Queueing Systems 42, 377–400, 2002
© 2002 Kluwer Academic Publishers. Manufactured in The Netherlands.

Two $M/M/1$ Queues with Transfers of Customers

QI-MING HE
Department of Industrial Engineering, Dalhousie University, Halifax, NS, Canada B3J 2X4

MARCEL F. NEUTS
Department of Systems and Industrial Engineering, University of Arizona, Tucson, AZ 85721, USA

Received 18 January 2002; Revised 23 June 2002

Abstract. We study a system consisting of two $M/M/1$ queues with transfers of customers. In that system, when the difference of the queue lengths reaches L (> 0), a batch of K ($0 < K < L$) customers is transferred from the longer to the shorter queue. A quasi birth-and-death Markov process describes the queueing process of the two queues. By using the mean-drift method, a simple condition for system stability is found. For the stationary distribution of the queueing process, a matrix geometric solution is obtained. It is shown that the stationary distribution of the total number of customers in the system decays exponentially. The decay rate is found explicitly. Results on the busy periods of the queueing system are also obtained. By using these theoretical results, we numerically explore the optimal design of such queueing systems in terms of customer transfer rates, the batch size of transfer, throughputs, arrival rates, and service rates. In particular, we observe that a balanced queueing system has a total transfer rate that is either the smallest or close to the smallest. We also observe that for system optimization with respect to system descriptors such as the total transfer rate, the mean total queue length, or the system idle probability, the choice of the batch size K has much to do with the difference of the relative traffic intensities of the two $M/M/1$ queues.

Keywords: matrix analytic methods, queueing theory, Markov process, mean-drift method, exponential decay

1. Introduction

For many queueing systems with multiple queues, transfer of customers is a natural way to reduce the total number or the waiting times of customers. Such queueing systems find application in manufacturing, computer networks, health care systems, telecommunication systems, and vehicle traffic control. For instance, two clinics provide the same type of specialized service to patients in two cities that are geographically not too far from each other. To utilize their facilities efficiently and to cut patient waiting times, the two clinics may coordinate their services and transfer patients to each other. To be cost effective, the clinics must decide when to transfer patients and, if a transfer occurs, how many patients to be transferred. Queueing models with transfers of customers can be developed for studying such problems. Because of such applications, there are many studies of queuing systems with transfers of customers.

The literature on queueing systems with customer transfers (customer jockeying or switching) concentrates on optimal scheduling, routing, and performance analysis. For optimal scheduling and routing, it studies how to route arriving customers [4,7, 11,13,14], etc. For many cases, the shortest queue policy is optimal. Yet, for some objective functions, the optimal routing policy can be complicated. For instance, for the queueing model in [14], the optimal policy is of threshold type. Performance analysis was carried out for many queueing systems with customer transfers [1,3,5,12,15,16], etc. Earlier work mainly focuses on stability conditions and on the stationary analysis of the queue length(s). As a multidimensional stochastic process is involved, the analysis is usually complicated. Therefore, asymptotic analysis was also carried out. Most earlier work considers queueing systems in which one customer is transferred each time. Zhao and Grassmann [15,16] use matrix analytic methods to analyze queueing models with jockeying. The main differences between this paper and [15,16] are the transfer batch size and the issues of interest.

In this paper, we study a system with two $M/M/1$ queues and batch transfers of customers. To study such systems, we apply matrix analytic methods [6,8,10]. Introducing a level independent quasi birth-and-death (QBD) process for the queue lengths, we obtain a matrix-geometric solution for the stationary distribution of the queue. Efficient algorithms are developed for computing various performance descriptors. Using these, we study design issues for such systems. We concentrate on customer transfer rates, on the mean total queue length, on the system idle probability, and on throughputs.

The contributions of this paper are two-fold. First, a level independent QBD process is introduced for the queueing process. That immediately allows the use of matrix analytic methods in the analysis of the system. Second, a detailed analysis of design issues is carried out. Using the results obtained in the first part, we analyze the relationship between system parameters and system descriptors. Numerically, a near-optimal solution is found for minimizing the mean total transfer rate. The relative traffic intensities of the $M/M/1$ queues play important roles in system design of such models.

The remainder of the paper is organized as follows. In section 2, the queueing system of interest is introduced and a simple system stability condition is obtained. Section 3 introduces a QBD process for the queueing system. In section 4, a matrix geometric solution to the QBD process is obtained and performance descriptors are defined in terms of the matrix geometric solution. In section 5, it is shown that the tail of the distribution of the total number of customers in the system decays exponentially and the decay rate is explicitly obtained. Section 6 presents an analysis of the busy period and busy cycle. In section 7, numerical examples are given. Some design issues for the queueing system are numerically explored.

2. The queueing system of interest

We consider a system consisting of two $M/M/1$ queues with transfers of customers. We distinguish the two queues as queue 1 and queue 2. The customer arrivals of queue i form a Poisson process with parameter λ_i, $i = 1, 2$. The service times of queue i have a common exponential distribution with parameter μ_i, $i = 1, 2$. We assume that the service times and the arrival processes are mutually independent. The two queues are independent except for the transfers of customers between them. Whenever the lengths of the two queues differ by L, K customers are transferred from the longer to the shorter queue, where L and K are positive integers with $K < L$.

Let $q_i(t)$ be the length of queue i, $i = 1, 2$. Since the arrival processes are Poisson and service times are exponential, $\{(q_1(t), q_2(t)),\ t \geqslant 0\}$ is a Markov process. The state space of the Markov process is $\{(n_1, n_2)\colon\ n_1 \geqslant 0,\ n_2 \geqslant 0, \text{ and } |n_1 - n_2| < L\}$. The Markov process is irreducible since every state communicates with the state $(0, 0)$. For the remainder of this paper, we study the Markov process $\{(q_1(t), q_2(t)),\ t \geqslant 0\}$.

We begin our analysis of $\{(q_1(t), q_2(t)),\ t \geqslant 0\}$ by finding a simple stability condition for the queueing system. The queueing system is stable if and only if the Markov process $\{(q_1(t), q_2(t)),\ t \geqslant 0\}$ is positive recurrent. Set

$$\rho = \frac{\lambda_1 + \lambda_2}{\mu_1 + \mu_2}. \tag{2.1}$$

Theorem 2.1. The queueing system is stable if and only if $\rho < 1$.

Proof. First, we prove sufficiency of the condition. The proof is based on Foster's criterion (see Cohen [2, chapter I.3]). The key is to construct a Lyapunov function (test function). For any state (n_1, n_2), we define

$$f(n_1, n_2) = n_1 + n_2. \tag{2.2}$$

The function $f(n_1, n_2)$ goes to infinity when n_1 or n_2 goes to infinity. Let $Q = (Q_{(n_1,n_2)(n_1',n_2')})$ be the infinitesimal generator of the Markov process $\{(q_1(t), q_2(t)),\ t \geqslant 0\}$. It is readily seen that

$$Q_{(n_1,n_2)(n_1,n_2)} = -(\lambda_1 + \lambda_2 + \mu_1 + \mu_2), \quad \text{for } n_1 > 0,\ n_2 > 0. \tag{2.3}$$

Thus, for any state (n_1, n_2) with $n_1 > 0$ and $n_2 > 0$, the mean drift based on the Lyapunov function $f(n_1, n_2)$ is calculated as follows:

$$\begin{aligned}
&\sum_{(n_1',n_2')\colon\ (n_1',n_2')\neq(n_1,n_2)} \frac{Q_{(n_1,n_2)(n_1',n_2')}}{(-Q_{(n_1,n_2)(n_1,n_2)})} f\left(n_1', n_2'\right) - f(n_1, n_2)\\
&= \frac{(n_1 + n_2 + 1)(\lambda_1 + \lambda_2) + (n_1 + n_2 - 1)(\mu_1 + \mu_2)}{\lambda_1 + \lambda_2 + \mu_1 + \mu_2} - (n_1 + n_2)\\
&= \frac{(\lambda_1 + \lambda_2) - (\mu_1 + \mu_2)}{\lambda_1 + \lambda_2 + \mu_1 + \mu_2} \equiv -\varepsilon < 0,
\end{aligned} \tag{2.4}$$

where $\varepsilon = (1 - \rho)/(1 + \rho)$ and the last inequality is obtained by $\rho < 1$. Note that a transfer of customers does not change the total number of customers in the queueing system directly. Since L is finite, it is easy to see that equation (2.4) holds for all but a finite number of states. For the states for which (2.4) may not hold, we must have

$n_1 < L$ and $n_2 < L$. Thus, for these states, the expression (2.4) (the mean drift) is finite. According to Foster's criteria, the Markov process $\{(q_1(t), q_2(t)),\ t \geqslant 0\}$ must be positive recurrent. Therefore, the queueing system is stable.

Next, we prove necessity by using the mean-value method. Clearly, the average total number of arrivals in the time interval $(0, t)$ is given by $(\lambda_1 + \lambda_2)t$. On the other hand, if both servers are working without breaks, the average total number of customers served in $(0, t)$ is given by $(\mu_1 + \mu_2)t$. If the queueing system is stable, it must have the capacity to serve all customers. Therefore, we must have $(\lambda_1 + \lambda_2)t < (\mu_1 + \mu_2)t$ for some $t > 0$, which implies $\rho < 1$. This completes the proof of theorem 2.1. □

From now on, we assume $\rho < 1$ so that the queueing system is stable.

3. The QBD Markov process $\{(X(t), J(t)),\ t \geqslant 0\}$

The process $\{(q_1(t), q_2(t)),\ t \geqslant 0\}$ is a two-dimensional Markov process. Since both $q_1(t)$ and $q_2(t)$ can be any nonnegative integer, it is inconvenient to analyze $\{(q_1(t), q_2(t)),\ t \geqslant 0\}$ directly. Fortunately, the difference between $q_1(t)$ and $q_2(t)$ cannot exceed $L - 1$. That property can be used to rearrange the state space of $\{(q_1(t), q_2(t)),\ t \geqslant 0\}$ to obtained a QBD process that can be easily dealt with by matrix analytic methods. A natural way to do so is to use the total number of customers in the system as the *level* variable and the difference of the two queue lengths as the *phase* variable. Let

$$q(t) = q_1(t) + q_2(t), \qquad J(t) = q_1(t) - q_2(t), \quad t \geqslant 0. \tag{3.1}$$

Then $q(t)$ is the total number of customers in the system and $J(t)$ represents the difference of the two queue lengths. Clearly, $\{(q_1(t), q_2(t)),\ t \geqslant 0\}$ and $\{(q(t), J(t)),\ t \geqslant 0\}$ determine each other uniquely. In fact, we have

$$q_1(t) = \frac{q(t) + J(t)}{2}; \qquad q_2(t) = \frac{q(t) - J(t)}{2}, \quad t \geqslant 0. \tag{3.2}$$

It is readily seen that $\{(q(t), J(t)),\ t \geqslant 0\}$ is a Markov process. Since the state of the queueing system changes only at customer arrival or service completion epochs, $q(t)$ changes its value by one at its transition epochs. The variable $J(t)$ takes finite values $\{L-1, L-2, \ldots, 1, 0, -1, \ldots, -(L-2), -(L-1)\}$. Therefore, $\{(q(t), J(t)),\ t \geqslant 0\}$ is a QBD process.

Note. The Markov process $\{(q_1(t), q_2(t)),\ t \geqslant 0\}$ can also be transformed into an $M/G/1$ type Markov process $\{(I(t), J(t)),\ t \geqslant 0\}$ by setting $I(t) = \min\{q_1(t), q_2(t)\}$ and $J(t) = q_1(t) - q_2(t)$. Methods were developed in [8,10] for analyzing such Markov processes.

The state space of the Markov process $\{(q(t), J(t)),\ t \geqslant 0\}$ can be divided into levels according to the value of $q(t)$. The states in each level are explicitly given as follows.

(i) For $0 \leqslant n \leqslant L-1$, the level n has $n+1$ states: (n, n), $(n, n-2), \ldots, (n, -(n-2))$, and $(n, -n)$.

(ii) For $n \geqslant L$, there are two cases:

(ii.1) If $n \geqslant L$ and $n - L$ is odd, the level n has L states: $(n, L-1)$, $(n, L-3)$, $\ldots, (n, -(L-3))$, and $(n, -(L-1))$.

(ii.2) If $n \geqslant L$ and $n - L$ is even, the level n has $L-1$ states: $(n, L-2)$, $(n, L-4)$, $\ldots, (n, -(L-4))$, and $(n, -(L-2))$.

Based on the above arrangement of states, the infinitesimal generator of the Markov process $\{(q(t), J(t)),\ t \geqslant 0\}$ is given as

$$Q_1 = \begin{pmatrix} A_{0,1} & A_{0,0} & & & & & & \\ A_{1,2} & A_{1,1} & A_{1,0} & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & A_{L-1,2} & A_{L-1,1} & A_{L-1,0} & & & \\ & & & A_{L,2} & A_{L,1} & A_{L,0} & & \\ & & & & A_{L+1,2} & A_{L+1,1} & A_{L+1,0} & \\ & & & & & \ddots & \ddots & \ddots \end{pmatrix}, \quad (3.3)$$

where $A_{L+2n,i} = A_{L,i}$ and $A_{L+2n+1,i} = A_{L+1,i}$ for $i = 0, 1$ and 2, and $n \geqslant 0$. All the blocks in Q_1 are given explicitly in appendix A. The organization of the states in each level is also shown explicitly in appendix A.

The construction of Q_1 shows that the QBD process $\{(q(t), J(t)),\ t \geqslant 0\}$ is level dependent. The analysis of level dependent QBD processes is, in general, not straightforward. Fortunately, we can reorganize the state space of $\{(q(t), J(t)),\ t \geqslant 0\}$ to generate a level independent QBD.

According to (ii.1) and (ii.2), for $n \geqslant 0$, the level $L + 2n$ has $L - 1$ states and the level $L+2n+1$ has L states. If we regroup the states in the levels $L+2n$ and $L+2n+1$, we obtain a new set with $2L - 1$ states for all $n \geqslant 0$. We call the new set the level $L + n$, whose states are arranged as: $(L + 2n, L - 2)$, $(L + 2n, L - 4), \ldots, (L + 2n, -(L - 4))$, $(L + 2n, -(L - 2))$, $(L + 2n + 1, L - 1)$, $(L + 2n + 1, L - 3), \ldots, (L + 2n + 1, -(L - 3))$, and $(L + 2n + 1, -(L - 1))$. After regrouping the states, the resulting QBD process is level independent. Corresponding to the new partition of the state space, we introduce a new Markov process $\{(X(t), J(t)),\ t \geqslant 0\}$ as

$$X(t) = \begin{cases} q(t), & \text{if } q(t) \leqslant L-1, \\ L + \left\lfloor \dfrac{q(t) - L}{2} \right\rfloor, & \text{if } q(t) \geqslant L, \end{cases} \quad (3.4)$$

and $J(t)$ given in equation (3.1), where $\lfloor x \rfloor$ is the largest integer that is equal to or smaller than x. Clearly, the Markov process $\{(X(t), J(t)),\ t \geqslant 0\}$ is irreducible and level independent. The infinitesimal generator Q_2 of $\{(X(t), J(t)),\ t \geqslant 0\}$ is given by

$$Q_2 = \begin{pmatrix} A_{0,1} & A_{0,0} & & & & & & \\ A_{1,2} & A_{1,1} & A_{1,0} & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & A_{L-1,2} & A_{L-1,1} & A^*_{L-1,0} & & & \\ & & & A^*_{L,2} & A_1 & A_0 & & \\ & & & & A_2 & A_1 & A_0 & \\ & & & & & A_2 & A_1 & A_0 \\ & & & & & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{3.5}$$

All the blocks in Q_2 are given explicitly in appendices A and B.

In the rest of the paper, we focus on the Markov process $\{(X(t), J(t)),\ t \geqslant 0\}$. To make it easier to explain results obtained, we give the relationship between the three Markov processes we have introduced:

$$\begin{aligned} &q_1(t) - q_2(t) = J(t); && \\ &q(t) = q_1(t) + q_2(t) = X(t), && \text{if } X(t) < L; \\ &q(t) = q_1(t) + q_2(t) = 2X(t) - L, && \text{if } X(t) \geqslant L,\ J(t) - L \text{ is even}; \\ &q(t) = q_1(t) + q_2(t) = 2X(t) - L + 1, && \text{if } X(t) \geqslant L,\ J(t) - L \text{ is odd}. \end{aligned} \tag{3.6}$$

4. Stationary distribution of $\{(X(t), J(t)),\ t \geqslant 0\}$

Since $\{(X(t), J(t)),\ t \geqslant 0\}$ is a level independent QBD Markov process, its stationary distribution (if exists) has a matrix geometric solution. We refer to [8] for details about the matrix geometric solution of QBD Markov processes. Let R be the minimal nonnegative (matrix) solution to the equation

$$A_0 + RA_1 + R^2A_2 = 0. \tag{4.1}$$

The matrix R plays an important role in matrix analytic methods. We refer to Neuts [8,10] and Latouche and Ramaswami [6] for more details about R. If the Markov process $\{(X(t), J(t)),\ t \geqslant 0\}$ is positive recurrent, then $\mathrm{sp}(R) < 1$, where $\mathrm{sp}(R)$ is the Perron–Frobenius eigenvalue of the matrix R (i.e., the eigenvalue with the largest modulus). Define matrices $\{R_L, R_{L-1}, \ldots, R_1\}$ as

$$\begin{aligned} R_L &= -A^*_{L-1,0}(A_1 + RA_2)^{-1}; \\ R_{L-1} &= -A_{L-2,0}\bigl(A_{L-1,1} + R_L A^*_{L,2}\bigr)^{-1}; \\ R_n &= -A_{n-1,0}(A_{n,1} + R_{n+1}A_{n+1,2})^{-1}, \quad 1 \leqslant n \leqslant L-2. \end{aligned} \tag{4.2}$$

The invertibility of the (inverse) matrices in equation (4.2) can be verified routinely [8].

Denote by $\boldsymbol{\pi}_n$ the stationary probability vector of the Markov process $\{(X(t), J(t)), t \geqslant 0\}$ corresponding to the level n. Note that the elements of the vector $\boldsymbol{\pi}_n$ are indexed according to the actual value of $J(t)$(e.g., $\boldsymbol{\pi}_n = (\pi_{n,L-2}, \pi_{n,L-4}, \ldots, \pi_{n,-(L-4)}, \pi_{n,-(L-2)}, \pi_{n,L-1}, \pi_{n,L-3}, \ldots, \pi_{n,-(L-3)}, \pi_{n,-(L-1)})$, for $n \geqslant L$). According to Neuts [8], the stationary probability vectors are given as

$$\begin{aligned} \pi_0 &= \left(1 + R_1\mathbf{e} + R_1R_2\mathbf{e} + \cdots + R_1R_2\cdots R_{L-1}\mathbf{e} + R_1R_2\cdots R_L(I-R)^{-1}\mathbf{e}\right)^{-1}; \\ \boldsymbol{\pi}_n &= \pi_0 R_1 R_2 \cdots R_n, \qquad 1 \leqslant n \leqslant L; \qquad (4.3) \\ \boldsymbol{\pi}_n &= \pi_0 R_1 R_2 \cdots R_L R^{n-L}, \quad n \geqslant L+1, \end{aligned}$$

where $\mathbf{e}$ is the column vector with all elements being one and I is the identity matrix. The matrix $I - R$ is invertible since $\mathrm{sp}(R) < 1$. Based on the stationary distribution, a number of descriptors for a stable queueing system can be derived.

(1) The mean total number of customers in the queueing system: We partition the vector $\boldsymbol{\pi}_{L+n}$ as $\boldsymbol{\pi}_{L+n} = (\boldsymbol{\pi}_{L+n}(0), \boldsymbol{\pi}_{L+n}(1))$, where $\boldsymbol{\pi}_{L+n}(0)$ corresponds to the first $L-1$ states $\{(L+n, L-2), (L+n, L-4), \ldots, (L+n, -(L-4)), (L+n, -(L-2))$ of the level $L+n$ of Q_2. Denote by q_{mean} the mean total queue length in steady state. By equation (3.6), we have

$$\begin{aligned} q_{\mathrm{mean}} &= \sum_{n=0}^{L-1} n\boldsymbol{\pi}_n\mathbf{e} + \sum_{n=0}^{\infty}\left((L+2n)\boldsymbol{\pi}_{L+n}(0)\mathbf{e} + (L+2n+1)\boldsymbol{\pi}_{L+n}(1)\mathbf{e}\right) \\ &= \sum_{n=0}^{L-1}(n-L)\boldsymbol{\pi}_n\mathbf{e} + L + \boldsymbol{\pi}_L(I-R)^{-1}\left[2R(I-R)^{-1}\mathbf{e} + \begin{pmatrix}\mathbf{0}_{L-1}\\ \mathbf{e}\end{pmatrix}\right], \qquad (4.4) \end{aligned}$$

where $\mathbf{0}_{L-1}$ is a column vector of dimension $L-1$ with all elements being zero.

(2) Let $\{p_n(i), n \geqslant 0\}$ be the distribution of the queue length of queue i in steady state, $i = 1, 2$. $\{p_n(i), n \geqslant 0\}$ is also a marginal distribution of $\{\boldsymbol{\pi}_n, n \geqslant 0\}$. Then we have

$$p_{n_1}(1) = \sum_{n_2=\max\{0,n_1-L+1\}}^{\max\{0,L-n_1-1\}} \pi_{n_1+n_2,n_1-n_2} + \sum_{n_2=\max\{0,L-n_1-1\}+1}^{n_1+L-1} \pi_{L+\lfloor(n_1+n_2-L)/2\rfloor,n_1-n_2}, \quad n_1 \geqslant 0. \qquad (4.5)$$

Similar formulas can be obtained for $\{p_n(2), n \geqslant 0\}$.

(3) The probability that the system is empty is π_0. Systems with a larger π_0 is preferred since a larger π_0 means that the system resource is utilized more effectively.

(4) The probability that server 1 is idle is given as $\pi_0 + \sum_{n=1}^{L-1}\pi_{n,-n}$ and the probability that server 2 is idle is given as $\pi_0 + \sum_{n=1}^{L-1}\pi_{n,n}$.

(5) The probability that at least one server is idle is given as

$$\pi_0 + \sum_{n=1}^{L-1} (\pi_{n,n} + \pi_{n,-n}).$$

(6) Let r_i be the rate at which jobs are completed by server i per unit time, $i = 1, 2$, then

$$r_1 = \mu_1\left(1 - \pi_0 - \sum_{n=1}^{L-1} \pi_{n,-n}\right); \qquad r_2 = \mu_2\left(1 - \pi_0 - \sum_{n=1}^{L-1} \pi_{n,n}\right). \tag{4.6}$$

r_i is also called the throughput in the queue i, $i = 1, 2$. If $\rho < 1$, the total throughput is $r_1 + r_2 = \lambda_1 + \lambda_2$, which is useful for accuracy check in computation.

(7) The (average) transfer rate of customers from queue 1 to queue 2 (or from queue 2 to queue 1) is an interesting descriptor of the queueing system. In applications, there are often costs associated with transfers of customers. Thus, the queueing system should be designed to minimize the number of transfers. We define $T_{R,1\to 2}$ ($T_{R,2\to 1}$) as the average transfer rate from queue 1 to queue 2 (queue 2 to queue 1), and T_R the average total transfer rate. Then these transfer rates can be calculated as follows:

$$\begin{aligned} T_{R,1\to 2} &= \pi_{L-1,L-1}\lambda_1 + \sum_{n=0}^{\infty} \pi_{L+n,L-1}(\lambda_1 + \mu_2) \\ &= \pi_{L-1,L-1}\lambda_1 + \left(\boldsymbol{\pi}_L(I - R)^{-1}\right)_{L-1}(\lambda_1 + \mu_2); \\ T_{R,2\to 1} &= \pi_{L-1,-(L-1)}\lambda_2 + \left(\boldsymbol{\pi}_L(I - R)^{-1}\right)_{-(L-1)}(\lambda_2 + \mu_1). \end{aligned} \tag{4.7}$$

The total transfer rate can be calculated by $T_R = T_{R,1\to 2} + T_{R,2\to 1}$. If $\rho < 1$, the queueing system can reach its steady state and, consequently, the total arrival rate to queue 1 (or 2) must be equal to the total departure rate of queue 1 (or 2). Thus, the following relationship holds for $\{r_1, r_2, \lambda_1, \lambda_2, T_{R,1\to 2}, T_{R,2\to 1}, K\}$:

$$\begin{aligned} \lambda_1 + KT_{R,2\to 1} &= KT_{R,1\to 2} + r_1; \\ \lambda_2 + KT_{R,1\to 2} &= KT_{R,2\to 1} + r_2. \end{aligned} \tag{4.8}$$

The equations in (4.8) are useful for accuracy check in computation.

(8) Since there are exactly K customers involved in each transfer, the average total number of customers transferred per unit time is $T_R K$. Since the total arrival rate of customers is $\lambda_1 + \lambda_2$, the mean number of transfers per customer N_{TR} (during its sojourn time in the queueing system) is given by

$$N_{TR} = \frac{T_R K}{\lambda_1 + \lambda_2}. \tag{4.9}$$

In section 7, we study some of these descriptors numerically.

5. Asymptotics of the stationary distribution

Since the stationary distribution of $\{(X(t), J(t)),\ t \geqslant 0\}$ has a matrix geometric distribution (4.3), it decays exponentially. According to equation (3.5), the distribution of the total number of customers in the queueing system also decays exponentially. In this section, we give an explicit expression for the decay rate and we obtain asymptotic expressions for the stationary distributions of the queue lengths.

According to Neuts [9], the decay rate of the matrix geometric distribution (4.3) is $\mathrm{sp}(R)$. That is: $\boldsymbol{\pi}_n\mathbf{e} \approx C(\mathrm{sp}(R))^n$ for large n, where C is a constant. For our system, an explicit characterization of $\mathrm{sp}(R)$ is given in the following lemmas.

Lemma 5.1. If $\rho < 1$, $\mathrm{sp}(R) = \rho^2$. If $\rho \geqslant 1$, $\mathrm{sp}(R) = 1$.

Proof. Define $A^*(z) = A_0 + zA_1 + z^2A_2$. Let $\chi(z)$ be the eigenvalue of $A^*(z)$ with the largest modulus. According to Neuts [8], $\mathrm{sp}(R)$ is the smallest nonnegative solution to $\chi(z) = 0$ (note that $A^*(R) = 0$). Furthermore, the solution to $\chi(z) = 0$ in the interval $(0, 1)$ is unique if it exists. Note that, since $A^*(1)\mathbf{e} = 0$, $z = 1$ is always a solution to $\chi(z) = 0$.

If $\rho < 1$, the Markov process $\{(X(t), J(t)),\ t \geqslant 0\}$ is positive recurrent. Then $\mathrm{sp}(R)$ is the unique solution to $\chi(z) = 0$ in $(0, 1)$. Therefore, we only need to find the unique solution to $\chi(z) = 0$ in the interval $(0, 1)$. For any $0 < z < 1$, let $\mathbf{u}(z) = (\mathbf{u}_1(z), \mathbf{u}_2(z))$ be a left eigenvector of $A^*(z)$ corresponding to the eigenvector $\chi(z)$, i.e., $\mathbf{u}(z)A^*(z) = \mathbf{u}(z)\chi(z)$. If $\chi(z) = 0$, we must have $\mathbf{u}(z)A^*(z) = 0$, which is equivalent to

$$\begin{aligned} z\mathbf{u}_1(z)A_{L,1} + \mathbf{u}_2(z)(A_{L+1,0} + zA_{L+1,2}) &= 0;\\ \mathbf{u}_1(z)\big(zA_{L,0} + z^2A_{L,2}\big) + z\mathbf{u}_2(z)A_{L+1,1} &= 0, \end{aligned} \tag{5.1}$$

where $A_{L,0}$, $A_{L,1}$, $A_{L,2}$, $A_{L+1,0}$, $A_{L+1,1}$, and $A_{L+1,2}$ are given in appendix A. Equation (5.1) leads to

$$\begin{aligned} z\phi\mathbf{u}_1(z) &= \mathbf{u}_2(z)(A_{L+1,0} + zA_{L+1,2});\\ \phi\mathbf{u}_2(z) &= \mathbf{u}_1(z)(A_{L,0} + zA_{L,2}), \end{aligned} \tag{5.2}$$

where $\phi = \lambda_1 + \lambda_2 + \mu_1 + \mu_2$. Substituting $\mathbf{u}_2(z)$ in the first equation by the second equation in (5.2), yields

$$z\phi^2\mathbf{u}_1(z) = \mathbf{u}_1(z)(A_{L,0} + zA_{L,2})(A_{L+1,0} + zA_{L+1,2}). \tag{5.3}$$

It can be verified that $(A_{L,0} + zA_{L,2})(A_{L+1,0} + zA_{L+1,2})\mathbf{e} = (\lambda_1 + \lambda_2 + z(\mu_1 + \mu_2))^2\mathbf{e}$. Thus, the matrix $(A_{L,0} + zA_{L,2})(A_{L+1,0} + zA_{L+1,2})$ has a positive eigenvalue $(\lambda_1 + \lambda_2 + z(\mu_1 + \mu_2))^2$ with a positive eigenvector $\mathbf{e}$. Choose $z_0 = \rho^2 = ((\lambda_1 + \lambda_2)/(\mu_1 + \mu_2))^2$. It is easy to verify that $(\lambda_1 + \lambda_2 + z_0(\mu_1 + \mu_2))^2 = z_0\phi^2$. Therefore, at $z_0 = \rho^2$, $\rho^2\phi^2$ is an eigenvalue of the matrix $(A_{L,0} + \rho^2A_{L,2})(A_{L+1,0} + \rho^2A_{L+1,2})$. Then $\mathbf{u}_1(\rho^2)$ exists as the left eigenvector corresponding to eigenvalue $\rho^2\phi^2$. If we choose $\mathbf{u}_2(\rho^2)$ satisfying the second equation in (5.2), then $\mathbf{u}(\rho^2) = (\mathbf{u}_1(\rho^2), \mathbf{u}_2(\rho^2))$ satisfies the equation $\mathbf{u}(\rho^2)A^*(\rho^2) = 0$. Since $0 < \rho^2 < 1$, we have $\mathrm{sp}(R) = \rho^2$.

If $\rho \geqslant 1$, the Markov process $\{(X(t), J(t)),\ t \geqslant 0\}$ is null recurrent or transient. According to Neuts [8], $\mathrm{sp}(R) = 1$. This completes the proof of lemma 5.1. □

In [15,16], the decay rate of the matrix geometric distribution was obtained explicitly as well. Next, we characterize the matrix R and its eigenvectors corresponding to $\mathrm{sp}(R)$.

Lemma 5.2. The minimal nonnegative solution R to equation (4.1) has the structure

$$R = \begin{pmatrix} 0 & 0 \\ R_{2,1} & R_{2,2} \end{pmatrix}, \tag{5.4}$$

where $R_{2,1}$ and $R_{2,2}$ are the minimal nonnegative solutions to the equations:

$$\begin{aligned} &\left(-A_{L+1,0}A_{L,1}^{-1}A_{L,0}\right) + R_{2,2}\left[\left(-A_{L+1,2}A_{L,1}^{-1}A_{L,0}\right) + A_{L+1,1} + \left(-A_{L+1,0}A_{L,1}^{-1}A_{L,2}\right)\right] \\ &\quad + R_{2,2}^2\left(-A_{L+1,2}A_{L,1}^{-1}A_{L,2}\right) = 0; \qquad (5.5) \\ &R_{2,1} = -\left(A_{L+1,0} + R_{2,2}A_{L+1,2}\right)A_{L,1}^{-1}. \end{aligned}$$

The matrix $R_{2,2}$ is irreducible and $\mathrm{sp}(R_{2,2}) = \mathrm{sp}(R) = \rho^2$. Let $\mathbf{u}$ be a left eigenvector of R corresponding to the eigenvalue ρ^2. Let $\mathbf{v}$ be a right eigenvector of R corresponding to the eigenvalue ρ^2. Then $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2)$, where $\mathbf{u}_2$ is the left eigenvector of $R_{2,2}$ corresponding to the eigenvalue ρ^2, i.e., $\mathbf{u}_2 R_{2,2} = \rho^2 \mathbf{u}_2$, and $\mathbf{u}_1 = \mathbf{u}_2 R_{2,1}$. The vector $\mathbf{v}$ is given as

$$\mathbf{v} = \begin{pmatrix} 0 \\ \mathbf{v}_2 \end{pmatrix}, \qquad R_{2,2}\mathbf{v}_2 = \rho^2 \mathbf{v}_2. \tag{5.6}$$

The vectors $\mathbf{u}_2$ and $\mathbf{v}_2$ are positive vectors.

Proof. The structure of R (see equation (5.4)) is obtained immediately from equation (B.1) and the fact that R is the minimal nonnegative solution to equation (4.1). All other results are obvious from the theory of nonnegative matrices. This completes the proof of lemma 5.2. □

Denote by

$$q_1 = \lim_{t\to\infty} q_1(t), \qquad q_2 = \lim_{t\to\infty} q_2(t), \qquad \text{and} \quad q = \lim_{t\to\infty} q(t).$$

By lemmas 5.1 and 5.2, asymptotic expressions for these steady state random variables can be obtained.

Theorem 5.3. For the queueing system defined in section 2, we assume that $\rho^2 < 1$. The vectors $\mathbf{u}$ and $\mathbf{v}$, defined in the proof of lemma 5.2, are normalized by $\mathbf{ue} = 1$ and $\mathbf{uv} = 1$. For large n, n_1, and n_2 (with $-L < n_1 - n_2 < L$) we have

$$\boldsymbol{\pi}_n = \pi_0 R_1 R_2 \cdots R_L \mathbf{vu}\rho^{2(n-L)} + \mathrm{o}\big(\rho^{2(n-L)}\big), \tag{5.7}$$

$$P\{q = n\} = \begin{cases} \pi_0 R_1 R_2 \cdots R_L \mathbf{vu}_1\mathbf{e}\rho^{2\lfloor(n-L)/2\rfloor} + \mathrm{o}\big(\rho^{2\lfloor(n-L)/2\rfloor}\big), \\ \quad \text{if } n - L \text{ is even;} \\ \pi_0 R_1 R_2 \cdots R_L \mathbf{vu}_2\mathbf{e}\rho^{2\lfloor(n-L)/2\rfloor} + \mathrm{o}\big(\rho^{2\lfloor(n-L)/2\rfloor}\big), \\ \quad \text{if } n - L \text{ is odd,} \end{cases} \tag{5.8}$$

$$\begin{aligned} P\{q_1 = n_1, q_2 = n_2\} = \pi_0 R_1 R_2 \cdots R_L \mathbf{v}(\mathbf{u})_{n_1-n_2}\rho^{2\lfloor(n_1+n_2-L)/2\rfloor} \\ + \mathrm{o}\big(\rho^{2\lfloor(n_1+n_2-L)/2\rfloor}\big). \end{aligned} \tag{5.9}$$

Proof. It is well known (see [9]) that, for large n, R^{n-L} has an approximation $R^{n-L} = \mathbf{vu}\rho^{2(n-L)} + \mathrm{o}(\rho^{2(n-L)})$. By equation (4.3), the asymptotic formula (5.7) for $\boldsymbol{\pi}_n$ is immediately obtained.

By (3.5), we have that

$$P\{q(t) = n\} = \begin{cases} \displaystyle\sum_{j=-(L-2):\ j-L \text{ is even}}^{L-2} P\left\{X(t) = L + \left\lfloor \frac{n-L}{2} \right\rfloor,\ J(t) = j\right\}, \\ \text{if } n - L \text{ is even;} \\ \displaystyle\sum_{j=-(L-1):\ j-L \text{ is odd}}^{L-1} P\left\{X(t) = L + \left\lfloor \frac{n-L}{2} \right\rfloor,\ J(t) = j\right\}, \\ \text{if } n - L \text{ is odd.} \end{cases} \tag{5.10}$$

The asymptotic expression (5.8) for $P\{q = n\}$ is obtained from equation (5.10). The expression in (5.9) is obtained by using equation (3.6) in a similar way. This completes the proof of theorem 5.3. □

In summary, we have shown that the decay rate of the stationary distribution of the queue lengths is ρ with respect to the total queue length, the queue length of queue 1, or the queue length of queue 2. It is interesting to see that the decay rate is independent of L and K.

6. Analysis of first passage times

In this section, we investigate busy periods, busy cycles, and transfers of customers within busy periods. The basic approach is to consider the first passage time from the level n to the level $n - 1$ for the Markov process $\{(X(t), J(t)),\ t \geqslant 0\}$.

For the level n ($> L$), let $g^*_{j,j'}(s, z_1, z_2)$ be the joint transform of the length of the first passage time from the level n to the level $n - 1$, the number of transfers from queue 1 to queue 2 during that first passage time, the number of transfers from queue 2 to queue 1 during the first passage time, and the difference of the numbers of customers in the two queues is j' at the end of the first passage time, given that, at the beginning of the first passage time, the difference of the numbers of customers in the two queues is j. Let $G^*(s, z_1, z_2)$ be a $(2L - 1) \times (2L - 1)$ matrix with elements $g^*_{j,j'}(s, z_1, z_2)$ with j

and j' taking values $\{L-2, L-4, \dots, -(L-4), -(L-2), L-1, L-3, \dots, -(L-3), -(L-1)\}$. According to Neuts [8,10], for $s > 0$, $0 < z_1, z_2 < 1$, the matrix $G^*(s, z_1, z_2)$ is the minimal nonnegative solution to the equation

$$A_2 + \big(sI + A_1(z_1, z_2)\big)G^*(s, z_1, z_2) + A_0(z_1, z_2)\big(G^*(s, z_1, z_2)\big)^2 = 0, \tag{6.1}$$

where

$$\begin{aligned} A_1(z_1, z_2) &= A_1 + (z_1 - 1)B_{1,1} + (z_2 - 1)B_{2,1}; \\ A_0(z_1, z_2) &= A_0 + (z_1 - 1)B_{1,0} + (z_2 - 1)B_{2,0}, \end{aligned} \tag{6.2}$$

and $B_{1,1}$, $B_{2,1}$, $B_{1,0}$, and $B_{2,0}$ are $(2L-1) \times (2L-1)$ nonnegative matrices with only one positive elements: $(B_{1,1})_{L-1,L-2K} = \mu_2$, $(B_{2,1})_{-(L-1),2K-L} = \mu_1$, $(B_{1,0})_{L-1,L-2K} = \lambda_1$, and $(B_{2,0})_{-(L-1),2K-L} = \lambda_2$. By using equation (6.1), the mean length of the first passage time and the mean numbers of transfers of customers can be obtained.

Let $G = \lim_{s\to 0, z_1\to 1, z_2\to 1} G^*(s, z_1, z_2)$, then G is the minimal nonnegative solution to the equation

$$A_2 + A_1G + A_0G^2 = 0. \tag{6.3}$$

The matrix G plays an important role in matrix analytic methods. See [8,10] and [6] for more details about the matrix G.

Let $u_{s,j}$ be the mean first passage time from the level n to the level $n-1$ $(n > L)$, given that the first passage time started in the state (n, j). Let $\mathbf{u}_s$ be a row vector of dimension $2L - 1$ with elements $u_{s,j}$. Let $u_{1,j}$ be the mean number of transfers from queue 1 to queue 2 during the first passage time from the level n to the level $n-1$ $(n > L)$, given that the first passage time started in the state (n, j). Let $\mathbf{u}_1$ be a row vector of dimension $2L-1$ with elements $u_{1,j}$. Let $u_{2,j}$ be the mean number of transfers from queue 2 to queue 1 during the first passage time from the level n to the level $n-1$ $(n > L)$, given that the first passage started in the state (n, j). Let $\mathbf{u}_2$ be a row vector of dimension $2L - 1$ with elements $u_{2,j}$. By routine calculations, from equation (6.1), we obtain

$$\begin{aligned} \mathbf{u}_s &= -\frac{\partial G^*(s, z_1, z_2)\mathbf{e}}{\partial s}\bigg|_{s=0, z_1=1, z_2=1} = -\big[A_1 + A_0(I+G)\big]^{-1}\mathbf{e}; \\ \mathbf{u}_1 &= \frac{\partial G^*(s, z_1, z_2)\mathbf{e}}{\partial z_1}\bigg|_{s=0, z_1=1, z_2=1} = -\big[A_1 + A_0(I+G)\big]^{-1}(B_{1,1} + B_{1,0})\mathbf{e}; \\ \mathbf{u}_2 &= \frac{\partial G^*(s, z_1, z_2)\mathbf{e}}{\partial z_2}\bigg|_{s=0, z_1=1, z_2=1} = -\big[A_1 + A_0(I+G)\big]^{-1}(B_{2,1} + B_{2,0})\mathbf{e}. \end{aligned} \tag{6.4}$$

Thus, if the matrix G is known, all these performance measures can be computed in a straightforward way.

Similar to the level $n > L$, we can analyze the first passage time from the level L to the level $L-1$, $L-1$ to $L-2, \dots, 2$ to 1, or 1 to 0. Similar to $G^*(s, z_1, z_2)$, G, $\mathbf{u}_s$, $\mathbf{u}_1$, and $\mathbf{u}_2$, we define matrices $G_n^*(s, z_1, z_2)$ and G_n, vectors $\mathbf{u}_s(n)$, $\mathbf{u}_1(n)$, and $\mathbf{u}_2(n)$ for the first passage time from the level n to the level $n-1$, $0 < n \leqslant L$. We also define

$G_0^*(s, z_1, z_2)$, $u_s(0)$, $u_1(0)$, and $u_2(0)$ for the first passage time from the level zero to the level zero. Equations and expressions for these matrices and vectors can be established and are given in appendix C.

With these vectors $\{\mathbf{u}_s, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_s(n), \mathbf{u}_1(n), \mathbf{u}_2(n),\ 0 \leqslant n \leqslant L\}$, we can derive some descriptors for the queueing systems. For instance, $u_s(0)$ is the mean length of a busy cycle (starting from the time the system becomes empty until the system becomes empty again); $u_1(0)$ is the total number of transfers from queue 1 to queue 2 during a busy cycle; $u_2(0)$ is the total number of transfers from queue 2 to queue 1 during a busy cycle. Then the ratio $u_1(0)/u_s(0)$ is the average transfer rate from queue 1 to queue 2 and the ratio $u_2(0)/u_s(0)$ is the average transfer rate from queue 2 to queue 1. Thus, the first passage time analysis not only provides information about the queueing process in the busy cycle, but also information complementary to the steady-state analysis in section 4.

7. Numerical examples and system design

Using the results obtained in sections 4–6, in this section, we study some design issues related to customer transfer rates, throughputs, system idle probability, and the mean total queue length. For instances, we wish to find out under what circumstance the system idle probability is maximized, the mean total queue length is minimized, the throughputs are balanced, or the total transfer rate is minimized. We consider three scenarios.

Scenario #1 (Example 7.1). For systems with fixed parameters $\{L, K, \lambda_1, \lambda_2\}$ and a fixed total service rate μ $(= \mu_1 + \mu_2)$, we like to find how to allocate the service capacity to queue 1 (μ_1) and queue 2 (μ_2) to optimize system descriptors introduced in section 4. Apparently, no single allocation of service capacity can optimize all the system descriptors.

Scenario #2 (Example 7.2). For systems with fixed parameters $\{L, K, \mu_1, \mu_2\}$ and a fixed total arrival rate λ $(= \lambda_1 + \lambda_2)$, we like to find how to route customers to queue 1 (λ_1) and queue 2 (λ_2) to optimize system descriptors.

Scenario #3 (Examples 7.3–7.5). For systems with fixed parameters $\{L, \lambda_1, \lambda_2, \mu_1, \mu_2\}$, we wish to find the batch size K of a customer transfer to optimize system descriptors.

Let $\rho_1 = \lambda_1/\mu_1$ and $\rho_2 = \lambda_2/\mu_2$. We call ρ_1 and ρ_2 the *relative traffic intensities* of queue 1 and queue 2 (without customer transfers), respectively. We call a system *balanced* if $\rho_1 = \rho_2$. The following examples show that the above design issues are closely related to ρ_1 and ρ_2.

Example 7.1. Consider queueing systems with $L = 5$, $K = 3$, $\lambda_1 = 1$, $\lambda_2 = 2$, and $\mu_1 + \mu_2 = 4$. Since $\rho = 0.75 < 1$, these queueing systems are stable. First, we consider

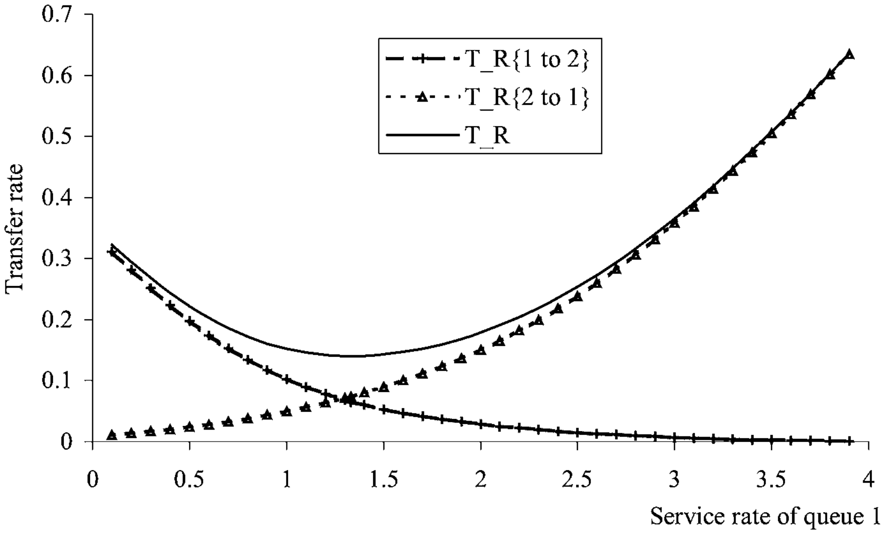

Figure 1. Transfer rates for example 7.1.

the relationship between service capacity allocation and customer transfer rates. The results are presented in figure 1. The horizontal axis of figure 1 is for the service rate μ_1 of queue 1. The vertical axis of figure 1 is for the transfer rates. To generate figure 1, we calculated the three transfer rates for $\mu_1 = 0.1n$, for $1 \leqslant n \leqslant 39$, and $\mu_1 = 4/3$.

It is shown in figure 1 that the transfer rate $T_{R,1\to 2}$ from queue 1 to queue 2 (or $T_{R,2\to 1}$) is monotone in μ_1, which is quite intuitive. The total transfer rate is a unimodal function (a function with a single minimum or maximum) or even a convex function. This is also intuitive since, if μ_1 is close to zero or 4, one of the queues is unstable (if it operates independently). That implies that one of the two queues has a long queue. Therefore, transfers of customers must occur frequently. An interesting issue is when the total transfer rate is minimized. As shown in figure 1, the total transfer rate is minimized at $\mu_1^* = 4/3$, where $\rho_1 = \rho_2 = \rho$ or $\mu_1^* = \lambda_1/\rho$ and $\mu_2^* = \lambda_2/\rho$. For this case, the two $M/M/1$ queues are balanced in terms of the queue length distribution. (Note that for an $M/M/1$ queue with a traffic intensity ρ, the stationary distribution of the queue length is $\{1-\rho, (1-\rho)\rho, (1-\rho)\rho^2, \ldots\}$.) Thus, this example shows that, in order to reduce the total number of transfers, the queueing system should be designed such that the two $M/M/1$ queues have the same relative traffic intensity, i.e., $\rho_1 = \rho_2$. This observation is held true for many examples we have tested. In general, we found that the total transfer rate is minimized at an allocation of service rate for which $\rho_1 \approx \rho_2$, i.e., when the system is balanced or close to be balanced.

The results for the π_0, q_{mean}, r_1 and r_2 are plotted in figure 2. The horizontal axis of figure 2 is still for the service rate μ_1 of server 1. Since π_0 is relatively small, we plot $100\pi_0$ in figure 2.

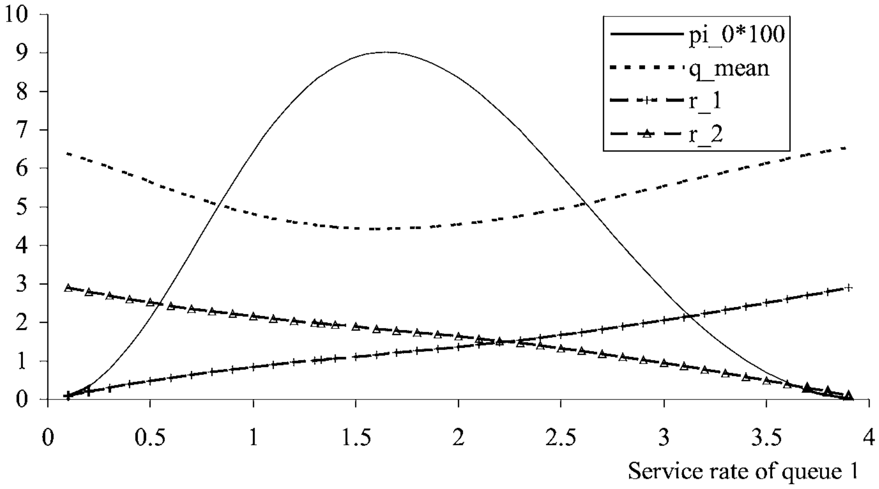

Figure 2. Idle probability, queue length and departure rates for example 7.1.

In figure 2, it is easy to see that the throughputs (r_1 and r_2) are monotone in μ_1, which is intuitive. But the system idle probability π_0 and the mean total queue length q_{mean} are unimodal in μ_1. When is the mean total queue length minimized? When is the system idle probability maximized? Figure 2 shows that the two descriptors are minimized or maximized at some point μ_1 that is larger than λ_1/ρ (approximately at $\mu_1 = 1.6$). That is, the queue with a smaller input rate gets, proportional to the arrival rates, more service capacity in order to reduce the total queue length (or increase system idle time), which is counterintuitive. The reason is the following: queue 2 has larger input and service rates, so the changes in its queue length are more frequent than in queue 1. Thus, more service capacity (proportional to the arrival rates) to queue 1 means that the service capacity is more efficiently utilized. Unfortunately, our examples do not provide enough information for us to find a simple, explicit allocation of (μ_1, μ_2) to minimize the mean total queue length or to maximize the system idle probability.

If the two $M/M/1$ queues operate independently, the probability that both queues are empty at the same time is $\max\{0, 1 - \rho_1\} \cdot \max\{0, 1 - \rho_2\}$. Simple calculations show that systems with transfers have a larger system idle probability, i.e., $\pi_0 \geqslant \max\{0, 1 - \rho_1\} \cdot \max\{0, 1 - \rho_2\}$. This observation is true for all numerical examples. It implies that customer transfers greatly increase the efficiency of the queueing system, which is intuitive.

If the two $M/M/1$ queues operate independently, the mean total queue length is $\rho_1/\max\{0, 1 - \rho_1\} + \rho_2/\max\{0, 1 - \rho_2\}$. Simple calculations show that customer transfers greatly decrease the mean total queue length. This observation is true for all numerical examples. That observation, too, implies that customer transfers greatly increase the efficiency of the queueing system.

Table 1
Transfer rates for example 7.2 with $L = 5$, $K = 3$, $\mu_1 = 1.5$, and $\mu_2 = 2.5$.

λ_1	0.1	0.3	0.5	1	1.125	1.5	2	2.5
$T_{R,1\to 2}$	0.0059	0.0108	0.0182	0.0529	0.0671	0.1258	0.2380	0.3754
$T_{R,2\to 1}$	0.2654	0.2184	0.1755	0.0897	0.0739	0.0393	0.0152	0.0044
T_R	0.2713	0.2292	0.1937	0.1426	0.1410	0.1651	0.2532	0.3798

Table 2
Idle probability, queue length and departure rates for example 7.2 with $L = 5$, $K = 3$, $\mu_1 = 1.5$, and $\mu_2 = 2.5$.

λ_1	0.1	0.3	0.5	1	1.125	1.5	2	2.5
π_0	0.0942	0.0954	0.0955	0.0888	0.0857	0.0741	0.0582	0.0452
q_{mean}	4.3492	4.2976	4.2611	4.2868	4.3240	4.4925	4.7700	5.0238
r_1	0.8785	0.9228	0.9719	1.1104	1.1454	1.2404	1.3315	1.3870
r_2	2.1215	2.0772	2.0281	1.8896	1.8546	1.7596	1.6684	1.6130

Example 7.2. Consider queueing systems with $L = 5$, $K = 3$, $\mu_1 = 1.5$, $\mu_2 = 2.5$, and $\lambda_1 + \lambda_2 = 3$. Since $\rho = 0.75 < 1$, these queueing systems are stable. Table 1 shows that transfer rates for systems with $\lambda_1 = 0.1, 0.3, 0.5, 1, 1.125, 1.5, 2$, and 2.5, respectively. It is clear to see that the transfer rate from queue 1 to queue 2 (or queue 2 to queue 1) is monotone in λ_1. But the total transfer rate is unimodal in λ_1. In fact, this example and many other examples we have tested indicate that T_R is minimized at $\lambda_1^* = \rho\mu_1$ and $\lambda_2^* = \rho\mu_2$, where the two $M/M/1$ queues are balanced with $\rho_1 = \rho_2 = \rho$. In general, the total transfer rate is minimized at an allocation of the arrival rates for which the system is either balanced or close to be balanced.

Numerical results for π_0, q_{mean}, r_1, and r_2 are shown in table 2. The functions r_1 and r_2 are monotone in λ_1, which is intuitive. The functions π_0 and q_{mean} are unimodal in λ_1. In fact, π_0 is maximized around $\lambda_1 = 0.4$, where the queue 2 is unstable, and q_{mean} is minimized around $\lambda_1 = 0.6$. It implies that, if the transfer policy is given, then it can be optimal (with respect to system idle probability or the mean total queue length) to route customers to queue 2 even if queue 2 is in heavy traffic (while queue 1 is still in light traffic). This observation is counterintuitive and is quite interesting in the design of such queueing systems.

Next, we consider the impact of the parameters $\{L, K\}$ of the transfer policy on system descriptors. When L increases, intuitively, the capability to adjust the use of the two servers is down. Thus, when L increases, transfer rates are decreasing, the total mean queue length is increasing, and the idle probability is decreasing. The impact of the parameter K on system descriptors is not easy to predict. We consider three cases based on the relationship between ρ_1 and ρ_2: $0 < \rho_1 \approx \rho_2 < 1$ (example 7.3); $0 < \rho_1 < 1 < \rho_2$ (example 7.4); and $0 < \rho_1, \rho_2 < 1$ and ρ_1 is significantly smaller than ρ_2 (example 7.5).

Table 3
Transfer rates for example 7.3 with $L = 9$, $\lambda_1 = 1$, $\lambda_2 = 2$, $\mu_1 = 1.5$, and $\mu_2 = 2.5$.

K	1	2	3	4	5	6	7	8
$T_{R,1\to 2}$	0.0084	0.0057	0.0051	0.0053	0.0060	0.0076	0.0111	0.0218
$T_{R,2\to 1}$	0.0490	0.0291	0.0228	0.0200	0.0189	0.0190	0.0212	0.0304
T_R	0.0574	0.0348	0.0279	0.0253	0.0249	0.0266	0.0323	0.0522
N_{TR}	0.0191	0.0232	0.0279	0.0337	0.0415	0.0533	0.0753	0.1393

Table 4
Idle probability, queue length and departure rates for example 7.3 with $L = 9$, $\lambda_1 = 1$, $\lambda_2 = 2$, $\mu_1 = 1.5$, and $\mu_2 = 2.5$.

K	1	2	3	4	5	6	7	8
π_0	0.0735	0.0743	0.0747	0.0746	0.0742	0.0735	0.0726	0.0717
q_{mean}	5.0803	5.0250	4.9942	4.9873	5.0036	5.0408	5.0956	5.1641
r_1	1.0406	1.0469	1.0531	1.0591	1.0644	1.0684	1.0702	1.0686
r_2	1.9594	1.9531	1.9469	1.9409	1.9356	1.9316	1.9298	1.9314

Example 7.3. Consider queueing systems with $L = 9$, $\lambda_1 = 1$, $\lambda_2 = 2$, $\mu_1 = 1.5$, $\mu_2 = 2.5$, and $1 \leqslant K \leqslant 8$. Since $\rho = 0.75 < 1$, these queueing systems are stable. In fact, the two $M/M/1$ queues are also stable (if they operate independently) with moderate relative traffic intensities: $\rho_1 = 0.6667$ and $\rho_2 = 0.8$, respectively.

Table 3 shows that the average number of transfers per customer N_{TR} is monotone in K, which is intuitive. None of the transfer rates is monotone. Instead, these functions are unimodal. The total transfer rate is minimized at 5. Note that all the descriptors have a relatively big change in value when K goes from 1 to 2 or from 7 to 8. The reason is that, if $K = 1$ or 8, a transfer of customers only makes the two $M/M/1$ queues a little bit more balanced. This observation shows that K should not be chosen too close to 1 nor L if ρ_1 and ρ_2 are close. Usually, $K \approx L/2$ is a good choice for this case.

Nonetheless, *all our examples* show that the average number of transfers per customer (N_{TR}) is always increasing in K. Thus, a small transfer batch size implies that an arbitrary customer will experience fewer transfers, though the total transfer rate can be larger.

Table 4 shows that the impact of K on π_0, q_{mean}, r_1, and r_2, is insignificant. For r_1 and r_2, that observation is *generally* true. For π_0 and q_{mean}, the impact can be significant if the difference of the relative traffic intensities of the two $M/M/1$ queues is significant. Table 4 shows that a moderate K ($\approx L/2$) is a good choice for maximizing π_0 or minimizing q_{mean}.

As the impact of K on $T_{R,1\to 2}$, $T_{R,2\to 1}$, N_{TR}, r_1, and r_2 is either predictable or insignificant, in the next two examples, we concentrate on T_R, π_0, and q_{mean}.

Example 7.4. Consider queueing systems with $L = 9$, $\lambda_1 = 1$, $\lambda_2 = 2$, $\mu_1 = 3.5$, $\mu_2 = 0.5$, and $1 \leqslant K \leqslant 8$. Since $\rho = 0.75 < 1$, these queueing systems are stable.

Table 5

Numerical results for example 7.4 with $L = 9$, $\lambda_1 = 1$, $\lambda_2 = 2$, $\mu_1 = 3.5$, and $\mu_2 = 0.5$.

K	1	2	3	4	5	6	7	8
T_R	1.5000	0.7500	0.5000	0.3750	0.3001	0.25033	0.2168	0.2163
π_0	0.00000	0.00001	0.00003	0.00010	0.00029	0.00077	0.00170	0.00295
q_{mean}	10.5838	10.1979	9.8556	9.5400	9.2435	8.9636	8.7064	8.5102

Table 6

Numerical results for example 7.5 with $L = 9$, $\lambda_1 = 1$, $\lambda_2 = 2$, $\mu_1 = 10$, and $\mu_2 = 2.1$.

K	1	2	3	4	5	6	7	8
T_R	0.18006	0.09675	0.06939	0.05612	0.04857	0.04396	0.04118	0.04176
π_0	0.12002	0.12578	0.13207	0.13904	0.14678	0.1551	0.16326	0.1692
q_{mean}	3.8330	3.6352	3.4551	3.2917	3.1472	3.0253	2.9314	2.8780

Since $\rho_1 = 2/7$ and $\rho_2 = 4$, queue 1 is stable and queue 2 is unstable (if they operate independently). The difference of the relative traffic intensities is about 3.7, which is significant. The queue length of queue 2 tends to grow rapidly. Thus, we expect frequent transfers of customers from queue 1 to queue 2. We also expect that a larger transfer batch size K will make the two queues more balanced (i.e., reduce the total number of transfers). These conjectures are all confirmed in table 5.

Table 5 shows that T_R, π_0, and q_{mean} are all monotone in K and a large transfer batch size $K = L - 1$ is preferred for T_R, π_0, and q_{mean}. The changes in these system descriptors are significant with respect to K. A possible interpretation is that if the two $M/M/1$ queues are dramatically different (since ρ_2 is significantly larger than ρ_1), a large transfer batch size is necessary to keep the two queues more balanced.

In example 7.4, queue 2 is unstable. Next, we consider an example with $0 < \rho_1$, $\rho_2 < 1$. For this example, both $M/M/1$ queues are stable, but ρ_1 is significantly smaller than ρ_2.

Example 7.5. Consider queueing systems with $L = 9$, $\lambda_1 = 1$, $\lambda_2 = 2$, $\mu_1 = 10$, $\mu_2 = 2.1$, and $1 \leqslant K \leqslant 8$. Since $\rho \approx 0.25 < 1$, these queueing systems are stable. Since $\rho_1 = 0.1$ and $\rho_2 \approx 0.95$, both $M/M/1$ queues are stable, but queue 1 has a light traffic and queue 2 has a (relatively) heavy traffic (if they operate independently). The difference of the traffic intensities is about 0.85. Table 6 presents the results for the queueing system.

Table 6 shows that π_0 and q_{mean} are all monotone in K. Thus, a larger transfer batch size $K = L - 1$ is preferred for π_0 and q_{mean}. As for T_R, it is decreasing until $K = 8$, where its value increases a little bit. Since the increase is small, it is reasonable to say that a large $K = L - 1$ is preferred for T_R as well. The conclusions we draw from this example is consistent with that of example 7.4. Therefore, combining with example 7.3, it is reasonable to say that the selection of K depends largely on the difference of the relative traffic intensities ρ_1 and ρ_2.

In summary, the design of the queueing system has much to do with the relative traffic intensities $\{\rho_1, \rho_2\}$ of the two $M/M/1$ queues. The analysis on the service capacity allocation shows that, for many of the descriptors of interest, $\mu_1^* = \lambda_1/\rho$ and $\mu_2^* = \lambda_2/\rho$ is a good choice. For this allocation, the two $M/M/1$ queues have the same traffic intensity and are balanced. In fact, this allocation is near-optimal for the average total transfer rate. However, for some other system descriptors (such as the system idle probability and the mean total queue length), a balanced system may not be the best choice. The analysis on customer routing shows that, in order to minimize the total transfer rate, we should have a balanced or close to a balanced queueing system. However, in order to optimize some other system descriptors, an unbalanced system may be preferred. Furthermore, the system idle probability can be maximized or the mean total queue length can be minimized even when one of the two $M/M/1$ queues is unstable (if it operates independently). The analysis on K shows that a good choice of K depends largely on the difference of the relative traffic intensities ρ_1 and ρ_2. If the difference is significant, a large $K(\approx L-1)$ is preferred; otherwise, a moderate K $(\approx L/2)$ is preferred. In addition, a system with two $M/M/1$ queues with customer transfers, if compared to a system with two $M/M/1$ queues without customer transfers, always has a larger system idle probability and a smaller mean total queue length.

Appendix A. Transition blocks in Q_1

First, note that the matrices are indexed by $J(t) = q_1(t) - q_2(t)$, which has the values $\{L-2, L-4, \ldots, -(L-4), -(L-2)\}$ and $\{L-1, L-3, \ldots, -(L-3), -(L-1)\}$. The matrices given below represent the transition rates between individual levels.

For the level 0 of Q_1, we have $A_{0,1} = -(\lambda_1 + \lambda_2)$ and $A_{0,0} = (\lambda_1, \lambda_2)$.

For the level n of Q_1 $(1 \leqslant n \leqslant L-2)$, no transition triggers a transfer. The matrix $A_{n,2}$ is an $(n+1) \times n$ matrix with $(A_{n,2})_{n-k,n-k-1} = \mu_1$ and $(A_{n,2})_{n-k-2,n-k-1} = \mu_2$ for $k = 0, 2, \ldots, 2n-2$, and all other elements zero:

$$A_{n,2} = \begin{array}{cccccc} & n-1 & n-3 & \ldots & -(n-3) & -(n-1) \\ n & \mu_1 & & & & \\ n-2 & \mu_2 & \mu_1 & & & \\ \vdots & & \ddots & \ddots & & \\ \vdots & & & \ddots & \ddots & \\ -(n-2) & & & & \mu_2 & \mu_1 \\ -n & & & & & \mu_2 \end{array}. \tag{A.1}$$

The matrix $A_{n,1}$ is an $(n+1) \times (n+1)$ matrix with $(A_{n,1})_{n,n} = -\phi + \mu_2$, $(A_{n,1})_{-n,-n} = -\phi + \mu_1$, $(A_{n,1})_{n-k,n-k} = -\phi$ for $k = 2, 4, \ldots, 2n-2$, and all other elements zero, where $\phi = \lambda_1 + \lambda_2 + \mu_1 + \mu_2$. Note that queue 1 is empty if $J(t) = -n$ and queue 2 is empty if $J(t) = n$.

$$A_{n,1} = \begin{array}{c} \\ n \\ n-2 \\ \vdots \\ -(n-2) \\ -n \end{array} \begin{array}{c} \begin{array}{ccccc} n & n-2 & \dots & -(n-2) & -n \end{array} \\ \left(\begin{array}{ccccc} -\phi+\mu_2 & & & & \\ & -\phi & & & \\ & & \ddots & & \\ & & & -\phi & \\ & & & & -\phi+\mu_1 \end{array} \right) \end{array}. \tag{A.2}$$

The matrix $A_{n,0}$ is an $(n+1) \times (n+2)$ matrix with $(A_{n,0})_{n-k,n-k+1} = \lambda_1$ and $(A_{n,0})_{n-k,n-k-1} = \lambda_2$ for $k = 0, 2, 4, \dots, 2n$, and all other elements zero:

$$A_{n,0} = \begin{array}{c|cccccc} & n+1 & n-1 & \dots & \dots & -(n-1) & -(n+1) \\ \hline n & \lambda_1 & \lambda_2 & & & & \\ n-2 & & \lambda_1 & \ddots & & & \\ \vdots & & & \ddots & \ddots & & \\ -(n-2) & & & & \ddots & \lambda_2 & \\ -n & & & & & \lambda_1 & \lambda_2 \end{array}. \tag{A.3}$$

Note that, since $X(t) < L - 1$, the difference of the two queues is less than $L - 1$. Therefore, there is no transfer after an arrival or a service completion.

For the level $L - 1$ of Q_1, the matrices $A_{L-1,2}$ and $A_{L-1,1}$ are obtained by letting $n = L - 1$ in (A.1) and (A.2). The matrix $A_{L-1,0}$ is different from that of (A.3) since in this case an arrival may trigger a transfer. Note that $A_{L-1,0}$ is an $L \times (L - 1)$ matrix since the level L (of Q_1) has only $L - 1$ states. The elements of $A_{L-1,0}$ are given as: $(A_{L-1,0})_{L-k,L-k+1} = \lambda_1$ for $k = 3, 5, \dots, 2L - 1$; $(A_{L-1,0})_{L-k,L-k-1} = \lambda_2$ for $k = 1, 3, \dots, 2L - 3$; $(A_{L-1,0})_{L-1,L-2K} = \lambda_1$ (if $K = 1$, add λ_1 to the element $(A_{L-1,0})_{L-1,L-2}$); $(A_{L-1,0})_{-(L-1),2K-L} = \lambda_2$ (if $K = 1$, add λ_2 to the element $(A_{L-1,0})_{-(L-1),2-L}$); and all other elements are zero:

$$A_{L-1,0} = \begin{array}{c|cccccc} & L-2 & L-4 & \dots & \dots & -(L-4) & -(L-2) \\ \hline L-1 & \lambda_2 & & & \lambda_1 & & \\ L-3 & \lambda_1 & \lambda_2 & & & & \\ \vdots & & \ddots & \ddots & & & \\ \vdots & & & \ddots & \ddots & & \\ \vdots & & & & \ddots & \ddots & \\ -(L-3) & & & & & \lambda_1 & \lambda_2 \\ -(L-1) & & & \lambda_2 & & & \lambda_1 \end{array}. \tag{A.4}$$

For the level L of Q_1, neither the arrival of a customer nor the completion of a service will trigger a transfer of customers. The matrix $A_{L,2}$ is an $(L - 1) \times L$ matrix with $(A_{L,2})_{L-k,L-k+1} = \mu_2$ and $(A_{L,2})_{L-k,L-k-1} = \mu_1$ for $k = 2, 4, \dots, 2L-2$ and all other elements are zero:

$$A_{L,2} = \begin{array}{c} \\ L-2 \\ L-4 \\ \vdots \\ -(L-4) \\ -(L-2) \end{array} \begin{array}{c} \begin{array}{cccccc} L-1 & L-3 & \dots & \dots & -(L-3) & -(L-1) \end{array} \\ \left(\begin{array}{cccccc} \mu_2 & \mu_1 & & & & \\ & \mu_2 & \ddots & & & \\ & & \ddots & \ddots & & \\ & & & \mu_2 & \mu_1 & \\ & & & & \mu_2 & \mu_1 \end{array}\right) \end{array}. \qquad (A.5)$$

The matrix $A_{L,1}$ is an $(L-1) \times (L-1)$ matrix with $(A_{L,1})_{L-k,L-k} = -\phi$ for $k = 2, 4, \dots, 2L-2$ and all other elements are zero:

$$A_{L,1} = \begin{array}{c} \\ L-2 \\ L-4 \\ \vdots \\ -(L-4) \\ -(L-2) \end{array} \begin{array}{c} \begin{array}{ccccc} L-2 & L-4 & \dots & -(L-4) & -(L-2) \end{array} \\ \left(\begin{array}{ccccc} -\phi & & & & \\ & -\phi & & & \\ & & \ddots & & \\ & & & -\phi & \\ & & & & -\phi \end{array}\right) \end{array} = -\phi I. \qquad (A.6)$$

The matrix $A_{L,0}$ is an $(L-1) \times L$ matrix with $(A_{L,0})_{L-k,L-k+1} = \lambda_1$ and $(A_{L,0})_{L-k,L-k-1} = \lambda_2$ for $k = 0, 2, 4, \dots, 2L-2$, and all other elements are zero:

$$A_{L,0} = \begin{array}{c} \\ L-2 \\ L-4 \\ \vdots \\ -(L-4) \\ -(L-2) \end{array} \begin{array}{c} \begin{array}{cccccc} L-1 & L-3 & \dots & \dots & -(L-3) & -(L-1) \end{array} \\ \left(\begin{array}{cccccc} \lambda_1 & \lambda_2 & & & & \\ & \lambda_1 & \lambda_2 & & & \\ & & \ddots & \ddots & & \\ & & & \lambda_1 & \lambda_2 & \\ & & & & \lambda_1 & \lambda_2 \end{array}\right) \end{array}. \qquad (A.7)$$

For the level $L+1$ of Q_1, both the arrival of a customer and the completion of a service may trigger a transfer. The matrix $A_{L+1,2}$ is an $L \times (L-1)$ matrix with $(A_{L+1,2})_{L-k-2,L-k-1} = \mu_2$ and $(A_{L+1,2})_{L-k,L-k-1} = \mu_1$ for $k = 1, 3, \dots, 2L-3$; $(A_{L+1,2})_{L-1,L-2K} = \mu_2$ (if $K = 1$, add μ_2 to $(A_{L+1,2})_{L-1,L-2}$); $(A_{L+1,2})_{-(L-1),2K-L} = \mu_1$ (if $K = 1$, add μ_1 to $(A_{L+1,2})_{-(L-1),2-L}$); and all other elements are zero:

$$A_{L+1,2} = \begin{array}{c} \\ L-1 \\ L-3 \\ \vdots \\ \vdots \\ \vdots \\ -(L-3) \\ -(L-1) \end{array} \begin{array}{c} \begin{array}{cccccc} L-2 & L-4 & \dots & \dots & -(L-4) & -(L-2) \end{array} \\ \begin{pmatrix} \mu_1 & & & \mu_2 & & \\ \mu_2 & \mu_1 & & & & \\ & \ddots & \ddots & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & \mu_2 & \mu_1 \\ & & \mu_1 & & & \mu_2 \end{pmatrix} \end{array}. \tag{A.8}$$

The matrix $A_{L+1,1}$ is an $L \times L$ matrix with $(A_{L+1,1})_{L-k,L-k} = -\phi$ for $k = 1, 3, \dots, 2L-1$ and all other elements are zero:

$$A_{L+1,1} = \begin{array}{c} \\ L-1 \\ L-3 \\ \vdots \\ -(L-3) \\ -(L-1) \end{array} \begin{array}{c} \begin{array}{ccccc} L-1 & L-3 & \dots & -(L-3) & -(L-1) \end{array} \\ \begin{pmatrix} -\phi & & & & \\ & -\phi & & & \\ & & \ddots & & \\ & & & -\phi & \\ & & & & -\phi \end{pmatrix} \end{array} = -\phi I. \tag{A.9}$$

The matrix $A_{L+1,0}$ is an $L \times (L-1)$ matrix and $A_{L+1,0} = A_{L-1,0}$.

For $n \geqslant 0$ and $i = 0, 1, 2$, we have $A_{L+2n,i} = A_{L,i}$ and $A_{L+2n+1,i} = A_{L+1,i}$.

Appendix B. Transition blocks in Q_2

Since the levels of the Markov process $\{(X(t), J(t)),\ t \geqslant 0\}$ are obtained by grouping the levels of $\{(q(t), J(t)),\ t \geqslant 0\}$, the blocks in Q_2 can be constructed from that of Q_1 in the following way:

$$\begin{aligned} &A^*_{L-1,0} = (A_{L-1,0}, 0); \qquad A^*_{L,2} = \begin{pmatrix} A_{L,2} \\ 0 \end{pmatrix}; \\ &A_2 = \begin{pmatrix} 0 & A_{L,2} \\ 0 & 0 \end{pmatrix}; \qquad A_1 = \begin{pmatrix} A_{L,1} & A_{L,0} \\ A_{L+1,2} & A_{L+1,1} \end{pmatrix}; \qquad A_0 = \begin{pmatrix} 0 & 0 \\ A_{L+1,0} & 0 \end{pmatrix}. \end{aligned} \tag{B.1}$$

Note that there are $2L-1$ states in the level $L+n$ of Q_2 for $n \geqslant 0$.

Appendix C. Formulas for the first passage time analysis

First, we have

$$\begin{aligned} &A^*_{L,2} + \big(sI + A_1(z_1, z_2)\big)G^*_L(s, z_1, z_2) + A_0(z_1, z_2)G^*(s, z_1, z_2)G^*_L(s, z_1, z_2) = 0; \\ &A_{L-1,2} + (sI + A_{L-1,1})G^*_{L-1}(s, z_1, z_2) + A^*_{L-1,0}(z_1, z_2)G^*_L(s, z_1, z_2)G^*_{L-1}(s, z_1, z_2) \\ &= 0; \end{aligned}$$

$$\begin{aligned}&A_{n,2} + (sI + A_{n,1})G_n^*(s, z_1, z_2) + A_{n,0}G_{n+1}^*(s, z_1, z_2)G_n^*(s, z_1, z_2) = 0,\\ &\quad 1 \leqslant n \leqslant L-2;\\ &G_0^*(s, z_1, z_2) = -(sI + A_{0,1})^{-1}A_{0,0}G_1^*(s, z_1, z_2),\end{aligned} \tag{C.1}$$

where $A^*_{L-1,0}(z_1, z_2) = A^*_{L-1,0} + (z_1 - 1)B_{L-1,1,0} + (z_2 - 1)B_{L-1,2,0}$, $B_{L-1,1,0}$ and $B_{L-1,2,0}$ are $L \times (2L-1)$ nonnegative matrices with only one positive element: $(B_{L-1,1,0})_{L-1,L-2K} = \lambda_1$ and $(B_{L-1,2,0})_{-(L-1),2K-L} = \lambda_2$.

By routine calculations, the matrices $\{G_n\}$ can be obtained as follows:

$$\begin{aligned}&G_L = -(A_1 + A_0G)^{-1}A^*_{L,2};\\ &G_{L-1} = -(A_{L-1,1} + A^*_{L-1,0}G_L)^{-1}A_{L-1,2};\\ &G_n = -(A_{n,1} + A_{n,0}G_{n+1})^{-1}A_{n,2}, \quad 1 \leqslant n \leqslant L-2;\\ &G_0 = -A_{0,1}^{-1}A_{0,0}G_1.\end{aligned} \tag{C.2}$$

The vectors $\mathbf{u}_s(n)$, $\mathbf{u}_1(n)$, and $\mathbf{u}_2(n)$ can be calculated as follows:

$$\begin{aligned}&\mathbf{u}_s(L) = \mathbf{u}_s;\\ &\mathbf{u}_s(L-1) = -\left(A_{L-1,1} + A^*_{L-1,0}G_L\right)^{-1}\left(\mathbf{e} + A^*_{L-1,0}\mathbf{u}_s(L)\right);\\ &\mathbf{u}_s(n) = -(A_{n,1} + A_{n,0}G_{n+1})^{-1}\left(\mathbf{e} + A_{n,0}\mathbf{u}_s(n+1)\right), \quad 1 \leqslant n \leqslant L-2;\\ &u_s(0) = -(A_{0,1})^{-1}\left(1 + A_{0,0}\mathbf{u}_s(1)\right);\\ &\mathbf{u}_1(L) = -(A_1 + A_0G)^{-1}\left[(B_{1,1} + B_{1,0})\mathbf{e} + A_0\mathbf{u}_1\right];\\ &\mathbf{u}_1(L-1) = -\left(A_{L-1,1} + A^*_{L-1,0}G_L\right)^{-1}\left[B_{L-1,1,0}\mathbf{e} + A^*_{L-1,0}\mathbf{u}_1(L)\right];\\ &\mathbf{u}_1(n) = -(A_{n,1} + A_{n,0}G_{n+1})^{-1}A_{n,0}\mathbf{u}_1(n+1), \quad 1 \leqslant n \leqslant L-2;\\ &u_1(0) = -(A_{0,1})^{-1}A_{0,0}\mathbf{u}_1(1);\\ &\mathbf{u}_2(L) = -(A_1 + A_0G)^{-1}\left[(B_{2,1} + B_{2,0})\mathbf{e} + A_0\mathbf{u}_2\right];\\ &\mathbf{u}_2(L-1) = -\left(A_{L-1,1} + A^*_{L-1,0}G_L\right)^{-1}\left[B_{L-1,2,0}\mathbf{e} + A^*_{L-1,0}\mathbf{u}_2(L)\right];\\ &\mathbf{u}_2(n) = -(A_{n,1} + A_{n,0}G_{n+1})^{-1}A_{n,0}\mathbf{u}_2(n+1), \quad 1 \leqslant n \leqslant L-2;\\ &u_2(0) = -(A_{0,1})^{-1}A_{0,0}\mathbf{u}_2(1).\end{aligned} \tag{C.3}$$

Acknowledgements

This work was supported in part by NSF Grant Nr. DMI-9988749 (MFN) and by NSERC (Canada).

References

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

[2] W. Cohen, *Single Server Queues* (North-Holland, Amsterdam, 1980).

[3] R.D. Foley and D.R. McDonald, Join the shortest queue: Stability and exact asymptotics (1998) submitted for publication.

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

[5] I.A. Kurkova and Yu.M. Suhov, Malyshev's theory and JS-queues: Asymptotics of stationary probabilities (2001) submitted for publication.

[6] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modelling* (ASA/SIAM, Philadelphia, PA, 1999).

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

[8] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach* (Johns Hopkins Univ. Press, Baltimore, MD, 1981).

[9] M.F. Neuts, The caudal characteristic curve of queues, Adv. in Appl. Probab. 18 (1986) 221–254.

[10] M.F. Neuts, *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications* (Marcel Dekker, New York, 1989).

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

[12] S.H. Xu and H. Chen, On the asymptote of the optimal routing policy for two service stations, IEEE Trans. Automat. Control 38 (1993) 187–189.

[13] S.H. Xu, R. Righter and J.G. Shanthikumar, Optimal dynamic assignment of customers to heterogeneous servers in parallel, Oper. Res. 40 (1992) 1126–1138.

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

[15] Y. Zhao and W.K. Grassmann, The shortest queue model with jockeying, Naval Res. Logistics 37 (1990) 773–783.

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