The Discrete Time *MMAP*[*K*]/*PH*[*K*]/1/LCFS-GPR Queue and Its Variants

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

Abstract: In this paper, we study a discrete time queueing system with multiple types of customers and a last-come-first-served general preemptive resume (LCFS-GPR) service discipline (*MMAP*[*K*]/*PH*[*K*]/1/LCFS-GPR). When the waiting space is infinite, matrix analytic methods are used to find a system stability condition, to derive the distributions of the busy periods and sojourn times, and to obtain a matrix geometric solution of the queue string. The results lead to efficient algorithms for computing various performance measures at the level of individual types of customers. Using those algorithms, the impact of the LCFS-GPR service discipline on the corresponding queueing system can be analyzed. When the waiting space is finite, the Gaussian elimination method is used to develop an efficient algorithm for computing the stationary distribution of the queue string. The relationship between the loss probabilities of individual types of customers and the size of the waiting space is explored. This paper also serves as a brief survey of the study of the *MMAP*[*K*]/*PH*[*K*]/1 queue and its related queueing models.

1. Introduction

In this paper, we study a class of queues that involve a multiple class of customers, with each class having different service time requirements. These types of queues have major applications in the design and analysis of many manufacturing, telecommunications, and service systems. Our interest is in the last-come-first-served (LCFS) case that is very common in telecommunication systems. We focus on the discrete time analysis of the queueing systems. Such queueing systems have received considerable attention from researchers and practitioners recently.

The basic model under consideration in this paper can be described as follows. Customers arrive in the queueing system according to a discrete time

Markov arrival process with marked transitions (*MMAP*[*K*]). Customers are distinguished into *K* types. All customers join a single queue and are served on a last-come-first-served (LCFS) basis, i.e., any new arrival pushes the customer in service (if any) out of the server and starts its service immediately. A general preemptive resume (GPR) rule, which includes the well-known preemptive resume and preemptive repeat service disciplines as its special cases, is applied to determine the service time when a customer reenters the server. The service times of different classes of customers can be different and have phase type (*PH*) distributions. We distinguish queueing systems with an infinite waiting space (buffer) from that with a finite waiting space. When the waiting space is finite, we assume that a customer who finds a full queue cannot enter the queueing system and is lost forever. We denote this queueing system as *MMAP*[*K*]/*PH*[*K*]/1/LCFS-GPR.

Theoretically, the study of the queueing systems of interest was made possible by the developments in the study of Markov chains with a tree structure (Takine, Sengupta, and Yeung [20], Yeung and Alfa [21], and Yeung and Sengupta [22]). The approach has been proven to be a success in the study of such queueing systems. For instance, using the approach, HE [6] and HE and Alfa [9] studied the *MMAP*[*K*]/*PH*[*K*]/1/LCFS non-preemption queue and preemptive resume or repeat queue, respectively.

In this paper, we shall use matrix analytic methods, based on the recent developments in the study of *GI/M*/1 type or *M/G*/1 type Markov chains with a tree structure, to study the queueing systems of interest. First, an *M/G*/1 type Markov chain with a tree structure will be introduced to represent the queue string process in such a queueing system. Second, a method is introduced to analyze the stability of the queueing system. Third, distributions of the busy periods and sojourn times are found. Fourth, the queue string process is reformulated into a *GI/M*/1 type Markov chain with a tree structure and a matrix geometric solution of the queue string is obtained. These results lead to efficient algorithms for computing various performance measures at the level of individual types of customers. The impact of the LCFS-GPR service discipline on the queueing process can then be analyzed. For queueing systems with a finite waiting space, a Gaussian elimination method is used to develop an algorithm for computing the distribution of the queue string. The relationship between the loss probabilities of individual types of customers and the size of the waiting space is investigated. A numerical example is presented to gain insights into the queueing systems of interest.

This paper serves as a survey of the study of the continuous and discrete time *MMAP*[*K*]/*PH*[*K*]/1 queues. For that purpose, a brief review of the recent developments in the study of the *MMAP*[*K*]/*PH*[*K*]/1 queue as well as its related areas is given in Section 2 of this paper.

The rest of the paper is organized as follows. In Section 2, a brief literature review on the recent developments in the study of the *MMAP*[*K*]/*PH*[*K*]/1 queue and related areas is given. Section 3 introduces the discrete time

MMAP[*K*]/*PH*[*K*]/1/LCFS-GPR queue and investigates various performance measures. The discrete time *MMAP*[*K*]/*PH*[*K*]/1/*c*/LCFS-GPR queue with a finite waiting space is studied in Section 4. Finally, in Section 5, we briefly discuss a few directions for future research.

2. Literature review

In this section, we briefly review recent developments in the study of queueing systems with multiple types of customers. More specifically, we shall discuss the study of Markov arrival processes (*MAP*), Markov arrival processes with marked arrivals (*MMAP*[*K*]), Markov chains with a tree structure, the *MMAP*[*K*]/*G*[*K*]/1 queue, the *MMAP*[*K*]/*PH*[*K*]/1/LCFS queue, and a few variants of the *MMAP*[*K*]/*PH*[*K*]/1/LCFS queue. To be brief, we focus on research works closely related to matrix analytic methods.

Markov arrival process (*MAP*), also called Neuts process, was introduced in Neuts [13] as a generalization of the Poisson processes. That formation of arrival process has been widely accepted because of its capability in modelling input processes and the tractability of its corresponding stochastic models, especially in queueing theory (Ramaswami [16], Lucantoni, Hellerstern, Neuts [12], and Neuts [15]). A generalization to multiple types of correlated arrivals - Markov arrival processes with marked arrivals (*MMAP*[*K*]) - was introduced in Asmussen and Koole [2], HE [5], and HE and Neuts [11]. Similar ideas can be found in the extensive literature related to Markov modulated Poisson processes (*MMPP*). *MMAP*[*K*] can capture not only the correlation between different types of arrivals, but also the arrival pattern among all types of arrivals (see HE and Alfa [9] and Example 3.2.1 in this paper). Primarily, we are interested in queueing systems with a Markov arrival process with marked arrivals in this paper.

An extensive study has been carried out on queueing systems with multiple types of arrivals when the arrival process is a Poisson process (Takagi [18]). Since the Poisson process is not flexible enough to model various input processes in practice, queueing models with *MMAP*[*K*] are introduced. HE [5] and Takine and Hasegawa [19] studied the *MMAP*[*K*]/*G*[*K*]/1 queue. A number of results were obtained for the fundamental periods, waiting times, and the queue length. HE [7, 8] identified conditions for ergodicity of such queueing models with a work conserving or a non work conserving service discipline. In summary, significant progress has been made on the study of the *MMAP*[*K*]/*G*[*K*]/1/FCFS queue.

Queueing systems with multiple types of customers are frequent occurrences in telecommunication systems. The LCFS service discipline has been recognized in the telecommunication systems as a way of increasing throughput in a system where some customers have a time threshold for waiting. This application motivated most of the research in the multiclass LCFS queueing models. With the interest in

maximizing throughput in queues with customer waiting time threshold, Schreiber [17] introduced a hybrid FCFS/LCFS queue discipline for the *M/M*/1 queue. Later Doshi [3] considered the *M/G*/1 FCFS/LCFS queue. Using a much simpler analysis technique Alfa and Fitzpatrick [1] developed a computationally efficient approach for the *Geo/D*/1 FCFS/LCFS queue. All these systems are for a single class of customers ($K = 1$).

The study of LCFS queueing systems with multiple types of customers ($K>1$) continues in HE [6, 7, 8], HE and Alfa [9, 10], and this paper. An important feature of these studies is that the results allow us to look at the queueing behaviors of individual types of customers. In HE [6], the *MMAP*[*K*]/*PH*[*K*]/1/LCFS non-preemptive queue was introduced and studied. In that queueing model, the service times have *PH*-distributions for different types of customers. Results obtained in HE [7, 8] can be used to identify ergodicity conditions for LCFS queueing systems. In HE and Alfa [9], the *MMAP*[*K*]/*PH*[*K*]/1/LCFS queue with a preemptive resume or repeat service discipline was introduced and studied. In HE and Alfa [10], the *MMAP*[*K*]/*PH*[*K*]/1 with a hybrid FCFS and LCFS service discipline was introduced and studied in detail. In general, the queueing processes of these queueing systems can be formulated into *GI/M*/1 or *M/G*/1 type Markov chains with a tree structure and an analysis of the stationary distribution of queue strings, busy periods, and sojourn times can be carried out. Thus, Markov chains with a tree structure play an important role in our research, as was mentioned in Section 1.

To end this section, we would like to point out that the analysis of the continuous and discrete time *MMAP*[*K*]/*PH*[*K*]/1/LCFS queues is essentially the same. However, the discrete time queueing systems are usually more complicated in formulation and sometimes the structure of the corresponding Markov chains changes (see the Notes in Section 3.1). Thus, the results obtained in this paper are not straightforward extensions of their continuous time counterparts.

3. Discrete time *MMAP*[*K*]/*PH*[*K*]/1/LCFS-GPR queue

We first introduce the discrete time *MMAP*[*K*]/*PH*[*K*]/1/LCFS-GPR queue and an *M/G*/1 type Markov chain with a tree structure for its queueing process in Section 3.1. In Section 3.2 we develop a method for analyzing the stability of the queueing system and its relationship with the LCFS-GPR rule. The fundamental periods, busy periods, busy cycles, and sojourn times are investigated in Section 3.3. In Section 3.4, a *GI/M*/1 type (or QBD type) Markov chain with a tree structure is introduced for the queue string process. In steady state, a matrix geometric solution of the queue string is obtained.

3.1 The model

The arrival process of the queueing system of interest is a discrete Markov arrival process with marked transitions (*MMAP*[*K*]). Customers of the arrival process are distinguished into *K* types. The *MMAP*[*K*] is defined by a set of $m \times m$ matrices $\{D_k, 0 \le k \le K\}$, where *m* is a positive integer. The matrices D_k, $0 \le k \le K$, are nonnegative. The matrix $I - D_0$ is assumed to be non-singular, where *I* is the identity matrix. Let

$$D = \sum_{k=0}^{K} D_k \ . \tag{3.1.1}$$

Then the matrix *D* is the transition matrix of the underlying Markov chain of the arrival process, which has *m* phases. Consequently, the matrix *D* is a stochastic matrix. Let $I(n)$ be the phase of the underlying Markov chain at time *n*, $1 \le I(n) \le m$. An arrival is called a type *k* customer if it is marked by *k*. The (matrix) marking rate of type *k* customer is D_k. Let θ be the stationary probability vector of the matrix *D*. The stationary arrival rate of type *k* arrival is given by $\lambda_k = \theta D_k \mathbf{e}$, $1 \le k \le K$, where $\mathbf{e}$ is the column vector with all components one.

The service times have phase-type distributions. The service times of type *k* customers have a common phase-type distribution (*PH*-distribution) function with a matrix representation (m_k, α_k, T_k), where m_k is positive integer, α_k is an m_k-dimension nonnegative vector with $\alpha_k \mathbf{e} = 1$, and T_k is an $m_k \times m_k$ substochastic matrix. Let $\mathbf{T}^0{}_k = \mathbf{e} - T_k \mathbf{e}$. The mean service time is given by $1/\mu_k = \alpha_k (I - T_k)^{-1} \mathbf{e}$. Then μ_k is the average service rate of type *k* customers. For more details about *PH*-distribution, see Chapter 2 in Neuts [14]. We assume that the service process and the arrival process are independent.

All customers are served on an LCFS-GPR service discipline. When a customer of type *k* arrives, it pushes the customer in service (if any) out of the server and starts its service with service time (m_k, α_k, T_k). For the outgoing customer, its current service phase is recorded (say *i*) and its future service phase is chosen according to the probability distribution $\mathbf{q}_{k,i} = (q_{k,i,1}, \cdots, q_{k,i,m_k})$ at the epoch it is pushed out. If the future service phase is *j*, then the distribution of its service time is $(m_k, \mathbf{e}(j), T_k)$ when the customer reenters the server, where $\mathbf{e}(j)$ is the row vector for which the *j*th element is one and all others zero. Let Q_k be an $m_k \times m_k$ matrix with elements $q_{k,i,j}$. Then matrix Q_k is a stochastic matrix and it specifies the service phases for interrupted services. It is worth to point out that some well-known service disciplines are special cases of the LCFS-GPR. For instance, when $Q_k = I$,

customers are served on an LCFS preemptive resume basis. When $Q_k = \mathbf{e}\alpha_k$, customers are served on an LCFS preemptive repeat basis. It is easy to see that the service disciplines for different types of customers can be made different through the matrices $\{Q_k, 1 \le k \le K\}$. Finally, when the server becomes available to customers in queue, the customer who arrived last gets the server.

Note 3.1.1: The LCFS-GPR introduced here is slightly different from that in previous papers. For instance, in Yeung and Alfa [21], the phase of service is determined according to $\mathbf{q}_{k,i} = (q_{k,i,1}, \cdots, q_{k,i,m_k})$ at the epoch a customer reenters the server, if the customer was pushed out of the server in phase i. Nonetheless, that definition and our definition both imply that if a customer was pushed out of the server when its service phase is i, its service time has the *PH*-distribution $(m_k, \mathbf{q}_{k,i}, T_k)$ when it reenters the server. Therefore, with regard to the queueing process, the two definitions are equivalent. But their corresponding Markov chains of the queue strings can be different. Our definition is suitable for analyzing the fundamental periods, busy periods, busy cycles, and sojourn times.

For each customer in the queueing system, a pair (k, j) is used to represent its status, where k is the type of the customer and j is the phase of the service time of the customer. The phase j is either the phase of the service time when the customer resumes its service if the customer is in the queue or the current service phase if the customer is in service. Let $q(n)$ be the *queue string* consisting of the status of all the customers in the queueing system at the beginning of time n - a string of the pair (k, j) where k is between 1 and K and j is between 1 and m_k. For instance, when $q(n) = (k_1, j_1)\ldots(k_t, j_t)$, there are t customers in the queueing system at this moment. The service phase of the customer currently in service is j_t, the (future) service phase of the first customer in queue is j_{t-1}, i.e., when the customer reenters the server, its service starts in phase j_{t-1}, …, and the last (oldest) customer in the queue is of type k_1 and its future service phase is j_1. We assume that the change of the phase of the arrival process or the service process occurs at the end of each unit time.

It is easy to see that $(q(n), I(n))$ is an irreducible and aperiodic Markov chain. The state space of the Markov chain is $\Omega \times \{1, 2, \ldots, m\}$, where $\Omega = \{0\} \cup \{J: J = (k_1, j_1)(k_2, j_2)\ldots(k_n, j_n), 1 \le k_t \le K, 1 \le j_t \le m_{k_t}, 1 \le t \le n, n \ge 1\}$. Let $|J|$ be defined as the number of pairs of integers in J. Hence $|J|$ is the length of the string J. The level n of queue strings consists of all the strings with $|J| = n$. It is easy to see that after each transition, the level of the Markov chain $(q(n), I(n))$ (with respect to $q(n)$) can increase or decrease at most by one. Define the addition operation "+" in Ω as follows: for $J = (k_1, j_1)(k_2, j_2)\ldots(k_n, j_n)$ in Ω, $J+(k, j) = (k_1, j_1)(k_2, j_2)\ldots(k_n, j_n)(k, j)$. Assume that the Markov chain is in node $J+(k, j) \in \Omega$ at time unit n, i.e., $q(n) = J + (k, j)$. The transition blocks of the Markov chain are given as follows.

a) When a new customer arrives and there is no service completion,

$$
\begin{aligned}
&A_0((k, j),(k, j')(k_1, j_1)) \\
&\equiv (\mathbf{P}\{q(n+1) = J + (k, j')(k_1, j_1),\, I(n+1) = i' \mid q(n) = J + (k, j),\, I(n) = i\}) \\
&= (T_k Q_k)_{j,j'} (\alpha_{k_1})_{j_1} D_{k_1}, \quad 1 \le k, k_1 \le K, 1 \le j, j' \le m_k, 1 \le j_1 \le m_{k_1},
\end{aligned} \tag{3.1.2}
$$

where "≡" means definition. Note that in equation (3.1.2) and equations (3.1.3) to (3.1.5), $1 \le i, i' \le m$. Also note that $(T_k Q_k)_{j,j'}$ represents the (j, j')th element of the matrix $T_k Q_k$ and $(\alpha_{k_1})_{j_1}$ the j_1th element of the vector α_{k_1}.

b) When a service is completed and a new customer arrives, or when there is no service completion and no new arrival, $1 \le k, k_1 \le K$,

$$
\begin{aligned}
&A_1((k, j),(k_1, j_1)) \\
&\equiv (\mathbf{P}\{q(n+1) = J + (k_1, j_1),\, I(n+1) = i' \mid q(n) = J + (k, j),\, I(n) = i\}) \\
&= \begin{cases}
(\mathbf{T}_k^0)_j (\alpha_{k_1})_{j_1} D_{k_1} + (T_k)_{j,j_1} D_0, & k = k_1, 1 \le j, j_1 \le m_k; \\
(\mathbf{T}_k^0)_j (\alpha_{k_1})_{j_1} D_{k_1}, & k \ne k_1, 1 \le j \le m_k, 1 \le j_1 \le m_{k_1}.
\end{cases}
\end{aligned} \tag{3.1.3}
$$

c) When a service is completed and there is no new arrival,

$$
\begin{aligned}
&A_2(k, j) \equiv (\mathbf{P}\{q(n+1) = J,\, I(n+1) = i' \mid q(n) = J + (k, j),\, I(n) = i\}) \\
&= (\mathbf{T}_k^0)_j D_0, \quad 1 \le j \le m_k, 1 \le k \le K.
\end{aligned} \tag{3.1.4}
$$

d) When no customer is in the queueing system, i.e., $q(n) = J = 0$,

$$
\begin{aligned}
A_0(0,(k_1, j_1)) &\equiv (\mathbf{P}\{q(n+1) = (k_1, j_1),\, I(n+1) = i' \mid q(n) = 0,\, I(n) = i\}) \\
&= (\alpha_{k_1})_{j_1} D_{k_1}, \quad 1 \le k_1 \le K, 1 \le j_1 \le m_{k_1}; \\
A_1(0,0) &\equiv (\mathbf{P}\{q(n+1) = 0, I(n+1) = i' \mid q(n) = 0, I(n) = i\}) = D_0.
\end{aligned} \tag{3.1.5}
$$

The matrices $\{A_0((k, j), (k, j')(k_1, j_1)),\ A_1((k, j), (k_1, j_1)),\ A_2(k, j),\ A_1(0, 0),\ A_0(0, (k_1, j_1))\}$ determine the one step transition of the Markov chain and thus play an important role in the analysis. A tree structure can be introduced in Ω in a way similar to that in HE and Alfa [9]. We call J in Ω a node in a tree and $J = 0$ is the

root of the tree. It is clear that every node J in the tree has $m_1+ m_2+ ...+m_K$ children and (except node 0) one parent. In any node, the Markov chain ($q(n)$, $I(n)$) can only transit to its children, itself, its parent (except node 0), its parent's children, and its parent's grandchildren in one transition. Therefore, ($q(n)$, $I(n)$) is an *M/G/*1 type Markov chain with a tree structure (Takine, Sengupta, and Yeung [20]), where $I(n)$ is the auxiliary variable. In the next two sections, the theory about the *M/G/*1 type Markov chains with a tree structure will be utilized to study ($q(n)$, $I(n)$).

Note 3.1.2: The LCFS-resume case and the LCFS-repeat case were dealt with differently in HE and Alfa [9]. Since the LCFS-resume and LCFS-repeat are two special cases of LCFS-GPR, this section shows that the two cases can be treated uniformly.

Note 3.1.3: If the service phase of a reentering customer is determined when the customer reenters the server, ($q(n)$, $I(n)$) is no longer a Markov chain. In that case, the technique that was developed in HE and Alfa [9] in dealing with the preemptive repeat case can be used, if the objective is to analyze the busy periods and sojourn times. If the objective is to analyze the stationary distribution, the formulation approach introduced in Section 3.4 can be used.

2 Stability issues

Let $\rho = \lambda_1/\mu_1 + \ldots+\lambda_K/\mu_K$. It has been proved in HE [8] that the queueing system of interest is stable if $\rho < 1$ when a work conserving service discipline is applied. However, for some $\{Q_k,\ 1\le k\le K\}$, the service discipline may not be work conserving. This brings up two interesting issues: 1) for a given set of $\{Q_k,\ 1\le k\le K\}$, is the queueing system stable? 2) how do the matrices $\{Q_k,\ 1\le k\le K\}$ influence the stability of the queueing system? This section shows that the results obtained in HE [8] can be used to answer the two questions.

First, the following set of matrices is introduced. Denote by $\{\overline{G}(k, j),\ 1\le k\le K,\ 1\le j\le m_k\}$ a set of stochastic matrices that satisfy the following equations,

$$\begin{aligned}\overline{G}(k, j) = A_2(k, j) &+ \sum_{k_1=1}^{K}\sum_{j_1=1}^{m_{k_1}} A_1((k, j),(k_1, j_1))\overline{G}(k_1, j_1) \\ &+ \sum_{j'=1}^{m_k}\sum_{k_1=1}^{K}\sum_{j_1=1}^{m_{k_1}} A_0((k, j),(k, j')(k_1, j_1))\overline{G}(k_1, j_1)\overline{G}(k, j').\end{aligned} \tag{3.2.1}$$

It has been proved (see HE [8]) that the matrix set $\{\overline{G}(k, j),\ 1\le j\le m_k,\ 1\le k\le K\}$ exists, but may not be unique. It has also been proved that the matrix set

$\{\overline{G}(k, j),\ 1\le j\le m_k,\ 1\le k\le K\}$ is unique and all these matrices are stochastic when the Markov chain $(q(n), I(n))$ is positive recurrent. In fact, $\{\overline{G}(k, j),\ 1\le j\le m_k,\ 1\le k\le K\}$ is the minimal nonnegative solution to equation (3.2.1) when the Markov chain is positive recurrent. From the matrix set $\{\overline{G}(k, j),\ 1\le j\le m_k,\ 1\le k\le K\}$, we introduce the following $m\times m$ matrices, $1\le j\le m_k$, $1\le j_1\le m_{k_1}$, $1\le k, k_1\le K$,

$$
\begin{aligned}
&p((k, j),(k_1, j_1))\\
&=\begin{cases}
A_1((k, j),(k_1, j_1)) + \sum_{j'=1}^{m_k} A_0((k, j),(k, j')(k_1, j_1)), & \text{if} \quad k \ne k_1;\\
A_1((k, j),(k, j_1)) + \sum_{j'=1}^{m_k} A_0((k, j),(k, j')(k, j_1))\\
\quad + A_0((k, j),(k, j_1)(k, j_1))\overline{G}(k, j_1)\\
\quad + \sum_{k'=1}^{K} \sum_{j'=1:\,(k', j')\ne(k, j_1)}^{m_{k'}} A_0((k, j),(k, j_1)(k', j'))\overline{G}(k', j'), & \text{if} \quad k = k_1,
\end{cases}
\end{aligned}
\tag{3.2.2}
$$

and

$$
P=\begin{pmatrix}
p((1,1),(1,1)) & p((1,1),(1,2)) & \cdots & p((1,1),(K,m_K))\\
p((1,2),(1,1)) & p((1,2),(1,2)) & \cdots & p((1,2),(K,m_K))\\
\vdots & \vdots & \vdots & \vdots\\
p((K,m_K),(1,1)) & p((K,m_K),(1,2)) & \cdots & p((K,m_K),(K,m_K))
\end{pmatrix}. \tag{3.2.3}
$$

Denote by $sp(P)$ the Perron-Frobenius eigenvalue (the eigenvalue with the largest modulus) of the matrix P. The stability results of the queueing system are summarized in the following theorem. See Theorem 3.2 in HE [8] for a proof.

Theorem 3.1 The queueing system introduced in Section 3.1 is stable if and only if $sp(P) < 1$. More specifically, the Markov chain $(q(n), I(n))$ is

a) positive recurrent if and only if $sp(P) < 1$;
b) null recurrent if and only if $sp(P) = 1$;
c) transient if and only if $sp(P) > 1$.

When $Q_k = I$, $1\le k\le K$, i.e., the service discipline is preemptive resume (and hence work-conserving), $sp(P)$ is equivalent to ρ in classifying the corresponding Markov chain. ⌘

Next, we use Theorem 3.1 to study the impact of the matrices $\{Q_k, 1 \le k \le K\}$ on the stability of the queueing system through a numerical example.

Example 3.2.1 Consider an *MMAP*[2]/*PH*[2]/1/LCFS-GPR queue. For the arrival process: $K = 2$, $m = 2$,

$$D_0 = \begin{pmatrix} 0.4 & 0.15 \\ 0.3 & 0.3 \end{pmatrix}, \; D_1 = \begin{pmatrix} 0.15 & 0.3 \\ 0 & 0 \end{pmatrix}, \; D_2 = \begin{pmatrix} 0 & 0 \\ 0.4 & 0 \end{pmatrix}.$$

It is interesting to see that each type 2 customer is likely to be followed immediately by a type 1 customer. Numerical results show that the arrival pattern has much influence on the queueing process. The service times are given as follows:

$$m_1 = 2, \quad \alpha_1 = (0.5, 0.5), \quad T_1 = \begin{pmatrix} 0.7 & 0.2 \\ 0 & 0.2 \end{pmatrix}, \quad Q_1 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix};$$

$$m_2 = 2, \quad \alpha_2 = (0.4, 0.6), \quad T_2 = \begin{pmatrix} 0.2 & 0.2 \\ 0.2 & 0.3 \end{pmatrix}, \quad Q_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

It can be obtained that $sp(P) = 0.909421$. Thus, the queueing system is stable. However, if we switch the two GPRs $\{Q_1, Q_2\}$ between the two types of customers, $sp(P)$ becomes 1.013271, i.e., the queueing system becomes unstable. This example shows that, when multiple types of customers are present, the impact of the service discipline on the queueing process is significant.

To learn more about the impact of the GPR on the stability of the queueing system, we look at the preemptive resume and repeat cases. First, the classical traffic intensity of this queueing system is given as $\rho = \lambda_1/\mu_1 + \lambda_2/\mu_2 = 0.2739/0.369276 + 0.156532/0.541666 = 1.030811$. Thus, the queueing system with any work conserving service discipline is unstable. Therefore, the preemptive resume queueing system ($Q_1 = Q_2 = I$) is unstable. It can be obtained that $sp(P) = 1.011458$ for that case. On the other hand, $sp(P) = 0.980069$ for the preemptive repeat case ($Q_1 = \mathbf{e}\alpha_1$, $Q_2 = \mathbf{e}\alpha_2$), i.e., the queueing system is stable.

For the rest of Section 3, we assume that $sp(P) < 1$, i.e., the queueing system is stable. We shall study the busy periods, sojourn times, and the stationary distribution of the queue string. The impact of the LCFS-GPR on these performance measures is discussed numerically.

3 The fundamental periods, busy periods, and sojourn times

In general, a fundamental period is defined as the first passage time during which the (total) queue length decreases by one. Define $\mathbf{N} = \{\mathbf{n} = (n_1, ..., n_K): n_k \geq 0, 1 \leq k \leq K\}$. Similar to the classical QBD case (see Neuts [14]) and the LCFS repeat case (HE and Alfa [9]), we define, for $J = (k_1, j_1)(k_2, j_2)\ldots(k_t, j_t) \in \Omega$, $1 \leq k \leq K$, $1 \leq j \leq m_k$, $1 \leq i, i' \leq m$, and $\mathbf{n} = (n_1, ..., n_K) \in \mathbf{N}$,

$g_{i, i'}(k, j, x, \mathbf{n})$: the taboo probability that the Markov chain $(q(n), I(n))$ reaches node J for the first time in state (J, i') in no more than x units of time and there are n_1 type 1, n_2 type 2, …, and n_K type K customers served during this time, given that the Markov chain started in $(J+(k, j), i)$.

Let $G(k, j, x, \mathbf{n})$ be an $m \times m$ matrix with elements $g_{i,i'}(k, j, x, \mathbf{n})$, $1 \leq i, i' \leq m$. Because of the special structure of the $M/G/1$ type Markov chain, $G(k, j, x, \mathbf{n})$ does not depend on the node J. $G(k, j, x, \mathbf{n})$ is defined for a busy period with a customer (k, j) initially, for $0 \leq j \leq m_k$, $1 \leq k \leq K$. In a similar way, define $G(0, x, \mathbf{n})$ for a busy cycle (from the beginning of an idle period to the beginning of the next idle period). Let $G^*(k, j, \omega, \mathbf{z})$ be the joint probability generating function of $G(k, j, x, \mathbf{n})$ with respect to x and $\mathbf{n}$, i.e.,

$$G^*(k, j, \omega, \mathbf{z}) = \sum_{x=1}^{\infty} \sum_{\mathbf{n} \in \mathbf{N}} \omega^x \Delta_x G(k, j, x, \mathbf{n}) \prod_{l=1}^{K} z_l^{n_l}, \quad 0 < \omega, z_l < 1. \tag{3.3.1}$$

where $\mathbf{z} = (z_1, ..., z_K)$ and $\Delta_x G(k, j, x, \mathbf{n}) = G(k, j, x, \mathbf{n}) - G(k, j, x-1, \mathbf{n})$. Then it can be proved that $\{G^*(k, j, \omega, \mathbf{z}), 1 \leq j \leq m_k, 1 \leq k \leq K\}$ are the minimal nonnegative solutions to the equations:

$$G^*(0, \omega, \mathbf{z}) = [\omega I - A_1(0,0)]^{-1} \sum_{k=1}^{K} \sum_{j=1}^{m_k} A_0(0, (k, j)) G^*(k, j, \omega, \mathbf{z}); \tag{3.3.2}$$

$$\begin{aligned} & G^*(k, j, \omega, \mathbf{z}) \\ & = \left[\omega I - A_1^*((k, j), (k, j), z_k)\right]^{-1} \Bigg[z_k A_2(k, j) \\ & \quad + \sum_{k_1=1}^{K} \sum_{j_1=1:\,(k_1, j_1) \neq (k, j)}^{m_{k_1}} A_1^*((k, j), (k_1, j_1), z_k) G^*(k_1, j_1, \omega, \mathbf{z}) \\ & \quad + \sum_{j'=1}^{m_k} \sum_{k_1=1}^{K} \sum_{j_1=1}^{m_{k_1}} A_0((k, j), (k, j')(k_1, j_1)) G^*(k_1, j_1, \omega, \mathbf{z}) G^*(k, j', \omega, \mathbf{z}) \Bigg], \end{aligned} \tag{3.3.3}$$

where

$$A_1^*((k,j),(k',j'),z_k) = \begin{cases} z_k(\mathbf{T}_k^0)_j(\alpha_{k'})_{j'}D_{k'} + (T_k)_{j,j'}D_0, & k = k', 1 \le j, j' \le m_k; \\ z_k(\mathbf{T}_k^0)_j(\alpha_{k'})_{j'}D_{k'}, & k \ne k', 1 \le j \le m_k, 1 \le j' \le m_{k'}. \end{cases} \tag{3.3.4}$$

Denote by $G(k, j) = G^*(k, j, 1\text{-}, \mathbf{1}\text{-})$, $0 \le j \le m_k$, $1 \le k \le K$, and $G(0) = G^*(0, 1\text{-}, \mathbf{1}\text{-})$, where $\mathbf{1}\text{-} = (1\text{-}, \ldots, 1\text{-})$. When the $M/G/1$ type Markov chain is positive recurrent, matrices $G(k, j)$ and $G(0)$ are stochastic matrices. Matrices $\{G(k, j), 1 \le j \le m_k, 1 \le k \le K\}$ are the minimal nonnegative solutions to equation (3.2.1). When the Markov chain is positive recurrent, $G(k, j) = \overline{G}(k, j)$, $1 \le j \le m_k$, $1 \le k \le K$. By equation (3.3.2), the matrix $G(0)$ is obtained as

$$G(0) = [I - A_1(0,0)]^{-1} \sum_{k=1}^{K} \sum_{j=1}^{m_k} A_0(0,(k, j))G(k, j). \tag{3.3.5}$$

The moments of the number of customers served in a busy period (busy cycle) and the moments of the length of a busy period (busy cycle) can be derived using equations (3.3.2) and (3.3.3). For instance, let

$$\mathbf{u}(k, j) = \left.\frac{\partial G^*(k, j, \omega, \mathbf{z})\mathbf{e}}{\partial \omega}\right|_{\omega=1-,\, \mathbf{z}=\mathbf{1}-}, \quad \mathbf{u}(0) = \left.\frac{\partial G^*(0, \omega, \mathbf{z})\mathbf{e}}{\partial \omega}\right|_{\omega=1-,\, \mathbf{z}=\mathbf{1}-}, \tag{3.3.6}$$

for $1 \le k \le K$ and $1 \le j \le m_k$. The term $(\mathbf{u}(k, j))_i$ is the mean length of a busy period started with a type k customer with initial service phase j, and the initial phase of the underlying Markov process D is i, $1 \le i \le m$. Similar interpretations go to $\mathbf{u}(0)$. Simple but lengthy calculations lead to the following expressions, for $1 \le k \le K$, $1 \le j \le m_k$,

$$\mathbf{u}(0) = [I - A_1(0,0)]^{-1} \left[\mathbf{e} + \sum_{k=1}^{K} \sum_{j=1}^{m_k} A_0(0,(k, j))\mathbf{u}(k, j) \right]; \tag{3.3.7}$$

$$\begin{aligned} \mathbf{u}(k, j) = \mathbf{e} &+ \sum_{k_1=1}^{K} \sum_{j_1=1}^{m_{k_1}} A_1((k, j),(k_1, j_1))\mathbf{u}(k_1, j_1) \\ &+ \sum_{j'=1}^{m_k} \sum_{k_1=1}^{K} \sum_{j_1=1}^{m_{k_1}} A_0((k, j),(k, j')(k_1, j_1))[\mathbf{u}(k_1, j_1) + G(k_1, j_1)\mathbf{u}(k, j')]. \end{aligned} \tag{3.3.8}$$

Performance measures of interest can be obtained by solving equations (3.3.2), (3.3.3), (3.3.5), (3.3.7), and (3.3.8). Although the formulas look formidable, the actual programming is not difficult to implement. In fact, equations (3.3.7) and (3.3.8) can be solved as linear equations or using an iteration method. The space complexity of the algorithm is $O(m^2(m_1+\ldots+m_K)^2)$.

The probability that the queueing system is idle at an arbitrary time can be obtained as follows. Let **g** be the left invariant vector of the matrix $G(0)$. Then **g** is the stationary distribution of the underlying Markov chain D at the beginning of an arbitrary idle period (busy cycle). The mean length of an arbitrary busy cycle is thus given by $\mathbf{gu}(0)$ and the mean length of an arbitrary idle period is given by $\mathbf{g}[I-A_1(0,0)]^{-1}\mathbf{e}$. Then it has

$$\pi(-1)\mathbf{e} = \frac{\mathbf{g}[I - A_1(0,0)]^{-1}\mathbf{e}}{\mathbf{gu}(0)}, \tag{3.3.9}$$

where $\pi(-1)\mathbf{e}$ represents the probability that the queueing system is idle at an arbitrary time (see Section 3.4 for the definition of the vector $\pi(-1)$).

Finally, in this section, we study the ***sojourn time***. The sojourn time of a customer is defined as the total time that the customer stays in the queueing system. Since all customers are served on a last-come-first-served basis, the sojourn time can be obtained directly from the fundamental periods. In steady state, let w_k be the sojourn time of a type k customer. Let $(\mathbf{w}^*_k(\omega))_i$ be the probability generating function of w_k when the phase of the Markov arrival process at the departure epoch of the customer of interest is i, $1\le i\le m$. It is easy to obtain, for $1\le k\le K$,

$$\begin{aligned}
&\mathbf{w}^*_k(\omega) = \frac{1}{\lambda_k}\theta D_k\left(\sum_{j=1}^{m_k}(\alpha_k)_j G^*(k, j, \omega, \mathbf{1}\text{-})\right);\\
&\mathbf{E}w_k = \frac{1}{\lambda_k}\theta D_k\left(\sum_{j=1}^{m_k}(\alpha_k)_j \mathbf{u}(k, j)\right).
\end{aligned} \tag{3.3.10}$$

The sojourn time of an arbitrary customer has a probability generating function given as $\mathbf{w}^*(\omega) = [\lambda_1\mathbf{w}^*_1(\omega) + \ldots + \lambda_K\mathbf{w}^*_K(\omega)]/(\lambda_1+\ldots+\lambda_K)$ and its mean can be obtained easily from equation (3.3.10).

Example 3.3.1 For the queueing system introduced in Example 3.2.1, it can be obtained that the mean busy cycle is 13.765331 and the mean idle period is

2.315323. Thus, the probability that the queue system is idle at an arbitrary time is $\pi(-1)\mathbf{e} = 0.1682$. That is: approximately 17% of the time the queueing system is idle. The mean sojourn time of type 1 customers is $\mathbf{E}w_1 = 12.53449$ and $\mathbf{E}w_2 = 23.946754$ for type 2 customers. Since each type 2 customer is likely to be followed by a type 1 customer who has a longer service time, the mean sojourn time of a type 2 customer is much longer than that of a type 1 customer. As was shown in Example 3.2.1, a simple switch of the GPRs $\{Q_1, Q_2\}$ can prolong service so much so that, in the long term, the system will always be busy and the sojourn time increases to infinity.

4 Matrix geometric solution of the queue string

According to Yeung and Sengupta [22], a matrix geometric solution exists for a positive recurrent *GI/M*/1 type Markov chain with a tree structure. Unfortunately, the Markov chain ($q(n)$, $I(n)$) can go from a node to one of its parent's grandchildren in one transition (e.g., from $J+(k, j)$ to $J+(k, j')(k_1, j_1)$ where $j' \neq j$). Thus, ($q(n)$, $I(n)$) is not a *GI/M*/1 type Markov chain with a tree structure. However, the special structure possessed by ($q(n)$, $I(n)$) leads to a method to formulate the queueing process into a *GI/M*/1 type Markov chain with a tree structure. The idea is to consider the status of the customer in service as a part of the auxiliary variable, rather than a part of the queue string. The same idea was used in HE [6] for the LCFS non-preemption case. Specifically, we consider the Markov chain ($q_W(n)$, $I(n)$, $k(n)$, $J(n)$) defined as follows:

$q_W(n)$: the string consisting of the states of customers waiting in the queue at time n;
$I(n)$: the phase of the *MMAP*[*K*] at time n (introduced in Section 3.1);
$k(n)$: the type of the customer in service (if any) at time n;
$J(n)$: the phase of the service time (if any) at time n.

Random variable $q_W(n)$ takes value –1 when there is no customer in the queueing system, 0 when there is one customer in the queueing system. Thus, $q_W(n)$ takes values in $\{-1\}\cup\Omega$, where Ω was defined in Section 3.1. Consider ($I(n)$, $k(n)$, $J(n)$) as an auxiliary variable. The one step transitions of ($q_W(n)$, $I(n)$, $k(n)$, $J(n)$) are determined by its transition blocks given as follows.

a) $q_W(n)$ goes (in one step) from –1 to –1 and from –1 to 0:

$$\hat{A}_0(-1,0) = \begin{pmatrix} D_1 \otimes \alpha_1 & \cdots & D_K \otimes \alpha_K \end{pmatrix}; \qquad \hat{A}_1(-1,-1) = D_0, \tag{3.4.1}$$

where $\otimes$ represents the Kronecker product of matrices (see Gantmacher [4]).

b) $q_W(n)$ goes from 0 to –1:

$$\hat{A}_2(0,-1) = D_0 \otimes \begin{pmatrix} \mathbf{T}_1^0 \\ \vdots \\ \mathbf{T}_K^0 \end{pmatrix}. \tag{3.4.2}$$

Matrices for transitions from 0 to 0 and from 0 to (k, j) are given in c).

c) $q_W(n)$ goes from $J + (k, j)$ to $J + (k, j)(k_1, j_1)$, $J + (k, j)$, or J (J in Ω):

$$\hat{A}_0(k_1, j_1) = \sum_{t=1}^{K} D_t \otimes \left[\begin{pmatrix} 0 \\ (T_{k_1} Q_{k_1})_{j_1} \alpha_t \\ 0 \end{pmatrix} (0,\cdots,0,I_{m_t \times m_t},0,\cdots,0) \right]; \tag{3.4.3}$$

$$\begin{aligned} &\hat{A}_1(k, j)(= \hat{A}_1) \\ &\quad = D_0 \otimes \begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_K \end{pmatrix} + \sum_{t=1}^{K} D_t \otimes \left[\begin{pmatrix} \mathbf{T}_1^0 \alpha_t \\ \vdots \\ \mathbf{T}_K^0 \alpha_t \end{pmatrix} (0,\cdots,0,I_{m_t \times m_t},0,\cdots,0) \right]; \end{aligned} \tag{3.4.4}$$

$$\hat{A}_2(k, j) = D_0 \otimes \left[\begin{pmatrix} \mathbf{T}_1^0 \mathbf{e}(j) \\ \vdots \\ \mathbf{T}_K^0 \mathbf{e}(j) \end{pmatrix} (0,\cdots,0,I_{m_k \times m_k},0,\cdots,0) \right], \tag{3.4.5}$$

where $(T_{k_1} Q_{k_1})_{j_1}$ represents the j_1th column of the matrix $T_{k_1} Q_{k_1}$. Note that $\hat{A}_1(k, j)$ is independent of (k, j). Matrices $\hat{A}_0(k, j)$, $\hat{A}_1$, and $\hat{A}_2(k, j)$ are of the dimension $m(m_1+\ldots+m_K)$.

Note 3.4.1: This formulation approach can be used to handle the case when the service phase of a customer reentering the server is chosen at the epoch it reenters. The transition matrices are obtained by simply moving $\{Q_k, 1 \le k \le K\}$ from equation (3.4.3) to equation (3.4.5) and putting them in the appropriate positions. Details are omitted.

Clearly, $(q_W(n), I(n), k(n), J(n))$ is a Markov chain with a tree structure. Since $(q_W(n), I(n), k(n), J(n))$ does not transit in one step from any node to its parent's children (its siblings) nor to its parent's grandchildren, it is a *GI/M*/1 type Markov chain with a tree structure. In fact, it is a simple QBD Markov chain with a

tree structure. According to Yeung and Sengupta [22], a matrix geometric solution can be found for the stationary distribution of the Markov chain. Define, for $J \in \Omega$,

$$\begin{aligned} &\pi(J,i,k,j) = \lim_{n\to\infty} \mathbf{P}\{(q_w(n), I(n), k(n), J(n)) = (J,i,k,j)\}; \\ &\pi(-1,i) = \lim_{n\to\infty} \mathbf{P}\{(q_w(n), I(n)) = (-1,i)\}. \end{aligned} \tag{3.4.6}$$

Note that the initial condition is not shown explicitly in equation (3.4.6) since the limits are independent of it when the Markov chain is positive recurrent. Denote by

$$\begin{aligned} &\pi(J,i,k) = (\pi(J,i,k,1),\cdots,\pi(J,i,k,m_k)); \ \pi(J,i) = (\pi(J,i,1),\cdots,\pi(J,i,K)); \\ &\pi(J) = (\pi(J,1),\cdots,\pi(J,m)), \quad J \neq -1; \qquad \pi(-1) = (\pi(-1,1),\cdots,\pi(-1,m)). \end{aligned} \tag{3.4.7}$$

Theorem 3.2 When the queueing system of interest is stable, the stationary distribution of $(q_W(n), I(n), k(n), J(n))$ is given by

$$\begin{aligned} &\pi(J+(k,j)) = \pi(J)R(k,j), \quad \text{for} \quad J \in \Omega, \ 1 \le k \le K, 1 \le j \le m_k; \\ &\pi(0) = \pi(-1)\hat{A}_0(-1,0) + \pi(0)\hat{A}_1(0,0) + \sum_{1\le k\le K} \sum_{1\le j\le m_k} \pi(0)R(k,j)\hat{A}_2(k,j); \\ &\pi(-1) = \pi(-1)\hat{A}_1(-1,-1) + \pi(0)\hat{A}_2(0,-1); \\ &\pi(-1)\mathbf{e} + \pi(0)(I-R)^{-1}\mathbf{e} = 1, \end{aligned} \tag{3.4.8}$$

where $R = \sum_{1\le k\le K} \sum_{1\le j\le m_k} R(k,j)$ and the matrices $\{R(k,j), 1\le j\le m_k, 1\le k\le K\}$ are defined as the minimal nonnegative solutions to the matrix equations:

$$R(k,j) = \hat{A}_0(k,j) + R(k,j)\hat{A}_1 + R(k,j) \sum_{1\le k_1\le K} \sum_{1\le j_1\le m_{k_1}} R(k_1,j_1)\hat{A}_2(k_1,j_1). \tag{3.4.9}$$

The computation of $\{R(k,j), 1\le j\le m_k, 1\le k\le K\}$ can be carried out using an algorithm given in Yeung and Alfa [21] and an algorithm given in Yeung and Sengupta [22].

Note 3.4.2: For the continuous time *MMAP*[*K*]/*PH*[*K*]/1/LCFS-GPR queue, the Markov chain introduced in Section 3.1 does not transit from a node to its parent's grandchildren. Therefore, that is a QBD Markov chain with a tree structure and a matrix geometric solution exists. This implies that there is truly a uniform approach for the continuous time queue. This reveals one of the differences between continuous and discrete time stochastic models.

With the matrix geometric solution of $(q_W(n), I(n), k(n), J(n))$, an efficient

algorithm can be developed for computing the queue string distribution and the mean queue length. For brevity, all the details are omitted.

Example 3.4.1 Consider the queueing system introduced in Example 3.2.1. First, the mean number of customers in the queueing system is 2.7927. The mean number of waiting customers is 1.9645. The average queue length is surprisingly small. We can also have a look at the composition of the queue. For instance, when there are five customers in the queue, the probabilities of some combinations of customers and their corresponding service phases are shown in Table 3.4.1.

Table 3.4.1 Stationary distribution of the queue string

Queue string $J = (k_1, j_1)\ldots(k_5, j_5)$	Probability of J
(1, 2)(1, 2)(1, 2)(1, 2)(1, 2)	0.00320
(2, 2)(2, 2)(2, 2)(2, 2)(2, 2)	0.00003
(1, 2)(2, 2)(1, 2)(2, 2)(1, 2)	0.00085
(2, 2)(1, 2)(2, 2)(1, 2)(2, 2)	0.00029

Note that the service phases of all the waiting customers are 2. It is so chosen since type 1 customer's phase is always 2 when waiting in the queue. Table 3.4.1 demonstrates clearly that the occurrences of these combinations are dramatically different. J = (2, 2)(2, 2)(2, 2)(2, 2)(2, 2) has the smallest probability since every type 2 customer is likely to be followed by a type 1 customer. Thus, the probability of five type 2 customers in a row is small.

4. Discrete time *MMAP*[*K*]/*PH*[*K*]/1/*c*/LCFS-GPR queue

In this section, we consider the discrete time *MMAP*[*K*]/*PH*[*K*]/1/*c*/LCFS-GPR queue. This queueing system is the same as the queueing system introduced in Section 3.1 except that there are only total c waiting space. Thus, there can be at most $1+c$ customers in the queueing system at any time. When a customer arrives and there are already $1+c$ customers in the system, the customer does not enter the system and is lost forever.

In Section 4.1, we introduce a Markov chain to represent the queueing process in this queueing system. In Section 4.2, the Gaussian elimination method is used to develop an efficient algorithm for computing the stationary distribution of the queue string. Finally, in Section 4.3, the loss probabilities are given.

4.1 The Markov chain

In order to analyze the queue length, we use the same notation introduced in Section 3.4. It is easy to see that $(q_W(n), I(n), k(n), J(n))$ is still an irreducible and aperiodic Markov chain. The state space of $q_W(n)$ of the Markov chain is $\Omega_C \times \{1, 2, \ldots, m\}$, where $\Omega_C = \{J: J \in \Omega \cup \{-1\} \text{ and } |J| \le c\}$. Furthermore, $\{(q_W(n), I(n), k(n), J(n)), n \ge 0\}$ can be coupled with the QBD Markov chain with a tree structure introduced in Section 3.4 until the total number of customers in the queueing system reaches $1 + c$. The transition matrices of the Markov chain are the same as those given in Section 3.4 except for the nodes with c customers waiting in the queue. When there are c customer waiting in the queue, no new arrival can enter the queueing system. Therefore, the transition matrices of $\{(q_W(n), I(n), k(n), J(n)), n \ge 0\}$ are given by equations (3.4.1) – (3.4.5) and,

d) for $|J+(k, j)| = c$, when the service is completed and a new customer arrives, there is no service completion and no new arrival, or a new customer arrives and there is no service completion, i.e., from $J+(k, j)$ to $J+(k, j)$:

$$\hat{A}_{1,c}(k, j)(= \hat{A}_{1,c})$$
$$= D \otimes \begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_K \end{pmatrix} + \sum_{t=1}^{K} D_t \otimes \left[\begin{pmatrix} \mathbf{T}_1^0 \alpha_t \\ \vdots \\ \mathbf{T}_K^0 \alpha_t \end{pmatrix} (0, \cdots, 0, I_{m_t \times m_t}, 0, \cdots, 0) \right]. \quad (4.1.1)$$

It is worth to point out that if the objective is to analyze the busy periods or the sojourn times, it might be better to use the formulation introduced in Section 3.1.

4.2 Distribution of the queue string: a Gaussian elimination method

Since the Markov chain $\{(q_W(n), I(n), k(n), J(n)), n \ge 0\}$ now has finite states, there is no simple matrix geometric solution for its stationary distribution. Nonetheless, the special tree structure can be used again for developing algorithms for computing the stationary distribution of the queue string. In this section, the Gaussian elimination method is utilized. Similar to Section 3.4, we use $\{\pi(J), J \in \Omega_C\}$ to denote the stationary distribution of the Markov chain. By definition, $\{\pi(J), J \in \Omega_C\}$ satisfy the following equations:

$$
\begin{aligned}
&\pi(-1) = \pi(-1)\hat{A}_1(-1,-1) + \pi(0)\hat{A}_2(0,-1);\\
&\pi(0) = \pi(-1)\hat{A}_0(-1,0) + \pi(0)\hat{A}_1 + \sum_{k=1}^{K}\sum_{j=1}^{m_k}\pi(k,j)\hat{A}_2(k,j);\\
&\pi(J+(k,j)) = \pi(J)\hat{A}_0(k,j) + \pi(J+(k,j))\hat{A}_1\\
&\qquad + \sum_{1\le k_1\le K}\sum_{1\le j_1\le m_{k_1}}\pi(J+(k,j)(k_1,j_1))\hat{A}_2(k_1,j_1),\\
&\qquad\qquad 1\le k\le K, 1\le j\le m_k, |J+(k,j)|< c;
\end{aligned}
\tag{4.2.1}
$$

$$
\begin{aligned}
&\pi(J+(k,j)) = \pi(J)\hat{A}_0(k,j) + \pi(J+(k,j))\hat{A}_{1,c}, \quad |J+(k,j)|= c;\\
&\sum_{J\in\Omega_c}\pi(J)\mathbf{e} = 1,
\end{aligned}
$$

The stationary distribution right after an arbitrary customer is given by $\{\pi(J)(D-D_0)/\lambda,\ J\in\Omega_c\}$. The stationary distribution right after an arbitrary type k customer is given by $\{\pi(J)D_k/\lambda_k,\ J\in\Omega_c\}$, $1\le k\le K$.

Since no matrix geometric solution exists, we shall develop a computational method for computing the stationary distribution using the Gaussian elimination method. The idea is to eliminate all $\pi(J+(j,k))$ with $|J+(j,k)| = c$ by expressing them in terms of $\pi(J)$. Then inductively eliminate other subsets of nodes with $|J| = n$ for $1\le n<c$. Eventually, we solve an equation for $\pi(-1)$. Then we calculate all $\pi(J)$ backwards. Details are given as follows. For $|J+(k,j)| = c$, the fourth expression in equation (4.2.1) leads to

$$
\pi(J+(k,j)) = \pi(J)\hat{A}_0(k,j)[I-\hat{A}_{1,c}]^{-1} \equiv \pi(J)\hat{R}_c(c,k,j). \tag{4.2.2}
$$

Notice that $\hat{R}_c(c,k,j)$ is independent of the string J. Suppose that $\pi(J+(k,j)) = \pi(J)\hat{R}_c(n,k,j)$ holds for all $n = |J+(k,j)| < c+1$. With equations (4.2.1), we obtain

$$
\begin{aligned}
\pi(J+(k,j)) &= \pi(J)\hat{A}_0(k,j) + \pi(J+(k,j))\hat{A}_1\\
&\quad + \sum_{1\le k_1\le K}\sum_{1\le j_1\le m_{k_1}}\pi(J+(k,j))\hat{R}_c(|J|+2,k_1,j_1)\hat{A}_2(k_1,j_1).
\end{aligned}
\tag{4.2.3}
$$

Then the relationship among $\{\hat{R}_c(n,k,j)\}$ are given as follows, for $1\le n < c$,

$$\hat{R}_c(n,k,j) = \hat{A}_0(k,j)\left[I - \hat{A}_1 - \sum_{k_1=1}^{K}\sum_{j_1=1}^{m_{k_1}} \hat{R}_c(n+1,k_1,j_1)\hat{A}_2(k_1,j_1) \right]^{-1}. \qquad (4.2.4)$$

For $J = 0$, $\pi(0) = \pi(-1)\hat{A}_0(-1,0) + \pi(0)\hat{A}_1(0,0) + \sum_{k=1}^{K}\sum_{j=1}^{m_k} \pi(0)\hat{R}_c(1,k,j)\hat{A}_2(k,j)$. Then we have $\pi(0) = \pi(-1)\hat{R}_c(0)$, where

$$\hat{R}_c(0) = \hat{A}_0(-1,0)\left[I - \hat{A}_1 - \sum_{k=1}^{K}\sum_{j=1}^{m_k} \hat{R}_c(1,k,j)\hat{A}_2(k,j) \right]^{-1}. \qquad (4.2.5)$$

Finally, the vector $\pi(-1)$ satisfies the equation

$$\pi(-1) = \pi(-1)[\hat{A}_1(-1,-1) + \hat{R}_c(0)\hat{A}_2(0,-1)]. \qquad (4.2.6)$$

We summarize the results in the following theorem.

Theorem 4.1 For the discrete time *MMAP*[*K*]/*PH*[*K*]/1/*c*/LCFS-GPR queue, we have, for $J = (k_1, j_1)\ldots(k_n, j_n)$ and $1 \le n \le c$,

$$\begin{aligned} &\pi(0) = \pi(-1)\hat{R}_c(0); \\ &\pi(J) = \pi(-1)\hat{R}_c(0)\hat{R}_c(1,k_1,j_1)\cdots\hat{R}_c(n,k_n,j_n); \qquad (4.2.7) \\ &\pi(-1)[I + \hat{R}_c(0) + \hat{R}_c(0)\hat{R}_c(1) + \cdots + \hat{R}_c(0)\hat{R}_c(1)\cdots\hat{R}_c(c)]\mathbf{e} = 1, \end{aligned}$$

where $\hat{R}_c(n) = \sum_{k=1}^{K}\sum_{j=1}^{m_k} \hat{R}_c(n,k,j)$ for $1 \le n \le c$. The stationary distribution of the queue string can be obtained by 1) calculating $\{\hat{R}_c(n,k,j),\ 1 \le n \le c,\ 1 \le k \le K,\ 1 \le j \le m_k\}$ and $\hat{R}_c(0)$; 2) solving equation (4.2.6) for $\pi(-1)$; 3) normalizing $\pi(-1)$ using the last expression in equation (4.2.7); 4) calculating string probabilities using equation (4.2.7). The space complexity of the algorithm is $O(cm^2(m_1+\ldots+m_K)^3)$, which increases linearly in c. ⌘

Equation (4.2.7) shows that the stationary distribution has a product form and is close to a matrix geometric solution. In fact, numerical results show that $\hat{R}_c(n,k,j)$ converges to $R(k,j)$ when c goes to infinity for any fixed n, k, and j. Thus, when c goes to infinity, equation (4.2.7) reduces to the matrix geometric

solution given in Theorem 3.2.

4.3 The loss probabilities

In this section, we focus on the loss probability of an arbitrary customer and the loss probability of an arbitrary type k customer. This makes it possible to investigate the relationship between the loss probability and the size of the waiting space c. Denote by P_{loss} the probability that an arriving customer is lost and $P_{loss}(k)$ the probability that an arriving type k customer is lost. It is easy to see that,

$$P_{loss} = \frac{1}{\lambda}\left(\sum_{J\in\Omega_c:\,|J|=c}\pi(J)\right)\left((D-D_0)\otimes\begin{pmatrix}T_1 & & \\ & \ddots & \\ & & T_K\end{pmatrix}\right)\mathbf{e} \tag{4.3.1}$$

$$P_{loss}(k) = \frac{1}{\lambda_k}\left(\sum_{J\in\Omega_c:\,|J|=c}\pi(J)\right)\left(D_k\otimes\begin{pmatrix}T_1 & & \\ & \ddots & \\ & & T_K\end{pmatrix}\right)\mathbf{e}\,. \tag{4.3.2}$$

Apparently, the relationship $P_{loss} = [\lambda_1 P_{loss}(1)+\ldots+\lambda_K P_{loss}(K)]/\lambda$ holds. We use the following example to illustrate the impact of c on the loss probabilities.

Example 4.3.1 Consider the queueing system introduced in Example 3.2.1. The loss probabilities (as a function of c) are given in Table 4.3.1.

Table 4.3.1 The loss probabilities

	$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 20$	$c = 30$
P_{loss}	0.266679	0.160926	0.037527	0.005462	0.000159	0.000005
$P_{loss}(1)$	0.268417	0.171606	0.039895	0.005787	0.000168	0.000005
$P_{loss}(2)$	0.263636	0.145797	0.034157	0.005003	0.000145	0.000004

Table 4.3.1 shows that when the size of the waiting space is moderate, the loss probabilities of the two types of customers can be significantly different. For other cases, these probabilities are close to each other. Another observation is that the loss probabilities decrease exponentially with respect to c. The reason is that the stationary distribution has a product form and $\hat{R}_c(n,k,j)$ converges to $R(k,j)$ when c goes to infinity for any fixed n, k, and j.

5. Extensions

We now briefly discuss a few directions for future research.

Multiple server queues Generalization to the multiple server case is not simple. Since the service discipline is LCFS, the system has to determine which customer in service should be pushed out when a new customer arrives and all servers are occupied at the moment. Naturally, the oldest customer in service can be chosen and be pushed to the queue. In this case, one has to record the sequence of all customers in service in order to construct a Markov chain for an analysis. Certain mechanism must be proposed for a mathematical formulation, which can be quite complicated, especially for the discrete time case.

For the finite waiting space case, the queueing system also has to determine which customer to be lost when a new customer finds that the system is full. In Section 4.1, it was assumed that the new arrival is lost. But it is possible that the new arrival is accepted and the oldest in the queue is pushed out of the queue and lost. The mathematical formulation becomes much involved for the case.

Batch arrival cases When customers may arrive in batches of any kind, the queueing process can be formulated into an *M/G/*1 type Markov chain with a tree structure. The stability of the queueing system can be determined using the results in HE [8]. An analysis of the busy period can also be carried out. An interesting special case is that when each batch consists of at most one customer from each type. An arrival process of this kind has wide application in modelling telecommunication networks (see Alfa and Fitzpatrick [1]).

A hybrid FCFS&LCFS-GPR service discipline Suppose that the service discipline is FCFS when there are N or less than N customers in the queueing system; otherwise, the LCFS-GRP service discipline is applied. We call this service discipline the hybrid FCFS&LCFS-GPR. For queueing systems with such service disciplines, a Markov chain with a network structure can be constructed. Algorithms can be developed for computing performance measures. But those algorithms may not be efficient (HE and Alfa [10]). The development of more efficient algorithms for such queueing system is a challenging problem.

REFERENCES

[1] Alfa, A.S. and G. J. Fitzpatrick, Waiting time distribution of a FIFO/LIFO *Geo*/*D*/1 queue, *INFOR*, **Vol 39**, No 1, 149-159, 1999.

[2] Asmussen, S. and G. Koole, Marked point processes as limits of Markovian arrival streams, *J. Appl. Prob*., **Vol 30**, 365-372, 1993.

[3] Doshi, B.T., An *M/G/*1 queue with a hybrid discipline, *AT&T Tech. Journal*,

Vol 62, #5, 1251-1271, 1983.
[4] Gantmacher, F.R., *The theory of matrices*, New York: Chelsea, 1959.
[5] HE, Q-M, Queues with marked customers, *Adv. Appl. Prob*. **Vol 28**, 567-587, 1996.
[6] HE, Q-M, Quasi-birth-and-death Markov processes with a tree structrue and a detailed analysis of the *MMAP*[*K*]/*PH*[*K*]/1 LCFS non-preemptive queue, *European Journal of Operations Research* , **Vol 120**/3, 641-656, 2000.
[7] HE, Q-M, Classification of Markov processes of *M/G/*1 type with a tree structure and its applications to queueing systems, *O.R. Letters*, **Vol 25**, 2000.
[8] HE, Q-M, Classification of Markov processes of matrix *M/G/*1 type with a tree structure and its applications to the *MMAP*[*K*]/*PH*[*K*]/1 queue, (1998) (submitted for publication)
[9] HE, Q-M and A.S. Alfa, The *MMAP*[*K*]/*PH*[*K*]/1 queues with a last-come-first-served preemptive service discipline, *Queueing Systems*, **Vol 28**, 269-291, 1998.
[10] HE, Q-M and A.S. Alfa, A Computational Approach for the Analysis of *MMAP*[*K*]/*PH*[*K*]/1 Queues with a Mixed FCFS and LCFS Service Discipline, (1999) (submitted for publication)
[11] HE, Q-M and Neuts, M.F., Markov arrival processes with marked transitions, *Stochastic Process and their Applications*, **Vol 74**/1, 37-52, 1998.
[12] Lucantoni, D.M., K.S. Meier-Hellstern, K.S., and M.F. Neuts, A single-server queue with server vacations and a class of non-renewal arrival processes, *Adv. in Appl. Prob*., **Vol 22**, 676-705, 1990.
[13] Neuts, M.F., A versatile Markovian point process, *J. Appl. Prob*., **Vol 16**, 764-779, 1979.
[14] Neuts, M.F., *Matrix-Geometric Solutions in Stochastic Models: An algorithmic Approach*, The Johns Hopkins University Press, Baltimore, 1981.
[15] Neuts, M.F., *Structured Stochastic Matrices of M/G/*1 *type and Their Applications*, Marcel Dekker, New York, 1989.
[16] Ramaswami, V., the *N/G/*1 queue and its detailed analysis, *Adv. in Appl. Prob*., **Vol 12**, 222-261, 1980.
[17] Schreiber, F., Eine Familie von warteprozeduren zweichen FIFO und LIFO, Archiv fur Elektronk und Uebertragungstechnik, **Vol. 30**, # 12, 497-501, 1976.
[18] Takagi, H., Queueing Analysis: *A Foundation of Performance Evaluation*, **Vol 1***: Vacation and Priority Systems, Part* 1, Elsevier Science Publisher, B.V., Amsterdam, 1990.
[19] Takine, T. and T. Hasegawa, The workload in the *MAP/G/*1 queue with state-dependent services its application to a queue with preemptive resume priority, *Stochastic Models*, **Vol 10**, 183-204, 1994.
[20] Takine, T., B. Sengupta, and R.W. Yeung, A generalization of the matrix *M/G/*1 paradigm for Markov chains with a tree structure, *Stochastic Models*, **Vol 11**, 411-421, 1995.

[21] Yeung, R.W. and A. S. Alfa, The quasi-birth-death type Markov chain with a tree structure, *Stochastic Models*, **Vol 15**/4, 639-659, 1999.
[22] Yeung, R.W. and B. Sengupta, Matrix product-form solutions for Markov chains with a tree structure, *Adv. Appl. Prob.*, **Vol 26**, No. 4, 965-987, 1994.