Performance Analysis of Service Systems with Upgrade of Priorities

Abstract In this paper, we study the performance of service systems with priority upgrades. We model the service system as a single-server two-class priority queue, with queue 1 as the normal queue and queue 2 as the priority queue. The queueing model of interest has various applications in healthcare service, perishable inventory and project management. We give a comprehensive study on the system stationary distribution, computational algorithm design and sensitivity analysis. We observe that when queue 2 is large, the conditional distribution of queue 1 approximates a Poisson distribution. The tail probability of queue 2 decays geometrically, while the tail probability of queue 1 decays much faster than queue 2's. This helps us to design an algorithm to compute the stationary distribution. Finally, by using the algorithm, we do sensitivity analysis on various system parameters, i.e., the arrival rates, service rates and the upgrading rate. The numerical study provides helpful insights on designing such service systems.

Keywords: OR in service; priority upgrade; performance analysis; finite truncation

1. Introduction

We study a service system serving two types of customers: type-1 and type-2 customers. Type-2 customers have priorities over type-1 customers. That is, a type-1 customer is served only when there is no type-2 customer waiting in the queue. If a type-2 customer enters the system and finds that the server is busy serving a type-1 customer, then the serving type-1 customer would be pushed back to the queue and the server begins to serve this type-2 customer. The service of that type-1 customer will be resumed if the server is available for a type-1 customer. In addition, while waiting in the queue, the priority level of the type-1 customers could be upgraded. If this happens, a type-1 customer becomes a type-2 customer. The service time of a customer depends on the current class of this customer.

The system of interest can find many applications, such as call center operations, perishable inventory control and healthcare services (Down & Lewis, 2010; Deniz et al., 2010; Akan et al., 2011; Wang, 2004). For example, in a call center, customers can access the service either by phone or email. The customers requiring service by email have lower priorities than those requiring immediate service by phone. However, a customer waiting for email reply will become impatient and call the service center, leading to a change of this customer's service type. Another example is that, in an emergency medical system, patients are categorized into critical and non-critical groups. The condition of a patient in the non-critical group may deteriorate while waiting, and become critical. This patient will then be transferred to the critical group. The distinguishing feature of such systems is that low priority customers may upgrade their priorities and transfer from their current class to the more important class. To better design such service systems, we have to carefully model the system and analyze system performances accurately and efficiently.

In this paper, we model the service system of interest as a single-server two-class queueing model, where low priority customers may be upgraded to the high priority class after they have been in queue for some time. The randomness of upgrading time is captured by an exponential random variable. We focus on performance analysis of such systems, and provide a computation algorithm such that system performance measurements (e.g., system delays, proportion of upgrades) can be computed when parameters (i.e. arrival rate, service rate etc.) are given. To achieve that, we make effort to study the system stationary distribution, which is the fundamental element of system performance.

Our study is closely related to queueing systems with dynamic priorities and queueing systems with customer transfers (e.g., Gómez-Corral et al., 2005; He & Neuts, 2002; He et al. 2012; Maertens et al., 2006; Wang, 2004; Xie et al., 2008, 2009). Different from these existing papers, we are interested in the asymptotics and computational study of system stationary distribution (see e.g. Phung-Duc and Kawanishi, 2014). We showed that the stationary distribution has an asymptotic product-form solution. Furthermore, we found that the tail probability of the stationary distribution of the high priority queue decays exactly geometrically, while the tail probability of the stationary distribution of the low priority queue decays faster than any geometric distribution. Based on this result, we truncated the capacity of low priority queue and designed an algorithm to calculate the steady-state probability (Bini et al., 2012). Finally, we analyzed the impact of system parameters on the average queue lengths (AQLs). We observed that improvement of service rate for both types of customers can reduce system delay (queue length) for both types of customers. Another interesting observation is that the AQL of the low priority customer is not monotonic decreasing with the transfer rate. This implies that it does not always help the system effectiveness when promoting the upgrades.

The contribution of this paper is mainly twofold. First, the service systems of interest are common in the industry, and the performance analysis can help better design such systems. For example, if we know the tail decay rate of the queue, then we can design the proper buffer size. If we know the sensitivity of system delay on all system parameters, then we know how to change or control the system parameters to reduce system delay. Second, for the theoretical aspect, we are among those few papers that discuss the computation of two-dimension queueing systems by using the finite truncation and the matrix-analytic method. The discussion of convergence of finite truncation may be useful and helpful in analyzing other systems. The designed numerical algorithm may also be useful in other problems.

This article is organized as follows. In Section 2, the queueing model and its continuous-time Markov chain (CTMC) representation is introduced. We study the asymptotics of the tails of the stationary distributions of both queues in Section 3. In Section 4, a finite truncation algorithm is designed to calculate the steady-state probability. In Section 5, we analyze the impact of system parameters on the AQLs. Conclusions are made in Section 6.

2. Queueing model

The queueing model of interest consists of a single server serving two types of customers: type-1 and type-2 customers, which form two queues: queue 1 and queue 2, respectively. Type-1 and 2 customers arrive to the system according to two independent Poisson processes with parameter λ_1 and λ_2, respectively. The service times of type-1 and 2 customers are exponentially distributed with parameters μ_1 and μ_2, respectively. The arrival processes and service times are mutually independent. Moreover, the type-2 customers have higher service priority that the server serves the type-1 customer only when there is no type-2 customer in the system. If a type-2 customer arrives when the server is serving a type-1 customer, the type-1 customer is pushed back to queue 1 and the server begins to serve the type-2 customer. The service of this type-1 customer will be resumed if the server is available to serve type-1 customers. Due to the memoryless property of the exponential distribution, the service time of this type-1 customer is the same as other type-1 customers. Furthermore, while waiting in queue, a type-1 customer may upgrade to a type-2 customer after an exponential time with parameter λ_T.

Define $q_j(t)$ as the number of type j customers in system at time t, which consists of those in service and those waiting to be served, j = 1, 2. A CTMC can be defined by $\{(q_2(t), q_1(t)), t \geq 0\}$ with a state space $\{(q_2, q_1), q_2 \geq 0, q_1 \geq 0\}$. It is noticeable that if $q_2 \neq 0$, the server is serving a type-2 customer, while if $q_2 = 0$ and $q_1 \neq 0$, the server is serving a type-1 customer. If all system parameters are positive, it is easy to see that this Markov chain is irreducible. As it will be shown later, using (q_2, q_1) rather than (q_1, q_2) as the state can simplify the vector representation and facilitate readability. Denote by Q the infinitesimal generator of the Markov chain. Then we have, for $(q_2, q_1) \neq (y_2, y_1)$,

$$Q_{(q_2,q_1)(y_2,y_1)} = \begin{cases} \lambda_1, & \text{if } y_1 = q_1+1, y_2 = q_2; \\ \lambda_2, & \text{if } y_1 = q_1, y_2 = q_2+1; \\ \mu_1, & \text{if } y_1 = q_1 - 1 \geq 0, y_2 = q_2 = 0; \\ \mu_2, & \text{if } y_1 = q_1, y_2 = q_2 - 1 \geq 0; \\ q_1\lambda_T, & \text{if } y_1 = q_1 - 1 \geq 0, y_2 = q_2 + 1 > 1; \\ (q_1-1)\lambda_T, & \text{if } y_1 = q_1 - 1 > 0, y_2 = q_2 + 1 = 1; \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

We say that the system is stable if the Markov chain $\{(q_2(t), q_1(t)), t \geq 0\}$ is ergodic (irreducible and positive recurrent). Define $\rho = (\lambda_1+\lambda_2)/\mu_2$. It has been shown that the Markov chain $\{(q_2(t), q_1(t)), t \geq 0\}$ is ergodic if $\rho < 1$ (Xie et al., 2008). For such a system, we are interested in its stationary distribution and performance measures.

3. Stationary distribution

Assuming that the CTMC $\{(q_2(t), q_1(t)), t \geq 0\}$ is ergodic, denote by $\boldsymbol{\pi} = (\boldsymbol{\pi}(q_2,q_1))$ its stationary distribution (i.e. $\boldsymbol{\pi} Q = 0$). Let $\boldsymbol{\pi}_n = (\boldsymbol{\pi}(n,0), \boldsymbol{\pi}(n,1), \ldots)$, for $n \geq 0$. Then, we have $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots)$. Ordering Q lexicographically, it has the quasi-birth-death (QBD) form

$$Q = \begin{pmatrix} C_0 & C_1 & & & \\ Q_{-1} & Q_0 & Q_1 & & \\ & Q_{-1} & Q_0 & Q_1 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{3.1}$$

where

$$C_0 = \begin{pmatrix} \Sigma_0^C & \lambda_1 & & & \\ \mu_1 & \Sigma_1^C & \lambda_1 & & \\ & \mu_1 & \Sigma_2^C & \lambda_1 & \\ & & \mu_1 & \Sigma_3^C & \lambda_1 \\ & & & \ddots & \ddots & \ddots \end{pmatrix}, C_1 = \begin{pmatrix} \lambda_2 & & & & \\ & \lambda_2 & & & \\ & \lambda_T & \lambda_2 & & \\ & & 2\lambda_T & \lambda_2 & \\ & & & \ddots & \ddots \end{pmatrix}, \tag{3.2}$$

$$Q_{-1} = \begin{pmatrix} \mu_2 & & & \\ & \mu_2 & & \\ & & \mu_2 & \\ & & & \ddots \end{pmatrix}, Q_0 = \begin{pmatrix} \Sigma_0^Q & \lambda_1 & & \\ & \Sigma_1^Q & \lambda_1 & \\ & & \Sigma_2^Q & \lambda_1 \\ & & & \ddots & \ddots \end{pmatrix}, Q_1 = \begin{pmatrix} \lambda_2 & & & \\ \lambda_T & \lambda_2 & & \\ & 2\lambda_T & \lambda_2 & \\ & & \ddots & \ddots \end{pmatrix}, \tag{3.3}$$

and

$$\begin{cases} \Sigma_0^C = -\lambda_1 - \lambda_2, \\ \Sigma_i^C = -\lambda_1 - \lambda_2 - \mu_1 - (i-1)\lambda_T, \quad i = 1, 2, \ldots, \end{cases} \tag{3.4}$$

$$\Sigma_i^Q = -\lambda_1 - \lambda_2 - \mu_2 - i\lambda_T, \quad i = 0, 1, \ldots. \tag{3.5}$$

The asymptotic solution of stationary distribution $\boldsymbol{\pi}_n$ is given in Theorem 3.1 (see Appendix A for the proof).

Theorem 3.1 Assume that all system parameters $\{\lambda_1, \lambda_2, \mu_1, \mu_2, \lambda_T\}$ are positive and the system is stable (i.e. $\rho < 1$), we have

$$\lim_{n\to\infty} \rho^{-n} \boldsymbol{\pi}_n = \alpha \mathbf{c}, \tag{3.6}$$

where α is a positive constant, and $\mathbf{c} = (c_0, c_1, \ldots)$ is a probability vector of a Poisson distribution with parameter λ_1/λ_T, where

$$c_i = \frac{(\lambda_1/\lambda_T)^i \exp(-\lambda_1/\lambda_T)}{i!}, \quad i = 0, 1, \ldots. \tag{3.7}$$

Theorem 3.1 indicates that, for large enough n, $\boldsymbol{\pi}_n$ has a product-form asymptotic solution $\boldsymbol{\pi}_n \approx \alpha \mathbf{c} \rho^n$, which is a product of a vector of Poisson distribution with parameter λ_1/λ_T and the kernel of a geometric distribution with parameter $1-\rho$.

From theorem 3.1, we also see that the tail probability of the stationary distribution of queue 2 (i.e. $\boldsymbol{\pi}_n \mathbf{e}$, where $\mathbf{e}$ is a column vector of all ones) decreases geometrically with rate ρ. Fig. 1 displays an example of this decay, where the parameters are $\{\lambda_1, \lambda_2, \mu_1, \mu_2, \lambda_T\} = \{8, 2, 10, 10.5, 0.5\}$. On the other hand, the conditional distribution of queue 1 given queue 2 converges to a Poisson distribution. We use the example above to demonstrate this convergence in Fig. 2. In the following, we will show that the marginal distribution of the queue 1 decays faster than any geometric distribution.

To study the tail asymptotic distribution of queue 1, we design two auxiliary queues. Note that there are two possible scenarios for the first customer (if there is one) in queue 1. If there is no type-2 customer in the system, this type-1 customer is in service mode; otherwise, he or she is in transfer mode. Let $s_1 = \max\{\mu_1,\lambda_T\}$ and $s_2 = \min\{\mu_1,\lambda_T\}$. Thus, $s_1 \geq s_2$. Design two modified queues which are the same as queue 1 except that their first customers in queue are always in service mode with service rate s_1 and s_2, respectively. The modified queues are both birth and death processes. Denote by $\boldsymbol{\eta}_i$ the stationary distribution of modified queue with service rate s_i (i =1, 2), then we have

$$\begin{cases} \boldsymbol{\eta}_i(k) = \boldsymbol{\pi}_i(0)\lambda_1^k \prod_{i=0}^{k-1} \frac{1}{s_i + i\lambda_T}, k \geq 1 \\ \boldsymbol{\eta}_i(0) = \left(1 + \sum_{k=1}^{\infty} \lambda_1^k \prod_{i=0}^{k-1} \frac{1}{s_i + i\lambda_T}\right)^{-1} \end{cases} \tag{3.8}$$

If $\mu_1 < \lambda_T$, then $s_1 = \lambda_T$. Thus, the modified queue with service rate s_1 has a Poisson distribution on its queue length; If $\mu_1 > \lambda_T$, then $s_2 = \lambda_T$. Thus, the modified queue with service rate s_2 has a Poisson distribution on its queue length; If $\mu_1 = \lambda_T$, then both modified queues have Poisson distributions on their queue lengths. Therefore, at least one of the modified queues has a Poisson distribution on its queue length, which is the stationary distribution of the queue length in an $M/M/\infty$ queue with arrival rate λ_1 and service rate λ_T. From Eq.(3.8), it is easy to see that $\boldsymbol{\eta}_1(0) \geq \boldsymbol{\eta}_2(0)$, where equality holds when $\mu_1 = \lambda_T$. Moreover, we have the following results (see Appendix B for the corresponding proofs).

Theorem 3.2 For large enough k, $\boldsymbol{\eta}_1(k) \leq \boldsymbol{\eta}_2(k)$; and both $\boldsymbol{\eta}_1(k)$ and $\boldsymbol{\eta}_2(k)$ approach to 0 faster than any geometric decay.

Denote by $L_i(t)$ the number of customers in modified queue with service rate s_i, and $N_1(t)$ the number of customers in queue 1, at time $t \geq 0$. Then we have the following stochastic order relationships.

Lemma 3.1 Assuming that all systems are empty initially, we have

$$L_1(t) \leq_{st} N_1(t) \leq_{st} L_2(t) . \tag{3.9}$$

Assume all systems are stable, let $L_1 = L_1(\infty)$, $L_2 = L_2(\infty)$ and $N_1 = N_1(\infty)$. Then we have $L_1 \leq_{st} N_1 \leq_{st} L_2$, by taking t to infinity, where the "$\leq_{st}$" stands for "stochastically less" which is a stochastic order. The bounds of tail distribution of queue 1 are given as follows.

Theorem 3.3 Assume that all system parameters $\{\lambda_1, \lambda_2, \mu_1, \mu_2, \lambda_T\}$ are positive. If the system is stable (i.e. $\rho < 1$), we have

$$\sum_{k=n+1}^{\infty} \boldsymbol{\eta}_1(k) \leq \sum_{k=n+1}^{\infty} \boldsymbol{\pi}(\cdot,k) \leq \sum_{k=n+1}^{\infty} \boldsymbol{\eta}_2(k), \ n \geq 0. \tag{3.10}$$

In addition, given any $\gamma > 0$, there exists k^*, such that $(1-\gamma)\boldsymbol{\eta}_1(k) \leq \boldsymbol{\pi}(\cdot,k) \leq (1+\gamma)\boldsymbol{\eta}_2(k)$, for $k > k^*$.

From Theorem 3.2, we know that for large enough k, $\boldsymbol{\eta}_1(k) \leq \boldsymbol{\eta}_2(k)$, and both $\boldsymbol{\eta}_1(k)$ and $\boldsymbol{\eta}_2(k)$ approach to 0 faster than any geometric decay. Theorem 3.3 shows that the tail probability of queue

1 is bounded by $\boldsymbol{\eta}_1(k)$ and $\boldsymbol{\eta}_2(k)$. Thus, the tail decay of queue 1 is faster than any geometric decay. Up to now, we have a quite clear picture of the stationary asymptotic distribution (see Fig. 3). In the direction of queue 2, the stationary distribution decays exactly geometrically, and has an asymptotic product-from solution. Given the length of queue 2, queue 1 has an asymptotic Poisson distribution. In the direction of queue 1, the stationary distribution decays faster than any geometric distribution. However, we are not clear about the exact distribution, and specially the boundary distribution. In the next section, we conduct a complete computational study on the stationary distribution.

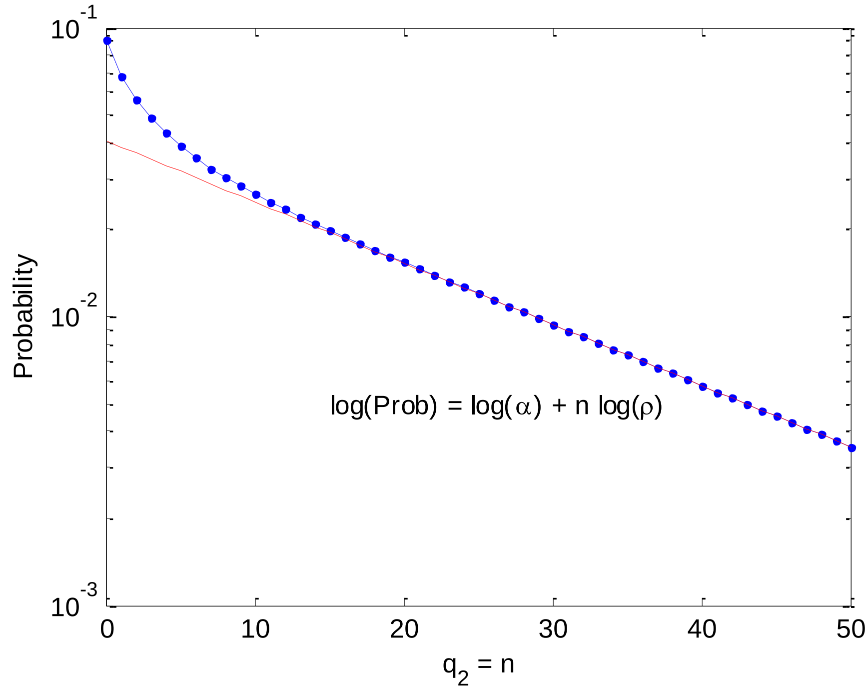

Fig. 1 Decay of the tail probability of the stationary distribution of queue 2

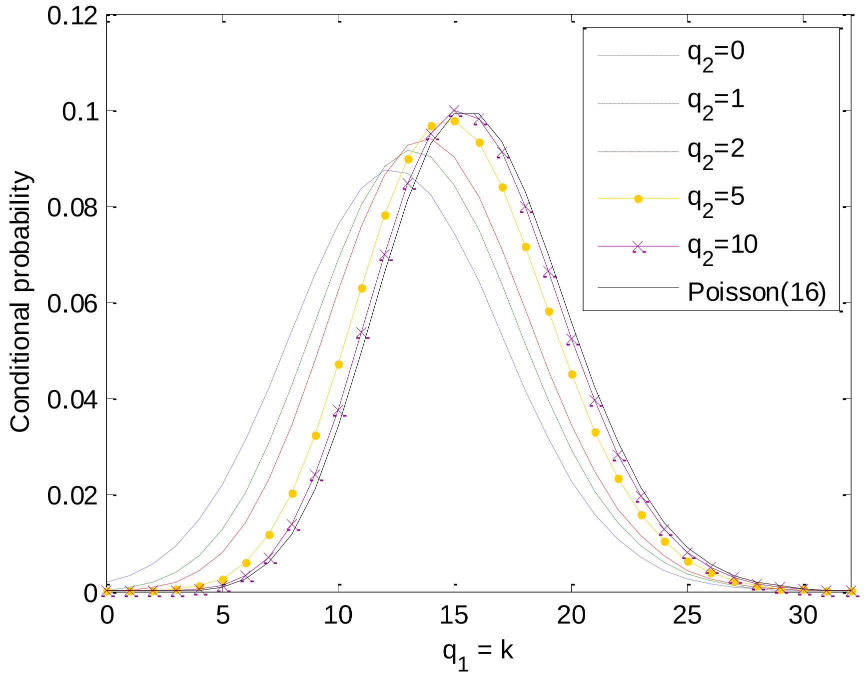

Fig. 2 Convergence of the conditional distribution of queue 1 given queue 2

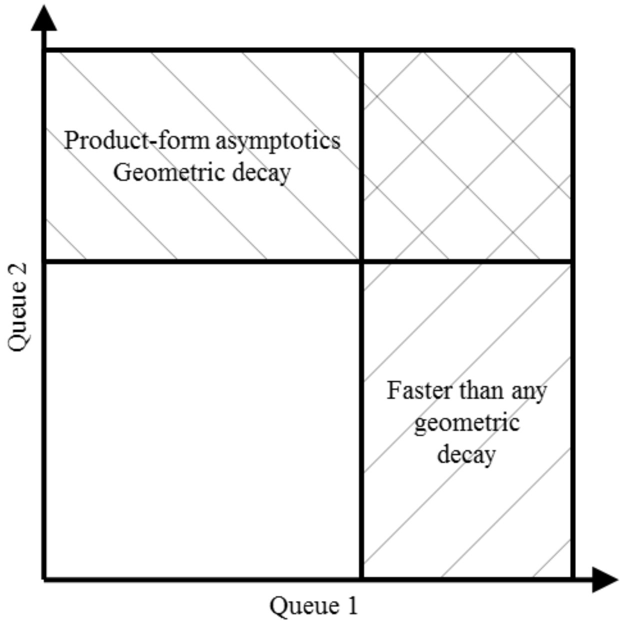

Fig. 3 Structure property of the stationary distribution

4. Computational study

In general, it is rather difficult to find the explicit stationary distribution of a queueing system with multiple types of customers. In order to compute various performance measures, we need to design an efficient algorithm. One intuitive way is to assume finite buffers for both queues, which is referred as *finite truncation*. In this case, the state space is finite and the steady-state probability can be calculated numerically. The remained question is to determine appropriate truncation sizes for both queues. If we truncate too much such that the queue buffer is very small, some queue lengths whose stationary probabilities are significant nonzero in the original system cannot be represented in the truncated system, leading to a big difference between these two systems. On the other hand, if we truncate too less such the queue buffer is very large, we still face the computational difficulty on multiple dimensions. The study of tail probability can help us to choose proper truncation sizes. If the tail decays (faster than) geometrically, then a sufficient large truncation size can achieve almost zero loss.

According to Neuts (1981), instead of truncating both queues, it is sufficient to truncate only one queue and apply the matrix-analytic method. The steady-state distribution of the other queue, which is not truncated, can be computed iteratively. As stated in Section 3, the tail probability of the stationary distribution of queue 1 decays much faster than that of queue 2, so it is better to truncate the capacity of queue 1 by a finite number K. We expect that the truncated model can approximate the original system well for a large K. Before we show this, let's elaborate the truncated model. The corresponding CTMC for the truncated model has a Q-matrix of QBD form as follows:

$$\bar{Q} = \begin{pmatrix} \bar{C}_0 & \bar{C}_1 & & & \\ \bar{Q}_{-1} & \bar{Q}_0 & \bar{Q}_1 & & \\ & \bar{Q}_{-1} & \bar{Q}_0 & \bar{Q}_1 & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \tag{4.1}$$

where

$$\bar{C}_0 = \begin{pmatrix} \bar{\Sigma}_0^C & \lambda_1 & & & \\ \mu_1 & \bar{\Sigma}_1^C & \lambda_1 & & \\ & \mu_1 & \bar{\Sigma}_2^C & \ddots & \\ & & \ddots & \ddots & \lambda_1 \\ & & & \mu_1 & \bar{\Sigma}_K^C \end{pmatrix}, \bar{C}_1 = \begin{pmatrix} \lambda_2 & & & & \\ & \lambda_2 & & & \\ & \lambda_T & \lambda_2 & & \\ & & \ddots & \ddots & \\ & & & (K-1)\lambda_T & \lambda_2 \end{pmatrix}, \tag{4.2}$$

$$\bar{Q}_{-1} = \begin{pmatrix} \mu_2 & & & \\ & \mu_2 & & \\ & & \ddots & \\ & & & \mu_2 \end{pmatrix}, \bar{Q}_0 = \begin{pmatrix} \bar{\Sigma}_0^Q & \lambda_1 & & \\ & \bar{\Sigma}_1^Q & \ddots & \\ & & \ddots & \lambda_1 \\ & & & \bar{\Sigma}_K^Q \end{pmatrix}, \bar{Q}_1 = \begin{pmatrix} \lambda_2 & & & \\ \lambda_T & \lambda_2 & & \\ & \ddots & \ddots & \\ & & K\lambda_T & \lambda_2 \end{pmatrix}, \tag{4.3}$$

and

$$\begin{cases} \bar{\Sigma}_0^C = -\lambda_1 - \lambda_2, \\ \bar{\Sigma}_i^C = -\lambda_1 - \lambda_2 - \mu_1 - (i-1)\lambda_T, \quad 1 \le i \le K-1, \\ \bar{\Sigma}_K^C = -\lambda_2 - \mu_1 - (K-1)\lambda_T, \end{cases} \tag{4.4}$$

$$\begin{cases} \bar{\Sigma}_i^Q = -\lambda_1 - \lambda_2 - \mu_2 - i\lambda_T, \quad 0 \le i \le K-1, \\ \bar{\Sigma}_K^Q = -\lambda_2 - \mu_2 - K\lambda_T. \end{cases} \tag{4.5}$$

The following theorem implies that the truncated model is stable if the original model is stable.

Theorem 4.1 Assume that system parameters $\{\lambda_1, \lambda_2, \mu_1, \mu_2, \lambda_T\}$ are positive, for any given $K > 0$, the truncated system is stable if $\lambda_1+\lambda_2<\mu_2$.

Proof: It is well known that the QBD process is stable if and only if (Neuts, 1981)

$$\boldsymbol{\pi}_A \bar{Q}_1 \mathbf{e} < \boldsymbol{\pi}_A \bar{Q}_{-1} \mathbf{e}\,, \tag{4.6}$$

where $\boldsymbol{\pi}_A$ is the steady-state probability vector of generator matrix $A = \bar{Q}_{-1} + \bar{Q}_0 + \bar{Q}_1$, and $\mathbf{e}$ is a K+1 dimensional column vector of all ones. By regular computation, Eq. (4.6) can be simplified as

$$\lambda_1(1-\pi_{A,K}) + \lambda_2 < \mu_2\,, \tag{4.7}$$

where

$$\pi_{A,i} = \frac{(\lambda_1/\lambda_T)^i}{i!}\left[\sum_{i=0}^{K}\frac{(\lambda_1/\lambda_T)^i}{i!}\right]^{-1}, \quad i = 0,1,\ldots,K. \tag{4.8}$$

Since $0 < \pi_{A,K} < 1$ for any $K > 0$, the system is stable if $\lambda_1 + \lambda_2 < \mu_2$. □

Denote by $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots)$ the stationary distribution of the truncated system, where $\boldsymbol{\theta}_i = (\theta_{i0}, \theta_{i1}, \ldots, \theta_{iK})$, $i \geq 0$. For the stationary distribution, we have the well-known geometric matrix form

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_1 G^{i-1}, \; i > 0, \tag{4.9}$$

where G is the minimal nonnegative solution of

$$\overline{Q}_1 + G\overline{Q}_0 + G^2\overline{Q}_{-1} = 0. \tag{4.10}$$

With the equilibrium equation $\boldsymbol{\theta}\bar{Q} = 0$ and normalization condition $\boldsymbol{\theta}\mathbf{e} = 1$, we can obtain the stationary distribution numerically (see the algorithm in Appendix C). For the truncated system, the AQLs for queue 1 and 2 are given respectively by

$$\begin{aligned} \text{AQL}_1 &= \sum_{i\geq 0, 0\leq j\leq K} j\theta_{ij}, \\ \text{AQL}_2 &= \sum_{i\geq 0, 0\leq j\leq K} i\theta_{ij}. \end{aligned} \tag{4.11}$$

When a type-1 customer arrives at the truncated system, he or she will not enter the system if queue 1 is full. By the Poisson arrivals see time averages (PASTA) property of Poisson arrivals (Wolf 1982), the loss probability that the type-1 customer cannot enter the truncated system equals to the probability that queue 1 is full. Specifically, this loss probability is

$$p_{loss}(K) = \sum\nolimits_{i\geq 0}\theta_{iK}. \tag{4.12}$$

This loss probability will become 0 when K tends to infinity, which indicating that the truncated model will approximate the original model well as K is large enough. We demonstrate this in Fig. 4 and 5 with the same example used in Fig. 1. Fig. 4 displays the decay of the loss probability as the truncation size K increases. In Fig. 5, we compute the AQLs of both queues and compare them with the simulated AQLs. The detail of the simulation will be listed later. We observed that the computed AQLs converge to the simulated AQLs as K increases. Furthermore, by comparing Fig. 4 with Fig. 5, we see that this convergency happens as the loss probability becomes small.

To further study the truncated system, the AQLs of both queues are computed by the matrix-analytic method for various combinations of parameters. These parameters include the arrival rates λ_1 and λ_2, the service rates μ_1 and μ_2, and the transfer rate λ_T. In order to make the results comparable, we keep the sum of the arrival rates fixed at 10. We set two levels for each parameter, leading to a total of 16 combinations, and summarize the detail of the levels in Table 1. For each of the 16 combinations, we apply the matrix-analytic method to compute the AQLs and compare them with the simulated AQLs.

One important issue in the matrix-analytic method is to appropriately determine the truncation size, K. From Fig. 2, we know that the conditional distribution of the queue length of queue 1 give the queue length of queue 2 converges to a Poisson distribution with parameter λ_1/λ_T. Based on this result we may make an initial guess by finding a K so that the cumulative probability of the Poisson distribution is close to 1. Specifically, this K can be the minimum value that satisfies $P(X \leq K) > 1 - \varepsilon$, where X is a Poisson random variable and ε is a predetermined error. Then we apply the matrix-analytic method with this K and compute the loss probability. If the loss probability is greater than a given tolerance δ, then we set $K = K + 1$ and continue this procedure; otherwise, we have found an appropriate K. In this study, we set both ε and δ at $2^{-52} \approx 2.22\times10^{-16}$, which is the relative accuracy of the double floating-point number. The initial guess K and the actual truncation size K^* are listed in Table 2. We can see that the initial guesses are greater but close to the actual values. This implies that with these initial guesses, we only need compute once without wasting much computational resources. Another notable point is that when the transfer rate λ_T is small, a larger K is required, as well as more computation time. The approximation method may not work well when λ_T is very small. However, these cases can be approximated by typical two-class priority queue without transfers between queues.

For the simulation study, we consider the original system without state space truncation. We generate 1,000,000 events, which include customer arrivals and departures for both queues and priority changes, and compute the transition matrix and the AQLs of both queues. To make this result comparable, we repeat this procedure 100 times for each combination of parameters. Table 2 reports the mean and standard deviation (in brackets) of these 100 AQLs. It can be seen that the computed AQLs do not significantly differ from the simulated AQLs. However, it takes about one hour to simulate the AQLs in Table 2, while it takes about 3 seconds to obtain the computational result.

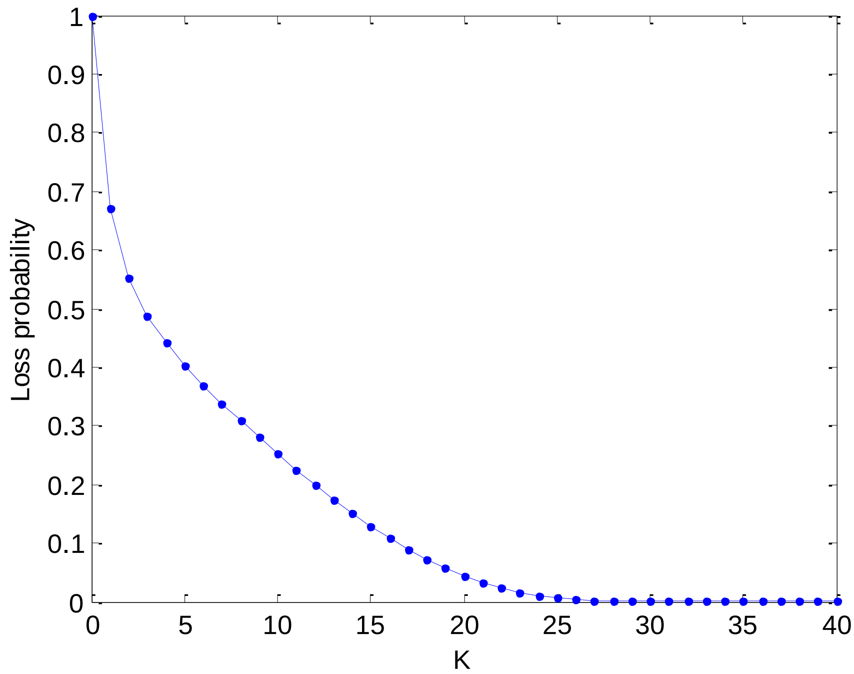

Fig. 4 Loss probability of the truncated model

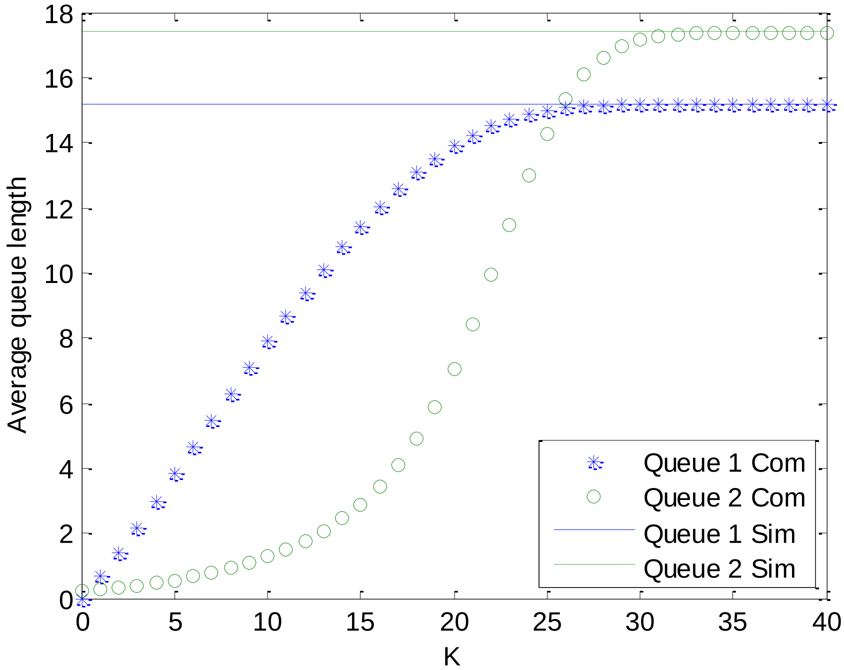

Fig. 5 Convergence of the finite truncation method on the AQLs

Table 1 Levels of parameters

Level	$\{\lambda_1, \lambda_2\}$	μ_1	μ_2	λ_T
Low	{8, 2}	5	10.5	0.5
High	{6, 4}	15	12.5	2.5

Table 2 Computation and simulation results on the AQLs

					Computation Result				**Simulation Result**	
λ_1	λ_2	μ_1	μ_2	λ_T	K	K^*	**Queue 1**	**Queue 2**	**Queue 1**	**Queue 2**
8	2	5	10.5	0.5	58	57	15.1848	17.3856	15.1786 (0.0068)	17.4406 (0.1383)
6	4	5	10.5	0.5	49	49	11.2004	17.4064	11.1970 (0.0063)	17.4381 (0.1625)
8	2	15	10.5	0.5	58	56	4.8715	3.5704	4.8841 (0.0228)	3.5790 (0.0768)
6	4	15	10.5	0.5	49	48	5.8161	7.8007	5.8492 (0.0190)	7.9919 (0.1045)
8	2	5	12.5	0.5	58	56	13.0155	2.5850	13.0200 (0.0063)	2.5860 (0.0057)
6	4	5	12.5	0.5	49	48	9.0925	2.6018	9.0906 (0.0053)	2.5981 (0.0052)
8	2	15	12.5	0.5	58	52	2.2834	0.3759	2.2830 (0.0027)	0.3751 (0.0010)
6	4	15	12.5	0.5	49	45	2.2783	0.8266	2.2780 (0.0029)	0.8257 (0.0017)
8	2	5	10.5	2.5	26	26	3.1286	19.1160	3.1300 (0.0011)	19.2198 (0.1285)
6	4	5	10.5	2.5	23	23	2.3393	19.2020	2.3401 (0.0008)	19.3655 (0.1211)
8	2	15	10.5	2.5	26	26	2.6133	15.3272	2.6109 (0.0027)	15.1671 (0.1386)
6	4	15	10.5	2.5	23	23	2.0578	16.8639	2.0569 (0.0020)	16.8152 (0.1243)
8	2	5	12.5	2.5	26	26	2.9314	3.3593	2.9321 (0.0009)	3.3647 (0.0063)
6	4	5	12.5	2.5	23	23	2.1691	3.4198	2.1680 (0.0008)	3.4132 (0.0056)
8	2	15	12.5	2.5	26	26	1.6960	1.7321	1.6961 (0.0012)	1.7325 (0.0047)
6	4	15	12.5	2.5	23	23	1.3850	2.2695	1.3835 (0.0009)	2.2684 (0.0049)

5. Performance analysis

In this section, we perform sensitivity analysis of the system parameters on the AQLs. In particular, we consider the case $\{\lambda_1, \lambda_2, \mu_1, \mu_2, \lambda_T\} = \{8, 2, 5, 10.5, 0.5\}$ and change the arrival rates $\{\lambda_1, \lambda_2\}$, the service rate of the type-1 customers μ_1, the service rate of the type-2 customers μ_2, and the transfer rate λ_T.

1) Sensitivity analysis of the arrival rates $\{\lambda_1, \lambda_2\}$

We study the effect of the arrival rates on the AQLs for λ_1 from 0 to 10 while keeping $\lambda_1 + \lambda_2 = 10$. Fig. 6 shows the AQLs of both queues against different values of λ_1. As can be seen, as λ_1 increases, AQL_1 increases but AQL_2 decreases. This result is because the proportion of two types of customers has been changed. However, AQL_1 may have different trends. In Fig 7, we represent another case where the parameters are $\{\mu_1, \mu_2, \lambda_T\} = \{15, 10.5, 0.5\}$. We observe that AQL_2 decreases constantly, but AQL_1 increases when $\lambda_1 < 6$ and decreases when $\lambda_1 > 6$.

2) Sensitivity analysis of the service rate of type-1 customers μ_1

We compute the AQLs for μ_1 from 0 to 30 and plot them in Fig. 8. According to Fig. 8, we can see that the AQLs of both queues decrease as μ_1 increases. This is not a surprising result for queue 1. A possible reason for the queue 2 could be that AQL_1 decreases when μ_1 increases, and hence less customers are transferred to queue 2. For the same reason, we can see that AQL_2 is similar to AQL_1 in Fig. 8. Moreover, we observed that the AQLs of both queues fall steeply for μ_1 between 5 and 15. This provides insights on designing service systems. For example, when designing the service rate of queue 1, we only need to consider μ_1 in a certain range (e.g. [5, 15]) instead of all possible values.

3) Sensitivity analysis of the service rate of type-2 customers μ_2

We study the effect of the service rate of type-2 customers on AQLs for μ_2 from 10.1 to 13. Fig. 9 demonstrates these AQLs. From Fig. 9, we observe that the AQLs decrease as μ_2 increases. This result is intuitive for queue 2. For queue 1, one explanation is that when μ_2 increases, the server has more time to serve the type-1 customers. We can also see that the impacts of μ_2 on both queues are different. AQL_2 falls dramatically in the beginning and then tends to be flat afterward, while AQL_1 has small but consistent decrease rate.

4) Sensitivity analysis of the transfer rate λ_T

We perform the sensitivity analysis of the transfer rate on the AQLs for λ_T from 0 to 3. Fig. 10 presents these AQLs. It can be seen that as λ_T increases, AQL_1 decreases but AQL_2 increases. We should notice that when λ_T equals zero, the system is not stable because $\lambda_1 = 8 < \mu_1 = 5$, which indicates that AQL_1 is infinity. When λ_T is small, AQL_1 is large, and a slight increment of λ_T could lead to a large amount of priority changes. Thus, we can see a steep fall of AQL_1 in Fig. 10. This example presents an interesting fact that even with a small transfer rate the system with priority changes could be very different from the system without priority changes.

However, AQL_1 does not decrease like that in Fig. 10. We consider another combination of parameters $\{\lambda_1, \lambda_2, \mu_1, \mu_2\} = \{8, 2, 15, 10.5\}$ and plot the AQLs in Fig. 11. As can be seen, AQL_1 increases when λ_T is small and decreases when $\lambda_T > 0.7$. The reason could be as follows. If λ_T equals zero, the system is stable and AQL_1 is finite. When $\lambda_T > 0$ is small, some type-1 customers are transferred to queue 2, so the server spends less time serving the type-1 customers because there are more type-2 customers. As a result, the AQLs of both queue increase. But this trend would change as λ_T becomes larger. If AQL_1 and λ_T are large, many type-1 customers would be transferred to the queue 2. Eventually, the rate of this transformation (consider the queue length of the queue 1 and λ_T) would exceed the arrival rate of the type-1 customer. Therefore, AQL_1 would decrease if λ_T is large.

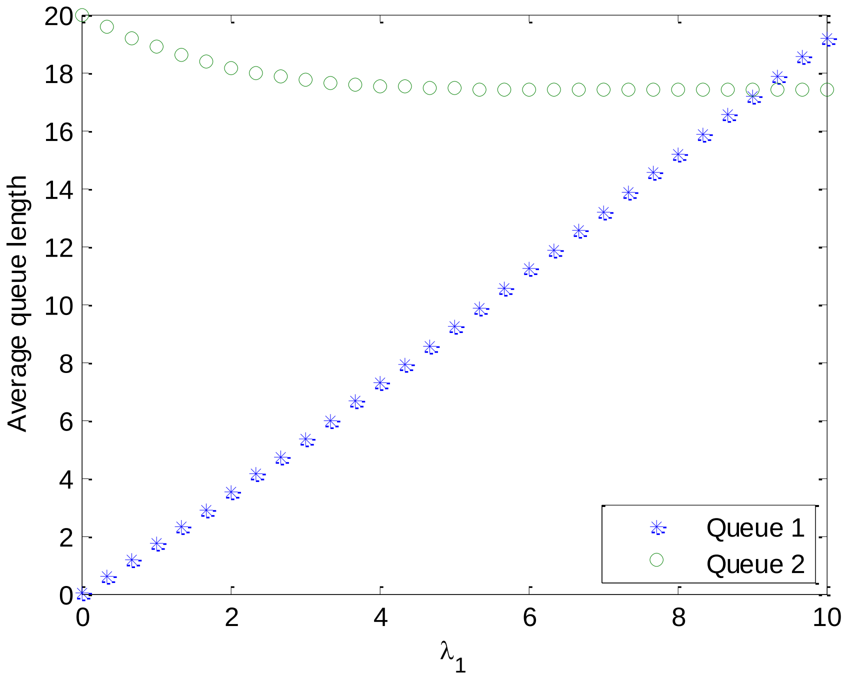

Fig. 6 AQLs versus the arrival rate of the low priority customer

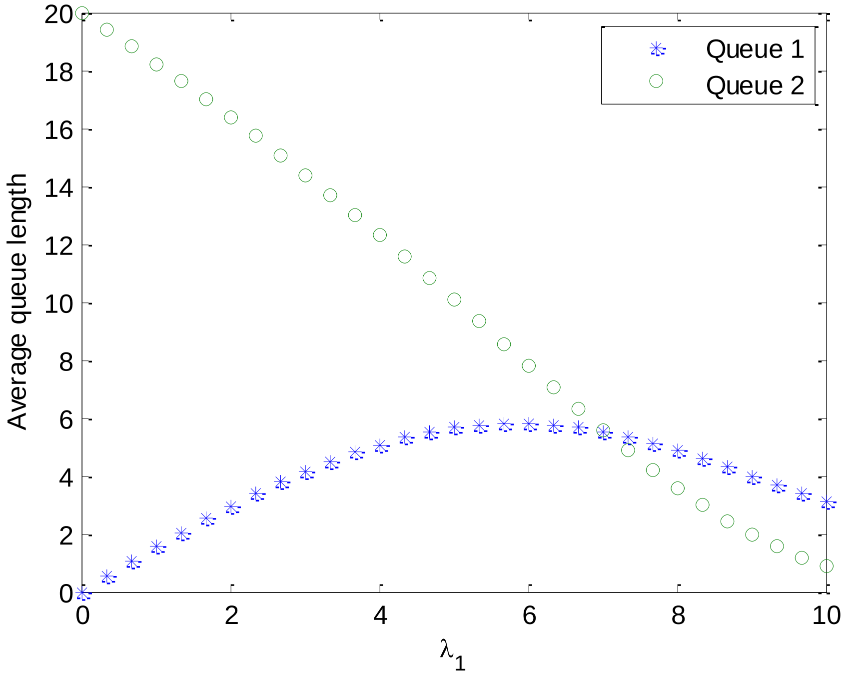

Fig. 7 AQLs versus the arrival rate of the low priority customer with $\{\mu_1, \mu_2, \lambda_T\} = \{15, 10.5, 0.5\}$

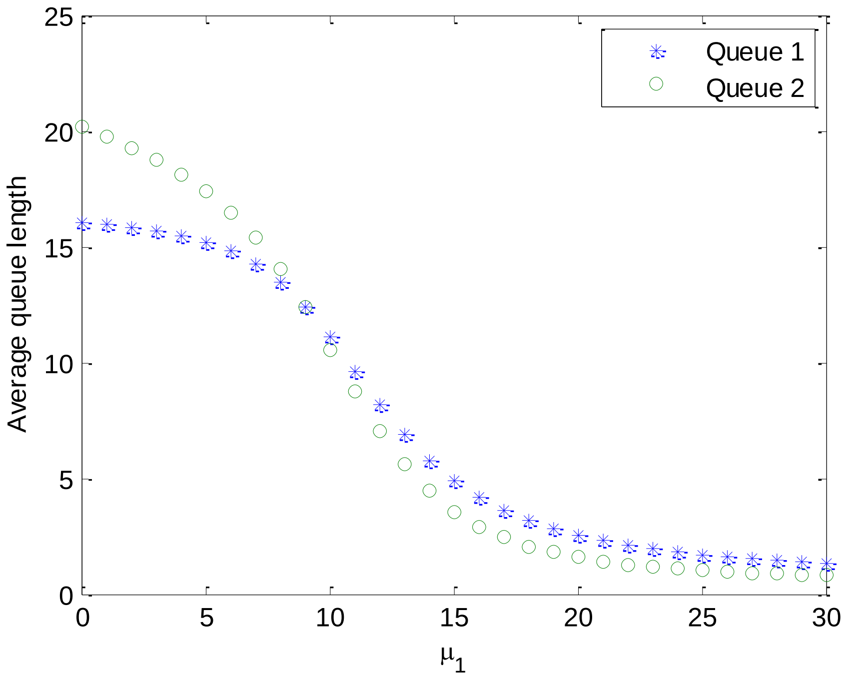

Fig. 8 AQLs versus the service rate of the type-1 customers

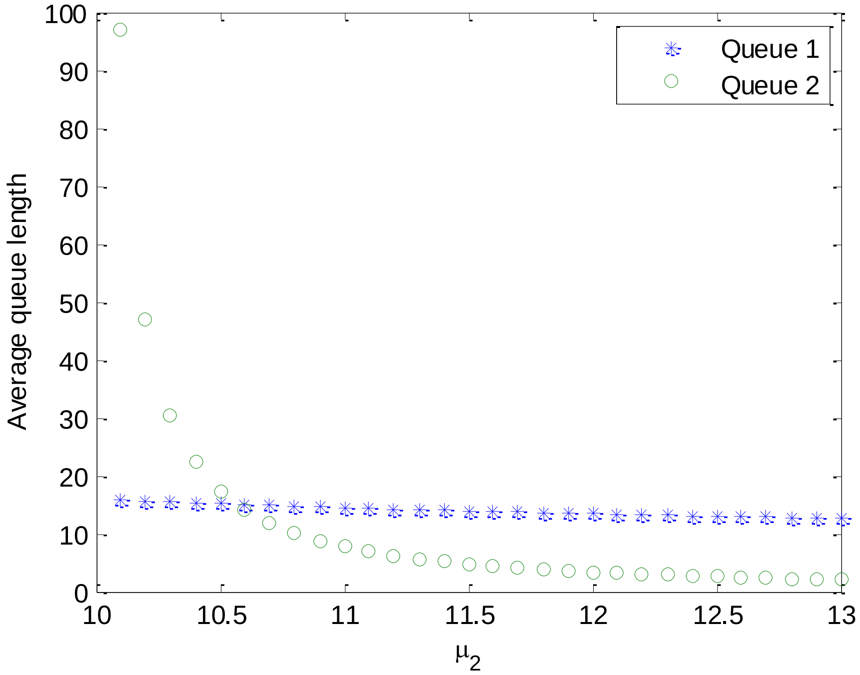

Fig. 9 AQLs versus the service rate of the type-2 customers

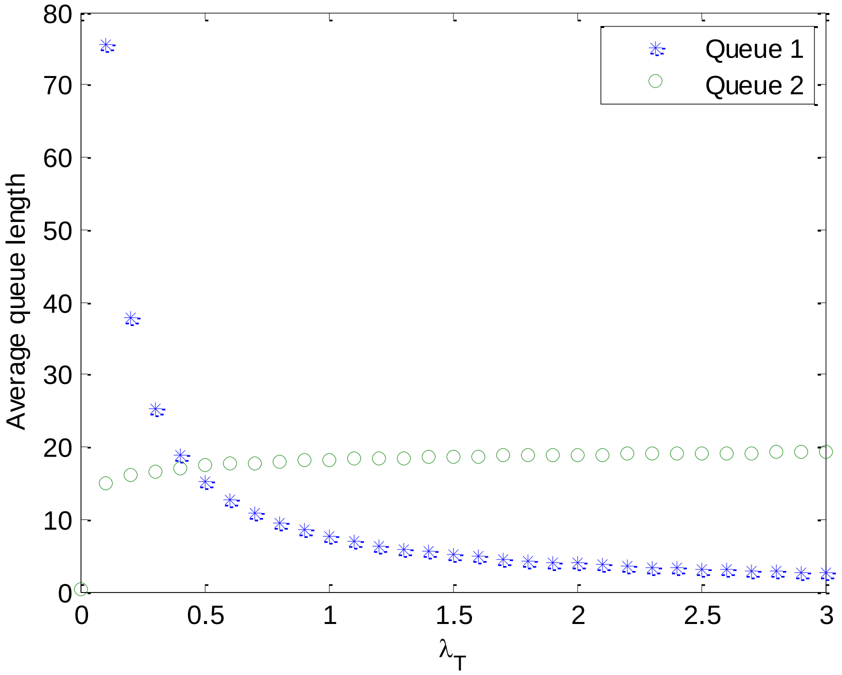

Fig. 10 AQLs versus the transfer rate

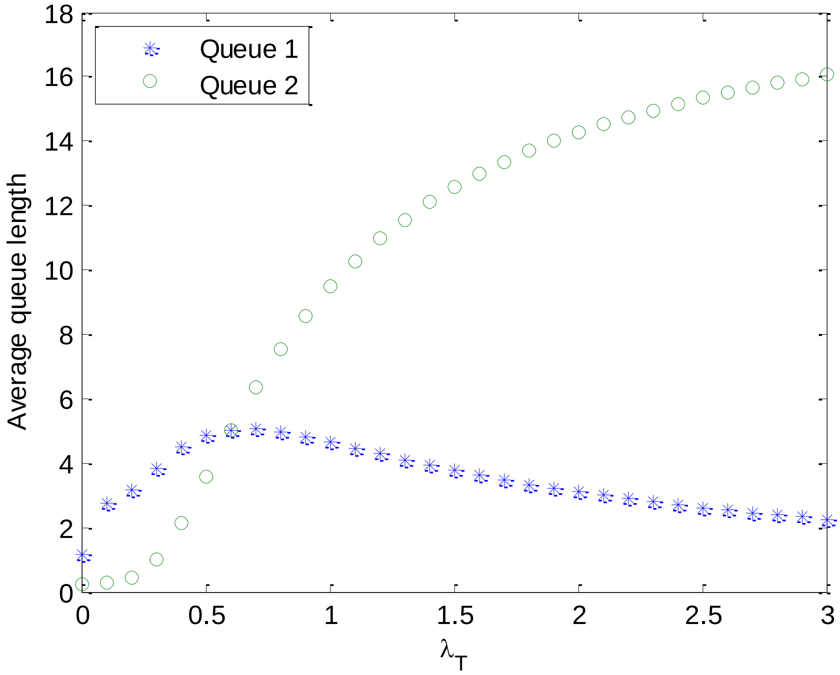

Fig. 11 AQLs versus transfer rate with $\{\lambda_1, \lambda_2, \mu_1, \mu_2\} = \{8, 2, 15, 10.5\}$

6. Conclusion

This paper studies a two-class service system, where low priority customers may upgrade to the high priority class after they have been waiting in queue for some time. The randomness of upgrading process is characterized by a stochastic process. To help better designing such systems, we make effort to analyze system performance. We first study the aymptotics of system stationary distribution, and then design an algorithm to calculate the stationary distribution. Finally, we analyze the impact of system parameters on system performance measures. In the future research, it may be interesting to evaluate the performance of service systems with non-homogeneous arrivals and multiple servers.

Acknowledgement

The authors gratefully acknowledge that this work was supported by NSF of China under grant 71201154.

Appendix A

The current theory for tail types and asymptotics of the stationary distributions is mainly for discrete-time processes. We first review the basic sufficient conditions for the discrete-time QBD process to have a stationary distribution whose tail decays geometrically, and then tailor the theory to the continuous-time process.

The discrete-time QBD process is introduced as follows. Let $\{(X_n, Y_n), n = 0, 1, \dots\}$ be a discrete-time Markov chain with countable state space S. Assume that X_n is nonnegative integer valued, and Y_n has the state space S_0 if $X_n = 0$, and the state space S_1 if $X_n = 1$, etc. Thus, $S=(\{0\}\times S_0) \cup(\{1\}\times S_1)$. We refer to X_n and Y_n as level and background process, respectively. The transition probability matrix P of the Markov process is given by

$$P = \begin{pmatrix} B_0 & B_1 & & & \\ B_{-1} & A_0 & A_1 & & \\ & A_{-1} & A_0 & A_1 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{A.1}$$

where the block size may be finite or infinite.

Lemma A.1 (Neuts, 1981) Suppose that P defined in Eq. (A.1) is ergodic. Let $\mathbf{v} = (\mathbf{v}_n, n \geq 0)$ be its stationary distribution. There exists a minimal nonnegative solution R of the matrix equation:

$$R = R^2 A_{-1} + RA_0 + A_1, \tag{A.2}$$

and the stationary distribution has the following matrix geometric form

$$\mathbf{v}_n = \mathbf{v}_1 R^{n-1}, \quad n > 1, \tag{A.3}$$

$$\mathbf{v}_0 B_0 + \mathbf{v}_1 B_{-1} = \mathbf{v}_0, \tag{A.4}$$

$$\mathbf{v}_0 B_1 + \mathbf{v}_1 \left(A_0 + RA_{-1}\right) = \mathbf{v}_1. \tag{A.5}$$

Lemma A.2 Under the assumption of Lemma A.1, we define the matrix generating function $A^*(z) = z^{-1}A_1 + A_0 + zA_{-1}$. If there exist a positive row vector $\boldsymbol{x}$, a positive column vector $\boldsymbol{y}$, and a real number $z \in (0, 1)$ satisfying the following conditions

$$\mathbf{x}A^*(z) = \mathbf{x}, \tag{A.6}$$

$$A^*(z)\mathbf{y} = \mathbf{y}, \tag{A.7}$$

$$\mathbf{xy} < \infty, , \tag{A.8}$$

$$\mathbf{v}_1\mathbf{y} < \infty, \tag{A.9}$$

then we have the finite limitation

$$\lim_{n\to\infty} z^{-n}\mathbf{v}_n = \frac{\mathbf{v}_1\mathbf{r}}{z\mathbf{xr}}\mathbf{x}, , \tag{A.10}$$

where $\mathbf{r} = (I - A_0 - RA_{-1} - zA_{-1})\mathbf{y}$.

Proof: Lemma A.2 follows by Theorem 2.1.1 and 2.2.1 in Sakuma & Miyazawa (2005).□

Proof of Theorem 3: Denote by P the transition probability matrix of its corresponding embedded Markov chain. The transition probability matrix P has QBD form of Eq. (A.1). We have

$$P = \begin{pmatrix} D_0^{-1} & & & \\ & D^{-1} & & \\ & & D^{-1} & \\ & & & \ddots \end{pmatrix} Q + I, \tag{A.11}$$

where $D_0 = -diag\{C_0\}$ and $D = -diag\{Q_0\}$; and I is the identity matrix. Let $\boldsymbol{v} = (\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots)$ be the stationary distribution of the embedded Markov chain, i.e., $\boldsymbol{v}P = \boldsymbol{v}$ and $\boldsymbol{ve} = 1$. Then, we have $\boldsymbol{\pi}_n = \beta^{-1}\boldsymbol{v}_n D^{-1}$, for all $n \geq 1$, where

$$\beta = \mathbf{v}_0 D_0^{-1}\mathbf{e} + \sum_{n=1}^{\infty} \mathbf{v}_n D^{-1}\mathbf{e} = \mathbf{v}_0 D_0^{-1}\mathbf{e} + \mathbf{v}_1 (I - R)^{-1} D^{-1}\mathbf{e}. \tag{A.12}$$

From the assumption, P is positive recurrent and its invariant vector $\boldsymbol{v}$ is given by Lemma 1. Let $z = \rho$, and assume that $x_0 = 1$ and $y_0 = 1$. By Eq.(A.6), we have

$$\frac{x_0}{\lambda_1 + \lambda_2 + \mu_2}(\lambda_2 z^{-1} + \mu_2 z) + \frac{x_1}{\lambda_1 + \lambda_2 + \mu_2 + \lambda_T}\lambda_T z^{-1} = x_0, \tag{A.13}$$

$$\frac{x_{i-1}\lambda_1}{\lambda_1 + \lambda_2 + \mu_2 + (i-1)\lambda_T} + \frac{x_i(\lambda_2 z^{-1} + \mu_2 z)}{\lambda_1 + \lambda_2 + \mu_2 + i\lambda_T} + \frac{x_{i+1}(i+1)\lambda_T z^{-1}}{\lambda_1 + \lambda_2 + \mu_2 + (i+1)\lambda_T} = x_i, \quad i > 0. \tag{A.14}$$

From Eqs. (A.13) and (A.14), we have

$$x_i = \frac{\lambda_1 + \lambda_2 + \mu_2 + i\lambda_T}{\lambda_1 + \lambda_2 + \mu_2} \times \frac{\lambda_1^i}{\prod_{k=1}^{i} k\lambda_T}, \quad i \geq 0. \tag{A.15}$$

By Eq. (A.7), we have

$$(\lambda_2 z^{-1} + \mu_2 z)y_0 + \lambda_1 y_1 = (\lambda_1 + \lambda_2 + \mu_2)y_0, \tag{A.16}$$

$$i\lambda_T z^{-1} y_{i-1} + (\lambda_2 z^{-1} + \mu_2 z)y_i + \lambda_1 y_{i+1} = (\lambda_1 + \lambda_2 + \mu_2 + i\lambda_T)y_i, \quad i > 0. \tag{A.17}$$

From Eqs. (A.16) and (A.17), we have $y_i = \rho^{-i}$, for $i \geq 0$.

$$y_i = \rho^{-i}, \quad i \geq 0.. \tag{A.18}$$

It's easy to verify Eq. (A.8). To verify Eq. (A.9), we use theorem 14.3.7 in [4]. Define the following function for $q_1 \geq 0$ and $q_2 \geq 0$,

$$V(q_2,q_1)=q_1\lambda_T\rho^{-q_1}\,. \tag{A.19}$$

There are three cases to be considered:
Case 1: Initial state (q_2, q_1) with $q_1 \geq 0$ and $q_2 \geq 1$. We have

$$\begin{aligned}
&\sum_{(y_2,y_1)} P_{(q_2,q_1)(y_2,y_1)}V(y_2,y_1)-V(q_2,q_1)\\
&=\frac{1}{\lambda_1+\lambda_2+\mu_2+q_1\lambda_T}\begin{pmatrix}\lambda_1V(q_2,q_1+1)+\lambda_2V(q_2+1,q_1)+\mu_2V(q_2-1,q_1)\\ +q_1\lambda_TV(q_2+1,q_1-1)-(\lambda_1+\lambda_2+\mu_2+q_1\lambda_T)V(q_2,q_1)\end{pmatrix}\\
&=\frac{1}{\rho^{q_1}}\frac{1}{\lambda_1+\lambda_2+\mu_2+q_1\lambda_T}\left(\lambda_1\lambda_T(\frac{q_1+1}{\rho}-q_1)+q_1\lambda_T^2((q_1-1)\rho-q_1)\right)\\
&=\frac{1}{\rho^{q_1}}\frac{-\lambda_T}{\lambda_1+\lambda_2+\mu_2+q_1\lambda_T}\left(q_1^2\lambda_T(1-\rho)+q_1(\lambda_T\rho-\lambda_1(\frac{1}{\rho}-1))-\frac{\lambda_1}{\rho}\right).
\end{aligned} \tag{A.20}$$

Case 2: Initial state (q_2, q_1) with $q_1 \geq 1$ and $q_2 = 0$. We have

$$\begin{aligned}
&\sum_{(y_2,y_1)} P_{(0,q_1)(y_2,y_1)}V(y_2,y_1)-V(0,q_1)\\
&=\frac{1}{\lambda_1+\lambda_2+\mu_1+(q_1-1)\lambda_T}\begin{pmatrix}\lambda_1V(0,q_1+1)+\lambda_2V(1,q_1)+(q_1-1)\lambda_TV(1,q_1-1)\\ +\mu_1V(0,q_1-1)-(\lambda_1+\lambda_2+\mu_i+(q_1-1)\lambda_T)V(0,q_1)\end{pmatrix}\\
&=\frac{\rho^{-q_1}}{\lambda_1+\lambda_2+\mu_1+(q_1-1)\lambda_T}\left(\lambda_1\lambda_T(\frac{q_1+1}{\rho}-q_1)+(q_1-1)\lambda_T^2((q_1-1)\rho-q_1)\right).
\end{aligned} \tag{A.21}$$

Case 3: Initial state (q_2, q_1) with $q_1 =0$ and $q_2 = 0$. We have

$$\begin{aligned}
&\sum_{(y_2,y_1)} P_{(0,0)(y_1,y_2)}V(y_2,y_1)-V(0,0)\\
&=\frac{1}{\lambda_1+\lambda_2}\left(\lambda_1V(0,1)+\lambda_2V(1,0)-(\lambda_1+\lambda_2)V(0,0)\right)=\frac{\lambda_1\lambda_T}{\rho(\lambda_1+\lambda_2)}.
\end{aligned} \tag{A.22}$$

Define functions

$$g_1(q_1)=\frac{\lambda_T}{\lambda_1+\lambda_2+\mu_2+q_1\lambda_T}\left(q_1^2\lambda_T(1-\rho)+q_1(\lambda_T\rho-\lambda_1(\frac{1}{\rho}-1))-\frac{\lambda_1}{\rho}\right), \tag{A.23}$$

$$g_2(q_1)=\frac{-1}{\lambda_1+\lambda_2+\mu_1+(q_1-1)\lambda_T}\left(\lambda_1\lambda_T(\frac{q_1+1}{\rho}-q_1)+(q_1-1)\lambda_T^2((q_1-1)\rho-q_1)\right). \tag{A.24}$$

Let

$$q_1^*=\min\left\{q_1;g_i(x)\geq g_i(q_1)>0,\ \forall x\geq q_1, i=1,2\right\}, \tag{A.25}$$

$$c=\min\left\{g_1(q_1^*),g_2(q_1^*)\right\}>0, \tag{A.26}$$

$$M = \max_{q_1 < q_1^*} \left\{ \sum_{(y_2, y_1)} P_{(q_2, q_1)(y_2, y_1)} V(y_2, y_1) - V(q_2, q_1) + \frac{c}{\rho^{q_1}} \right\}. \tag{A.27}$$

Then we have

$$\sum_{(y_2, y_1)} P_{(q_2, q_1)(y_2, y_1)} V(y_2, y_1) - V(q_2, q_1) < -c\rho^{-q_1} + M\ . \tag{A.28}$$

By Theorem 14.3.7 in Meyn & Tweedie (1993), we have

$$\sum_{(q_2, q_1)} v_{(q_2, q_1)} \frac{c}{\rho^{q_1}} \le \sum_{(q_2, q_1)} v_{(q_2, q_1)} M = M\ . \tag{A.29}$$

Hence

$$\mathbf{v}_1 \mathbf{y} = \sum_{q_1 \ge 0} v_{(1\ \ q_1)} \rho^{-q_1} < \sum_{(q_2, q_1)} v_{(q_2, q_1)} \rho^{-q_1} \le c^{-1} \mathrm{M} < \infty\ . \tag{A.30}$$

Therefore, by Lemma 2, we have

$$\lim_{n \to \infty} \rho^{-n} \mathbf{v}_n = \frac{\mathbf{v}_1 \mathbf{r}}{\rho \mathbf{x} \mathbf{r}} \mathbf{x}\,. \tag{A.31}$$

Since $\boldsymbol{\pi}_n = \beta^{-1} \mathbf{v}_n D^{-1}$, we have

$$\lim_{n \to \infty} \rho^{-n} \boldsymbol{\pi}_n = \frac{\mathbf{v}_1 \mathbf{r}}{\beta \rho \mathbf{x} \mathbf{r}} \mathbf{x} D^{-1}\,. \tag{A.32}$$

The constants $\mathbf{v}_1\mathbf{r}$ and $\mathbf{xr}$ are positive and finite. Denote by $\mathbf{c} = (c_0, c_1, \ldots)$ the Poisson distribution with parameter λ_1/λ_T. Define

$$\alpha = \frac{\mathbf{v}_1 \mathbf{r}}{\mathbf{x} \mathbf{r}} \frac{\mu_2^{-1} e^{\lambda_1/\lambda_T}}{\beta \rho (\rho + 1)} > 0\,. \tag{A.33}$$

Finally, we have $\lim_{n \to \infty} \rho^{-n} \boldsymbol{\pi}_n = \alpha \mathbf{c}$.□

Appendix B

Proof of Theorem 3.2:

1) Apparently $\boldsymbol{\eta}_1(k) = \boldsymbol{\eta}_2(k)$ if $\mu_1 = \lambda_T$. Otherwise we have $s_1 > s_2$. The following equation holds:

$$\lim_{k\to\infty}\frac{\boldsymbol{\eta}_1(k)}{\boldsymbol{\eta}_2(k)} = \lim_{k\to\infty}\frac{\boldsymbol{\eta}_1(0)\lambda_1^k\prod_{i=0}^{k-1}\frac{1}{s_1+i\lambda_T}}{\boldsymbol{\eta}_2(0)\lambda_1^k\prod_{i=0}^{k-1}\frac{1}{s_2+i\lambda_T}} = \lim_{k\to\infty}\frac{\boldsymbol{\eta}_1(0)\prod_{i=0}^{k-1}(s_2+i\lambda_T)}{\boldsymbol{\eta}_2(0)\prod_{i=0}^{k-1}(s_1+i\lambda_T)}. \\ = \frac{\boldsymbol{\eta}_1(0)}{\boldsymbol{\eta}_2(0)}\lim_{k\to\infty}\prod_{i=0}^{k-1}\frac{s_2+i\lambda_T}{s_1+i\lambda_T} \qquad \text{(B.1)}$$

If $\mu_1 > \lambda_T$, let $\mu_1 = (1+\alpha)\lambda_T$, $\alpha > 0$,

$$\lim_{k\to\infty}\prod_{i=0}^{k-1}\frac{s_1+i\lambda_T}{s_2+i\lambda_T} = \lim_{k\to\infty}\prod_{i=0}^{k-1}\frac{(1+\alpha)\lambda_T+i\lambda_T}{\lambda_T+i\lambda_T} = \lim_{k\to\infty}\prod_{i=1}^{k}\frac{i+\alpha}{i} \\ = \lim_{k\to\infty}\prod_{i=1}^{k}(1+\frac{\alpha}{i}) \ge \lim_{k\to\infty}\left(1+\sum_{i=1}^{k}\frac{\alpha}{i}\right) \to +\infty \qquad \text{(B.2)}$$

Note that numerator and denominator are interchanged from formula (B.1) to formula (B.2), as well as (B.3). If $\mu_1 < \lambda_T$, let $\mu_1 = \beta\lambda_T$, $0<\beta<1$,

$$\lim_{k\to\infty}\prod_{i=0}^{k-1}\frac{s_1+i\lambda_T}{s_2+i\lambda_T} = \lim_{k\to\infty}\prod_{i=0}^{k-1}\frac{\lambda_T+i\lambda_T}{\beta\lambda_T+i\lambda_T} = \lim_{k\to\infty}\prod_{i=0}^{k-1}\frac{1+i}{\beta+i} = \lim_{k\to\infty}\prod_{i=0}^{k-1}(1+\frac{1-\beta}{\beta+i}) \\ \ge \lim_{k\to\infty}\left(1+\sum_{i=0}^{k-1}\frac{1-\beta}{\beta+i}\right) \ge \lim_{k\to\infty}\left(1+\sum_{i=1}^{k}\frac{1-\beta}{i}\right) \to +\infty. \qquad \text{(B.3)}$$

Therefore,

$$\lim_{k\to\infty}\frac{\boldsymbol{\eta}_1(k)}{\boldsymbol{\eta}_2(k)} = 0. \qquad \text{(B.4)}$$

2) For any small $\theta > 0$, we have

$$\frac{\boldsymbol{\eta}_1(k+1)}{\boldsymbol{\eta}_1(k)} = \frac{\boldsymbol{\eta}_1(0)\lambda_1^{k+1}\prod_{i=0}^{k}\frac{1}{s_1+i\lambda_T}}{\boldsymbol{\eta}_1(0)\lambda_1^{k}\prod_{i=0}^{k-1}\frac{1}{s_1+i\lambda_T}} = \frac{\lambda_1}{s_1+k\lambda_T}. \qquad \text{(B.5)}$$

From (B.5) we know that when k goes to infinity, $\boldsymbol{\eta}_1(k+1)/\boldsymbol{\eta}_1(k)$ goes to zero which is smaller than any positive values. Therefore there exists a $k^* = \max\{0, \lceil(\lambda_1/\theta - s_1)/\lambda_T\rceil\}$ such that for any $k > k^*$, we have

$$\frac{\boldsymbol{\eta}_1(k+1)}{\boldsymbol{\eta}_1(k)} < \theta . \tag{B.6}$$

Therefore $\boldsymbol{\eta}_1(k)$ approaches to 0 faster than any geometric decay. Proof for the modified queue with service rate s_2 follows similarly. □

Proof of Lemma 3.1: If $\mu_1 = \lambda_T$, systems are identical. Then equalities in Eq. (3.9) hold. If $\mu_1 > \lambda_T$, we have $s_1 = \mu_1$ and $s_2 = \lambda_T$. Then the modified queue with service rate s_1 can be considered as a queue with two collaborative servers, with service rate λ_T and $\mu_1 - \lambda_T$, i.e., the original queue 1 can be considered as a queue with one fixed server with service rate λ_T and one flexible server with service rate $\mu_1 - \lambda_T$. The flexible server collaborates with the fixed server according to a certain stochastic process. For any sample path ω, we sort customers arrived before time t into the following categories:

1) Customers leaves the queue without accepting any services: these customers do not belong to $N_1(t)$, neither $L_2(t)$.
2) Customer served by the server with service rate λ_T: these customers do not belong to $N_1(t)$, neither $L_2(t)$.
3) Customers served by the server with service rate $\mu_1 - \lambda_T$: these customers do not belong to $N_1(t)$, but may belong to $L_2(t)$.

Therefore, for any time t and sample path ω, we have $N_1(t, \omega) \le L_2(t, \omega)$. This implies $N_1(t) \le_{st} L_2(t)$. The inequality $L_1(t) \le_{st} N_1(t)$ can be proved analogously. If $\mu_1 < \lambda_T$, Eq. (3.9) still holds by similar discussions.

Proof of Theorem 3.3:

By Lemma 3.1, it is easy to see that, for any $n \ge 0$,

$$\Pr\left\{L_1 > n\right\} \pounds \Pr\left\{N_1 > n\right\} \pounds \Pr\left\{L_2 > n\right\} \cdot \tag{B.7}$$

Then, we have

$$\begin{aligned}&\sum_{k=n+1}^{\infty} \boldsymbol{\pi}(\cdot,k) = \Pr\{N_1 > n\} \\ &\le \Pr\{L_2 > n\} = \sum_{k=n+1}^{\infty} \boldsymbol{\eta}_2(k) = \left(1 + \sum_{k=1}^{\infty} \lambda_1^k \prod_{i=0}^{k-1} \frac{1}{s_2 + i\lambda_T}\right)^{-1} \sum_{k=n+1}^{\infty} \lambda_1^k \prod_{i=0}^{k-1} \frac{1}{s_2 + i\lambda_T}\end{aligned} \tag{B.8}$$

The other direction can be proved in a similar manner.

Given any $\gamma > 0$, let $k^* = \max\{0, \lceil (\lambda_1/\gamma - s_2)/\lambda_T \rceil\}$. For any $k > k^*$, we have

$$\sum_{j=k+1}^{\infty} \boldsymbol{\eta}_2(j) = \boldsymbol{\eta}_2(k) \sum_{j=1}^{\infty} \prod_{i=1}^{j} \frac{\lambda_1}{s_2 + (i+k)\lambda_T} \le \boldsymbol{\eta}_2(k) \sum_{j=1}^{\infty} \gamma^j = \boldsymbol{\eta}_2(k) \frac{\gamma}{1-\gamma} \le \gamma \boldsymbol{\eta}_2(k). \tag{B.9}$$

It follows that

$$\boldsymbol{\pi}(\cdot,k) < \sum_{j=k}^{\infty} \boldsymbol{\pi}(\cdot,j) \le \sum_{j=k}^{\infty} \boldsymbol{\eta}_2(j) \le (1+\gamma)\boldsymbol{\eta}_2(k)\,. \tag{B.10}$$

Proof for the other half of the theorem can be completed in a similar manner. □

Appendix C

1) *Deriving G iteratively by successive substitution*:

This method, described by Neuts (1981), makes use of

$$G_{(n+1)} = -\left(\bar{Q}_1 + G_{(n)}^2 \bar{Q}_{-1}\right)\bar{Q}_0^{-1}, \quad n \geq 0, \tag{C.1}$$

which is derived from Eq.(4.10). Starting with $G_{(0)} = 0$, successive approximations of G can be obtained by using Eq. (C.1). The iteration is repeated until two consecutive iterates of G differ by less than a predefined tolerance ε:

$$\left\| G_{(n+1)} - G_{(n)} \right\| < \varepsilon, \tag{C.2}$$

where $\| \bullet \|$ is an appropriate matrix norm. The sequence $\{G_{(n)}\}$ is entry-wise non-decreasing which can be proven by induction:

$$G_{(1)} = -\left(\bar{Q}_1 + G_{(0)}^2 \bar{Q}_{-1}\right)\bar{Q}_0^{-1} = -\bar{Q}_1 \bar{Q}_0^{-1} \geq 0 = G_{(0)}. \tag{C.3}$$

The matrices $-\bar{Q}_0^{-1}$ and $\bar{Q}_{-1}$ are non-negative. For $\bar{Q}_{-1}$, this is readily seen considering the structure of $\bar{Q}$. $\bar{Q}_0^{-1}$ is non-positive because $\bar{Q}_0^{-1}$ is diagonally dominant with negative diagonal and non-negative off-diagonal elements.

If $G_{(n+1)} \geq G_{(n)}$, we have

$$G_{(n+2)} = -\left(\bar{Q}_1 + G_{(n+1)}^2 \bar{Q}_{-1}\right)\bar{Q}_0^{-1} \geq -\left(\bar{Q}_1 + G_{(n)}^2 \bar{Q}_{-1}\right)\bar{Q}_0^{-1} = G_{(n+1)} \tag{C.4}$$

The monotone convergence of $\{G_{(n)}\}$ towards G is shown by Neuts (1981).

2) *Deriving* $\boldsymbol{\theta}_0$ *and* $\boldsymbol{\theta}_1$:

Taking the boundary balance equations and normalization condition $\boldsymbol{\theta}\mathbf{e} = 1$, we have:

$$\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1\right)\begin{pmatrix} \bar{C}_0 & (\bar{C}_1)^* & \mathbf{e} \\ \bar{Q}_{-1} & (\bar{Q}_0 + G\bar{Q}_{-1})^* & (\mathbf{I} - G)^{-1}\mathbf{e} \end{pmatrix} = \left(0, \ldots, 0, 1\right). \tag{C.5}$$

where $(\cdot)^*$ indicates that the last column of the included matrix is removed to avoid linear dependency. The removed column is replaced by the normalization condition. Therefore, Eq. (C.5) is solved for computing $\boldsymbol{\theta}_0$ *and* $\boldsymbol{\theta}_1$.

$$\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1\right) = \left(0, \ldots, 0, 1\right)\begin{pmatrix} \bar{C}_0 & (\bar{Q}_1)^* & \mathbf{e} \\ \bar{Q}_{-1} & (\bar{Q}_0 + G\bar{Q}_{-1})^* & (\mathbf{I} - G)^{-1}\mathbf{e} \end{pmatrix}^{-1}. \tag{C.6}$$

3) *Deriving* $\boldsymbol{\theta}$:

The steady-state probability vectors $\boldsymbol{\theta}_i$ can be obtained quite easily by using Eq.(4.9). Of course not all $\boldsymbol{\theta}_i$ can be computed due to their infinite number, but the elements of $\boldsymbol{\theta}_i$ converge towards 0 for increasing i since $sp(G) < 1$.

Reference

[1] Akan, M., Alagoz, O., Ata, B., Erenay, F. S., Said, A. (2011). A Broader View of Designing the Liver Allocation System, Operations Research (In press). (Accepted in 2011)

[2] Bini D., B. Meini, S. Steff, J. F. Prez, B. Van Houdt. (2012). SMCSolver and Q-MAM: tools for matrix-analytic methods. SIGMETRICS Performance Evaluation Review, 39 46–46.

[3] Deniz, B., Karaesmen, I. and Sheller-Wolf, A. (2010). Managing perishables with substitution: inventory issuance and replenishment heuristics. *Manufacturing & Service Operations Management*, 12(2), 319-329.

[4] Down, D. G. and Lewis, M. E. (2010). The N-Network Model with Upgrades. *Probability and the Engineering and Informational Sciences*, 24, 171-200.

[5] Gómez-Corral, A., Krishnamoorthy, A., Narayanan, V. C. (2005). The impact of self-generation of priorities on multi-server queues with finite capacity. *Stochastic Models*, 21, 427-447.

[6] He, Q.-M., Neuts, M. F. (2002). Two M/M/1 queues with transfers of customers. *Queueing Systems*, 42, 377–400.

[7] He, Q.-M., Xie, J., Zhao, X. (2012). Priority queue with customer upgrades. *Naval Research Logistics*, 59(5):362-375.

[8] Maertens, T., Walraevens, J., Bruneel, H. (2006). On priority queues with priority jumps. *Performance Evaluation*, 1235–1252.

[9] Meyn, S. P. and Tweedie, R. L. (1993). Markov Chains and Stochastic Stability. Control and Communication in Engineering, Berlin: Springer-Verlag.

[10] Neuts, M. (1981), Matrix-geometric solutions in stochastic models: an algorithmic approach, Baltimore, MD: The Johns Hopkins University Press.

[11] Phung-Duc, T., and Kawanishi, K. (2014), "An Efficient Method for Performance Analysis of Blended Call Centers with Redial," to appear in *Asia-Pacific Journal of Operational Research*.

[12] Sakuma, Y. and Miyazawa, M. (2005). On the effect of finite buffer truncation in a two-node Jackson network. *Journal of Applied Probability*, 42, 199-222.

[13] Wang, Q. (2004). Modeling and analysis of high risk patient queues. *European Journal of Operational Research*, 155, 502-515.

[14] Xie, J., Q.-M. He, and X. Zhao (2008). Stability of a priority queueing system with customer transfers. *Operations Research Letters*, 36(6), 705-709.

[15] Xie, J., Q.-M. He, and X. Zhao (2009). On the stationary distribution of queue lengths in a multi-class priority queueing system with customer transfers. Queueing Systems, 62(3), 255-277.