. A common feature of the systems is that customers can transfer and only transfer from a lower priority group to higher priority groups. This feature is important in the modeling, design, and analysis of such queueing systems. In fact, our analysis, both in Xie et al. [32, 33] and in this paper, takes advantage of this useful feature. Studies on similar queueing models with applications in healthcare management and call center management can be found in Argon et al. [2], Brandt and Brandt [4], Gomez-Corral et al. [12], and Wang [30].

This paper is related to the study of priority queueing systems and the study of queueing systems with customer transfers. Priority queueing systems have been studied extensively ([8, 28], and references therein). Queueing systems with customer transfers have also been studied extensively ([1, 10, 13, 31, 34–36], and references therein). Both types of queueing systems are classical queueing systems with applications in manufacturing, service, and telecommunications industries. The queueing system studied in this paper is a hybrid of those two types of queueing systems. To our knowledge, this type of queueing systems has not been analyzed, except for He et al. [14] and Xie et al. [32], where system stability/instability conditions are obtained.

While the system stability/instability conditions given in Xie et al. [32, 33] are simple and explicit, there is no explicit solution to the stationary distribution of the queue lengths for the queueing system. Thus, this paper focuses on bounds and tail asymptotics of the stationary distribution of queue lengths. In the first part of the paper, sample path upper bounds on the queue lengths are found through the introduction of an auxiliary queueing system. The auxiliary queueing system possesses an explicit product-form solution to its stationary distribution of queue lengths, which is a type of solution existing for many queueing networks [5, 15]. The explicit product-form solution of the auxiliary queueing systems provides upper bounds on the stationary distribution of queue lengths. In the second part of the paper, matrix-analytic methods are applied for finding the tail asymptotics of the stationary distribution of the queue lengths, if the queue with the highest priority is overloaded (i.e., the queue length is large). We refer to Kroses et al. [16], Latouche and Ramaswami [17], Miyazawa and Zhao [22], Neuts [24], and Takahashi et al. [29] for tail asymptotics and matrix-analytic methods. By using conditions given in Li et al. [18], it is shown that the tail asymptotics is exact geometric, the decay rate is found explicitly, and the limit can be found explicitly up to a constant multiplier. Those results are useful in understanding the behavior of the queueing system if one queue is overloaded. In the last part of the paper, all the results obtained in the first and second parts of the paper are generalized to a model in which the arrivals of customers from outside to the queue with the highest priority follow a Markov arrival process. Results indicate that the estimation of the decay rate for tail asymptotics can be improved by utilizing Markov arrival processes (*MAP*) in queueing analysis.

In addition to the main contributions of this paper on the analysis of the stationary distribution of queue lengths, there are two by-products we would like to mention. First, the explicit product-form solution obtained for the auxiliary queueing system is of its own importance in the analysis of the system. We also find necessary and sufficient conditions for the system to be stable. Second, the existing sufficient conditions for tail asymptotics given in Li et al. [18] are applied to the queueing system after a transformation of a Markov chain constructed for the queueing system. This demonstrates that the conditions for tail asymptotics in the existing literature can be applied to a much larger class of Markov chains.

The paper is organized as follows. The queueing system of interest is introduced in Sect. 2. In Sect. 3, an auxiliary queueing system is introduced and an explicit product-form solution is found for the stationary distribution of queue lengths. Relationships between the queue lengths of the original and the auxiliary queueing systems are established. It is shown that the stationary distribution of queue lengths of the auxiliary system provides upper bounds on the stationary distribution of queue lengths of the original queueing system. Section 4 shows that the tail asymptotics of the stationary distribution is exact geometric, if the queue with the highest priority is overloaded. In Sect. 5, all results are generalized to a queueing system with a Markovian arrival process for customers to the highest priority queue from outside. Section 6 concludes the paper.

2 The priority service queueing system

The queueing system introduced in this section is called a *priority service queueing system* and is an extension of the one defined in Xie et al. [32].

The queueing system consists of s identical servers serving N types of customers: type 1, type $2, \dots,$ and type N customers. Type $1, 2, \dots,$ and N customers form queue $1, 2, \dots,$ and N, respectively. Type N customers have the highest service priority, type $N-1$ the second highest service priority, $\dots$, and type 1 the lowest service priority. When a server is available, it chooses a customer from the non-empty queue of the highest priority and begins to serve it. If some servers are serving type j customers when a type k customer arrives from outside or is transferred from another queue, for $j < k$, there is no idle server, and type j customers are the lowest priority customers in service, then one of the type j customers in service is pushed back to queue j and the server begins to serve the type k customer. The type j customer will resume (or repeat) its service when a server is available to serve type j customers. Thus, higher priority customers preempt lower priority customers from service.

Type $1, 2, \dots,$ and N customers arrive (from outside of the queueing system) according to N independent Poisson processes with parameters $\lambda_1, \lambda_2, \dots,$ and λ_N, respectively. The service times of type $1, 2, \dots,$ and N customers (regardless of where they come from, outside or another queue) are exponentially distributed with parameters $\mu_1, \mu_2, \dots,$ and μ_N, respectively. The arrival processes and service times are independent. Since the service time of a type j customer is exponentially distributed, it does not make a difference to assume that its interrupted service will be repeated or resumed. For the same reason, if a server is available to serve type j customers, it does not matter (to system stability/instability or queue length) which waiting type j customer enters the server to receive service.

While waiting in queue j, a type j customer may change to a customer of higher priority after an exponential time with parameter $\lambda_{T,j}$, for $1 \le j \le N-1$. Upon transfer, a type j customer becomes a type k customer with probability $p_{j,k}$, for $j+1 \le k \le N$. Note that $p_{j,j+1} + p_{j,j+2} + \cdots + p_{j,N} = 1$ for $1 \le j \le N-1$. Since the time before transfer is exponentially distributed, it does not make a difference to assume that the clock until transfer is reset or continued, if a type j customer's service is interrupted. The times until transfer for individual customers are independent of each other, and are independent of the arrival and service processes. Note that a customer in service does not change its type.

Remark 2.1 The queueing system studied in Xie et al. [32] is a special case with $p_{j,j+1} = 1$ for $1 \le j \le N-1$.

Define $q_j(t)$ the number of type j customers in queue j at time t, which includes the type j customers in service (if there are type j customers in service), for $1 \le j \le N$. If all system parameters are positive, it is easy to see that $\{(q_1(t), q_2(t), \ldots, q_N(t)), t \ge 0\}$ is an irreducible continuous time Markov chain (CTMC) with state space $\{(q_1, q_2, \ldots, q_N), q_1 \ge 0, \ldots, q_N \ge 0\}$.

Denote by $Q = (Q_{(q_1,q_2,\ldots,q_N),(y_1,y_2,\ldots,y_N)})$ the infinitesimal generator of the Markov chain. In each state, there are possibly three types of transitions: (1) the arrival of a customer from outside; (2) the completion of a service; and (3) the transfer of a customer. It is easy to see that the arrival rates from outside are $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$. The service completion rates depend on how many customers are in service, which further depends on how many servers are available to serve a queue and how many customers in that queue. In state $(q_1, q_2, \ldots, q_N)$, the number of type j customers in service is given by $\max\{0, \min\{s - \sum_{k=j+1}^N q_k, q_j\}\}$. Thus, the total service rate in queue j is given by $\max\{0, \min\{s - \sum_{k=j+1}^N q_k, q_j\}\}\mu_j$. The transfer rates depend on how many customers are waiting in queues. In state $(q_1, q_2, \ldots, q_N)$, the number of customers waiting in queue j is given by $\min\{q_j, \max\{\sum_{i=j}^N q_i - s, 0\}\}$. Thus, the transfer rate from queue j to queue k is given by $\min\{q_j, \max\{\sum_{i=j}^N q_i - s, 0\}\}\lambda_{T,j} p_{j,k}$ for $k = j+1, \ldots, N$. In summary, we have, for $(q_1, q_2, \ldots, q_N) \ne (y_1, y_2, \ldots, y_N)$,

$$
\begin{aligned}
&Q_{(q_1,q_2,\ldots,q_N),(y_1,y_2,\ldots,y_N)} \\
&= \begin{cases}
\lambda_j, & \text{if } y_j = q_j + 1,\ y_i = q_i,\ i \ne j, \\
& 1 \le j \le N; \\
\max\{0, \min\{s - \sum_{k=j+1}^N q_k, q_j\}\}\mu_j, & \text{if } y_j = q_j - 1 \ge 0,\ y_i = q_i,\ i \ne j, \\
& 1 \le j \le N; \\
\min\{q_j, \max\{\sum_{i=j}^N q_i - s, 0\}\}\lambda_{T,j} p_{j,k}, & \text{if } y_j = q_j - 1 \ge 0,\ j < k, \\
& y_k = q_k + 1, y_i = q_i,\ i \ne j,k, \\
& j+1 \le k \le N,\ 1 \le j \le N-1; \\
0, & \text{otherwise,}
\end{cases}
\end{aligned}
\tag{2.1}
$$

and, for $(q_1, q_2, \ldots, q_N) = (y_1, y_2, \ldots, y_N)$,

$$
\begin{aligned}
&Q_{(q_1,q_2,\ldots,q_N),(q_1,q_2,\ldots,q_N)} \\
&\quad = -\left\{\sum_{j=1}^{N} \lambda_j + \sum_{j=1}^{N} \max\left\{0, \min\left\{s - \sum_{k=j+1}^{N} q_k, q_j\right\}\right\}\mu_j \right.\\
&\qquad \left. + \sum_{j=1}^{N-1} \min\left\{q_j, \max\left\{\sum_{k=j}^{N} q_k - s, 0\right\}\right\}\lambda_{T,j}\right\}. \qquad (2.2)
\end{aligned}
$$

The queueing system is stable if the Markov chain $\{(q_1(t), q_2(t), \ldots, q_N(t)), t \geq 0\}$ is ergodic (i.e., irreducible and positive recurrent). The queueing system is unstable if the Markov chain is non-ergodic. The ergodicity of the Markov chain $\{(q_1(t), q_2(t), \ldots, q_N(t)), t \geq 0\}$ is characterized in the following theorem.

Theorem 2.1 *Assume that all system parameters* $\{\lambda_1, \ldots, \lambda_N, \mu_1, \ldots, \mu_N, \lambda_{T,1}, \ldots, \lambda_{T,N-1}, s\}$ *are positive. Then the Markov chain* $\{(q_1(t), q_2(t), \ldots, q_N(t)), t \geq 0\}$ *is irreducible. The Markov chain is*: (1) *ergodic if* $\sum_{k=1}^{N} \lambda_k < s\mu_N$; *and* (2) *non-ergodic if* $\sum_{k=1}^{N} \lambda_k > s\mu_N$.

Proof The proof is the same as that of Theorem 1 in Xie et al. [32] except for a few minor changes. The proof is based on the mean-drift method [8, 9, 11, 20] and on the conditions for the ergodicity and non-ergodicity of continuous time Markov chains given in Chen [6] and Choi and Kim [7]. For details, see Xie et al. [32, 33]. We note that it is unnecessary to assume that all system parameters are positive. We make that assumption for convenience. □

Throughout this paper, we shall assume $\sum_{k=1}^{N} \lambda_k < s\mu_N$ so that the queueing system is stable and the corresponding Markov chain $\{(q_1(t), q_2(t), \ldots, q_N(t)), t \geq 0\}$ has a (unique) stationary distribution. Unfortunately, the stationary distribution cannot be obtained explicitly and is difficult to analyze directly. In the rest of the paper, we find bounds and tail asymptotics for the stationary distribution.

3 Designated service queueing system

In this section, an auxiliary system is introduced and analyzed. Some results (e.g., Corollary 3.4) are used in analyzing the stationary distribution of queue lengths of the priority service queueing system.

Consider a queueing system that is the same as the one defined in Sect. 2 except that servers only serve queue N. We call this queueing system a *designated service queueing system*. It is readily seen that, in the designated service queueing system, all customers are eventually transferred to type N customers before being served. For $j \leq N-1$, queue j looks like a $G/M/\infty$ queueing system with service rate $\lambda_{T,j}$.

Denote by $q_{u,1}(t), q_{u,2}(t), \ldots$, and $q_{u,N}(t)$ the queue lengths of type $1, 2, \ldots, N$ customers in the designated service queueing system at time t, respectively. It is easy to see that $\{(q_{u,1}(t), q_{u,2}(t), \ldots, q_{u,N}(t)), t \geq 0\}$ is a continuous time Markov chain with state space $\{(q_1, q_2, \ldots, q_N), q_1 \geq 0, \ldots, q_N \geq 0\}$. Denote by $\bar{Q} = (\bar{Q}_{(q_1,q_2,\ldots,q_N),(y_1,y_2,\ldots,y_N)})$ the infinitesimal generator of the Markov chain. We have, for $(q_1, q_2, \ldots, q_N) \neq (y_1, y_2, \ldots, y_N)$,

$$\bar{Q}_{(q_1,q_2,\ldots,q_N),(y_1,y_2,\ldots,y_N)} = \begin{cases} \lambda_j, & \text{if } y_j = q_j + 1,\ y_i = q_i,\ i \neq j,\ 1 \leq j \leq N; \\ \min\{s, q_N\}\mu_N, & \text{if } y_N = q_N - 1 \geq 0,\ y_j = q_j,\ 1 \leq j \leq N-1; \\ q_j \lambda_{T,j} p_{j,k}, & \text{if } y_j = q_j - 1 \geq 0,\ j < k,\ y_k = q_k + 1, \\ & y_i = q_i,\ i \neq j, k,\ j+1 \leq k \leq N,\ 1 \leq j \leq N-1; \\ 0, & \text{otherwise,} \end{cases} \tag{3.1}$$

and, for $(q_1, q_2, \ldots, q_N) = (y_1, y_2, \ldots, y_N)$,

$$\bar{Q}_{(q_1,q_2,\ldots,q_N),(q_1,q_2,\ldots,q_N)} = -\left\{ \sum_{j=1}^{N} \lambda_j + \min\{s, q_N\}\mu_N + \sum_{j=1}^{N-1} q_j \lambda_{T,j} \right\}. \tag{3.2}$$

First, we find the total customer arrival rate for each queue. Denote by $\bar{p}^*_{i,j}$ the probability that a type i customer will ever become a type j customer (before eventually becoming a type N customer in the designated service queueing system) for $i < j$. Denote by $\bar{\lambda}^*_j$ the total customer arrival rate to queue j in the designated service queueing system. It is clear that $\bar{\lambda}^*_j$ only includes customers arriving from outside to queues 1 to j. The following relationships can be proved easily.

Lemma 3.1 *For $\bar{p}^*_{i,j}$, we have $\bar{p}^*_{i,j} = \sum_{k=i+1}^{j} p_{i,k} \bar{p}^*_{k,j} = \sum_{k=i}^{j-1} \bar{p}^*_{i,k} p_{k,j}$ for $i < j$, and $\bar{p}^*_{j,j} = \bar{p}^*_{j,N} = 1$ for $1 \leq j \leq N$. For $\bar{\lambda}^*_j$, we have $\bar{\lambda}^*_j = \lambda_j + \sum_{k=1}^{j-1} \bar{\lambda}^*_k p_{k,j} = \sum_{k=1}^{j} \lambda_k \bar{p}^*_{k,j}$ for $1 \leq j \leq N$. In particular, we have $\bar{\lambda}^*_1 = \lambda_1$ and $\bar{\lambda}^*_N = \sum_{k=1}^{N} \lambda_k$.*

Denote by $\{\pi_u(q_1, q_2, \ldots, q_N), q_1 \geq 0, q_2 \geq 0, \ldots, q_N \geq 0\}$ the stationary distribution of $\{(q_{u,1}(t), q_{u,2}(t), \ldots, q_{u,N}(t)), t \geq 0\}$, if it exists, i.e.,

$$\pi_u(q_1, q_2, \ldots, q_N) = \lim_{t \to \infty} P\big\{q_{u,1}(t) = q_1, q_{u,2}(t) = q_2, \ldots, q_{u,N}(t) = q_N\big\}. \tag{3.3}$$

Theorem 3.2 *Assume that all system parameters $\{\lambda_1, \ldots, \lambda_N, \mu_N, \lambda_{T,1}, \ldots, \lambda_{T,N-1}, s\}$ are positive. The Markov chain $\{(q_{u,1}(t), q_{u,2}(t), \ldots, q_{u,N}(t)), t \geq 0\}$ is ergodic if and only if $\sum_{k=1}^{N} \lambda_k < s\mu_N$. If the stationary distribution exists, it is given by*

$$\pi_u(q_1, q_2, \ldots, q_N) = \begin{cases} \left(\prod_{j=1}^{N-1} \left(\frac{\exp\{-\bar{\rho}_{T,j}\}\bar{\rho}_{T,j}^{q_j}}{q_j!}\right)\right) \frac{\frac{(s\bar{\rho}_N)^{q_N}}{q_N!}}{\left(\sum_{j=0}^{s-1} \frac{(s\bar{\rho}_N)^j}{j!} + \frac{(s\bar{\rho}_N)^s}{s!(1-\bar{\rho}_N)}\right)}, & 0 \leq q_N < s; \\ \left(\prod_{j=1}^{N-1} \left(\frac{\exp\{-\bar{\rho}_{T,j}\}\bar{\rho}_{T,j}^{q_j}}{q_j!}\right)\right) \frac{\frac{(s\bar{\rho}_N)^s}{s!} \bar{\rho}_N^{q_N - s}}{\left(\sum_{j=0}^{s-1} \frac{(s\bar{\rho}_N)^j}{j!} + \frac{(s\bar{\rho}_N)^s}{s!(1-\bar{\rho}_N)}\right)}, & q_N \geq s, \end{cases} \tag{3.4}$$

where

$$\bar{\rho}_{T,j} = \left(\sum_{k=1}^{j} \lambda_k \bar{p}^*_{k,j}\right) \Bigg/ \lambda_{T,j} = \bar{\lambda}^*_j / \lambda_{T,j}, \quad 1 \le j \le N-1;$$
$$\bar{\rho}_N = \left(\sum_{j=1}^{N} \lambda_j\right) \Bigg/ (s\mu_N) = \bar{\lambda}^*_N / (s\mu_N). \tag{3.5}$$

Remark 3.1 We would like to point out that the designated service queueing system is different from the priority service queueing system defined in Sect. 2 with $\mu_1 = \cdots = \mu_{N-1} = 0$. The reason is that in the priority service queueing system with $\mu_1 = \cdots = \mu_{N-1} = 0$, a type j customer's transferring process may be interrupted if a server becomes available to serve the customer. Thus, Theorem 2.1 cannot be applied for the stability/instability of the designated service queueing system. Nonetheless, the stability conditions for both queueing systems are similar. The proof of the stability condition for the designated service queueing system is much simpler.

Proof Since all customers will transfer to type N customers, the total customer arrival rate to queue N is $\bar{\lambda}^*_N = \sum_{k=1}^{N} \lambda_k$. Thus, for system stability or ergodicity of the Markov chain, we must have $\sum_{k=1}^{N} \lambda_k < s\mu_N$, i.e., $\bar{\rho}_N < 1$. On the other hand, if $\bar{\rho}_N < 1$, we shall show that the joint probability distribution given in (3.4) is a stationary distribution of the Markov chain $\{(q_{u,1}(t), q_{u,2}(t), \ldots, q_{u,N}(t), t \ge 0\}$. For an irreducible Markov chain with a countable state space, that is equivalent to ergodicity of the Markov chain. In addition, it can be shown that $\boldsymbol{\pi}_u \mathrm{diag}(-\bar{Q}) < \infty$ holds for the stationary distribution $\boldsymbol{\pi}_u$ to be given next, where $\mathrm{diag}(-\bar{Q})$ is a matrix obtained by keeping all diagonal elements of $-\bar{Q}$ and setting all other elements to zero. Thus, the Markov chain is nonexplosive [3]. Therefore, $\bar{\rho}_N < 1$ is a necessary and sufficient condition for system stability.

If $\bar{\rho}_N < 1$, we verify that the joint probability distribution given in (3.4) satisfies the equation $\boldsymbol{\pi}_u \bar{Q} = 0$, where $\boldsymbol{\pi}_u = (\pi_u(q_1, \ldots, q_N))$ in which the probabilities $\{\pi_u(q_1, \ldots, q_N), q_1 \ge 0, \ldots, q_N \ge 0\}$ are arranged lexicographically. For convenience, let $\mathbf{q} = (q_1, \ldots, q_N)$ and $\mathbf{e}(j)$ be the row vector with all elements being zero except that the jth element is one. For all states with $q_N \ge s$, we need to check

$$\begin{aligned} 0 = &-\pi_u(\mathbf{q})\left(s\mu_N + \sum_{j=1}^{N} \lambda_j + \sum_{j=1}^{N-1} q_j \lambda_{T,j}\right) + \pi_u\big(\mathbf{q} + \mathbf{e}(N)\big)s\mu_N \\ &+ \sum_{j=1:q_j \ge 1}^{N} \pi_u\big(\mathbf{q} - \mathbf{e}(j)\big)\lambda_j \\ &+ \sum_{j=1}^{N-1} \sum_{k=j+1:q_k \ge 1}^{N} \pi_u\big(\mathbf{q} - \mathbf{e}(k) + \mathbf{e}(j)\big)(q_j + 1)\lambda_{T,j} p_{j,k}. \end{aligned} \tag{3.6}$$

Using expressions given in (3.4), the right-hand side of (3.6) becomes

$$\begin{aligned}
&\pi_u(\mathbf{q})\Bigg\{-\Bigg(s\mu_N+\sum_{j=1}^{N}\lambda_j+\sum_{j=1}^{N-1}q_j\lambda_{T,j}\Bigg)+\sum_{j=1}^{N}\lambda_j+\sum_{j=1:q_j\ge 1}^{N-1}\lambda_j\frac{\lambda_{T,j}q_j}{\bar{\lambda}_j^*}\\
&\quad+\lambda_N\frac{s\mu_N}{\sum_{j=1}^{N}\lambda_j}+\sum_{j=1}^{N-1}(q_j+1)\lambda_{T,j}p_{j,N}\frac{\bar{\lambda}_j^* s\mu_N}{\lambda_{T,j}(q_j+1)(\sum_{j=1}^{N}\lambda_j)}\\
&\quad+\sum_{j=1}^{N-1}(q_j+1)\lambda_{T,j}\sum_{k=j+1:q_k\ge 1}^{N-1}p_{j,k}\frac{\bar{\lambda}_j^*\lambda_{T,k}q_k}{\lambda_{T,j}(q_j+1)\bar{\lambda}_k^*}\Bigg\}\\
&=\pi_u(\mathbf{q})\Bigg\{-\Bigg(s\mu_N+\sum_{j=1}^{N-1}q_j\lambda_{T,j}\Bigg)+\lambda_N\frac{s\mu_N}{\bar{\lambda}_N^*}+\frac{s\mu_N}{\bar{\lambda}_N^*}\sum_{j=1}^{N-1}\bar{\lambda}_j^*p_{j,N}\\
&\quad+\sum_{j=1}^{N-1}\lambda_j\frac{\lambda_{T,j}q_j}{\bar{\lambda}_j^*}+\sum_{j=1}^{N-1}\sum_{k=j+1}^{N-1}p_{j,k}\frac{\bar{\lambda}_j^*\lambda_{T,k}q_k}{\bar{\lambda}_k^*}\Bigg\}\\
&=\pi_u(\mathbf{q})\Bigg\{-\Bigg(s\mu_N+\sum_{j=1}^{N-1}q_j\lambda_{T,j}\Bigg)+\frac{s\mu_N}{\bar{\lambda}_N^*}\Bigg(\lambda_N+\sum_{j=1}^{N-1}\bar{\lambda}_j^*p_{j,N}\Bigg)\\
&\quad+\sum_{j=1}^{N-1}\lambda_j\frac{\lambda_{T,j}q_j}{\bar{\lambda}_j^*}+\sum_{k=2}^{N-1}\frac{q_k\lambda_{T,k}}{\bar{\lambda}_k^*}\sum_{j=1}^{k-1}\bar{\lambda}_j^*p_{j,k}\Bigg\}\\
&=\pi_u(\mathbf{q})\Bigg\{-\Bigg(\sum_{j=1}^{N-1}q_j\lambda_{T,j}\Bigg)+\sum_{k=1}^{N-1}\frac{q_k\lambda_{T,k}}{\bar{\lambda}_k^*}\Bigg(\lambda_k+\sum_{j=1}^{k-1}\bar{\lambda}_j^*p_{j,k}\Bigg)\Bigg\}=0. \qquad (3.7)
\end{aligned}$$

Note that the relationships given in Lemma 3.1 are used in the above calculations. The case with $q_N < s$ can be verified similarly as follows:

$$\begin{aligned}
&-\pi_u(\mathbf{q})\Bigg(q_N\mu_N+\sum_{j=1}^{N}\lambda_j+\sum_{j=1}^{N-1}q_j\lambda_{T,j}\Bigg)+\pi_u\big(\mathbf{q}+\mathbf{e}(N)\big)(q_N+1)\mu_N\\
&\quad+\sum_{j=1:q_j\ge 1}^{N}\pi_u\big(\mathbf{q}-\mathbf{e}(j)\big)\lambda_j\\
&\quad+\sum_{j=1}^{N-1}\sum_{k=j+1:q_k\ge 1}^{N}\pi_u\big(\mathbf{q}-\mathbf{e}(k)+\mathbf{e}(j)\big)(q_j+1)\lambda_{T,j}p_{j,k}\\
&=\pi_u(\mathbf{q})\Bigg\{-\Bigg(q_N\mu_N+\sum_{j=1}^{N}\lambda_j+\sum_{j=1}^{N-1}q_j\lambda_{T,j}\Bigg)+\sum_{j=1}^{N}\lambda_j+\sum_{j=1:q_j\ge 1}^{N-1}\lambda_j\frac{\lambda_{T,j}q_j}{\bar{\lambda}_j^*}
\end{aligned}$$

$$
\begin{aligned}
&+\lambda_N \frac{q_N\mu_N}{\bar{\lambda}_N^*} + \sum_{j=1}^{N-1}(q_j+1)\lambda_{T,j}p_{j,N}\frac{\bar{\lambda}_j^* q_N\mu_N}{\lambda_{T,j}(q_j+1)\lambda_N^*} \\
&+\sum_{j=1}^{N-1}(q_j+1)\lambda_{T,j}\sum_{k=j+1:q_k\ge 1}^{N-1} p_{j,k}\frac{\bar{\lambda}_j^*\lambda_{T,k}q_k}{\lambda_{T,j}(q_j+1)\bar{\lambda}_k^*}\Bigg\} \\
&=\pi_u(\mathbf{q})\Bigg\{-\Bigg(q_N\mu_N+\sum_{j=1}^{N-1}q_j\lambda_{T,j}\Bigg)+\lambda_N\frac{q_N\mu_N}{\bar{\lambda}_N^*}+\frac{q_N\mu_N}{\bar{\lambda}_N^*}\sum_{j=1}^{N-1}\bar{\lambda}_j^*p_{j,N} \\
&+\sum_{j=1}^{N-1}\lambda_j\frac{\lambda_{T,j}q_j}{\bar{\lambda}_j^*}+\sum_{j=1}^{N-1}\sum_{k=j+1}^{N-1}p_{j,k}\frac{\bar{\lambda}_j^*\lambda_{T,k}q_k}{\bar{\lambda}_k^*}\Bigg\}=0. \qquad (3.8)
\end{aligned}
$$

This completes the proof of Theorem 3.2. □

Theorem 3.2 indicates that the stationary distribution of the designated service queueing system has the product-form of $N-1$ independent Poisson distributions and one (modified) geometric distribution. The marginal distribution of the queue length of queue j has a Poisson distribution for $1 \le j \le N-1$, which is the stationary distribution of the queue length in an $M/M/\infty$ queue with arrival rate $\bar{\lambda}_j^*$ and service rate $\lambda_{T,j}$. The marginal distribution of the queue length of queue N is identical to that of an $M/M/s$ queue with arrival rate $\bar{\lambda}_N^* = \sum_{j=1}^N \lambda_j$ and service rate μ_N per server.

The designated service queueing system provides bounds on the stationary distribution of the priority service queueing system. To obtain the bounds, we first establish some sample path relationships between $\{(q_{u,1}(t), q_{u,2}(t), \ldots, q_{u,N}(t)), t \ge 0\}$ and $\{(q_1(t), q_2(t), \ldots, q_N(t)), t \ge 0\}$ in Lemma 3.3.

Lemma 3.3 *Assume that both the priority service queueing system and the designated service queueing system are empty at time zero. For the queue length processes* $\{(q_1(t), q_2(t), \ldots, q_N(t)), t \ge 0\}$ *and* $\{(q_{u,1}(t), q_{u,2}(t), \ldots, q_{u,N}(t)), t \ge 0\}$, *we have*

(i) $q_j(t) \le q_{u,j}(t) + js$, *for* $1 \le j \le N$ *and* $t \ge 0$

(ii) $q_1(t)+q_2(t)+\cdots+q_j(t) \le q_{u,1}(t)+q_{u,2}(t)+\cdots+q_{u,j}(t)+js$, *for* $1 \le j \le N$ *and* $t \ge 0$.

Proof Since we are only interested in how many customers there are in the queues, the order of service does not affect the queue length distribution. Thus, we further assume that, for queues 1 to $N-1$, all customers (including transferred customers) are served on a first-come-first-served basis; for queue N, upon a service completion, a customer who is originally a type N customer (i.e., a customer arrives at the queueing system as a type N customer) leaves the system if there is such a customer in queue N.

Consider queue 1 at time t. Customers arrived in $[0, t)$ can be categorized as follows. For all customers that have been transferred without receiving service (in the priority service queueing system), they are no longer part of $q_1(t)$ (in the priority service queueing system) nor part of $q_{u,1}(t)$ (in the designated service queueing system). For all customers that have been transferred, but was ever attended by servers in the priority service queueing system before being transferred, they are no longer part of $q_1(t)$ nor part of $q_{u,1}(t)$. Note that transfer times are exponentially distributed so that we can assume that a transferring process is resumed after a service interruption. For all customers that have been served in the priority service queueing system, they are no longer part of $q_1(t)$ but can be part of $q_{u,1}(t)$ in the designated service queueing system. For all customers that are either in service or have been attended by a server(s) but still in the transferring process, they are part of $q_1(t)$ but may not be part of $q_{u,1}(t)$. 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 $q_1(t) \leq q_{u,1}(t) + s$.

For queue 2 at time t, customers still in the queue can be divided into two groups: (i) customers who arrive to queue 2 at the same epochs for both systems; and (ii) customers who arrive to queue 2 at different time epochs for the two systems. For customers in the first group, similarly to the above analysis for queue 1, the queue 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 transferred from queue 1 and have been attended by a server in queue 1 (but their services were never completed.) The reason is that queue 2 must have no waiting customer before a server can attend a customer in queue 1. These customers can be part of $q_1(t)$, but may not be part of $q_{u,1}(t)$. Therefore, we must have $q_2(t) \leq q_{u,2}(t) + 2s$.

In general, for queue j, among customers who arrive to queue 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 $(j-1)s$ transferred customers who arrive to queue j at different time epochs for the two systems—at most s from queue k $(<j)$ (those are customers in queue j who have received incomplete service in queue k), $k = 1, 2, \ldots, j-1$. Therefore, we must have $q_j(t) \leq q_{u,j}(t) + js$.

Part (ii) can be proved by the same argument that there can be at most s customers that have received incomplete service in queue 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 queues $j+1, j+2, \ldots,$ and N, before a server can move down to serve queue j (to possibly create another such customer in queue j). This completes the proof of Lemma 3.3. □

It is well known that the sample path order implies stochastically smaller/larger order (see [27] for more about stochastic orders of random variables). Thus, by Lemma 3.3, $q_j(t)$ is stochastically smaller than $q_{u,j}(t) + js$ for $j = 1, 2, \ldots, N$. Intuitively, under proper initial conditions, $q_j(t)$ should be stochastically smaller than $q_{u,j}(t)$ for $j = 1, 2, \ldots, N$. Unfortunately, that is not true in general. For example, consider a system with $N = 2, \lambda_1 = 1$, sufficiently small λ_2, sufficiently small μ_1, sufficiently large μ_2, and large $\lambda_{T,1}$ (relative to μ_1). For this case, queue 1 in the designated service queueing system is almost empty, but queue 1 in the priority service queueing system can have a customer in service for a significant amount of time. Consequently, on average, queue 1 in the priority queueing system could be longer than the queue 1 in the designated service queueing system.

Denote by $\{\pi(q_1, q_2, \ldots, q_N), q_1 \ge 0, \ldots, q_N \ge 0\}$ the stationary distribution of $\{(q_1(t), q_2(t), \ldots, q_N(t)), t \ge 0\}$. Lemma 3.3 leads to the following bounds on the stationary distribution.

Corollary 3.4 *Assume that system parameters* $\{\lambda_1, \ldots, \lambda_N, \mu_1, \ldots, \mu_N, \lambda_{T,1}, \ldots, \lambda_{T,N-1}, s\}$ *are positive and* $\sum_{k=1}^{N}\lambda_k < s\mu_N$. *We have, for* $\mathbf{q} = (q_1, \ldots, q_N)$ *with* $q_N > (N+1)s$,

$$\sum_{\mathbf{w}\ge\mathbf{q}}\pi(\mathbf{w}) \le \left(\prod_{j=1}^{N-1}\left(\frac{\bar{\rho}_{T,j}^{(q_j-js)^+}}{(q_j-js)^+!}\right)\right)\frac{\frac{(s\bar{\rho}_N)^s}{s!}}{(\sum_{j=0}^{s-1}\frac{(s\bar{\rho}_N)^j}{j!}+\frac{(s\bar{\rho}_N)^s}{s!(1-\bar{\rho}_N)})}\frac{\bar{\rho}_N^{(q_N-(N+1)s)}}{(1-\bar{\rho}_N)}, \tag{3.9}$$

where $(q_j - js)^+ = \max\{0, q_j - js\}$ *and* $\mathbf{w} = (w_1, w_2, \ldots, w_N)$.

Proof By Lemma 3.3, it is easy to see that, for $\mathbf{q} \ge 0$,

$$\big\{\mathbf{q}(t) \ge \mathbf{q}\big\} \subset \big\{\mathbf{q}_u(t) \ge \mathbf{q} - (s, 2s, \ldots, Ns)\big\}. \tag{3.10}$$

By Theorems 2.1, 3.2, and Lemma 3.3, the priority service queueing system and the corresponding designated service system are stable. By Theorem 3.2, (3.10), and routine calculations, we have

$$\begin{aligned}
\sum_{\mathbf{w}\ge\mathbf{q}}\pi(\mathbf{w}) &= \lim_{t\to\infty} P\big\{\mathbf{q}(t)\ge\mathbf{q}\big\} \le \lim_{t\to\infty} P\big\{\mathbf{q}_u(t) \ge \mathbf{q}-(s,2s,\ldots,Ns)\big\}\\
&= \sum_{\mathbf{w}\ge\mathbf{q}-(s,2s,\ldots,Ns)}\left(\prod_{j=1}^{N-1}\left(\exp\{-\bar{\rho}_{T,j}\}\frac{\bar{\rho}_{T,j}^{w_j^+}}{w_j^+!}\right)\right)\frac{\frac{(s\bar{\rho}_N)^s}{s!}\bar{\rho}_N^{w_N-s}}{(\sum_{j=0}^{s-1}\frac{(s\bar{\rho}_N)^j}{j!}+\frac{(s\bar{\rho}_N)^s}{s!(1-\bar{\rho}_N)})}\\
&= \left(\prod_{j=1}^{N-1}\left(\sum_{w_j=(q_j-js)^+}^{\infty}\exp\{-\bar{\rho}_{T,j}\}\frac{\bar{\rho}_{T,j}^{w_j}}{w_j!}\right)\right)\frac{\frac{(s\bar{\rho}_N)^s}{s!}(\sum_{w_N=q_N-Ns}^{\infty}\bar{\rho}_N^{w_N-s})}{(\sum_{j=0}^{s-1}\frac{(s\bar{\rho}_N)^j}{j!}+\frac{(s\bar{\rho}_N)^s}{s!(1-\bar{\rho}_N)})}\\
&= \left(\prod_{j=1}^{N-1}\left(\frac{\bar{\rho}_{T,j}^{(q_j-js)^+}}{(q_j-js)^+!}\left(\exp\{-\bar{\rho}_{T,j}\}\sum_{w_j=0}^{\infty}\frac{\bar{\rho}_{T,j}^{w_j}(q_j-js)^+!}{(w_j+(q_j-js)^+)!}\right)\right)\right)\\
&\quad\times\frac{\frac{(s\bar{\rho}_N)^s}{s!}\frac{\bar{\rho}_N^{(q_N-(N+1)s)}}{(1-\bar{\rho}_N)}}{(\sum_{j=0}^{s-1}\frac{(s\bar{\rho}_N)^j}{j!}+\frac{(s\bar{\rho}_N)^s}{s!(1-\bar{\rho}_N)})}\\
&\le \left(\prod_{j=1}^{N-1}\left(\frac{\bar{\rho}_{T,j}^{(q_j-js)^+}}{(q_j-js)^+!}\left(\exp\{-\bar{\rho}_{T,j}\}\sum_{w_j=0}^{\infty}\frac{\bar{\rho}_{T,j}^{w_j}}{w_j!}\right)\right)\right)\\
&\quad\times\frac{\frac{(s\bar{\rho}_N)^s}{s!}\frac{\bar{\rho}_N^{(q_N-(N+1)s)}}{(1-\bar{\rho}_N)}}{(\sum_{j=0}^{s-1}\frac{(s\bar{\rho}_N)^j}{j!}+\frac{(s\bar{\rho}_N)^s}{s!(1-\bar{\rho}_N)})},
\end{aligned}\tag{3.11}$$

which leads to (3.9). This completes the proof of Corollary 3.4. □

Remark 3.2 Similarly to the designated service queueing system, another queueing system—*all service queueing system*—can be introduced, which provides explicit lower bounds on the stationary distribution of queue lengths of the priority service queueing system. See Xie et al. [33] for details.

4 Tail asymptotics of the stationary distribution: queue N overloaded

We consider $\lim_{q_N \to \infty} \pi(q_1, q_2, \dots, q_N)$ in this section, i.e., the tail asymptotics of the stationary distribution, if queue N is overloaded. For the designated service queueing system, the stationary distribution has a product-form. Thus, the tail asymptotics, if queue N is overloaded, is exact geometric and the decay rate is $\bar{\rho}_N$. This implies that the tail asymptotics of queue N in the priority service queueing system can be geometric and the decay rate might be close to $\bar{\rho}_N$. In this section, we show that the tail asymptotics is exact geometric and the decay rate is in fact $\bar{\rho}_N$.

In order to apply matrix-analytic methods, we reorder the queue length variables as $(q_N(t), (q_1(t), \dots, q_{N-1}(t)))$. We call $q_N(t)$ the level variable and $(q_1(t), q_2(t), \dots, q_{N-1}(t))$ the (vector) phase variable. The states $\{(q_N, (q_1, \dots, q_{N-1})), 0 \le q_j < \infty, 1 \le j \le N\}$ are ordered lexicographically. Level n consists of states $\{(q_N, (q_1, \dots, q_{N-1})) \colon q_N = n, 0 \le q_j < \infty, 1 \le j \le N-1\}$. The infinitesimal generator associated with the Markov chain can be rewritten as

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & & & & & & \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & Q_{s-2,s-1} & Q_{s-1,s-1} & Q_{s-1,s} & & & \\ & & & Q_{s,s-1} & Q_0 & Q_1 & & \\ & & & & Q_{-1} & Q_0 & Q_1 & \\ & & & & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{4.1}$$

We partition the stationary distribution vector $\boldsymbol{\pi}$ according to the level variable $q_N(t)$ as $(\boldsymbol{\pi}[0], \boldsymbol{\pi}[1], \boldsymbol{\pi}[2], \dots)$, where the elements of the vector $\boldsymbol{\pi}[n]$ are probabilities $\{\pi(q_1, q_2, \dots, q_{N-1}, n), q_1 \ge 0, q_2 \ge 0, \dots, q_{N-1} \ge 0\}$ ordered lexicographically. By Theorem 1 in Miller [21], the stationary distribution of the Markov chain has a matrix-geometric solution given as

$$\begin{aligned} &\boldsymbol{\pi}[n] = \boldsymbol{\pi}[s] R^{n-s}, \quad n \ge s; \\ &\big(\boldsymbol{\pi}[0], \boldsymbol{\pi}[1], \dots, \boldsymbol{\pi}[s]\big) \begin{pmatrix} Q_{0,0} & Q_{0,1} & & & \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & & \\ & \ddots & \ddots & \ddots & \\ & & Q_{s-1,s-2} & Q_{s-1,s-1} & Q_{s-1,s} \\ & & & Q_{s,s-1} & Q_0 + RQ_{-1} \end{pmatrix} = 0; \\ &\sum_{n=0}^{s-1} \boldsymbol{\pi}[n]\mathbf{e} + \boldsymbol{\pi}[s](I - R)^{-1}\mathbf{e} = 1, \end{aligned} \tag{4.2}$$

where I is the identity matrix, $\mathbf{e}$ is the column vector of ones, and the matrix R is the minimal nonnegative solution to the equation:

$$Q_1 + RQ_0 + R^2 Q_{-1} = 0. \tag{4.3}$$

To find tail asymptotics for $\{\boldsymbol{\pi}[n], n \geq 0\}$, we apply Theorem 2.1 in Li et al. [18]. Unfortunately, for state $(q_N, (q_1, \ldots, q_{N-1}))$ with $q_N \geq s$, the absolute value of the diagonal element in Q_0 is $s\mu_N + \sum_{j=1}^{N} \lambda_j + \sum_{j=1}^{N-1} q_j \lambda_{T,j}$, which is unbounded since $\{q_1, \ldots, q_{N-1}\}$ can be arbitrarily large. In order to apply Theorem 2.1 in Li et al. [18], we consider a discrete time Markov chain with the following transition probability matrix:

$$P = D^{-1}Q + I \equiv \begin{pmatrix} A_{0,0} & A_{0,1} & & & & & & \\ A_{1,0} & A_{1,1} & A_{1,2} & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & A_{s-2,s-1} & A_{s-1,s-1} & A_{s-1,s} & & & \\ & & & A_{s,s-1} & A_0 & A_1 & & \\ & & & & A_{-1} & A_0 & A_1 & \\ & & & & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{4.4}$$

where D is a diagonal matrix for which the diagonal elements are the absolute values of the diagonal elements of the matrix Q (i.e., $D = -\mathrm{diag}(Q)$, where $\mathrm{diag}(Q)$ is a matrix obtained by keeping all diagonal elements of Q and setting all other elements in Q to zero). It is easy to see that P is an irreducible stochastic matrix. Let

$$\boldsymbol{\theta} = \boldsymbol{\pi} D / (\boldsymbol{\pi} D\mathbf{e}). \tag{4.5}$$

Lemma 4.1 *The vector $\boldsymbol{\theta}$ is the stationary distribution of Markov chain P if and only if $\boldsymbol{\pi}$ is the stationary distribution of the continuous time Markov chain Q.*

Proof If both $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ are finite, then it is easy to see that they are the stationary distribution of P and Q, respectively. Thus, we only need to show that $\boldsymbol{\pi} D\mathbf{e}$ and $\boldsymbol{\theta} D^{-1}\mathbf{e}$ are finite if $\boldsymbol{\pi}\mathbf{e}$ and $\boldsymbol{\theta}\mathbf{e}$ are finite, respectively. Note that the inverse D^{-1} is well defined since D is a diagonal matrix with positive diagonal elements. It is easy to see that $\boldsymbol{\theta} D^{-1}\mathbf{e}$ is finite if $\boldsymbol{\theta}\mathbf{e}$ is finite. If $\boldsymbol{\pi}\mathbf{e}$ is finite, (3.9) indicates that $\boldsymbol{\pi}$ is bounded from above by the product of some Poisson functions (up to a constant multiplier). By the definition of D, it is easy to verify that $\boldsymbol{\pi} D\mathbf{e}$ is finite. This completes the proof of Lemma 4.1. □

Define

$$A_*(z) = z^{-1}A_1 + A_0 + zA_{-1}, \quad \text{for } z > 0. \tag{4.6}$$

Further, define a row vector $\mathbf{x}$ with elements $\{x(q_1, \ldots, q_{N-1}), 0 \leq q_j < \infty, 1 \leq j \leq N-1\}$ ordered lexicographically and a column vector $\mathbf{y}$ with elements

$\{y(q_1,\dots,q_{N-1}), 0 \le q_j < \infty, 1 \le j \le N-1\}$ ordered lexicographically, where

$$\begin{aligned} x(q_1,\dots,q_{N-1}) &= \prod_{j=1}^{N-1}\left(\frac{\bar{\rho}_{T,j}^{q_j}}{q_j!}\exp\{-\bar{\rho}_{T,j}\}\right); \\ y(q_1,\dots,q_{N-1}) &= \bar{\rho}_N^{(-q_1-q_2-\cdots-q_{N-1})}. \end{aligned} \tag{4.7}$$

Note that the vector $\mathbf{x}$ represents a finite measure and we have normalized $\mathbf{x}$ to have a unit sum. Define $D_1 = -\mathrm{diag}(Q_0)$. The diagonal elements of D_1 are $s\mu_N + \sum_{j=1}^{N}\lambda_j + \sum_{j=1}^{N-1} q_j\lambda_{T,j}$.

Lemma 4.2 *The vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *satisfy* $\mathbf{x}D_1A_*(\bar{\rho}_N) = \mathbf{x}D_1$ *and* $A_*(\bar{\rho}_N)\mathbf{y} = \mathbf{y}$. *In addition, we have* $\mathbf{x}D_1\mathbf{y} < \infty$.

Proof Note that $A_{-1} = D_1^{-1}Q_{-1}$, $A_0 = D_1^{-1}Q_0 + I$, and $A_1 = D_1^{-1}Q_1$. We show that $\mathbf{x}(\bar{\rho}_N Q_{-1} + Q_0 + Q_1/\bar{\rho}_N) = 0$. For state $(q_1,\dots,q_{N-1})$, we have

$$\begin{aligned} & x(q_1,\dots,q_{N-1})\bar{\rho}_N s\mu_N - x(q_1,\dots,q_{N-1})\left(s\mu_N + \sum_{j=1}^{N}\lambda_j + \sum_{j=1}^{N-1} q_j\lambda_{T,j}\right) \\ &\quad + \sum_{j=1:q_j\ge 1}^{N-1} x(q_1,\dots,q_j-1,\dots,q_{N-1})\lambda_j \\ &\quad + \sum_{j=1}^{N-1}\sum_{k=j+1:q_k\ge 1}^{N-1} x(q_1,\dots,q_j+1,\dots,q_k-1,\dots,q_{N-1})(q_j+1)\lambda_{T,j}p_{j,k} \\ &\quad + x(q_1,\dots,q_{N-1})\frac{\lambda_N}{\bar{\rho}_N} + \sum_{j=1}^{N-1} x(q_1,\dots,q_j+1,\dots,q_{N-1})\frac{(q_j+1)\lambda_{T,j}p_{j,N}}{\bar{\rho}_N} \\ &= x(q_1,\dots,q_{N-1})\left\{-\left(s\mu_N + \sum_{j=1}^{N-1} q_j\lambda_{T,j}\right) + \sum_{j=1}^{N-1}\frac{q_j\lambda_{T,j}\lambda_j}{\bar{\lambda}_j^*}\right. \\ &\quad \left. + \sum_{j=1}^{N-1}\sum_{k=j+1}^{N-1} p_{j,k}\frac{q_k\lambda_{T,k}\bar{\lambda}_j^*}{\bar{\lambda}_k^*} + \frac{\lambda_N}{\bar{\rho}_N} + \sum_{j=1}^{N-1}\frac{p_{j,N}\bar{\lambda}_j^*}{\bar{\rho}_N}\right\} \\ &= x(q_1,\dots,q_{N-1})\left\{-\left(s\mu_N + \sum_{j=1}^{N-1} q_j\lambda_{T,j}\right) + \sum_{j=1}^{N}\frac{\lambda_j}{\bar{\rho}_N}\right. \\ &\quad \left. + \sum_{j=1}^{N-1}\frac{q_j\lambda_{T,j}}{\bar{\lambda}_j^*}\left(\lambda_j + \sum_{k=1}^{j-1}\bar{\lambda}_k^* p_{k,j}\right)\right\} \\ &= 0. \end{aligned} \tag{4.8}$$

Then the vector $\mathbf{x}D_1$ satisfies $\mathbf{x}D_1A_*(\bar{\rho}_N) = \mathbf{x}(\bar{\rho}_N Q_{-1} + Q_0 + Q_1/\bar{\rho}_N) + \mathbf{x}D_1 = \mathbf{x}D_1$.

For $A_*(\bar{\rho}_N)\mathbf{y} = \mathbf{y}$, equivalently, we show that $(\bar{\rho}_N Q_{-1} + Q_0 + Q_1/\bar{\rho}_N)\mathbf{y} = 0$. For state $(q_1, \dots, q_{N-1})$, we have

$$\begin{aligned}
&\bar{\rho}_N s\mu_N y(q_1, \dots, q_{N-1}) - \left(s\mu_N + \sum_{j=1}^{N} \lambda_j + \sum_{j=1}^{N-1} q_j \lambda_{T,j}\right) y(q_1, \dots, q_{N-1}) \\
&\quad + \sum_{j=1}^{N-1} y(q_1, \dots, q_j + 1, \dots, q_{N-1})\lambda_j \\
&\quad + \sum_{j=1:q_j \ge 1}^{N-1} \sum_{k=j+1}^{N-1} p_{j,k} y(q_1, \dots, q_j - 1, \dots, q_k + 1, \dots, q_{N-1}) q_j \lambda_{T,j} \\
&\quad + y(q_1, \dots, q_{N-1}) \frac{\lambda_N}{\bar{\rho}_N} + \sum_{j=1}^{N-1} p_{j,N} y(q_1, \dots, q_j - 1, \dots, q_{N-1}) \frac{q_j \lambda_{T,j}}{\bar{\rho}_N} \\
&= y(q_1, \dots, q_{N-1}) \Bigg\{ -\left(s\mu_N + \sum_{j=1}^{N-1} q_j \lambda_{T,j}\right) + \sum_{j=1}^{N-1} \frac{\lambda_j}{\bar{\rho}_N} \\
&\quad + \sum_{j=1}^{N-1} \sum_{k=j+1}^{N-1} p_{j,k} q_j \lambda_{T,j} + \frac{\lambda_N}{\bar{\rho}_N} + \sum_{j=1}^{N-1} p_{j,N} q_j \lambda_{T,j} \Bigg\} \\
&= y(q_1, \dots, q_{N-1}) \left\{ -\left(\sum_{j=1}^{N-1} q_j \lambda_{T,j}\right) + \sum_{j=1}^{N-1} q_j \lambda_{T,j} \left(p_{j,N} + \sum_{k=j+1}^{N-1} p_{j,k}\right) \right\} \\
&= 0.
\end{aligned} \tag{4.9}$$

Finally, we have

$$\begin{aligned}
\mathbf{x}D_1\mathbf{y} &= \sum_{(q_1, q_2, \dots, q_{N-1}) \ge 0} \exp\left\{ -\sum_{j=1}^{N-1} \bar{\rho}_{T,j} \right\} \left(\prod_{j=1}^{N-1} \frac{\bar{\rho}_{T,j}^{q_j}}{q_j!} \right) \bar{\rho}_N^{-(q_1 + \dots + q_{N-1})} \\
&\quad \times \left(s\mu_N + \bar{\lambda}_N^* + \sum_{j=1}^{N-1} q_j \lambda_{T,j} \right) \\
&= \exp\left\{ -\sum_{j=1}^{N-1} \bar{\rho}_{T,j} \right\} \sum_{(q_1, q_2, \dots, q_{N-1}) \ge 0} \left(\prod_{j=1}^{N-1} \frac{(\bar{\rho}_{T,j}/\bar{\rho}_N)^{q_j}}{q_j!} \right) \\
&\quad \times \left(s\mu_N + \bar{\lambda}_N^* + \sum_{j=1}^{N-1} q_j \lambda_{T,j} \right)
\end{aligned}$$

$$= \exp\left\{\sum_{j=1}^{N-1} \bar{\rho}_{T,j}\left(\frac{1}{\bar{\rho}_N} - 1\right)\right\}\left(s\mu_N + \bar{\lambda}_N^* + \sum_{j=1}^{N-1} \frac{\bar{\rho}_{T,j}}{\bar{\rho}_N}\lambda_{T,j}\right) < \infty. \tag{4.10}$$

This completes the proof of Lemma 4.2. □

Lemma 4.2 indicates that $\mathbf{x}D_1$ and $\mathbf{y}$ are a positive invariant measure and a positive invariant vector of the matrix $A_*(\bar{\rho}_N)$, respectively.

Theorem 4.3 *Assume that system parameters* $\{\lambda_1, \ldots, \lambda_N, \mu_1, \ldots, \mu_N, \lambda_{T,1}, \ldots, \lambda_{T,N-1}, s\}$ *are positive. If* $\sum_{k=1}^N \lambda_k < s\mu_N$, *we have*

$$\lim_{q_N\to\infty} \frac{\boldsymbol{\pi}[q_N]}{\bar{\rho}_N^{(q_N-s)}} = \frac{\boldsymbol{\pi}[s-1]D_2\mathbf{r}}{\mathbf{x}D_1\mathbf{r}}\mathbf{x}, \quad (\textit{component-wise}) \tag{4.11}$$

where $\mathbf{r} = (I - A_0 - D_1^{-1}RD_1A_{-1} - \bar{\rho}_N A_{-1})\mathbf{y}$ *and* $D_2 = -\mathrm{diag}(Q_{s-1,s-1})$. *The constants* $\boldsymbol{\pi}[s-1]D_2\mathbf{r}$ *and* $\mathbf{x}D_1\mathbf{r}$ *are positive and finite.*

Proof We consider the discrete time Markov chain P and its stationary distribution $\boldsymbol{\theta}$. Similarly to $\boldsymbol{\pi}$, we partition $\boldsymbol{\theta}$ into $(\boldsymbol{\theta}[0], \boldsymbol{\theta}[1], \boldsymbol{\theta}[2], \ldots)$. First, note that the matrix $D_1^{-1}RD_1$ is the minimal nonnegative solution to $A_1 + XA_0 + X^2A_{-1} = 0$. By Theorem 2.1 in Li et al. [18], to prove (4.11), we need to check the following conditions:

(i) The matrix $A_1 + A_0 + A_{-1}$ is irreducible and aperiodic
(ii) The Markov additive process with blocks $\{A_1, A_0, A_{-1}\}$ are 1-arithmetic
(iii) $0 < \bar{\rho}_N < 1$
(iv) $\mathbf{x}D_1\mathbf{y} < \infty$
(v) $\boldsymbol{\theta}[s-1]A_{s-1,s}\mathbf{y} < \infty$.

Condition (v) is a condition associated with the set of boundary states, which consists of states in levels 0 to $s-1$, i.e., $\{(q_1, \ldots, q_{N-1}, q_N), q_1 \geq 0, \ldots, q_{N-1} \geq 0, 0 \leq q_N \leq s-1\}$. Condition (v) is obtained from condition (v) in Li et al. [18] as follows:

$$\big(\boldsymbol{\theta}[0], \boldsymbol{\theta}[1], \ldots, \boldsymbol{\theta}[s-1]\big)\begin{pmatrix} 0 \\ \vdots \\ 0 \\ A_{s-1,s} \end{pmatrix}\mathbf{y} = \boldsymbol{\theta}[s-1]A_{s-1,s}\mathbf{y}. \tag{4.12}$$

Conditions (i) and (ii) can be checked in a straightforward manner. Condition (iii) is an assumption. Condition (iv) is from Lemma 4.1. We verify Condition (v) by using Corollary 3.4. Note that $\boldsymbol{\theta}[s-1] = \boldsymbol{\pi}[s-1]D_2/(\boldsymbol{\pi}D\mathbf{e})$ and $A_{s-1,s} = D_2^{-1}Q_{s-1,s}$. By (3.9), we obtain

$$\begin{aligned}\boldsymbol{\theta}[s-1]&A_{s-1,s}\mathbf{y}\\ &= \boldsymbol{\pi}[s-1]Q_{s-1,s}\mathbf{y}/(\boldsymbol{\pi}D\mathbf{e})\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\boldsymbol{\pi} D\mathbf{e}} \sum_{(q_1,q_2,\ldots,q_{N-1})\geq 0} \bigl(\lambda_N \pi(q_1,\ldots,q_{N-1},s-1) y(q_1,\ldots,q_{N-1})\bigr) \\
&\quad + \frac{1}{\boldsymbol{\pi} D\mathbf{e}} \sum_{(q_1,q_2,\ldots,q_{N-1})\geq 0} \Biggl(\sum_{j=1}^{N-1} \pi(q_1,\ldots,q_j+1,\ldots,q_{N-1},s-1) \\
&\quad \times (q_j+1)\lambda_{T,j} p_{j,N} y(q_1,\ldots,q_{N-1})\Biggr) \\
&\leq c(sN) + \frac{\lambda_N}{\boldsymbol{\pi} D\mathbf{e}} \sum_{q_j\geq Ns, 1\leq j\leq N-1} \Biggl(\Biggl(\prod_{j=1}^{N-1}\biggl(\frac{\bar{\rho}_{T,j}^{(q_j-js)}}{(q_j-js)!}\biggr)\Biggr) \bar{\rho}_N^{-(q_1+\cdots+q_{N-1})}\Biggr) \\
&\quad + \frac{1}{\boldsymbol{\pi} D\mathbf{e}} \sum_{j=1}^{N-1} \sum_{q_j\geq Ns, 1\leq j\leq N-1} \Biggl(\Biggl(\prod_{i=1}^{N-1}\biggl(\frac{\bar{\rho}_{T,j}^{(q_i-js)}}{(q_i-js)!}\biggr)\Biggr) \\
&\quad \times \bar{\rho}_{T,j}(q_j+1)\lambda_{T,j} p_{j,N} \bar{\rho}_N^{-(q_1+\cdots+q_{N-1})}\Biggr) \\
&< \infty,
\end{aligned}
\tag{4.13}
$$

where $c(sN)$ is the summation for items associated with states $\{(q_1,\ldots,q_{N-1},s-1), 0\leq q_j\leq sN, 0\leq j\leq N-1\}$. By Theorem 2.1 in Li et al. [18], we obtain

$$\lim_{q_N\to\infty} \bar{\rho}_N^{-(q_N-s)} \boldsymbol{\theta}[q_N] = \frac{\boldsymbol{\theta}[s-1]\mathbf{r}}{\mathbf{x}D_1\mathbf{r}} \mathbf{x}D_1. \tag{4.14}$$

Since $\boldsymbol{\pi} D/(\boldsymbol{\pi} D\mathbf{e}) = \boldsymbol{\theta}$ (which implies that $\boldsymbol{\pi}[s-1]D_2/(\boldsymbol{\pi} D\mathbf{e}) = \boldsymbol{\theta}[s-1]$ and $\boldsymbol{\pi}[n]D_1/(\boldsymbol{\pi} D\mathbf{e}) = \boldsymbol{\theta}[n]$ for $n\geq s$), (4.11) is obtained from (4.14). This completes the proof of Theorem 4.3. □

Theorem 4.3 confirms that the stationary distribution decays exponentially with decay rate $\bar{\rho}_N$ and the tail asymptotics is exact geometric, if queue N is overloaded. Theorem 4.3 has an intuitive interpretation. If queue N is long, all servers are busy in serving queue N. Then all other queues will evolve like $M/M/\infty$ queues for a long time and have a conditional distribution $\mathbf{x}$, which is confirmed by the vector $\mathbf{x}$ on the right-hand side of (4.11). Note that $\mathbf{x}$ is the marginal probability distribution of queues 1 to $N-1$ in the designated service queueing system (see Theorem 3.2).

Theorem 4.3 implies that, if queue N is overloaded, the priority service queueing system behaves similarly to the designated service queueing system, which is independent of the service rates of lower priority customers in the priority service queueing system. But that does not mean that the service rates of lower priority customers have no impact on the stationary distribution $\boldsymbol{\pi}$. In general, probability $\boldsymbol{\pi}[q_N]$ is affected strongly by the service rates if q_N is small. Even if q_N is large, $\boldsymbol{\pi}[q_N]$ is affected by the service rates as demonstrated below. We rewrite (4.11) as

$$\boldsymbol{\pi}[q_N] \approx c\mathbf{x}\bar{\rho}_N^{(q_N-s)}, \quad \text{if } q_N\to\infty, \tag{4.15}$$

where $c = \boldsymbol{\pi}[s-1]D_2\mathbf{r}/(\mathbf{x}D_1\mathbf{r})$. In (4.15), $\mathbf{x}\bar{\rho}_N^{(q_N-s)}$ does not depend on the service rates of lower priority customers, but the constant c does. Unfortunately, it is difficult to find an explicit expression for c. Alternatively, we use simulation to estimate c and to compare c with c_u, where c_u satisfies $\boldsymbol{\pi}_u[q_N] = c_u\mathbf{x}\bar{\rho}_N^{(q_N-s)}$ for the designated service queueing system (see Theorem 3.2) for $q_N \geq s$. Since $\mathbf{x}$ is a stochastic measure, (4.11) leads to $\boldsymbol{\pi}[q_N]\mathbf{e} \approx c\bar{\rho}_N^{(q_N-s)}$, if $q_N \to \infty$. We can use those relationships to compare queue N in the priority service queueing system and in the designated service queueing system. Some results are given in the following example.

Example 4.1 Consider a priority service queueing system with $\lambda_1 = 2$, $\lambda_2 = 1$, $\mu_1 = 5$, $\mu_2 = 3.5$, $\lambda_{T,1} = 0.5$, and $s = 1$. Simulation results indicate that $c/c_u \approx 0.41$. Thus, compared to the corresponding designated service queueing system, queue 2 becomes overloaded significantly less frequently in the priority service queueing system, which is a direct consequence of the fast service in queue 1 ($\mu_1 = 5 > \mu_2 = 3.5$). If the service rate at queue 1 is changed from $\mu_1 = 5$ to $\mu_1 = 0.1$, simulation results indicate that $c/c_u \approx 1.16$. Since the service in queue 1 is too slow ($\mu_1 = 0.1 < \mu_2 = 3.5$), queue 1 in the priority service queueing system actually holds customers longer in the system. Thus, queue 2 becomes overloaded more frequently in the priority service queueing system.

It is well known that the transitions at boundary states have significant influence on the decay rate (e.g., see [16]). Theorem 4.3 shows that, for the priority service queueing system, services at lower priority queues have no influence on the decay rate. However, Example 4.1 indicates that services at lower priority queues still have influence on the tail asymptotics through the constant c.

5 An extension

In this section, we consider a priority service queueing system with a Markov arrival process for type N customers. We shall keep and use most of the notation defined in Sects. 2 to 4 and explain any change in notation whenever it occurs.

In the priority service queueing system defined in Sect. 2, while keeping all other assumptions the same, we change the arrival process of type N customers from a Poisson process to a Markov arrival process (*MAP*). Type N customers arrive from outside of the system according to a *MAP* with matrix representation (C_0, C_1), where C_0 and C_2 are $m \times m$ matrices, m is a positive integer, C_1 has negative diagonal elements and nonnegative off-diagonal elements, and C_1 is a nonnegative matrix. The *MAP* of type N customers is independent of the arrival processes of other types of customers. Define $C = C_0 + C_1$. The matrix C is the infinitesimal generator of the underlying continuous time Markov chain of the arrival process. We assume that the underlying Markov chain is irreducible, which implies that the matrix C is irreducible. Denote by $I(t)$ the state of the underlying Markov chain at time t; $\boldsymbol{\alpha}$ the stationary distribution of $\{I(t), t \geq 0\}$ (i.e., $\boldsymbol{\alpha}$ satisfies $\boldsymbol{\alpha}C = 0$, $\boldsymbol{\alpha} \geq 0$, and $\boldsymbol{\alpha}\mathbf{e} = 1$). In fact, the elements of vector $\boldsymbol{\alpha}$ are all positive since C is irreducible. The (average) arrival rate of type N customers from outside is given by $\lambda_N = \boldsymbol{\alpha}C_1\mathbf{e}$. We refer readers to Neuts [23] and Lucantoni [19] for more about *MAP*.

The state of the extended priority service queueing system can be represented by $(q_1(t), q_2(t), \ldots, q_N(t), I(t))$, where $(q_1(t), q_2(t), \ldots, q_N(t))$ is defined in Sect. 2. Combining the proof of Theorem 1 in He et al. [14] and the proof of Theorem 2.1 in this paper, if system parameters $\{\lambda_1, \ldots, \lambda_{N-1}, \mu_1, \ldots, \mu_N, \lambda_{T,1}, \ldots, \lambda_{T,N-1}, s\}$ are positive and C is irreducible, it can be shown that the Markov chain $\{(q_1(t), q_2(t), \ldots, q_N(t), I(t)), t \geq 0\}$ is (1) ergodic if $\sum_{k=1}^{N} \lambda_k < s\mu_N$, and (2) non-ergodic if $\sum_{k=1}^{N} \lambda_k > s\mu_N$. Therefore, we assume $\sum_{k=1}^{N} \lambda_k < s\mu_N$ so that the extended priority service queueing system is stable.

The designated service queueing system defined in Sect. 3 can be extended accordingly. For the extended queueing system, Lemma 3.1 continues to hold. The sample path relationships between the queue lengths in the two queueing systems (Lemma 3.3) hold as well. For the stationary distributions of queue lengths, we need the following results.

Consider a continuous time Markov chain with an infinitesimal generator given as

$$\begin{pmatrix} C_0 - \bar{\lambda}_{(N-1)} I & C_1 + \bar{\lambda}_{(N-1)} I & & & & \\ \mu_N I & C_0 - (\mu_N + \bar{\lambda}_{(N-1)}) I & C_1 + \bar{\lambda}_{(N-1)} I & & & \\ & \ddots & \ddots & \ddots & & \\ & & s\mu_N I & C_0 - (s\mu_N + \bar{\lambda}_{(N-1)}) I & C_1 + \bar{\lambda}_{(N-1)} I & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{5.1}$$

where $\bar{\lambda}_{(N-1)} = \bar{\lambda}_N^* - \lambda_N$. It is readily seen that the Markov chain is a quasi birth-and-death (QBD) process. Denote by $(\boldsymbol{\psi}_u(0), \boldsymbol{\psi}_u(1), \ldots)$ the stationary distribution of the Markov chain, where $\boldsymbol{\psi}_u(n)$ is a vector of size m for $n \geq 0$.

Lemma 5.1 *Assume that C is irreducible and system parameters $\{\mu_N, \bar{\lambda}_{(N-1)}, s\}$ are positive. Under the condition $\sum_{k=1}^{N} \lambda_k < s\mu_N$, the stationary distribution $(\boldsymbol{\psi}_u(0), \boldsymbol{\psi}_u(1), \ldots)$ exists and is given as follows*:

$$\begin{aligned}
&\boldsymbol{\psi}_u(n) = \boldsymbol{\psi}_u(s) R_u^{n-s}, \quad \text{for } n \geq s;\\
&\boldsymbol{\psi}_u(0)(C_0 - \bar{\lambda}_{(N-1)} I) + \mu_N \boldsymbol{\psi}_u(1) = 0;\\
&\boldsymbol{\psi}_u(0)(C_1 + \bar{\lambda}_{(N-1)} I) + \boldsymbol{\psi}_u(1)(C_0 - \bar{\lambda}_{(N-1)} I + \mu_N I) + 2\mu_N \boldsymbol{\psi}_u(2) = 0;\\
&\vdots\\
&\boldsymbol{\psi}_u(s-1)(C_1 + \bar{\lambda}_{(N-1)} I) + \boldsymbol{\psi}_u(s)(C_0 - \bar{\lambda}_{(N-1)} I - s\mu_N I) + s\mu_N \boldsymbol{\psi}_u(s) R_u = 0;\\
&\big(\boldsymbol{\psi}_u(0) + \cdots + \boldsymbol{\psi}_u(s-1)\big)\mathbf{e} + \boldsymbol{\psi}_u(s)(I - R_u)^{-1}\mathbf{e} = 1,
\end{aligned} \tag{5.2}$$

where R_u is the minimal nonnegative solution to the following matrix equation:

$$C_1 + \bar{\lambda}_{(N-1)} I + R_u(C_0 - \bar{\lambda}_{(N-1)} I - s\mu_N I) + s\mu_N R_u^2 = 0. \tag{5.3}$$

Proof Under the condition $\sum_{k=1}^{N} \lambda_k < s\mu_N$, ergodicity of the Markov chain is obtained by applying Neuts' condition [24]. The stationary distribution is the standard matrix geometric solution for QBD process [24]. This completes the proof of Lemma 5.1. □

Remark 5.1 We would like to point out that the Markov chain defined by (5.1) is associated with the queue length process of a $MAP/M/s$ queue, where the arrival process is a MAP with matrix representation $(C_0 - \bar{\lambda}_{(N-1)}I, C_1 + \bar{\lambda}_{(N-1)}I)$, service times are exponentially distributed with parameter μ_N, and there are s servers.

For the extended designated service queueing system, denote by $\{\boldsymbol{\pi}_u(q_1, q_2, \ldots, q_N), q_1 \geq 0, q_2 \geq 0, \ldots, q_N \geq 0\}$ the stationary distribution of queue lengths and the state of the underlying Markov chain $\{I(t), t \geq 0\}$, where $\boldsymbol{\pi}_u(q_1, q_2, \ldots, q_N) = (\pi_u(q_1, q_2, \ldots, q_N, 1), \ldots, \pi_u(q_1, q_2, \ldots, q_N, m))$ is a row vector of size m. Then, it can be shown that (3.4) is generalized to

$$\boldsymbol{\pi}_u(q_1, q_2, \ldots, q_N) = \begin{cases} (\prod_{j=1}^{N-1}(\frac{\exp\{-\bar{\rho}_{T,j}\}\bar{\rho}_{T,j}^{q_j}}{q_j!}))\boldsymbol{\psi}_u[q_N], & 0 \leq q_N < s; \\ (\prod_{j=1}^{N-1}(\frac{\exp\{-\bar{\rho}_{T,j}\}\bar{\rho}_{T,j}^{q_j}}{q_j!}))\boldsymbol{\psi}_u[s]R_u^{q_N - s}, & q_N \geq s. \end{cases} \tag{5.4}$$

Equation (5.4) shows that the marginal distribution of queue N in the extended designated service queueing system is the same as that of the $MAP/M/s$ queue defined in Remark 5.1.

For the extended priority service queueing system, denote by $\{\boldsymbol{\pi}(q_1, q_2, \ldots, q_N), q_1 \geq 0, q_2 \geq 0, \ldots, q_N \geq 0\}$ the stationary distribution of the queue lengths and the underlying process $I(t)$, where $\boldsymbol{\pi}(q_1, q_2, \ldots, q_N) = (\pi(q_1, q_2, \ldots, q_N, 1), \ldots, \pi(q_1, q_2, \ldots, q_N, m))$ is a vector of size m. Using the relationship given in (3.11), an upper bound on $\sum_{\mathbf{w} \geq \mathbf{q}} \boldsymbol{\pi}(\mathbf{w})\mathbf{e}$ can be found as follows:

$$\sum_{\mathbf{w} \geq \mathbf{q}} \boldsymbol{\pi}(\mathbf{w})\mathbf{e} \leq \left(\prod_{j=1}^{N-1}\left(\frac{\bar{\rho}_{T,j}^{(q_j - js)^+}}{(q_j - js)^+!}\right)\right)\boldsymbol{\psi}_u[s]R_u^{q_N - s(N+1)}(I - R_u)^{-1}\mathbf{e},$$
$$\text{for } q_N \geq (N+1)s. \tag{5.5}$$

For tail asymptotics of the stationary distribution of queue lengths, Theorem 4.3 can be generalized in a straightforward manner. For brevity, we only present changes in notation and results. We begin with the transition blocks in the infinitesimal generator Q given in (4.1). For Markov chain $\{(q_1(t), q_2(t), \ldots, q_N(t), I(t)), t > 0\}$, the transition blocks in Q are changed as follows:

(1) Q_{-1} is replaced by $Q_{-1} \otimes I$
(2) Q_0 is replaced by $Q_0 \otimes I - \lambda_N I \otimes I + I \otimes C_0$
(3) Q_1 is replaced by $Q_1 \otimes I - \lambda_N I \otimes I + I \otimes C_1$

where "$\otimes$" stands for the Kronecker product of matrices. For the transformation of the infinitesimal generator Q to transition probability P in (4.4), we have

(4) D is replaced by $(D - \lambda_N I) \otimes I$
(5) D_1 is replaced by $(D_1 - \lambda_N I) \otimes I$
(6) D_2 is replaced by $(D_2 - \lambda_N I) \otimes I$.

Note that matrices Q_{-1}, Q_0, Q_1, D, D_1, and D_2 are defined in Sect. 4.

Denote by ρ_u the Perron–Frobenius eigenvalue (the eigenvalue with the largest modulus) of the matrix R_u (see (5.3)). Denote by $\boldsymbol{\beta}$ the normalized nonnegative eigenvector of R_u corresponding to eigenvalue ρ_u, i.e., $\boldsymbol{\beta}R_u = \rho_u\boldsymbol{\beta}, \boldsymbol{\beta} \geq 0$ and $\boldsymbol{\beta}\mathbf{e} = 1$ (see [26] for more about nonnegative matrices). It is well known that $\rho_u < 1$ if $\sum_{k=1}^{N}\lambda_k < s\mu_N$. Multiplying by $\boldsymbol{\beta}$ on both sides of (5.3), yields

$$\boldsymbol{\beta}\big(C_1 + \bar{\lambda}_{(N-1)}I + \rho_u(C_0 - \bar{\lambda}_{(N-1)}I - s\mu_N I) + s\mu_N\rho_u^2 I\big) = 0. \tag{5.6}$$

The matrix in (5.6) is an irreducible M-matrix. Thus, every element of $\boldsymbol{\beta}$ is positive. In addition, there exists a normalized nonnegative right eigenvector $\boldsymbol{\eta}$ satisfying

$$\big(C_1 + \bar{\lambda}_{(N-1)}I + \rho_u(C_0 - \bar{\lambda}_{(N-1)}I - s\mu_N I) + s\mu_N\rho_u^2 I\big)\boldsymbol{\eta} = 0. \tag{5.7}$$

Furthermore, the vector $\boldsymbol{\eta}$ can be so chosen that every element of $\boldsymbol{\eta}$ is positive, and $\boldsymbol{\eta}\mathbf{e} = 1$.

With positive vectors $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, $\{\mathbf{x}, \mathbf{y}, \bar{\rho}_N\}$ in Lemma 4.2 are changed as follows:

(7) $\mathbf{x}$ is replaced by $\mathbf{x} \otimes \boldsymbol{\beta}$
(8) $\mathbf{y}$ is replaced by $\mathbf{y} \otimes \boldsymbol{\eta}$ and
(9) $\bar{\rho}_N$ is replaced by ρ_u in (4.7).

Finally, Theorem 4.3 is generalized as follows.

Theorem 5.2 *Assume that system parameters* $\{\lambda_1, \ldots, \lambda_{N-1}, \mu_1, \ldots, \mu_N, \lambda_{T,1}, \ldots, \lambda_{T,N-1}, s\}$ *are positive and the underlying Markov process* $\{I(t), t \geq 0\}$ *is irreducible. If* $\sum_{k=1}^{N}\lambda_k < s\mu_N$, *then*

$$\lim_{q_N\to\infty}\frac{\boldsymbol{\pi}[q_N]}{\rho_u^{(q_N-s)}} = c\mathbf{x}\otimes\boldsymbol{\beta}, \quad (\textit{component-wise}) \tag{5.8}$$

where the elements of $\boldsymbol{\pi}[n]$ *are the stationary probabilities* $\{\pi(q_1, q_2, \ldots, q_{N-1}, n, i), q_1 \geq 0, q_2 \geq 0, \ldots, q_{N-1} \geq 0, 1 \leq i \leq m\}$ *ordered lexicographically and c is a constant. The decay rate* ρ_u *can be calculated by* $\rho_u = (\lambda_1 + \cdots + \lambda_{N-1} + \boldsymbol{\beta}C_1\mathbf{e})/(s\mu_N)$.

If $m = 1$ (i.e., the arrival process of type N customers is Poisson), we must have $\rho_u = \bar{\rho}_N$. For the general case, $\rho_u = \bar{\rho}_N$ may not hold. Intuitively, during the periods that queue N is overloaded, the stationary distribution of the underlying Markov chain $\{I(t), t \geq 0\}$ is changed from $\boldsymbol{\alpha}$ to $\boldsymbol{\beta}$, as indicated by the right-hand side of (5.8). Thus, the arrival rate of type N customers from outside is changed from $\lambda_N = \boldsymbol{\alpha}C_1\mathbf{e}$ to $\boldsymbol{\beta}C_1\mathbf{e}$. Consequently, the traffic intensity of queue N is changed from $\bar{\rho}_N$ to ρ_u.

Example 5.2 Consider a priority service queueing system with $N = 2$, $\lambda_1 = 1$, $\mu_2 = 4$, $s = 1$, and

$$C_0 = \begin{pmatrix} -10 & 1 \\ 0.1 & -0.1 \end{pmatrix}, \qquad C_1 = \begin{pmatrix} 9 & 0 \\ 0 & 0 \end{pmatrix}. \tag{5.9}$$

The arrival process of type 2 customers is a bursty process. By routine calculations, we obtain $\lambda_2 = 0.8182$, $\bar{\rho}_N = 0.4545$, $R_u = \begin{pmatrix} 0.8893 & 1.6107 \\ 0.0024 & 0.2476 \end{pmatrix}$, and $\rho_u = 0.8952$. Thus, ρ_u is significantly larger than $\bar{\rho}_N$, which is the decay rate if the arrivals of type 2 customers from outside follow a Poisson process. On the other hand, if

$$C_0 = \begin{pmatrix} -4 & 1 \\ 0.4 & -1 \end{pmatrix}, \qquad C_1 = \begin{pmatrix} 0 & 3 \\ 0 & 0.6 \end{pmatrix}, \tag{5.10}$$

we obtain $\lambda_2 = 0.8182$, $\bar{\rho}_N = 0.4545$, $R_u = \begin{pmatrix} 0.1689 & 0.8311 \\ 0.0222 & 0.3778 \end{pmatrix}$, and $\rho_u = 0.4447$. For this case, ρ_u is smaller than $\bar{\rho}_N$.

Example 5.2 indicates that the use of queueing models with Poisson inputs may either underestimate the decay rate ($\rho_u > \bar{\rho}_N$) or overestimate the decay rate ($\rho_u < \bar{\rho}_N$) significantly. For finite background state Markov chains associated with classical queueing models with versatile input processes, Neuts [25] offers an in-depth discussion on the decay rate of the stationary distribution.

6 Discussion

It is interesting to investigate how the queueing system behaves if a lower priority queue(s) is overloaded. Intuitively, if a lower priority queue is overloaded (e.g., $N = 2$), transfers to queue N become intensive and the distribution of queue N may be different from the (modified) geometric distribution for the $M/M/s$ queue. However, in the designated service queueing system, queue N has the same distribution as that of the $M/M/s$ queue if lower priority queues are overloaded. For the priority service queueing system, simulation results indicate that the tail asymptotics $\boldsymbol{\pi}(q_1, q_2) \approx cx(q_1)\bar{\rho}_2^{(q_2-s)}$ (if $N = 2$) may not hold if q_1 is large and q_2 is small (i.e., queue 1 is overloaded), which is different from the designated service queueing system. Finding the tail asymptotics for this case is not straightforward. Technically, the corresponding Markov chain Q, if q_1 is chosen as the level variable, becomes level-dependent after re-blocking. Thus, matrix-analytic methods and Markov additive method cannot be applied and another method has to be used for such cases. Tail asymptotics, if a lower priority queue is overloaded, is an interesting issue for future research.

Acknowledgements The authors would like to thank the associate editor and two anonymous referees for their valuable comments and suggestions. The research was supported by Natural Science Foundation of China under Grant 70532004 and partially supported by Natural Science Foundation of China under Grant 70871066.

References

1. Adan, I.J.B.F., Wessels, J., Zijm, W.H.M.: Analysis of the asymmetric shortest queue problem with threshold jockeying. Stoch. Models **7**, 615–628 (1991)
2. Argon, N.T., Ziya, S., Righer, R.: Scheduling impatient jobs in a clearing system with insights on patient triage in mass casualty incidents. Probab. Eng. Inf. Sci. **22**, 301–332 (2008)
3. Asmussen, S.: Applied Probability and Queues, 2nd edn. Springer, New York (2003)

4. Brandt, A., Brandt, M.: On a two-queue priority system with impatience and its application to a call center. Methodol. Comput. Appl. Probab. **1**, 191–210 (1999)
5. Chao, X., Miyazawa, M., Pinedo, M.: Queueing Networks, Customers, Signals and Production Form Solutions. Wiley, New York (1999)
6. Chen, M.F.: On three classical problems for Markov chains with continuous time parameters. J. Appl. Probab. **28**(2), 305–320 (1991)
7. Choi, B.D., Kim, B.: Non-ergodicity criteria for denumerable continuous time Markov processes. Oper. Res. Lett. **32**, 574–580 (2004)
8. Cohen, J.W.: The Single Server Queue. North-Holland, Amsterdam (1982)
9. Fayolle, G., Malyshev, V.A., Menshikov, M.V.: Topics in the Constructive Theory of Countable Markov Chains. Cambridge University Press, Cambridge (1995)
10. Foley, R.D., McDonald, D.R.: Join the shortest queue: stability and exact asymptotics. Ann. Appl. Probab. **11**, 569–607 (2001)
11. Foster, F.G.: On stochastic matrices associated with certain queueing processes. Ann. Math. Stat. **24**, 355–360 (1953)
12. Gomez-Corral, A., Krishnamoorthy, A., Narayanan, V.C.: The impact of self-generation of priorities on multi-server queues with finite capacity. Stoch. Models **21**, 427–447 (2005)
13. Hajek, B.: Optimal control of two interacting service stations. IEEE Trans. Automat. Contr. **AC-29**, 491–499 (1984)
14. He, Q.M., Xie, J.G., Zhao, X.B.: Stability conditions of a preemptive repeat priority $MMAP[N]/PH[N]/S$ queue with customer transfers (short version). In: ASMDA (Advanced Stochastic Models and Data Analysis) 2009 Conference Proceedings (accepted)
15. Jackson, J.R.: Networks of waiting lines. Oper. Res. **5**, 518–521 (1975)
16. Kroese, D.P., Scheinhardt, W.R.W., Taylor, P.G.: Spectral properties of the tandem Jackson network, seen as a quasi-birth-and-death process. Ann. Appl. Probab. **14**(4), 2057–2089 (2004)
17. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modelling. SIAM, Philadelphia (1999)
18. Li, H., Miyazawa, M., Zhao, Y.Q.: Geometric decay in a QBD process with countable background states with applications to a join-the-shortest-queue model. Stoch. Models **23**, 413–438 (2007)
19. Lucantoni, D.M.: New results on the single server queue with a batch Markovian arrival process. Stoch. Models **7**, 1–46 (1991)
20. Meyn, S.P., Tweedie, R.: Markov Chains and Stochastic Stability. Springer, Berlin (1993)
21. Miller, D.G.: Computation of steady-state probabilities for $M/M/1$ priority queues. Oper. Res. **29**, 945–958 (1981)
22. Miyazawa, M., Zhao, Y.Q.: The stationary tail asymptotics in the $GI/G/1$ type queue with countably many background states. Adv. Appl. Probab. **36**(4), 1231–1251 (2004)
23. Neuts, M.F.: A versatile Markovian point process. J. Appl. Probab. **16**, 764–779 (1979)
24. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models—An Algorithmic Approach. Johns Hopkins Press, Baltimore (1981)
25. Neuts, M.F.: The caudal characteristic curve of queues. Adv. Appl. Probab. **18**, 221–254 (1986)
26. Seneta, E.: Non-Negative Matrices: An Introduction to Theory and Applications. Wiley, New York (1973)
27. Stoyan, D.: Comparison Methods for Queues and Other Stochastic Models. Wiley, New York (1983)
28. Takagi, H.: Queueing Systems. Vacation and Priority Systems, vol. 1. Elsevier, Amsterdam (1991)
29. Takahashi, Y., Fujimoto, K., Makimoto, N.: Geometric decay of the steady-state probabilities in a quasi-birth-and-death number of phases. Stoch. Models **17**(1), 1–24 (2001)
30. Wang, Q.: Modeling and analysis of high risk patient queues. Eur. J. Oper. Res. **155**, 502–515 (2004)
31. Whitt, W.: Deciding which queue to join: some counterexamples. Oper. Res. **34**, 55–62 (1986)
32. Xie, J.G., He, Q.M., Zhao, X.B.: Stability of a priority queueing system with customer transfers. Oper. Res. Lett. **36**, 705–709 (2008)
33. Xie, J.G., He, Q.M., Zhao, X.B.: On the stationary distribution of the queue lengths in a multi-class priority queueing system with customer transfers. Working Paper #08-03, Department of Industrial Engineering, Dalhousie University (2008)
34. Xu, S.H., Chen, H.: On the asymptote of the optimal routing policy for two service stations. IEEE Trans. Automat. Contr. **38**, 187–189 (1990)
35. Xu, S.H., Zhao, Y.Q.: Dynamic routing and jockeying controls in a two-station queueing system. Adv. Appl. Probab. **28**, 1201–1226 (1996)
36. Zhao, Y., Grassmann, W.K.: Queueing analysis of a jockeying model. Oper. Res. **43**, 520–529 (1995)