Computational Analysis of $MMAP[K]/PH[K]/1$ Queues with a Mixed FCFS and LCFS Service Discipline

Qi-Ming HE,[1] Attahiru Sule Alfa[2]

[1] *Department of Industrial Engineering, DalTech, Dalhousie University, Halifax, Nova Scotia, B3J 2X4, Canada*

[2] *Department of Industrial and Manufacturing Systems Engineering, University of Windsor, Windsor, Ontario, N9B 3P4, Canada*

Received February 1999; September 1999; accepted 16 February 2000

Abstract: This paper studies a queueing system with a Markov arrival process with marked arrivals and PH-distribution service times for each type of customer. Customers (regardless of their types) are served on a mixed first-come-first-served (FCFS) and last-come-first-served (LCFS) nonpreemptive basis. That is, when the queue length is N (a positive integer) or less, customers are served on an FCFS basis; otherwise, customers are served on an LCFS basis. The focus is on the stationary distribution of queue strings, busy periods, and waiting times of individual types of customers. A computational approach is developed for computing the stationary distribution of queue strings, the mean of busy period, and the means and variances of waiting times. The relationship between these performance measures and the threshold number N is analyzed in depth numerically. It is found that the variance of the virtual (actual) waiting time of an arbitrary customer can be reduced by increasing N. © 2000 John Wiley & Sons, Inc. Naval Research Logistics 47: 399–421, 2000

Keywords: queueing theory; matrix analytic methods; tree structure; FCFS; LCFS; QBD Markov process; queue string; waiting time

1. INTRODUCTION

This paper studies a single server queueing system with a Markovian arrival process with K types of customers. The service times of different types of customers may have different probability distributions. Customers, regardless of their types, join the same queue according to the order they arrived. The server picks up customers to be served based on a mixed first-come-first-served (FCFS) and last-come-first-served (LCFS) service discipline. By this, the server attends to the queue according to FCFS when there are no more than N (a positive integer) waiting customers. However, if the number of waiting customers exceeds N, customers will be served according to the LCFS service discipline. This type of service discipline has potential applications in telecommunication systems.

In telecommunication systems, customers' reactions to large delays during the call setup phase, such as waiting for dial tone and post dialing delays, may lead to them abandoning their calls.

Correspondence to: Q.-M. HE

© 2000 John Wiley & Sons, Inc.

This often leads to a reduced throughput and inefficient use of the system because of the wasted work created by customers who abandoned their calls. In addition, some of these customers who abandoned their calls may reattempt the calls later, thereby inflating the overload. Different versions of the LCFS queueing disciplines have been adopted for controlling overload in some telephone switching systems where customers show this type of behavior. However, LCFS is known to produce high variance of waiting times. FCFS discipline, on the other hand, is fair and has low variance of waiting times when compared to LCFS. One of the earliest papers dealing with such problems is Forys [4], where an LCFS service discipline is proposed for the call processing mechanism of a telephone switching system in order to keep the throughput high. There has been a continued attempt to improve waiting time experienced by customers, in situations where customers have a threshold associated with waiting time. Several ideas have been tried to deal with this. One of them is the hybrid LCFS/FCFS system (Alfa and Fitzpatrick [1]—$Geo/D/1$, Doshi [3]—$M/G/1$, Schreiber [14]—$M/M/1$). However, all the existing literature focus on the case in which every customer goes through FCFS, and some get reshuffled according to LCFS before going into the FCFS queue. In this paper we consider the case where some customers go through LCFS only, and others face both LCFS and FCFS, i.e., a customer, while waiting, could see some customers getting served according to LCFS and FCFS. This is a major difference between our model and the ones in the literature. In addition we are considering K types of customers and it is an $MMAP[K]/PH/[K]/1$ system. Most of the work in the literature considers cases with only one type of customer.

One main application of this work is in the area of Advanced Intelligent Networks (AIN). Subscribers in an AIN are usually of different classes, such as free phone (800 and 888 number) users, credit card verifiers, etc. A subscriber submits a request for service (queues up), and information is requested of the subscriber when service begins. Some subscribers may abandon their request if response is not received within a short period of time. For those subscribers who do not abandon at this stage, they receive the first stage of service in terms of request for information. While the subscriber is preparing this information it is preempted for another subscriber. This may lead to another wait and the potential that this subscriber abandons service is still likely. If a subscriber abandons at this stage then the initial work carried out is wasted leading to an inefficient use of the resources. There could be second, third, ... stage services also, and a subscriber may abandon service during any of the waiting periods. The idea of a hybrid LCFS&FCFS queue discipline in the $MMAP[K]/PH[K]/1$ system may be used to determine how to control such systems effectively in order to minimize wasted work.

The main contributions of this paper are: (1) modeling: the queueing model studied in this paper is new; (2) a computation approach is developed for computing stationary distributions related to queue lengths (at an arbitrary time, right after an arbitrary arrival, or right after a type k arrival), mean busy periods, and means and variances of waiting times of different types of customers; (3) the relationship between the threshold number N and these performance measures. Of special interest to us in this paper is to see how the variance of the waiting time is affected by N. This way we can choose N to control the variance of the waiting time in a queueing system.

The rest of the paper is organized as follows. In Section 2, the queueing system of interest is introduced. In Section 3, a Markov process is constructed for the queueing system of interest. Section 4 presents results for the stationary distribution of queue string and develops an algorithm for computing the stationary distribution. Section 5 deals with fundamental periods and busy periods. Section 6 studies waiting times. The Laplace Stieltjes transforms, means, and variances of waiting times of individual types of customers and an arbitrary customer are obtained. In Section 7, a number of numerical examples are presented to show the relationship between N and various performance measures. In Section 8, we summarize the results obtained in this paper. In addition, blocks in transition matrices and a number of formulas are given in Appendices A, B, and C.

2. THE $MMAP[K]/PH[K]/1$/FCFS&LCFS QUEUE

This section introduces a single server queueing system with a Markov arrival process with marked transitions $(MMAP[K])$ and phase-type service times. Customers are distinguished into K types. The service times of different types of customers may have different probability distributions. All types of customers are served on a mixed FCFS and LCFS nonpreemptive basis. To define the queueing system of interest explicitly, the input process $MMAP[K]$ is introduced first and then the service time distributions and the service discipline are specified.

The Markov arrival process with marked transitions is defined by a set of $m \times m$ matrices $\{D_k, 0 \leq k \leq K\}$. The matrices $D_k, 1 \leq k \leq K$, are nonnegative. The matrix D_0 has negative diagonal elements and nonnegative off-diagonal elements. D_0 is assumed to be nonsingular. Let

$$D = D_0 + \sum_{k=1}^{K} D_k. \tag{1}$$

Then matrix D is the infinitesimal generator of the underlying Markov process. Let $I(t)$ be the phase of the underlying Markov process at time t. An arrival is called a type k arrival if the arrival is marked by k. The (matrix) marking rate of type k arrivals is D_k. Let θ be the stationary probability vector of the matrix D. The stationary arrival rate of type k arrivals is given by $\lambda_k = \theta D_k \mathbf{e}, 1 \leq k \leq K$. (See Asmussen and Koole [2], HE and Neuts [9], and Neuts [11] for more about $MMAP[K]$.) Customers join the queue according to the order they arrived.

The service times of type k customers have a common phase-type distribution (PH-distribution) function with a matrix representation (α_k, T_k), where α_k is an m_k-dimensional nonnegative vector with $\alpha_k \mathbf{e} = 1$, and T_k is an $m_k \times m_k$ matrix. Let $\mathbf{T}_k^0 = -T_k \mathbf{e}$. The mean service time is given by $1/\mu_k = -\alpha_k T_k^{-1} \mathbf{e}$. Then μ_k is the average service rate of type k customers. For (α_k, T_k), its associated Markov process has m_k phases. For more details about PH-distribution, see Chapter 2 in Neuts [12]. We assume that service times are independent of each other and are independent of the $MMAP[K]$.

The queue string and queue length: The queue string for the queueing system of interest consists of all the customers in the queue waiting for service (the customer in service is not included). Customers join the queue according to the order they arrived. Define the queue length as the number of customers in the queue string.

A mixed FCFS&LCFS nonpreemptive service discipline: When the queue length (number of waiting customers) is equal to or less than a threshold value N (a positive integer), customers are served on an FCFS basis. When the queue length is larger than N, customers are served on an LCFS nonpreemptive basis. In other words, when the queue length is N or less, the next customer to be served is the customer who arrived first. When the queue length becomes larger than N, the next customer to be served is the customer who arrived last.

Two special cases are of particular interest to application. When $N = 1$, the queueing system of interest becomes an LCFS nonpreemptive queueing system. When $N = \infty$, it is an FCFS queueing system. While the $N = 1$ case has been studied in HE [7], the $N = \infty$ case has been studied in HE [6]. For more information about various service disciplines of queueing systems, see Takagi [15].

The traffic intensity of the queueing system is defined as $\rho = \lambda_1/\mu_1 + \cdots + \lambda_K/\mu_K$. It is assumed that $\rho < 1$ throughout this paper to ensure the stability of the queueing system.

3. THE QBD MARKOV PROCESS WITH A NETWORK STRUCTURE

In this section, the queueing process associated with the $MMAP[K]/PH[K]/1$/FCFS&LCFS nonpreemptive queue is formulated into a quasi-birth-and-death (QBD) Markov process with a network structure. See HE [7], HE and Alfa [8], Takine, Sengupta, and Yeung [16], Yeung and Sengupta [17], and Yeung and Alfa [18] for more about Markov processes with a tree or network structure and their applications.

Let $\aleph = \{J\colon J = k_1k_2\cdots k_n, 1 \le k_i \le K, 1 \le i \le n, n \ge 1\} \cup \{-1, 0\}$. The length of a string J in $\aleph$ is defined as the number of integers in the string and is denoted by $|J|$. When $J = 0$ or -1, $|J| = 0$. The following two operations related to strings in $\aleph$ are used in this paper.

1. Addition operation: for $J = k_1\cdots k_n \in \aleph$ and $1 \le k \le K$, $J + k = k_1\cdots k_nk \in \aleph$;
2. Subtraction operation: for $J = k_1\cdots k_n \in \aleph$, $J - k_n = k_1\cdots k_{n-1} \in \aleph$;
   for $J = k_1\cdots k_n \in \aleph$, $-k_1 + J = k_2\cdots k_n \in \aleph$.

For example, $21 + 2 = 212, 212 - 2 = 21$, and $-2 + 2212 = 212$. As was defined in Section 2, in this paper, the queue is represented by a string at any time. An integer k in the string represents a customer of type k. The queueing system of interest is represented by the following four dimensional stochastic process:

$q(t)$: the string of customers in queue (exclude the one in server, if any), $q(t) \in \aleph$;
$I_1(t)$: the state of the underlying Markov process $D, 1 \le I_1(t) \le m$;
$I_2(t)$: the type of the customer in service (if any), $1 \le I_2(t) \le K$;
$I_3(t)$: the phase of the PH-distribution of the current service (if any), $1 \le I_3(t) \le m_{I_2(t)}$.

When there is no customer in the system at time t, let $q(t) = -1$. When there is one customer in the system at time $t, q(t) = 0$. When there are customers waiting at time $t, q(t)$ is a string in $\aleph$. The transitions of $q(t)$ are illustrated in Figure 1 for $K = 2$ and $N = 2$. For example, when $q(t) = 21$, there are 2 customers waiting in the system at time t: The customer who arrived first is of type 2 and the customer who arrived second is of type 1. When a new customer of type k arrives before the current service is completed, $q(t)$ becomes $21k$. When the current service is completed first, $q(t)$ becomes $q(t) = 1$ since $|q(t)| = 2 \le N = 2$ and the customer (of type 2) who arrived *first* enters service first. Notice that $|q(t)| = 2$ means that there are two customers waiting in the queue or there are three customers in the queueing system at time t. When $q(t) = 212$ and a new customer of type k arrives before the current service is completed, $q(t)$ becomes $212k$. When the current service is completed first, $q(t)$ becomes 21 since $|q(t)| = |212| = 3 > N = 2$ and the customer of type 2 who arrived *last* enters service first.

It is easy to see that $(q(t), I_1(t), I_2(t), I_3(t))$ is a Markov process with a state space: $\aleph \times \{1, 2, \ldots, m\} \times \{\cup_{k=1}^K \{(k, 1), (k, 2), \ldots, (k, m_k)\}\}$, where $(I_1(t), I_2(t), I_3(t))$ can be defined as an auxiliary random variable with $m\overline{m}$ states (where $\overline{m} = m_1 + \cdots + m_K$), except that when $q(t) = -1$, the auxiliary variable takes values $\{1, 2, \ldots, m\}$. The set of all the states with $|q(t)| = n$ is defined as level n, i.e., the set $\aleph_n = \{(J, i_1, i_2, i_3)\colon |J| = n, J \in \aleph$, and all possible $(i_1, i_2, i_3)\}$ is called the level n, for $n \ge 0$. Level -1 is defined as $\aleph_{-1} = \{(-1, 1), \ldots, (-1, m)\}$.

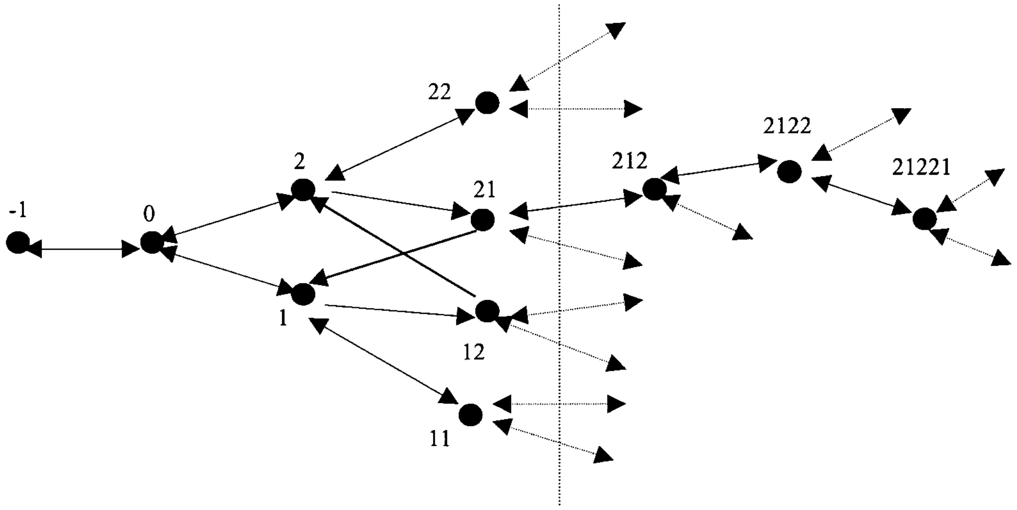

Figure 1. Transitions of $q(t)$ for $K = 2$ and $N = 2$.

The states in level n are arranged lexicographically. The transition law of Markov process $(q(t), I_1(t), I_2(t), I_3(t))$ can be defined at the node level using the transition probabilities (rates) between nodes in the network. The transition law of Markov process $(q(t), I_1(t), I_2(t), I_3(t))$ can also be defined at the aggregate level using transition probabilities between levels. In this paper, we shall use both and switch between them.

When transitions of $(q(t), I_1(t), I_2(t), I_3(t))$ among different aggregate levels are considered, $(q(t), I_1(t), I_2(t), I_3(t))$ can be a quasi-birth-and-death (QBD) Markov process with level dependent transitions and an infinitesimal generator

$$Q = \begin{pmatrix} \bar{A}_1(-1) & \bar{A}_0(-1) & & & & & & \\ \bar{A}_2(0) & \bar{A}_1(0) & \bar{A}_0(0) & & & & & \\ & \bar{A}_2(1) & \bar{A}_1(1) & \bar{A}_0(1) & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & \bar{A}_2(N) & \bar{A}_1(N) & \bar{A}_0(N) & & \\ & & & & \bar{A}_2(N+1) & \bar{A}_1(N+1) & \bar{A}_0(N+1) & \\ & & & & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{2}$$

Matrix blocks in Q are given explicitly in Appendix A. Since the service disciplines for $|q(t)| \leq N$ or $|q(t)| > N$ are different, the nature of transitions of the Markov process $(q(t), I_1(t), I_2(t), I_3(t))$ is different for $|q(t)| \leq N$ and $|q(t)| > N$, which has a major influence on the solution methods.

When $|q(t)| > N, (q(t), I_1(t), I_2(t), I_3(t))$ is a QBD Markov process with a tree structure since customers are served on a LCFS nonpreemptive basis, i.e., when $q(t) = k_1 \cdots k_n \in \aleph$ and $n > N$, if the current service completes before the next arrival, $q(t)$ becomes $q(t) = k_1 \cdots k_{n-1}$. Transitions of $(q(t), I_1(t), I_2(t), I_3(t))$, for $|q(t)| > N$, can be defined at the node level by matrices $\{A_0(k), A_1(k), A_2(k), 1 \leq k \leq K\}$ given in Appendix A and are shown in Figure 2. When $|q(t)| \leq N$, customers are served on an FCFS basis. Transitions of Markov process $(q(t), I_1(t), I_2(t), I_3(t))$ no longer have the tree structure shown in Figure 2. When

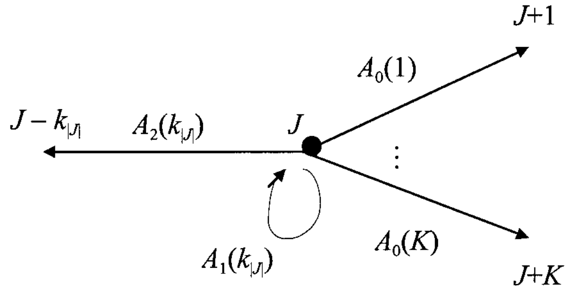

Figure 2. One step transitions when $|q(t)| = |J| > N$.

$q(t) = k_1 \cdots k_n \in \aleph$ and $n \leq N$, if the current service completes before the next arrival, $q(t)$ becomes $q(t) = k_2 \cdots k_n$. The transition law of the Markov process of interest becomes complicated. In this case, we say that the QBD Markov process $(q(t), I_1(t), I_2(t), I_3(t))$ has a network structure.

4. STATIONARY DISTRIBUTIONS

The computation of the stationary distribution of the Markov process $(q(t), I_1(t), I_2(t), I_3(t))$ is complicated since the transitions at $|q(t)| \leq N$ do not have a tree structure. Fortunately, a tree structure does exist when $|q(t)| > N$ and this property can be exploited to device feasible algorithms for computing the stationary distribution. In this section, an algorithm is proposed based on the matrix-geometric solution for QBD Markov processes with level dependent transitions and the matrix-geometric solution for QBD Markov processes with a tree structure.

Let $x_{i,k,j}(J)$ be the steady state probability of state (J, i, k, j) for J in $\aleph$. Let $\mathbf{x}_{i,k}(J) = (x_{i,k,1}(J), \ldots, x_{i,k,m_k}(J)), \mathbf{x}_i(J) = (\mathbf{x}_{i,1}(J), \ldots, \mathbf{x}_{i,K}(J))$, and $\mathbf{x}(J) = (\mathbf{x}_1(J), \ldots, \mathbf{x}_m(J))$, for all J in $\aleph$. Denote by $(\mathbf{x}_{-1}, \mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \cdots)$ the stationary distribution of $(q(t), I_1(t), I_2(t), I_3(t))$, where $\mathbf{x}_n, n \geq -1$, corresponding to states of level n. Then $\mathbf{x}_n = (\mathbf{x}(J)\colon |J| = n)$ which is arranged lexicographically. It is well known that $(\mathbf{x}_{-1}, \mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \cdots)$ satisfies equations:

$$\mathbf{x}_{-1}\bar{A}_1(-1) + \mathbf{x}_0\bar{A}_2(0) = 0,$$

$$\mathbf{x}_n\bar{A}_0(n) + \mathbf{x}_{n+1}\bar{A}_1(n+1) + \mathbf{x}_{n+2}\bar{A}_2(n+2) = 0, \qquad n \geq -1. \tag{3}$$

It is also well known that

$$\mathbf{x}_{n+1} = \mathbf{x}_n\bar{R}(n+1), \qquad n \geq -1, \tag{4}$$

where $\{\bar{R}(n), n \geq 0\}$ are $m\overline{m}K^n \times m\overline{m}K^{n+1}$ matrices and are the minimal nonnegative solutions to

$$0 = \bar{A}_0(n-1) + \bar{R}(n)\bar{A}_1(n) + \bar{R}(n)\bar{R}(n+1)\bar{A}_2(n+1), \qquad n \geq 0. \tag{5}$$

To develop an algorithm from Eqs. (4) and (5) for computing the stationary distribution, the vector $\mathbf{x}_{-1}$ and matrices $\{\bar{R}(n), n \geq 0\}$ must be found first. It is easy to see that the busy periods (or idle periods) of the queueing system with an FCFS&LCFS nonpreemptive service discipline are the same as that of the queueing system with a pure LCFS nonpreemptive service discipline.

Then vector $\mathbf{x}_{-1}$ is the same for both queueing systems. Then $\mathbf{x}_{-1}$ can be found by solving the following equation (see HE [7]):

$$(\mathbf{x}_{-1}, \mathbf{z}_0) \begin{pmatrix} \bar{A}_1(-1) & \bar{A}_0(-1) \\ \bar{A}_2(0) & \bar{A}_1(0) + \sum_{k=1}^{K} R(k)A_2(k) \end{pmatrix} = 0,$$

$$\mathbf{x}_{-1}\mathbf{e} + \mathbf{z}_0(\mathbf{I} - R)^{-1}\mathbf{e} = 0, \tag{6}$$

where $\mathbf{z}_0$ is a nonnegative vector of size $m\overline{m}K$, $\mathbf{I}$ is the identity matrix, $R = R(1) + \cdots + R(K)$, and $\{R(1), \ldots, R(K)\}$ are the minimal nonnegative solution to the following equations:

$$0 = A_0(k) + R(k)A_1(k) + R(k)\sum_{l=1}^{K} R(l)A_2(l), \qquad 1 \le k \le K. \tag{7}$$

For computation of $\{R(1), \ldots, R(K)\}$, see Latouche and Ramaswami [10], Yeung and Sengupta [17], and Yeung and Alfa [18]. For $\{\bar{R}(n), n \ge 0\}$, rewrite Eq. (5) as

$$\bar{R}(n) = -\bar{A}_0(n-1)[\bar{A}_1(n) + \bar{R}(n+1)\bar{A}_2(n+1)]^{-1}, \qquad n \ge 0. \tag{8}$$

Since Markov process $(q(t), I_1(t), I_2(t), I_3(t))$ has a tree structure when $|q(t)| = n > N$, it has

$$\bar{R}(n) = \begin{pmatrix} R(1) \; \cdots \; R(K) & & & \\ & R(1) \; \cdots \; R(K) & & \\ & & \vdots & \\ & & & R(1) \; \cdots \; R(K) \end{pmatrix}, \qquad n \ge N. \tag{9}$$

Thus, matrices $\{\bar{R}(n), n \ge 0\}$ can be found by using equations (7), (8), and (9).

While $\mathbf{x}_n$ with $n \le N$ can be computed using Eq. (4), the computation of $\mathbf{x}_n = (\mathbf{x}(J)$: $|J| = n)$ with $n > N$ is much simpler because of the tree structure. It has, for $|J| \ge N$,

$$\mathbf{x}(J + k) = \mathbf{x}(J)R(k), \qquad 1 \le k \le K. \tag{10}$$

In summary, the following scheme has been developed for computing the stationary distribution $(\mathbf{x}_{-1}, \mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \cdots)$.

Step 1: Input system parameters $\{D_k, 0 \le k \le K\}$ and $\{(\alpha_k, T_k), 1 \le k \le K\}$;
Step 2: Construction of transition blocks according to Appendix A;
Step 3: Compute matrices $\{R(k), 1 \le k \le K\}$ and $\mathbf{x}_{-1}$ using Eqs. (6) and (7);
Step 4: Compute matrices $\{\bar{R}(n), n > 0\}$ using Eqs. (8) and (9);
Step 5: Compute the queue string distribution using Eqs. (4) and (10).

Once the stationary distribution of queue string at an arbitrary time is found, the stationary distribution of queue string at some special epochs can be obtained. For instance, we can derive formulas for stationary distributions at arrival epochs.

The stationary distribution of queue string right after an arbitrary arrival: Define an embedded Markov chain $(q_n, I_{1,n}, I_{2,n}, I_{3,n})$ for which q_n is the queue string just before the nth arrival of an arbitrary customer (excluding the nth arrival), $I_{1,n}$ is the phase of the underlying Markov process right after the nth arrival, $I_{2,n}$ is the type of the customer in service right after the nth arrival (if any), and $I_{3,n}$ is the phase of the service (if any) right after the nth arrival. Similar to $\mathbf{x}(J)$, we denote by $\mathbf{y}(J)$ the stationary distribution of the embedded Markov chain $(q_n, I_{1,n}, I_{2,n}, I_{3,n})$ at the node level for all J in $\aleph$. It is easy to obtain

$$\mathbf{y}(J) = \frac{\mathbf{x}(J)(\bar{D}_1 \otimes \mathbf{I})}{\mathbf{x}_{-1}\bar{D}_1\mathbf{e} + \sum_{L\neq -1,|L|\leq N} \mathbf{x}(L)(\bar{D}_1 \otimes \mathbf{I})\mathbf{e} + \sum_{|L|=N} \mathbf{x}(L)R(\mathbf{I}-R)^{-1}(\bar{D}_1 \otimes \mathbf{I})\mathbf{e}}, \quad J \in \aleph. \quad (11)$$

where $\bar{D}_1 = D_1 + D_2 + \cdots + D_K$.

The stationary distribution of queue string right after a type k arrival: Similar to the embedded Markov chain at the nth arrival of an arbitrary customer, define an embedded Markov chain $(q_n(k), I_{1,n}(k), I_{2,n}(k), I_{3,n}(k))$ at the nth arrival of a type k customer. Denote by $\mathbf{y}(J,k)$ the stationary distribution of the embedded Markov chain $(q_n(k), I_{1,n}(k), I_{2,n}(k), I_{3,n}(k))$ at the node level for all J in $\aleph$ and $1 \leq k \leq K$. It is easy to obtain

$$\mathbf{y}(J,k) = \frac{\mathbf{x}(J)(D_k \otimes \mathbf{I})}{\mathbf{x}_{-1}D_k\mathbf{e} + \sum_{L\neq -1,|L|\leq N} \mathbf{x}(L)(D_k \otimes \mathbf{I})\mathbf{e} + \sum_{|L|=N} \mathbf{x}(L)R(\mathbf{I}-R)^{-1}(D_k \otimes \mathbf{I})\mathbf{e}}, \quad J \in \aleph. \quad (12)$$

The stationary distributions of these embedded Markov chains can be calculated once $(\mathbf{x}_{-1}, \mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \cdots)$ or $\{\mathbf{x}(J), J \in \aleph\}$ is obtained. They are useful for computing means and variances of the waiting times of different types of customers in Section 7.

5. FUNDAMENTAL PERIODS AND BUSY PERIODS

In this section, the fundamental periods of a queue with LCFS are studied first. The results will be used in Section 6 for the waiting times. Then the busy periods of a queue with FCFS&LCFS are considered.

In general, a fundamental period is defined as the first passage time during which the (total) queue length decreases by 1. A busy period is defined as the first passage time until the queue becomes empty, given that the server is busy initially. When $q(0) = 0$, the fundamental period and the busy period starting at this epoch are the same; otherwise, they are different.

The LCFS case $(N = 1)$: Similar to the classical QBD case (see Neuts [13]), define, for J in $\aleph$, $J = k_1k_2\cdots k_n, 1 \leq k, l, l' \leq K, 1 \leq i, i' \leq m, 1 \leq j \leq m_l, 1 \leq j' \leq m_{l'}$,

$g_{(i,l,j)(i',l',j')}(k,x)$: the taboo probability that the Markov process $(q(t), I_1(t), I_2(t), I_3(t))$ reaches node J for the first time in state (J, i', l', j') in less than x units of time, given that the Markov process started in $(J + k, i, l, j)$.

Let $G(k,x)$ be a matrix with elements $g_{(i,l,j)(i',l',j')}(k,x)$. Because of the special structure of the QBD Markov process, $G(k,x)$ does not depend on the node J, when $J > 0$. Let $G^*(k,\omega)$

be the joint Laplace Stieltjes transform of $G(k,x)$ with respect to x. Then it can be proved that $\{G^*(k,\omega), 1 \le k \le K\}$ are the minimal nonnegative solutions to the equations

$$G^*(k,\omega) = [\omega\mathbf{I} - A_1(k)]^{-1}\left[A_2(k) + \sum_{l=1}^{K} A_0(l)G^*(l,\omega)G^*(k,\omega)\right], \qquad 1 \le k \le K. \quad (13)$$

Let $G(k) = G^*(k, 0+), 1 \le k \le K$. When the QBD Markov process is ergodic, matrix $G(k)$ is a stochastic matrix. Matrices $\{G(k), 1 \le k \le K\}$ are the minimal nonnegative solutions to the matrix equations (Takine, Sengupta, and Yeung [16]):

$$0 = A_2(k) + A_1(k)G(k) + \sum_{l=1}^{K} A_0(l)G(l)G(k), \qquad 1 \le k \le K. \quad (14)$$

The moments of the length of a fundamental period can be derived using Eq. (13). For instance, let

$$\mathbf{u}^*(k,\omega) = G^*(k,\omega)\mathbf{e}, \mathbf{u}^{(1)}(k) = \left.\frac{\partial G^*(k,\omega)\mathbf{e}}{\partial\omega}\right|_{\omega=0+},$$

$$\mathbf{u}^{(2)}(k) = \left.\frac{\partial^2 G^*(k,\omega)\mathbf{e}}{\partial\omega^2}\right|_{\omega=0+}, \quad (15)$$

for $1 \le k \le K$, where $-\mathbf{u}^{(1)}(k)$ is the mean length of a fundamental period started in the node $J+k$. Since $A_1(1) = \cdots = A_1(K) = A_1$ and $A_2(1)\mathbf{e} = \cdots = A_2(K)\mathbf{e}$, we have $\mathbf{u}^*(1,\omega) = \cdots = \mathbf{u}^*(K,\omega) \equiv \mathbf{u}^*(\omega)$. Note that $\mathbf{u}^*(0) = \mathbf{e}$. Then we have

$$\mathbf{u}^*(\omega) = (\omega\mathbf{I} - A_1)^{-1}\left[A_2(k)\mathbf{e} + \sum_{l=1}^{K} A_0(l)G^*(l,\omega)\mathbf{u}^*(\omega)\right],$$

$$\mathbf{u}^{(1)} = \left[A_1 + \sum_{k=1}^{K} A_0(k)[\mathbf{I} + G(k)]\right]^{-1}\mathbf{e},$$

$$\mathbf{u}^{(2)} = 2\left[A_1 + \sum_{k=1}^{K} A_0(k)[\mathbf{I} + G(k)]\right]^{-1}\left[\mathbf{I} - \sum_{k=1}^{K} A_0(k)G^{*(1)}(k)\right]\mathbf{u}^{(1)}, \quad (16)$$

where $G^{*(1)}(k)$ can be obtained by solving the following equation iteratively:

$$G^{*(1)}(k) = A_1^{-1}\left\{G(k) - \sum_{l=1}^{K} A_0(l)[G^{*(1)}(l)G(k) + G(l)G^{*(1)}(k)]\right\}, \qquad 1 \le k \le K. \quad (17)$$

NOTE: Matrix $G^{*(1)}(k)$ can be obtained explicitly by applying direct-sum transformation (Gantmacher [5]) to Equation (17).

The FCFS&LCFS case $(1 < N < \infty)$: Let $b_{i,k,j}(J)$ be the mean length of the busy period, given that the queue string is $J = k_1 k_2 \cdots k_{|J|}$, and the state of the auxiliary variable is (i, k, j) initially. Let $\mathbf{b}(J)$ be a vector with components $b_{i,k,j}(J)$ arranged lexicographically. Note that $\{G^*(k, \omega), 1 \leq k \leq K\}$ and $\mathbf{u}^*(\omega)$ are still the Laplace Stieltjes transforms of the fundamental periods when $|q(t)| > N$. It is easy to see that vectors $\{\mathbf{b}(J)\}$ are obtained by solving the following equations:

$$\mathbf{b}(0) = -A_1^{-1}\left[\mathbf{e} + \sum_{k=1}^{K} A_0(k)\mathbf{b}(k)\right],$$

$$\mathbf{b}(J) = -A_1^{-1}\left[A_2(j_1)\mathbf{b}(-k_1 + J) + \mathbf{e} + \sum_{k=1}^{K} A_0(k)\mathbf{b}(J + k)\right], \qquad |J| > 0, |J| < N,$$

$$\mathbf{b}(J) = -A_1^{-1}\left[A_2(k_{|J|})\mathbf{b}(J - k_{|J|}) + \mathbf{e} + \sum_{k=1}^{K} A_0(k)[-\mathbf{u}^{(1)} + G(k)\mathbf{b}(J)]\right], \qquad |J| \geq N. \quad (18)$$

Computation of $\{\mathbf{b}(J), 0 \leq |J| \leq N\}$ can be done by using an iterative method. Then $\{\mathbf{b}(J), |J| > N\}$ can be computed. Another way to find $\{\mathbf{b}(J), J \in \aleph\}$ is by introducing vectors $\{\bar{\mathbf{b}}(n), n \geq 0\}$, where vector $\bar{\mathbf{b}}(n)$ is obtained by arranging vectors $\{\mathbf{b}(J), |J| = n\}$ lexicographically. In fact, it can be shown that explicit formulas can be obtained for $\{\bar{\mathbf{b}}(n), n \geq 0\}$ and $\{\mathbf{b}(J), J \in \aleph\}$. Details are given in Appendix B.

6. WAITING TIMES

In this section, waiting times of different types of customers are studied. We shall consider the virtual waiting time, the actual waiting time of an arbitrary customer, and the actual waiting time of a type k customer for $1 \leq k \leq K$.

We start with the virtual waiting time which is defined as the total time elapsed before a virtual customer enters the server. Note that the virtual waiting time may be different from the current total work-load in the system for $N < \infty$ (even for cases with $K = 1$). To find the distribution of the virtual waiting times, we consider the following two cases: The number of customers in queue is N or less and the number of customers in queue is more than N.

Let V_L be the conditional virtual waiting time at an arbitrary epoch, given that there are more than N waiting customers in the queue at this epoch. Note that V_L is different from the waiting time of an arbitrary customer in a queueing system with LCFS service discipline when $N \geq 1$. Since customers are served on a LCFS basis when $q(t) > N$, V_L is equivalent to the length of the current fundamental period and its Laplace Stieltjes transform can be obtained as

$$\mathbf{E}\exp\{-\omega V_L\} = \frac{\sum_{J:|J|>N} \mathbf{x}(J)\mathbf{u}^*(\omega)}{\sum_{J:|J|>N} \mathbf{x}(J)\mathbf{e}} = \frac{[\sum_{J:|J|=N} \mathbf{x}(J)R(\mathbf{I} - R)^{-1}]\mathbf{u}^*(\omega)}{\sum_{J:|J|=N} \mathbf{x}(J)R(\mathbf{I} - R)^{-1}\mathbf{e}} \equiv \frac{\mathbf{r}}{\mathbf{re}}\mathbf{u}^*(\omega), \quad (19)$$

where "≡" means definition and $\mathbf{r} = \sum_{J:|J|=N} \mathbf{x}(J)R(\mathbf{I} - R)^{-1}$.

Let V_F be the conditional virtual waiting time at an arbitrary epoch, given that there are N or less than N waiting customers in the queue at this epoch. To find the distribution of V_F, we introduce functions $\Phi^*(J,\omega)$ and $\Phi^*(J,L,\omega)$.

Let $\phi^*_{i,k,j}(J,\omega)$ be the Laplace Stieltjes transform of the total time elapsed before a virtual customer enters the server after all customers in J complete their service, given that the queue string is $J(|J| \leq N)$ and the auxiliary variable is in state (i,k,j) initially. Customers served during this period of time include the customer currently in service, all the customers in J, all LCFS customers who arrived before the last customer in J completes its service, and all LCFS customers who arrived during the last LCFS period (if any) associated with the last customer in J. Let $\Phi^*(J,\omega)$ be a column vector with components $\phi^*_{i,k,j}(J,\omega)$ arranged lexicographically. When the queue length is N or less, the virtual waiting time V_F has

$$\mathbf{E}\exp\{-\omega V_F\} = \frac{\sum_{J:|J|\leq N}\mathbf{x}(J)\Phi^*(J,\omega)}{\sum_{J:|J|\leq N}\mathbf{x}(J)\mathbf{e}} = \frac{1}{1-\mathbf{re}}\left[\sum_{J:|J|\leq N}\mathbf{x}(J)\Phi^*(J,\omega)\right]. \quad (20)$$

Note that for $J=-1$, $\Phi^*(-1,\omega)=1$.

Vectors $\{\Phi^*(J,\omega)\colon |J|<N+1\}$ can be obtained from function $\Phi^*(J,L,\omega)$. Let $\Phi^*(J,L,\omega)$ be the (column vector) Laplace Stieltjes transform of the total time elapsed before a virtual customer (already in the system) enters the server after the first J customers complete their service, given that the queue string is $J+L$ initially ($|J+L|\leq N$). Again, customers served during this period of time include the customer currently in service, all the customers in J, all LCFS customers who arrived before the last customer in J completes its service, and all LCFS customers who arrived during the last LCFS period (if any) associated with the last customer in $J(=k_1k_2\cdots k_{|J|})$. Then it is easy to establish the following equations: $\Phi^*(-1,0,\omega)=1$, and

$$\Phi^*(0,L,\omega) = (\omega\mathbf{I}-A_1)^{-1}\left[\bar{A}_2(0)\mathbf{e}+\sum_{k=1}^{K}A_0(k)\Phi^*(0,L+k,\omega)\right], \qquad 0\leq |L|<N,$$

$$\Phi^*(0,L,\omega) = (\omega\mathbf{I}-A_1)^{-1}\left[\bar{A}_2(0)\mathbf{e}+\sum_{k=1}^{K}A_0(k)G^*(k,\omega)\Phi^*(0,L,\omega)\right], \qquad |L|=N,$$

$$\Phi^*(J,L,\omega) = (\omega\mathbf{I}-A_1)^{-1}\left[A_2(k_1)\Phi^*(-k_1+J,L,\omega)+\sum_{k=1}^{K}A_0(k)\Phi^*(J,L+k,\omega)\right], \qquad |J|>0, |J+L|<N,$$

$$\Phi^*(J,L,\omega) = (\omega\mathbf{I}-A_1)^{-1}\left[A_2(k_1)\Phi^*(-k_1+J,L,\omega)+\sum_{k=1}^{K}A_0(k)G^*(k,\omega)\Phi^*(J,L,\omega)\right], \qquad |J|>0, |J+L|=N. \quad (21)$$

The second equality in Eq. (21) shows that when $J=0$ and $|L|=N$, an explicit formula for $\Phi^*(0,L,\omega)$ can be found in terms of transition blocks and $G^*(k,\omega)$. This implies that explicit formulas can be found for all $\{\Phi^*(J,L,\omega), |J+L|\leq N\}$. This direction will be explored further in Appendix C for the development of an efficient algorithm for computing the mean and variance of waiting times. Define, for $|J+L| \le N$,

$$\Phi^{(1)}(J,L) = \left.\frac{d\Phi^*(J,L,\omega)}{d\omega}\right|_{\omega=0} \quad \text{and} \quad \Phi^{(2)}(J,L) = \left.\frac{d^2\Phi^*(J,L,\omega)}{d\omega^2}\right|_{\omega=0}.$$

Using Eq. (21), the first two derivatives $\Phi^{(1)}(J,L)$ and $\Phi^{(2)}(J,L)$ of $\Phi^*(J,L,\omega)$ can be obtained as

$$\Phi^{(1)}(J,L) = A_1^{-1}\left[\mathbf{e} - A_2(k_1)\Phi^{(1)}(-k_1+J,L) - \sum_{k=1}^{K} A_0(k)\Phi^{(1)}(J,L+k)\right], \qquad |J|>0, |J+L|<N,$$

$$\Phi^{(1)}(J,L) = \left[A_1 + \sum_{k=1}^{K} A_0(k)G(k)\right]^{-1}\left[\mathbf{e} - A_2(k_1)\Phi^{(1)}(-k_1+J,L) - \sum_{k=1}^{K} A_0(k)\mathbf{u}^{(1)}\right], \qquad |J|>0, |J+L|=N, \tag{22}$$

$$\Phi^{(2)}(J,L) = A_1^{-1}\left[2\Phi^{(1)}(J,L) - A_2(k_1)\Phi^{(2)}(-k_1+J,L) - \sum_{k=1}^{K} A_0(k)\Phi^{(2)}(J,L+k)\right], \qquad |J|>0, |J+L|<N;$$

$$\Phi^{(2)}(J,L) = \left[A_1 + \sum_{k=1}^{K} A_0(k)G(k)\right]^{-1} \times \left\{2\Phi^{(1)}(J,L) - A_2(k_1)\Phi^{(2)}(-k_1+J,L) - \sum_{k=1}^{K} A_0(k)[\mathbf{u}^{(2)} + 2G^{*(1)}(k)\Phi^{(1)}(J,L)]\right\}, \qquad |J|>0, |J+L|=N. \tag{23}$$

Note that when $J=0$, similar formulas can be obtained by removing items associated with $A_2(k)$ in the above expressions. It is worth pointing out that explicit formulas can be found for $\{\Phi^{(1)}(0,L), |L|=N\}$ as well. Consequently, explicit formulas can be found for all the first two derivatives $\Phi^{(1)}(J,L)$ and $\Phi^{(2)}(J,L)$ of $\Phi^*(J,L,\omega)$ with $|J+L| \le N$.

By definition, it is clear that $\Phi^*(J,\omega) = \Phi^*(J,0,\omega)$. Let V be the virtual waiting time at an arbitrary epoch. Conditioning on the queue length and using Eqs. (19) and (20), we obtain

$$\begin{aligned}\mathbf{E}\exp\{-\omega V\} &= \left[\sum_{J:|J|\le N} \mathbf{x}(J)\mathbf{e}\right]\mathbf{E}\exp\{-\omega V_F\} + \left[\sum_{J:|J|>N} \mathbf{x}(J)\mathbf{e}\right]\mathbf{E}\exp\{-\omega V_L\} \\ &= \sum_{J:|J|\le N} \mathbf{x}(J)\Phi^*(J,0,\omega) + \mathbf{r}\mathbf{u}^*(\omega),\end{aligned} \tag{24}$$

where $\Phi^*(-1,0,\omega) = \Phi^*(-1,\omega) = \mathbf{e}$. The first two moments of the virtual waiting time can be obtained as

$$\mathbf{E}V = -\sum_{J:|J|\le N} \mathbf{x}(J)\Phi^{(1)}(J,0) - \mathbf{r}\mathbf{u}^{(1)},$$

$$\mathbf{E}V^2 = \sum_{J:|J|\le N} \mathbf{x}(J)\Phi^{(2)}(J,0) + \mathbf{r}\mathbf{u}^{(2)}. \tag{25}$$

Now, we consider the actual waiting times. Let W $(W(k))$ be the actual waiting time of an arbitrary (type k) customer, i.e., the time elapsed from the arrival of a customer until the customer enters the server. Similar to the virtual waiting time, the distribution of W $(W(k))$ can be found by conditioning on the queue length and the status of the auxiliary variable after an arbitrary (type k) arrival. The Laplace Stieltjes transforms of W and $W(k)$ are given as

$$\begin{aligned}\mathbf{E}\exp\{-\omega W\} = \frac{1}{p}\Bigg[\mathbf{x}_{-1}\left(\sum_{l=1}^{K} D_l\right)\mathbf{e} + \sum_{J\ge 0:|J|<N} \mathbf{x}(J)\left(\sum_{l=1}^{K}(D_l\otimes\mathbf{I})\Phi^*(J,l,\omega)\right)\\ + \hat{\mathbf{r}}\left(\sum_{l=1}^{K} D_l\otimes\mathbf{I}\right)\mathbf{u}^*(\omega)\Bigg],\end{aligned}$$

$$\begin{aligned}&\mathbf{E}\exp\{-\omega W(k)\}\\ &= \frac{1}{p(k)}\left[\mathbf{x}_{-1}D_k\mathbf{e} + \sum_{J\ge 0:|J|<N}\mathbf{x}(J)(D_k\otimes\mathbf{I})\Phi^*(J,k,\omega) + \hat{\mathbf{r}}(D_k\otimes\mathbf{I})\mathbf{u}^*(\omega)\right],\\ &\qquad 1\le k\le K, \end{aligned}\tag{26}$$

respectively, where

$$\hat{\mathbf{r}} = \sum_{J:|J|=N}\mathbf{x}(J)(\mathbf{I}-R)^{-1},$$

$$p = \mathbf{x}_{-1}\left(\sum_{l=1}^{K} D_l\right)\mathbf{e} + \sum_{J\ge 0:|J|<N}\mathbf{x}(J)\left(\sum_{l=1}^{K}(D_l\otimes\mathbf{I})\right)\mathbf{e} + \hat{\mathbf{r}}\left(\sum_{l=1}^{K} D_l\otimes\mathbf{I}\right)\mathbf{e},$$

$$p(k) = \mathbf{x}_{-1}D_k\mathbf{e} + \sum_{J\ge 0:|J|<N}\mathbf{x}(J)(D_k\otimes\mathbf{I})\mathbf{e} + \hat{\mathbf{r}}(D_k\otimes\mathbf{I})\mathbf{e}, \qquad 1\le k\le K. \tag{27}$$

The key difference between expressions in Eq. (26) and that of Eq. (24) is that $\Phi^*(J,0,\omega)$ is replaced by $\Phi^*(J,k,\omega)$ when the customer who just arrived is of type k. The reason is that an "*actual*" customer who just arrived may have impact on the service process during its waiting time, while a "*virtual*" customer does not. Similar to Eq. (25), based on Eq. (26), the first and second moments of the actual waiting time of an arbitrary customer or an arbitrary type k customer can be obtained.

In summary, the computation of the first and second moments of waiting times can be done in the following steps.

Step 1. Compute stationary distributions $(\mathbf{x}_{-1}, \mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \cdots)$.
Step 2. Compute $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}$, and $G^{*(1)}(k)$ using Eqs. (16) and (17).
Step 3. Compute $\Phi^{(1)}(J, L)$ and $\Phi^{(2)}(J, L)$ using Eqs. (22) and (23).
Step 4. Compute the moments of waiting times.

Similar to the busy periods, the waiting times can be dealt with at the aggregate level too. Denote by $\bar{\Phi}^*(n, l, \omega)$ the vector obtained by arranging vectors $\{\Phi^*(J, L, \omega), |J| = n, |L| = l\}$ lexicographically. Equations (and explicit formulas) for $\{\bar{\Phi}^*(n, l, \omega), 0 \le n + l \le N\}$ and their first and second moments can be obtained from Eq. (21) in terms of transition blocks in Q in Equation (2), $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}$, and $\{G(k), G^{*(1)}(k), 1 \le k \le K\}$. Details are given in Appendix C.

7. PERFORMANCE ANALYSIS

In this section, a number of numerical examples are presented. The main focus is on the relationship between the threshold number N and the waiting times (mean and variance). We also explore the relationship between N and the queue length. In order to do so, we shall compute the stationary distribution of queue length at an arbitrary time and the means and variances of the virtual waiting time and actual waiting times.

EXAMPLE 1: Consider a queueing system with two types of customers, i.e., $K = 2$. The input process is an $MMAP[2]$ with parameters:

$$D_0 = \begin{pmatrix} -4 & 0 \\ 0 & -3 \end{pmatrix}, \qquad D_1 = \begin{pmatrix} 3 & 1 \\ 0 & 1 \end{pmatrix}, \qquad D_2 = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}.$$

The service times of the two types of customers are exponentially distributed with parameters 6 and 4, respectively, i.e., $m_1 = 1, \alpha_1 = (1), T_1 = (-6)$ and $m_2 = 1, \alpha_2 = (1), T_2 = (-4)$. For the arrival process, every type 2 customer is followed immediately by type 1 customer(s). Numerical results are presented and discussed in four steps.

1. *Queue length distribution of all customers at an arbitrary time.* In Figure 3, the distributions of the queue length for $N = 1, N = 2, N = 3, N = 5$, and $N = 7$ are plotted. Note that when $N = 1$, all customers are served on an LCFS basis. To see the changes in queue length, we use $N = 1$ (LCFS) case as the base case and define the following function:

$$\mathit{Diff}(n, N) = [\mathbf{x}_n(1)\mathbf{e} - \mathbf{x}_n(N)\mathbf{e}] \times 10^6, \qquad n \ge -1, \tag{28}$$

where $\{\mathbf{x}_n(N), n \ge -1\}$ is the distribution of the queue length when the threshold value is N.

Figure 3 shows that when the service discipline changes from LCFS to the mixed FCFS&LCFS, i.e., N goes from 1 to $N > 1$, the distribution of the queue length changes significantly. When N increases, the probability $\mathbf{x}_0(N)\mathbf{e}$ that there is 1 customer in the queueing system increases, but the probability that there are more than 3 customers in the queueing system decreases (in general). This implies that the queue length becomes stochastically shorter when threshold N increases. Note that for the $K = 1$ case, the service time of all customers are stochastically equivalent, but it is not true for $K > 1$ cases. Therefore, service order of customers may change the queue length for $K > 1$ cases. Also note that $\mathit{Diff}(-1, N) = 0$ for all N since the queueing systems are work-conserving.

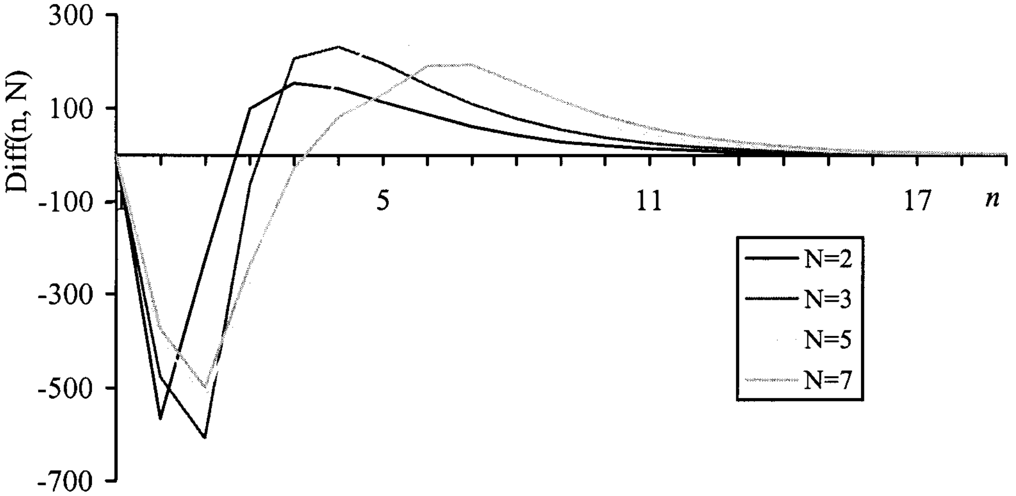

Figure 3. The difference between distributions of queue length.

2. *The virtual waiting time* V. We consider the means and variances associated with the virtual waiting times: $\mathbf{E}V_L$, Var(V_L), $\mathbf{E}V_F$, Var(V_L), $\mathbf{E}V$, and Var(V).

Table 1 shows that these performance measures change as a function of N. Naturally, the change in $\mathbf{E}V_F$ is most significant when N is not large. While the mean virtual waiting time $\mathbf{E}V$ stays more or less the same, its variance drops dramatically. The reason is that when N increases, the service discipline becomes closer to FCFS. This is consistent with the $K = 1$ case. Thus, FCFS reduces the variance of waiting times when multiple types of customers are present. In addition, in Table 1, the difference between the (conditional) waiting times of FCFS customers and LCFS customers is shown clearly.

3. *The actual waiting time* $W, W(1)$, *and* $W(2)$. We calculate the means and variance of the actual waiting times: $\mathbf{E}W(1)$, Var$(W(1))$, $\mathbf{E}W(2)$, Var$(W(2))$, $\mathbf{E}W$, and Var(W).

Table 2 shows that (1) the waiting times of different types of customers can be dramatically different and (2) means and variances of actual waiting times change with respect to N. While the mean waiting time $\mathbf{E}W$ goes down and up, the variance Var(W) decreases with respect to N. Therefore, on one hand, to reduce the variance of actual waiting times, FCFS is preferred. On the other hand, to reduce the mean actual waiting times, a mixed FCFS&LCFS service discipline should be chosen. For this example, it seems that $N = 5$ should be chosen so as to reduce the

Table 1. Mean and variance of the virtual waiting time.

	$\mathbf{E}V_L$	Var(V_L)	$\mathbf{E}V_F$	Var(V_F)	$\mathbf{E}V$	Var(V)
$N = 1$	0.567870	1.617281	0.298563	0.846358	0.382470	1.102110
$N = 2$	0.567887	1.617351	0.332636	0.836641	0.382098	1.009977
$N = 3$	0.567947	1.617567	0.351251	0.780004	0.382009	0.904609
$N = 4$	0.568039	1.617900	0.362328	0.708900	0.382052	0.799721
$N = 5$	0.568131	1.618231	0.369254	0.637237	0.382139	0.703190
$N = 6$	0.568208	1.618509	0.373717	0.571347	0.382234	0.618788
$N = 7$	0.568267	1.618723	0.376649	0.513847	0.382322	0.547612

Table 2. Mean and variance of actual waiting times.

	$\mathbf{E}W(1)$	Var($W(1)$)	$\mathbf{E}W(2)$	Var($W(2)$)	$\mathbf{E}W$	Var(W)
$N = 1$	0.390014	1.184061	0.356433	1.050861	0.384714	1.163189
$N = 2$	0.381649	1.093108	0.368726	1.049588	0.379610	1.086262
$N = 3$	0.376335	0.980451	0.375553	0.991266	0.376212	0.982158
$N = 4$	0.373776	0.866443	0.378957	0.902544	0.374594	0.872144
$N = 5$	0.372797	0.759920	0.380723	0.805221	0.374048	0.767078
$N = 6$	0.372635	0.665412	0.381713	0.711757	0.374068	0.672737
$N = 7$	0.372848	0.584623	0.382318	0.628254	0.374343	0.591521

mean actual waiting times. But there is a tradeoff between the mean waiting time and the variance of waiting time. In practice, a threshold value N should be chosen so that the mean actual waiting time is close to its minimum and the variance is not large. In this example, $N = 6$ and 7 might be good candidates.

Table 2 also shows that the mean waiting time of a type 2 customer is shorter than the mean waiting time of a type 1 customer when N is small. Intuitively, this has much to do with the arrival pattern of the two types of customers. When N is small, most of the customers are served on an LCFS basis. Since the service times of type 1 customers are shorter than that of type 2 customers on average, a type 2 customer who is followed by a type 1 customer before its service completion may soon reenter the server to complete its service.

Although some insights into the queueing systems with multiple types of customers can be gained from numerical examples, we like to show how complicated the behavior of a queueing system with multiple types of customers could be by comparing the following example to Example 1.

EXAMPLE 2: Consider another queueing system with the same input process of Example 1. The service times of the two types of customers are given as

$$m_1 = 2, \qquad \alpha_1 = (0.7, 0.3), \qquad T_1 = \begin{pmatrix} -6 & 1 \\ 0 & -8 \end{pmatrix},$$

$$m_2 = 2, \qquad \alpha_2 = (0.9, 0.1), \qquad T_2 = \begin{pmatrix} -35 & 2 \\ 1 & -1 \end{pmatrix}.$$

The two service times are special. For the service times of type 1 customers, they have the IFR (increased failure rate) property, i.e., the residual service time is shorter than the original service time stochastically. On the other hand, the service times of type 2 customers have the DFR (decreased failure rate) property. The idle probability of the queueing system is about 0.32, which is close to the idle probability of the queueing system in Example 1. But the system behavior and system reaction to the change in the threshold number N is different from what we have seen in Example 1.

1. *Queue length distribution of all customers at an arbitrary time.* Again, we look at the distributions of the queue length for $N = 1, N = 2, N = 3, N = 5$, and $N = 7$.

Figure 4 shows again that when the service discipline changes from LCFS to the mixed FCFS&LCFS, the distribution of the queue length changes significantly. But the changes in probability $\mathbf{x}_0\mathbf{e}$ and probability $\mathbf{x}_1\mathbf{e}$ are in the opposite directions when compared to Figure 3. Figure 3 shows that the LCFS can reduce the mean queue length, a fact that does not hold when

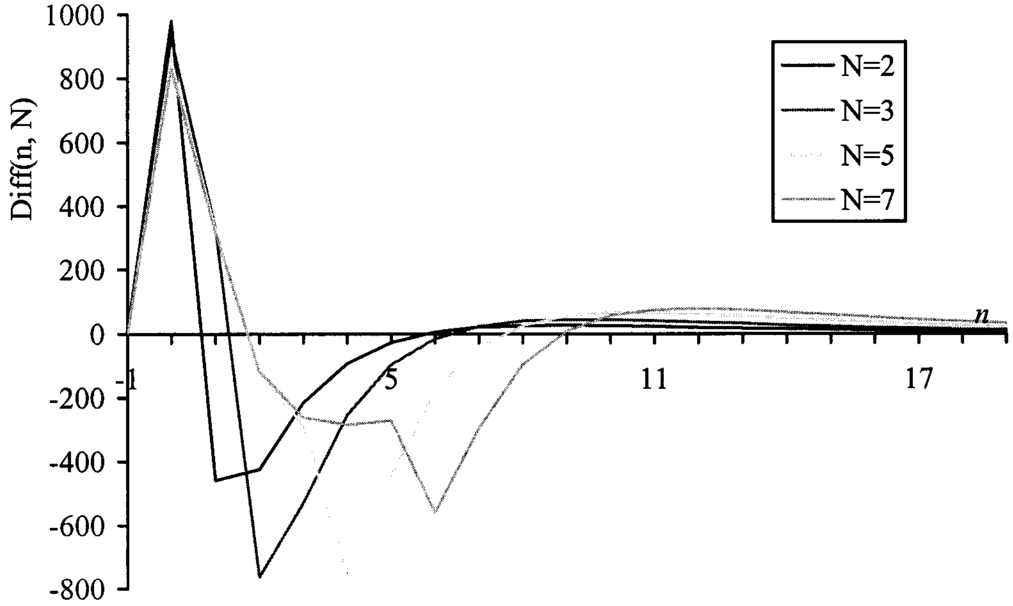

Figure 4. The difference between distributions of queue length.

there is only one class of customers. This reveals a fundamental difference between queueing systems with one class of customers $(K = 1)$ and multiple types of customers $(K > 1)$.

Figures 3 and 4 show that the relationship between the distribution of queue length and N is a complicated issue that needs further study.

2. *The virtual waiting time* V. We consider the means and variances associated with the virtual waiting times: $\mathbf{E}V_L$, Var(V_L), $\mathbf{E}V_F$, Var(V_F), $\mathbf{E}V$, and Var(V).

In Table 3, we see that the mean waiting times, the variances of V_L and V_F are increasing, but the variance of V is decreasing. The mean waiting time is increasing because of the special characteristics of the service times of type 2 customers. The variance of V is decreasing since we move from LCFS towards FCFS.

3. *The actual waiting time* $W, W(1)$, *and* $W(2)$. We consider the means and variance of the actual waiting times: $\mathbf{E}W(1)$, Var$(W(1))$, $\mathbf{E}W(2)$, Var$(W(2))$, $\mathbf{E}W$, and Var(W).

Table 3. Mean and variance of the virtual waiting time.

	$\mathbf{E}V_L$	Var(V_L)	$\mathbf{E}V_F$	Var(V_F)	$\mathbf{E}V$	Var(V)
$N = 1$	1.252612	8.022000	0.418372	2.482161	0.758930	4.911787
$N = 2$	1.307910	8.421603	0.497416	2.871062	0.759662	4.810788
$N = 3$	1.348626	8.711802	0.552340	3.085935	0.760222	4.676967
$N = 4$	1.377591	8.916156	0.593343	3.196526	0.760685	4.520211
$N = 5$	1.397635	9.056558	0.625123	3.237452	0.761086	4.348165
$N = 6$	1.411238	9.151377	0.650270	3.229622	0.761434	4.166918
$N = 7$	1.420348	9.214664	0.670422	3.187312	0.761736	3.981362

Table 4. Mean and variance of actual waiting times.

	$\mathbf{E}W(1)$	Var($W(1)$)	$\mathbf{E}W(2)$	Var($W(2)$)	$\mathbf{E}W$	Var(W)
$N = 1$	0.761490	5.027319	0.756683	4.693697	0.760748	4.975774
$N = 2$	0.756419	4.898750	0.767443	4.716237	0.758123	4.870565
$N = 3$	0.752162	4.736789	0.775536	4.688601	0.755774	4.729415
$N = 4$	0.749019	4.560807	0.781884	4.615539	0.754097	4.569405
$N = 5$	0.746764	4.375604	0.787090	4.507300	0.752994	4.396165
$N = 6$	0.745209	4.185241	0.791458	4.373268	0.752355	4.214573
$N = 7$	0.744190	3.993205	0.795158	4.221228	0.752065	4.028776

In Table 4, contrary to the virtual waiting time, the mean and the variance of the actual waiting times are decreasing, except $W(2)$. The variance of $W(1)$ changes dramatically because of the special arrival pattern.

In general, it can be concluded that the variance of the waiting time of an arbitrary customer decreases with respect to N, and the relationship between N and the mean waiting times is more complicated. We also found examples for which the variance of the waiting time of an arbitrary customer increases and then decreases with respect to N. One of such examples is given as follows.

EXAMPLE 3: Consider a queueing system with two types of customers, i.e., $K = 2$. The input process is an $MMAP[2]$ with parameters:

$$D_0 = \begin{pmatrix} -5 & 1 \\ 0 & -3 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 3 & 1 \\ 0 & 1 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}.$$

The service times of type 1 customers have the same distributions as that of type 2 customers in Example 2 and the service times of type 2 customers have the same distributions as that of type 1 customers in Example 2. The variances of the waiting times of an arbitrary customer are given in Table 5.

In Table 5, the variance of the virtual waiting time Var(V) is decreasing with respect to N (from 1 to 5). However, the variance of the actual waiting time of an arbitrary customer Var(W) is increased when N goes from $N = 1$ to $N = 2$. Immediately after $N = 2$, the Var(W) decreases with respect to N.

NOTE: The computation of the above numerical results is not an easy task. The main difficulty is the computer space needed to implement the algorithm developed in this paper, even though it looks like a straightforward thing to do. Further study is necessary to design more efficient algorithms for computing these performance measures.

The last example of this section is designed to demonstrate that the mixed FCFS&LCFS service discipline changes queueing behavior even for the $M/M/1$ case. This example has much to do with the famous PASTA property.

Table 5. The variances of the waiting times.

	$N = 1$	$N = 2$	$N = 3$	$N = 4$	$N = 5$
Var(V)	18.53412	18.40851	18.22485	17.98732	17.70100
Var(W)	18.97374	19.00324	18.93170	18.78720	18.57690

Table 6. The variances of the virtual and actual waiting times.

	$N=1$	$N=2$	$N=3$	$N=4$	$N=5$	$\cdots$	$N=50$
Var(V)	13.51111	12.35437	11.02342	9.70285	8.49716	$\cdots$	3.55555
Var(W)	14.22222	13.51111	12.35437	11.02342	9.70285	$\cdots$	3.55555

EXAMPLE 4: Consider a queueing system with only one type of customer $(K = 1)$, a Poisson input process with parameter $D_0 = -1$ and $D_1 = 1$, and exponential service times with parameters $m_1 = 1, \alpha_1 = 1$, and $T_1 = -1.5$. For $N = 1, 3, 5, 10, 20$, and 50, numerical results show that the mean waiting times (virtual and actual) are the same: $\mathbf{E}V = \mathbf{E}W = 4/3$ for all N. This is also intuitive according to PASTA. However, their corresponding variances are different. The results are presented in Table 6.

Table 6 shows that the distribution of the waiting times changes with respect to N. It also shows that the virtual and actual waiting times are different. The relationship Var$(V)(N)$ = Var$(W)(N+1)$ holds since for the actual waiting time case, the customer who just arrived adds one to the queue length.

8. CONCLUSIONS

In this paper, a model was developed for studying a queueing system with more than one type of customer. The idea of a hybrid service discipline was introduced to the system. The numerical examples have shown that, in a queueing system with multiple types of customers, it is possible to control both the mean and the variance of waiting times by introducing the mixed discipline of FCFS&LCFS, and then selecting the appropriate threshold level N. This would give system designers a lot more flexibility in meeting customer's expected quality of service, with given resources. This is one of the major findings in this paper. Further research needs to be carried out to see how to select N in order to achieve certain desired performance measures. From the viewpoint of traffic engineering, the probability that actual waiting times are greater than a threshold and the tail distribution of actual waiting times are important performance measures. But their analysis is much more complicated and lengthy. We leave them for future research.

APPENDIX A: TRANSITION MATRICES

For $n = -1, \bar{A}_1(-1) = D_0$ and $\bar{A}_0(-1) = (D_1 \otimes \alpha_1, \dots, D_K \otimes \alpha_K)$. For $n = 0$,

$$\bar{A}_1(0) = A_1(1), \bar{A}_2(0) = \mathbf{I}_{m\times m} \otimes \begin{pmatrix} \mathbf{T}_1^0 \\ \vdots \\ \mathbf{T}_K^0 \end{pmatrix},$$

$$A_0(k) = D_k \otimes \mathbf{I}_{\overline{m}\times\overline{m}} \qquad \text{(a type } k \text{ customer arrives)},$$

$$A_1(k) = D_0 \otimes \mathbf{I}_{\overline{m}\times\overline{m}} + \mathbf{I}_{m\times m} \otimes \begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_K \end{pmatrix} \qquad \text{(no service completed and no arrival)},$$

$$A_2(k) = \mathbf{I}_{m\times m} \otimes \left[\begin{pmatrix} \mathbf{T}_1^0 \alpha_k \\ \vdots \\ \mathbf{T}_K^0 \alpha_k \end{pmatrix} (0, \ldots, 0, \mathbf{I}_{m_k \times m_k}, 0, \ldots, 0) \right],$$

$$\text{(a service is completed and the next (last in queue) is of type } k), \qquad (29)$$

where "$\otimes$" represents the Kronecker product of matrices (see Gantmacher [5]). Transition blocks for $n > 0$,

$$\bar{A}_0(n) = \begin{pmatrix} (A_0(1), \ldots, A_0(K)) & & & & \\ & \cdots & & & \\ & & (A_0(1), \ldots, A_0(K)) & & \\ & & & \cdots & \\ & & & & (A_0(1), \ldots, A_0(K)) \end{pmatrix},$$

$$\bar{A}_1(n) = \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_1 \end{pmatrix},$$

$$\bar{A}_2(n) = \begin{pmatrix} \begin{pmatrix} A_2(1) & & \\ & \ddots & \\ & & A_2(1) \end{pmatrix} \\ \vdots \\ \begin{pmatrix} A_2(K) & & \\ & \ddots & \\ & & A_2(K) \end{pmatrix} \end{pmatrix}, \qquad 0 \le n \le N,$$

$$\bar{A}_2(n) = \begin{pmatrix} A_2(1) & & \\ \vdots & & \\ A_2(K) & & \\ & \ddots & \\ & & A_2(1) \\ & & \vdots \\ & & A_2(K) \end{pmatrix}, \qquad n > N,$$

where $\bar{A}_0(n)$ is an $(m\overline{m}K^n) \times (m\overline{m}K^{n+1})$ matrix, $\bar{A}_1(n)$ is an $(m\overline{m}K^n) \times (m\overline{m}K^n)$ matrix, $\bar{A}_2(n)$ is an $(m\overline{m}K^n) \times (m\overline{m}K^{n-1})$ matrix, and $\{A_2(k), A_1 = A_1(k), A_0(k), 1 \le k \le K\}$ are given in Eq. (29).

APPENDIX B

Define $\bar{\mathbf{b}}(n)$ as a vector obtained by arranging vectors $\{\mathbf{b}(J), |J| = n\}$ lexicographically. Vector $\bar{\mathbf{b}}(n)$ is the (conditional) mean busy period given that there are n waiting customers initially. Vector $\bar{\mathbf{b}}(n)$ can also be interpreted as the mean first passage time from level n to level -1 for the Markov process Q defined in Eq. (2). It is easy to see that Eq. (18) becomes

$$\bar{\mathbf{b}}(0) = -(\bar{A}_1(0))^{-1}[\mathbf{e} + \bar{A}_0(0)\bar{\mathbf{b}}(1)],$$

$$\bar{\mathbf{b}}(n) = -(\bar{A}_1(n))^{-1}[\bar{A}_2(n)\bar{\mathbf{b}}(n-1) + \mathbf{e} + \bar{A}_0(n)\bar{\mathbf{b}}(n+1)], \qquad 0 < n < N,$$

$$\bar{\mathbf{b}}(n) = -[\bar{A}_1(n) + \bar{A}_0(n)\bar{G}(n+1)]^{-1}[\bar{A}_2(n)\bar{\mathbf{b}}(n-1) + \mathbf{e} - \bar{A}_0(n)\bar{\mathbf{u}}^{(1)}(n+1)], \qquad n \geq N, \tag{30}$$

where, since LCFS is used when there are more than N waiting customers,

$$\bar{\mathbf{u}}^{(1)}(n) = \begin{pmatrix} \mathbf{u}^{(1)} \\ \vdots \\ \mathbf{u}^{(1)} \end{pmatrix}, \qquad \bar{G}(n) = \begin{pmatrix} G(1) & & \\ \vdots & & \\ G(K) & & \\ & \ddots & \\ & & G(1) \\ & & \vdots \\ & & G(K) \end{pmatrix}, \qquad n \geq N+1. \tag{31}$$

Equation (30) leads another way to compute the (conditional) mean busy period. In some cases, it is more convenient to use Eq. (30) than to use Eq. (18). It is clear that explicit formulas for $\{\bar{\mathbf{b}}(n), n > -1\}$ can be found by using Eq. (30).

APPENDIX C

Denote by $\bar{\Phi}^*(n, l, \omega)$ the vector obtained by sequencing $\{\Phi^*(J, L, \omega), |J| = n, |L| = l\}$ lexicographically. Vector $\bar{\Phi}^*(n, l, \omega), 0 \leq n + l \leq N$, is the Laplace Stieltjes transform of the length of time that the server finishes the current service, all the first n customers, and all LCFS customers who arrived before all the n customers complete their service, given that the queue length is $n + l$ initially. By Eq. (21), it is easy to obtain

$$\bar{\Phi}^*(0, l, \omega) = (\omega\mathbf{I} - \bar{A}_1(l))^{-1}[\bar{A}_2(l)\mathbf{e} + \bar{A}_0(l)\bar{\Phi}^*(0, l+1, \omega)], \qquad 0 \leq l < N,$$

$$\bar{\Phi}^*(0, N, \omega) = (\omega\mathbf{I} - \bar{A}_1(N))^{-1}[\bar{A}_2(N)\mathbf{e} + \bar{A}_0(N)\bar{G}^*(N+1, \omega)\bar{\Phi}^*(0, N, \omega)],$$

$$\bar{\Phi}^*(n, l, \omega) = (\omega\mathbf{I} - \bar{A}_1(n+l))^{-1}[\bar{A}_2(n+l)\bar{\Phi}^*(n-1, l, \omega) + \bar{A}_0(n+l)\bar{\Phi}^*(n, l+1, \omega)],$$
$$n > 0, n + l < N,$$

$$\bar{\Phi}^*(n, l, \omega) = (\omega\mathbf{I} - \bar{A}_1(N))^{-1}[\bar{A}_2(N)\bar{\Phi}^*(n-1, l, \omega) + \bar{A}_0(N)\bar{G}^*(N+1, \omega)\bar{\Phi}^*(n, l, \omega)],$$
$$n > 0, n + l = N, \tag{32}$$

where $\bar{G}^*(N+1, \omega)$ is given in terms of $\{G^*(k, \omega), 1 \leq k \leq K\}$ similar to $\bar{G}(N+1)$ given in Eq. (31). Define

$$\bar{\Phi}^{(1)}(n, l) = \left.\frac{d\bar{\Phi}^*(n, l, \omega)}{d\omega}\right|_{\omega=0} \quad \text{and} \quad \bar{\Phi}^{(2)}(n, l) = \left.\frac{d^2\bar{\Phi}^*(n, l, \omega)}{d\omega^2}\right|_{\omega=0}.$$

Then we have

$$\bar{\Phi}^{(1)}(0,l) = (\bar{A}_1(l))^{-1}[\mathbf{e} - \bar{A}_0(l)\bar{\Phi}^{(1)}(0,l+1)], \qquad l < N,$$

$$\bar{\Phi}^{(1)}(0,N) = [\bar{A}_1(N) + \bar{A}_0(N)\bar{G}(N+1)]^{-1}[\mathbf{e} - \bar{A}_0(N)\bar{\mathbf{u}}^{(1)}(N+1)], \qquad l = N,$$

$$\bar{\Phi}^{(1)}(n,l) = (\bar{A}_1(n+l))^{-1}[\mathbf{e} - \bar{A}_2(n+l)\bar{\Phi}^{(1)}(n-1,l) - \bar{A}_0(n+l)\bar{\Phi}^{(1)}(n,l+1)], \qquad n > 0, n+l < N,$$

$$\bar{\Phi}^{(1)}(n,l) = [\bar{A}_1(N) + \bar{A}_0(N)\bar{G}(N+1)]^{-1}[\mathbf{e} - \bar{A}_2(N)\Phi^{(1)}(n-1,l) - \bar{A}_0(N)\bar{\mathbf{u}}^{(1)}(N+1)], \qquad n > 0, n+l = N, \quad (33)$$

and

$$\bar{\Phi}^{(2)}(0,l) = (\bar{A}_1(l))^{-1}[2\bar{\Phi}^{(1)}(0,l) - \bar{A}_0(l)\bar{\Phi}^{(2)}(0,l+1)], \qquad l < N,$$

$$\begin{aligned}\bar{\Phi}^{(2)}(0,N) = {} & [\bar{A}_1(N) + \bar{A}_0(N)\bar{G}(N+1)]^{-1} \\ & \times \{2\bar{\Phi}^{(1)}(0,N) - \bar{A}_0(N)[\bar{\mathbf{u}}^{(2)}(N+1) + 2\bar{G}^{*(1)}(N+1)\bar{\Phi}^{(1)}(0,N)]\}, \qquad l = N,\end{aligned}$$

$$\bar{\Phi}^{(2)}(n,l) = (\bar{A}_1(n+l))^{-1}[2\bar{\Phi}^{(1)}(n,l) - \bar{A}_2(n+l)\bar{\Phi}^{(2)}(n-1,l) - \bar{A}_0(n+l)\bar{\Phi}^{(2)}(n,l+1)], \qquad n > 0, n+l < N,$$

$$\begin{aligned}\bar{\Phi}^{(2)}(n,l) = {} & [\bar{A}_1(N) + \bar{A}_0(N)\bar{G}(N+1)]^{-1}\{2\bar{\Phi}^{(1)}(n,l) - \bar{A}_2(N)\bar{\Phi}^{(2)}(n-1,l) \\ & - \bar{A}_0(N)[\bar{\mathbf{u}}^{(2)}(N+1) + 2\bar{G}^{*(1)}(N+1)\bar{\Phi}^{(1)}(n,l)]\}, \qquad n > 0, n+l = N, \quad (34)\end{aligned}$$

where $\bar{\mathbf{u}}^{(1)}(N+1)$ and $\bar{G}(N+1)$ are given in Eq. (31) and

$$\bar{\mathbf{u}}^{(2)}(N+1) = \begin{pmatrix} \mathbf{u}^{(2)} \\ \vdots \\ \mathbf{u}^{(2)} \end{pmatrix}, \qquad \bar{G}^{*(1)}(N+1) = \begin{pmatrix} G^{*(1)}(1) & & \\ \vdots & & \\ G^{*(1)}(K) & & \\ & \ddots & \\ & & G^{*(1)}(1) \\ & & \vdots \\ & & G^{*(1)}(K) \end{pmatrix}. \quad (35)$$

Note that explicit formulas for all the derivatives can be obtained from Eqs. (33) and (34). But the explicit formulas are complicated. On the other hand, Eqs. (33) and (34) are more convenient for the development of an iterative algorithm for computing the mean and variance of waiting times. In fact, numerical results presented in Section 7 are obtained by a computational procedure based on equations given in this appendix.

According to Eq. (25), the first two moments of the virtual waiting time are given by

$$\mathbf{E}V = -\sum_{n=0}^{N} \mathbf{x}_n \bar{\Phi}^{(1)}(n,0) - \mathbf{r}\mathbf{u}^{(1)},$$

$$\mathbf{E}V^2 = \sum_{n=0}^{N} \mathbf{x}_n \bar{\Phi}^{(2)}(n,0) + \mathbf{r}\mathbf{u}^{(2)}. \tag{36}$$

Similar equations can be established for the actual waiting times $\{W, W(k), 1 \leq k \leq K\}$, based on Eqs. (26) and (27). Details are omitted.

ACKNOWLEDGMENTS

The authors would like to thank three referees for their valuable comments and suggestions which greatly improved the quality of this paper.

REFERENCES

[1] A.S. Alfa and G.J. Fitzpatrick, Waiting time distribution of a FIFO/LIFO $Geo/D/1$ queue, INFOR, 39(1) (1999), 149–159.

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

[3] B.T. Doshi, An $M/G/1$ queue with a hybrid discipline, AT&T Tech J 62(5) (1983), 1251–1271.

[4] L.J. Forys, Performance analysis of a new overload strategy, ITC-10, 1982.

[5] F.R. Gantmacher, The theory of matrices, Chelsea, New York, 1959.

[6] Q.-M. HE, Queues with marked customers, Adv Appl Probab 28 (1996), 567–587.

[7] Q.-M. HE, Quasi-birth-and-death processes with a tree structure and a detailed analysis of the $MMAP[K]/PH[K]/1$ queue, Eur J Oper Res 120(3) (2000), 641–656.

[8] Q.-M. HE and A.S. Alfa, The $MMAP[K]/PH[K]/1$ queue with a last-come-first-served preemptive service discipline, Queueing Syst 28 (1998), 269–291.

[9] Q.-M. HE and M.F. Neuts, Markov arrival processes with marked transitions, Stochastic Process Appl 74 (1998), 37–52.

[10] G. Latouche and V. Ramaswami, A logarithmic reduction algorithm for quasi birth and death processes, J Appl Probab 30 (1993), 650–674.

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

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

[13] M.F. Neuts, Structured stochastic matrices of $M/G/1$ type and their applications, Marcel Dekker, New York, 1989.

[14] F. Schreiber, Eine Familie von warteprozeduren zweichen FIFO und LIFO, Arch Elektron Uebertragungstech 30(12) (1976), 497–501.

[15] H. Takagi, Queueing analysis: A foundation of performance evaluation, Volume 1: Vacation and priority systems, Part 1, Elsevier, Amsterdam, 1990.

[16] T. Takine, B. Sengupta, and R.W. Yeung, A generalization of the matrix $M/G/1$ paradigm for Markov chains with a tree structure, Stochastic Models 11 (1995), 411–421.

[17] R.W. Yeung and B. Sengupta, Matrix product-form solutions for Markov chains with a tree structure, Adv Appl Probab 26(4) (1994), 965–987.

[18] R.W. Yeung and A.S. Alfa, The quasi-birth-death type Markov chain with a tree structure, Stochastic Models, 15(4) (1999), 639–659.