European Journal of Operational Research 120 (2000) 641–656
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/orms

Theory and Methodology

Quasi-birth-and-death Markov processes with a tree structure and the *MMAP*[*K*]/*PH*[*K*]/*N*/*LCFS* non-preemptive queue

Qi-Ming He *

Department of Industrial Engineering, DalTech, Dalhousie University, Halifax, Nova Scotia, Canada B3J 2X4

Received 1 April 1998; accepted 1 September 1998

Abstract

This paper studies a multi-server queueing system with multiple types of customers and last-come-first-served (LCFS) non-preemptive service discipline. First, a quasi-birth-and-death (QBD) Markov process with a tree structure is defined and some classical results of QBD Markov processes are generalized. Second, the *MMAP*[*K*]/*PH*[*K*]/*N*/*LCFS* non-preemptive queue is introduced. Using results of the QBD Markov process with a tree structure, explicit formulas are derived and an efficient algorithm is developed for computing the stationary distribution of queue strings. Numerical examples are presented to show the impact of the correlation and the pattern of the arrival process on the queueing process of each type of customer. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Queueing theory; Matrix analytic methods; Tree structure; Last-come-first-served; Quasi-birth-and-death Markov process

1. Introduction

This paper studies a multi-server queueing system with multiple types of customers and a last-come-first-served (LCFS) non-preemptive service discipline. The subject of this paper is interesting since results obtained in this paper have potential applications, especially in broadband communications systems using ATM as transfer mode. In these systems, there are several classes of users (customers) with different service requirements. There is a need to assess performance measures for each class of user individually. Results obtained in this paper can be used for such a purpose.

A quasi-birth-and-death (QBD) Markov process with a tree structure is a generalization of classical QBD processes (see Refs. [7,13,15,20]). It is a special random walk on a set with a tree structure (see Ref. [10] and references therein) and a special case of the Markov process of matrix *GI*/*M*/1 type with a tree structure introduced in Ref. [22]. Yeung and Sengupta [22] obtained a matrix product-form solution for the stationary distribution of such a Markov process, which is a generalization from the *GI*/*M*/1 paradigm to a tree-like structure. They applied their results to

* Tel.: +1 902 494 6141; fax: +1 902 420 7858; e-mail: qmhe@newton.ccs.tuns.ca

0377-2217/00/$ – see front matter
PII: S 0 3 7 7 - 2 2 1 7 (9 8) 0 0 3 9 6 - 8

find the stationary distribution of a multi-class LCFS-GPR queue. A QBD Markov process with a tree structure is also a special case of the Markov process of matrix *M*/*G*/1 type with a tree structure introduced in Takine et al. [19]. The first part of this paper defines a discrete time QBD Markov process with a tree structure. Some classical results of classical QBD Markov processes are generalized. It is worth mentioning that Yeung and Alfa [21] also study QBD Markov processes with a tree structure. While Ref. [21] focuses on QBD Markov processes with a tree structure, this paper mainly deals with the *MMAP*[*K*]/*PH*[*K*]/*N*/*LCFS* non-preemptive queueing system.

When the arrival processes are independent Poisson processes and priorities among different types of customers present, queueing systems with multiple types of customers have been studied extensively (see Refs. [16,17] and references therein). The study of queueing systems with more general arrival processes and multiple types of customers is limited. Some classical results can be found in Ref. [14]. Takine et al. [19] and Yeung and Sengupta [22] considered queueing systems with Markov modulated Poisson processes and superposition processes of Markov arrival processes. In Refs. [4,18], a queueing system with dependent arrival processes and a first-come-first-served (FCFS) service discipline is studied. Some results are obtained for the fundamental periods, queue length, and waiting times. He and Alfa [5] studied the a single server queueing system with a Markov arrival process with marked arrivals, PH-distribution service times, and a LCFS preemptive resume or repeat service discipline.

The second part of this paper studies a queueing system with multiple types of customers, PH-distribution service times, and no priority among customers. Customers of all types are served on an LCFS non-preemptive basis. This queueing model is close to the one studied in Ref. [5], except that customers are served on a non-preemptive basis. The main contribution of this paper is the formulation of the queueing system of interest as a QBD Markov process with a tree structure. By using the results obtained for the QBD Markov process with a tree structure, an algorithm is developed for computing the queue string distribution, which is useful in finding out how often a particular pattern of queue appears. With carefully chosen parameters of the arrival process and service times, it is possible to discuss the impact of the correlation between arrival processes on the queueing process. In addition, the impact of the burstiness and the pattern of the arrival and service processes on the queueing process is explored as well. Since the formulation and solutions of the queueing systems with multiple servers are similar to that of the single server case, this paper shall focus on the single server case which is a better choice to present the ideas and the methodology utilized in this paper.

To model an arrival process with multiple types of customers, the superposition process of several independent Poisson processes (Example 3.1) is usually used [16,17]. But such a superposition process is not suitable for bursty arrival processes or arrival processes with a special arrival pattern. The Markov arrival process with marked arrivals (*MMAP*[*K*]) is a useful tool to model these complicated arrival processes. For instance, Example 3.2 presents an arrival process with two types of customers such that every type 2 customer is followed by a type 1 customer. The use of *MMAP*[*K*], plus the algorithm developed in this paper, enables us to gain more insights into the queueing system of interest.

The rest of the paper is organized as follows. Section 2 defines a QBD Markov process with a tree structure and develops an algorithm for computing the stationary distribution of the Markov process. Section 3 introduces the *MMAP*[*K*]/*PH*[*K*]/1/*LCFS* non-preemptive queue. An algorithm for computing the stationary distribution of the queue string is developed. In Section 4, several special cases are investigated. Detailed results are obtained for those special cases. Section 5 presents several numerical examples. Various issues of interest are discussed in detail so as to gain insights into the queueing processes of the queueing systems. In Section 6, the *MMAP*[*K*]/*PH*[*K*]/*N*/*LCFS* non-preemptive queue is introduced and formulated into a QBD Markov process with a tree structure. Finally, in Section 7, some discussion of the results obtained in this paper, is given.

2. The QBD Markov process with a tree structure

A QBD Markov process can be defined on a discrete time space or a continuous time space. Since the analyses of the two cases are analogue, details are presented for the discrete case only. An advantage of using discrete time QBD Markov chains is that a clear probabilistic interpretation can be given to matrices $\{R(k), 1 \leqslant k \leqslant K\}$ and $\{G(k), 1 \leqslant k \leqslant K\}$ directly. Nonetheless, in Sections 3 and 4, the continuous time analogue is applied.

Consider a discrete time two-dimensional Markov chain $\{(X_n, I_n), n \geqslant 0\}$ in which the values of X_n are represented by the nodes of a K-ary tree, and I_n takes integer values between 1 and m. X_n is referred to as the node and I_n is referred to as the auxiliary variable of the Markov chain at time n. To give a full description of the transitions of the Markov chain, the K-ary tree is defined first.

The K-ary tree of interest is a tree for which each node has K children, except the soil node that is denoted as -1. The soil node -1 is connected to the root that is denoted as 0. Strings of integers between 1 and K are used to represent nodes of the tree. For example, the kth child of the root has a representation of k. The lth child of node k has a representation of kl (see Fig. 1 for an example with $K=2$). Node kl is a child of node k and k is the parent of node kl. Let $\aleph = \{J: J = k_1k_2 \ldots k_n, 1 \leqslant k_i \leqslant K, 1 \leqslant i \leqslant n, n \geqslant 0\} \cup \{0\}$. Any string $J \in \aleph$ is a node in the K-ary tree. The length of string J is defined as the number of integers in the string and is denoted by $|J|$. When $J=0$, $|J|=0$. The following two operations related to strings in $\aleph$ are used in this paper:

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

For example, as shown in Fig. 1, 21+2+2 = 2122 and 22−2 = 2.

The Markov chain (X_n, I_n) takes values in $\{\aleph \times \{1, 2, \ldots, m\}\} \cup \{\{-1\} \times \{1, 2, \ldots, m_{-1}\}\}$. To be called a (homogenous) QBD Markov chain with a tree structure, (X_n, I_n) transits at each step to either the current node itself, one of its children, or its parent node. All possible transitions and their corresponding probabilities are given as follows. When $(X_n, I_n) = (J, i)$, the one step transition probabilities (see Fig. 2) are given as:

1. $(X_{n+1}, I_{n+1}) = (J + k, i')$ with probability $a_{0,(i,i')}(k)$ when $J > -1$;
2. $(X_{n+1}, I_{n+1}) = (J, i')$ with probability $a_{1,(i,i')}(k_{|J|})$ when $J > 0$;
3. $(X_{n+1}, I_{n+1}) = (J - k_{|J|}, i')$ with probability $a_{2,(i,i')}(k_{|J|})$ when $J > 0$;
4. $(X_{n+1}, I_{n+1}) = (J, i')$ with probability $a_{1,(i,i')}$ when $J=0$;
5. $(X_{n+1}, I_{n+1}) = (-1, i')$ with probability $b_{2,(i,i')}$ when $J=0$;
6. $(X_{n+1}, I_{n+1}) = (-1, i')$ with probability $b_{1,(i,i')}$ when $J=-1$;
7. $(X_{n+1}, I_{n+1}) = (0, i')$ with probability $b_{0,(i,i')}$ when $J=-1$.

In matrix form, transitions between nodes are represented by matrix blocks: (1) $A_0(k)$ is an $m \times m$ matrix with elements $a_{0,(i,i')}(k)$; (2) $A_1(k)$ is an $m \times m$ matrix with elements $a_{1,(i,i')}(k)$; (3) $A_2(k)$ is an $m \times m$ matrix with elements $a_{2,(i,i')}(k)$; (4) A_1 is an $m \times m$ matrix with elements $a_{1,(i,i')}$; (5) B_2 is an $m \times m_{-1}$ matrix with elements $b_{2,(i,i')}$; (6) B_1 is an $m_{-1} \times m_{-1}$ matrix with elements $b_{1,(i,i')}$; (7) B_0 is an $m_{-1} \times m$ matrix with elements $b_{0,(i,i')}$.

Notice that for $A_0(k)$, k is the child the Markov chain transits to; for $A_1(k)$ and $A_2(k)$, k is the last

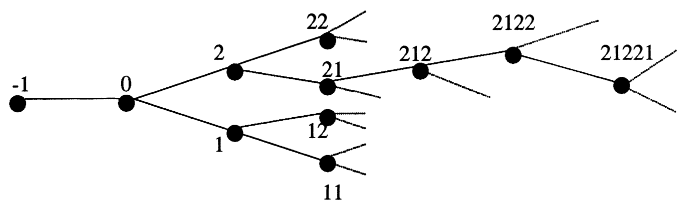

Fig. 1. A 2-ary tree with a soil node.

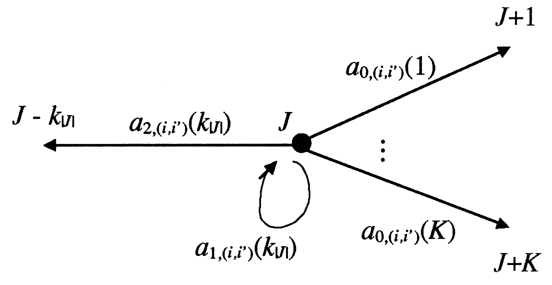

Fig. 2. Possible one step transitions when $J > -1$.

element of the string of the current node. According to the law of total probability, the matrix blocks satisfy the following equalities:

$$\left(\sum_{l=1}^{K} A_0(l) + A_1(k) + A_2(k)\right)\mathbf{e} = \mathbf{e}, \quad 1 \leqslant k \leqslant K,$$
$$\left(\sum_{l=1}^{K} A_0(l) + A_1\right)\mathbf{e} + B_2\mathbf{e} = \mathbf{e}; \quad B_0\mathbf{e} + B_1\mathbf{e} = \mathbf{e}, \tag{1}$$

where $\mathbf{e}$ is the vector with all components one.

Having defined the Markov chain of interest, the next step is to find the stationary distribution of the Markov chain. In order to do so, two sets of matrices, $\{R(k)\}$ and $\{G(k)\}$, are introduced (for more details, see Refs. [19,21,22]). A relationship between the two sets of matrices is shown in this section.

Matrices $\{R(k), 1 \leqslant k \leqslant K\}$: For $J \in \aleph, J > -1$, and $1 \leqslant k \leqslant K$, define the taboo probability ${}_JP^{(n)}_{(J,i)(J+k,i')}$ as the probability that the Markov chain (X_n, I_n) is in state $(J+k, i')$ after n transitions without visiting node J in between, given that the Markov chain started in (J, i). Because of the particular transition structure, ${}_JP^{(n)}_{(J,i)(J+k,i')}$ is independent of J. Let

$$r_{i,i'}(k) = \sum_{n=0}^{\infty} {}_JP^{(n)}_{(J,i)(J+k,i')}, \quad 1 \leqslant i, \ i' \leqslant m, \tag{2}$$

and $R(k)$ be an $m \times m$ matrix with elements $r_{i,i'}(k)$. It can be proved that $\{R(k), \ 1 \leqslant k \leqslant K\}$ are the minimal non-negative solutions to the equations

$$R(k) = A_0(k) + R(k)A_1(k) + \sum_{l=1}^{K} R(k)R(l)A_2(l),$$
$$1 \leqslant k \leqslant K. \tag{3}$$

When the Markov chain is irreducible and positive recurrent, the spectrum (the eigenvalue with the largest real part) of matrix $R = R(1) + \cdots + R(K)$ is less than one, i.e., $\mathrm{sp}(R) < 1$.

Matrices $\{G(k), \ 1 \leqslant k \leqslant K\}$: For $J \in \aleph$, $J > 0$, and $1 \leqslant k \leqslant K$, define the taboo probability $g_{i,i'}(k)$ as the probability that the Markov chain (X_n, I_n) reaches node J for the first time in state (J, i'), given that the Markov chain started in $(J + k, i)$. Let $G(k)$ be an $m \times m$ matrix with elements $g_{i,i'}(k)$. It can be proven that $\{G(k), \ 1 \leqslant k \leqslant K\}$ are the minimal non-negative solutions to the equations

$$G(k) = A_2(k) + A_1(k)G(k) + \sum_{l=1}^{K} A_0(l)G(l)G(k), \quad 1 \leqslant k \leqslant K. \tag{4}$$

When the Markov chain is irreducible and positive recurrent, matrix $G(k)$ is a stochastic matrix, i.e., $\mathrm{sp}(G(k)) = 1, \ 1 \leqslant k \leqslant K$.

When $K=1$, i.e., the classical QBD Markov chain case, a simple relationship between the matrices R and G was shown in Ref. [8]. A similar relationship holds when $K \geqslant 1$. Define $U_{i,i'}(k)$, for $1 \leqslant k \leqslant K$ and $1 \leqslant i, \ i' \leqslant m$, the probability that the Markov chain will eventually come back to node $J+k$ in state $(J+k, i')$, given that it started in $(J + k, i)$ and never visited its parent node J in between. Let $U(k)$ be an $m \times m$ matrix with elements $U_{i,i'}(k)$. When the Markov chain (X_n, I_n) is irreducible and positive recurrent, the following relationships hold for $R(k), G(k), U(k), \ 1 \leqslant k \leqslant K$:

$$R(k) = A_0(k)(\mathbf{I} - U(k))^{-1};$$
$$G(k) = (\mathbf{I} - U(k))^{-1}A_2(k);$$

$$R(k) = A_0(k)\left(\mathbf{I} - A_1(k) - \sum_{l=1}^{K} A_0(l)G(l)\right)^{-1};$$
$$G(k) = \left(\mathbf{I} - A_1(k) - \sum_{l=1}^{K} R(l)A_2(l)\right)^{-1} A_2(k), \tag{5}$$

where $\mathbf{I}$ is the identity matrix. In addition, $R(k)A_2(k) = A_0(k)G(k), \ 1 \leqslant k \leqslant K$.

The stationary distribution: Let

$$\pi(J, i) = \lim_{n\to\infty} \mathbf{P}\{(X_n, I_n) = (J, i)\},$$

$$\boldsymbol{\pi}(J) = (\pi(J,1),\ldots,\pi(J,m)),\ J > -1,$$

and

$$\boldsymbol{\pi}(-1) = (\pi(-1,1),\ldots,\pi(-1,m_{-1})).$$

The stationary distribution vectors $\{\boldsymbol{\pi}(J): J \in \aleph\}$ satisfy the following equation, for $J > 0$:

$$\boldsymbol{\pi}(J+k) = \boldsymbol{\pi}(J)A_0(k) + \boldsymbol{\pi}(J+k)A_1(k) + \sum_{l=1}^{K}\boldsymbol{\pi}(J+k+l)A_2(l), \tag{6}$$

which is useful in helping understand the following solution intuitively. According to Theorem 1 in Ref. [22], when the QBD Markov chain is irreducible and positive recurrent,

$$\boldsymbol{\pi}(J+k) = \boldsymbol{\pi}(J)R(k), \quad J \in \aleph,\ 1 \leqslant k \leqslant K;$$

$$(\boldsymbol{\pi}(-1),\boldsymbol{\pi}(0)) = (\boldsymbol{\pi}(-1),\boldsymbol{\pi}(0))\begin{pmatrix} B_1 & B_0 \\ B_2 & A_1 + \sum_{k=1}^{K} R(k)A_2(k) \end{pmatrix};$$

$$\boldsymbol{\pi}(-1)\mathbf{e} + \boldsymbol{\pi}(0)(\mathbf{I}-R)^{-1}\mathbf{e} = 1. \tag{7}$$

Notice that matrix R is the sum of $\{R(k),\ 1 \leqslant k \leqslant K\}$. Matrices $\{R(k),\ 1 \leqslant k \leqslant K\}$ can be calculated using the following simple algorithm. Let $R(k)[0] = 0,\ 1 \leqslant k \leqslant K$, and

$$R(k)[n+1] = A_0(k) + R(k)[n]A_1(k) + \sum_{l=1}^{K} R(k)[n]R(l)[n]A_2(l). \tag{8}$$

It can be proven that $\{R(k)[n],\ n \geqslant 0\}$ is a monotone sequence which converges to $R(k)$ from below, for $1 \leqslant k \leqslant K$. This algorithm is simple and easy to implement. A more complicated algorithm can be developed similar to that in Ref. [9] ([21]). Matrices $\{G(k),\ 1 \leqslant k \leqslant K\}$ can be computed in a similar way.

In summary, the stationary distribution of the Markov chain $(X_n,\ I_n)$ can be found using the following algorithm.

Algorithm I

Step 1: Data input: $m_{-1}, m, K,\ B_1, B_0, B_2, A_1$, and $\{A_0(k), A_1(k), A_2(k),\ 1 \leqslant k \leqslant K\}$.

Step 2: Computing matrices

$\{R(k), G(k),\ 1 \leqslant k \leqslant K\}$.

Step 3: Computing vectors

$\{\boldsymbol{\pi}(-1), \boldsymbol{\pi}(0)\}$.

Step 4: Computing string distribution

$\{\boldsymbol{\pi}(J), J \in \aleph\}$.

3. The *MMAP*[*K*]/*PH*[*K*]/1/LCFS queue

This section considers 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 distribution functions. All types of customers are served on an LCFS non-preemptive basis. To define the queueing systems of interest explicitly, the arrival process *MMAP*[*K*] is introduced first and then the service time distributions are specified.

The following definition of the *MMAP*[*K*] was given by Marcel Neuts (see Ref. [6]). The *MMAP*[*K*] was also introduced in Ref. [1]. A Markov arrival process with marked transitions is defined by a set of $m \times m$ matrices $\{D_k,\ 0 \leqslant k \leqslant K\}$, where m is a positive integer. The matrices $\{D_k,\ 1 \leqslant k \leqslant K\}$, are non-negative. The matrix D_0 has negative diagonal elements and non-negative off-diagonal elements. D_0 is assumed to be non-singular. Let

$$D = D_0 + \sum_{k=1}^{K} D_k. \tag{9}$$

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, $1 \leqslant I(t) \leqslant m$. 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 $\boldsymbol{\theta}$ be the stationary probability vector of the matrix D. The stationary arrival rate of type k arrivals is given by $\lambda_k = \boldsymbol{\theta} D_k \mathbf{e}$, $1 \leqslant k \leqslant K$. When $K=1$, the $MMAP[K]$ reduces to an MAP [11,12].

Example 3.1. The superposition process of K independent Poisson processes is an $MMAP[K]$. Suppose that the arrival rates of the K Poisson processes are $\{\lambda_1, \lambda_2, \ldots, \lambda_K\}$. Then the matrix representation of their superposition process is $D_0 = -(\lambda_1 + \cdots + \lambda_K), D_1 = \lambda_1, \ldots, D_K = \lambda_K$.

Example 3.2. Consider an $MMAP[K]$ with $K=2$, $m=2$, and

$$D_0 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0.5 & 0.5 \\ 0 & 0 \end{pmatrix},$$
$$D_2 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}. \tag{10}$$

In this point process, every type 2 customer is followed by a type 1 customer. Similar to this example, $MMAP[K]$s can be used to model a variety of arrival processes with special arrival patterns such as cyclic arrivals.

The service times of type k customers have a common phase-type distribution (PH-distribution) function with a matrix representation $(\boldsymbol{\alpha}_k, T_k)$, where $\boldsymbol{\alpha}_k$ is an m_k-dimension vector 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 more details about PH-distribution, see Ch. 2 in Ref. [13].

The traffic intensity of the system is defined as $\rho = \lambda_1/\mu_1 + \cdots + \lambda_K/\mu_K$. Throughout this paper, it is assumed that $\rho < 1$. Since it has been proved that the MAPs can be used to approximate any point process and PH-distributions can be used to approximate any non-negative distribution, the queueing system under consideration is a rather general queueing model.

The queueing system is represented by the following four-dimensional stochastic process:

$q(t)$: the string of customers in queue (exclude the one in server, if any), $q(t) \in \aleph \cup \{-1\}$;
$I(t)$: the state of the underlying Markov process $D, 1 \leqslant I(t) \leqslant m$;
$I_{1,1}(t)$: the type of the customer in service (if any) $1 \leqslant I_{1,1}(t) \leqslant m$;
$I_{1,2}(t)$: the phase of the PH-distribution of the current service (if any), $1 \leqslant I_{1,2}(t) \leqslant m_{I_{1,1}(t)}$.

When there is no customer in the system at time t, denote by $q(t)=-1$. 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 $\aleph$. For example (for $K=2$), $q(t)=122$ implies that there are 3 customers waiting in the system at time t: the customer who arrived first is of type 1; the customer who arrived second is of type 2; and the customer who arrived last is of type 2. When a new customer of type k arrives, $q(t)$ becomes $122k$. When the current service is completed, $q(t)$ becomes 122 since the customer (of type k) who arrived last enters service first.

It is easy to see that $(q(t), I(t), I_{1,1}(t), I_{1,2}(t))$ is a Markov process with a state space: $\aleph \times \{1, 2, \ldots, m\} \times \bigcup_{k=1}^{K} \{1, 2, \ldots, m_k\}$. This is a QBD Markov process with a tree structure when $(I(t), I_{1,1}(t), I_{1,2}(t))$ is defined as the auxiliary random variable with $m\bar{m}$ states (where $\bar{m} = m_1 + \cdots + m_K$), except that when $q(t)=-1$, the auxiliary variable takes values $\{1, 2, \ldots, m\}$. Furthermore, the infinitesimal generator of the QBD Markov process is defined by the following transition blocks.

For $J = k_1, \ldots, k_{n-1}k \in \aleph$ and $1 \leqslant k \leqslant K$,

$$A_0(k) = D_k \otimes \mathbf{I}_{\bar{m}\times\bar{m}}$$

(a type k customer arrives),

$$A_1(k) = D_0 \otimes \mathbf{I}_{\bar{m}\times\bar{m}} + \mathbf{I}_{m\times m} \otimes \begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_K \end{pmatrix}$$

(no service completed and no arrival),

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

(a service is completed and the next (last in queue) is of type k), (11) where $\otimes$ represents the Kronecker product of matrices (see Ref. [2]).

For $J=0$,

$$A_1 = A_1(1) \quad \text{(no service completed and no arrival)},$$

$$B_2 = \mathbf{I}_{m\times m} \otimes \begin{pmatrix} \mathbf{T}_1^0 \\ \vdots \\ \mathbf{T}_K^0 \end{pmatrix} \quad \text{(a service is completed)}. \tag{12}$$

For $J=-1$,

$$B_0 = (D_1 \otimes \boldsymbol{\alpha}_1 \cdots D_K \otimes \boldsymbol{\alpha}_K)$$
$$\text{(a customer arrives)},$$
$$B_1 = D_0 \quad \text{(no arrival and no service)}. \tag{13}$$

The QBD Markov process which describes the queueing system of interest is defined explicitly. The stationary distribution of this QBD Markov process is presented next. Let

$$\pi(J,i,k,j) = \lim_{t\to\infty} \mathbf{P}\{q(t), I(t), I_{1,1}(t), I_{1,2}(t)) = (J,i,k,j)\}; \tag{14}$$
$$\pi(-1,i) = \lim_{t\to\infty} \mathbf{P}\{(q(t), I(t)) = (-1,i)\},$$

and

$$\boldsymbol{\pi}(J,i,k) = (\pi(J,i,k,1),\ldots,\pi(J,i,k,m_k));$$
$$\boldsymbol{\pi}(J,i) = (\boldsymbol{\pi}(J,i,1),\ldots,\boldsymbol{\pi}(J,i,K));$$
$$\boldsymbol{\pi}(J) = (\boldsymbol{\pi}(J,1),\ldots,\boldsymbol{\pi}(J,m)),$$
$$J=-1; \qquad \boldsymbol{\pi}(-1) = (\pi(-1,1),,\ldots,\pi(-1,m)). \tag{15}$$

When the underlying Markov chain D of the arrival process and all the PH-distributions are irreducible, ($q(t)$, $I(t)$, $I_{1,1}(t)$, $I_{1,2}(t)$) is irreducible. Furthermore, a condition for positive recurrence of the Markov process (or the queueing system) and some elementary results about the stationary distribution are given in the following theorem.

Theorem 3.1. *For the queueing system of interest, when it is in steady state, for* $1 \leqslant k \leqslant K$,

(a) *the rate of starting to serve a type k customer is given by*

$$\sum_{t=1}^{m}\sum_{i=1}^{m}\pi(-1,t)(D_k)_{t,i} + \sum_{J\geqslant 0}\sum_{l=1}^{K}\sum_{i=1}^{m}\boldsymbol{\pi}(J+k,i,l)\mathbf{T}_l^0 = \lambda_k;$$

(b) *the probability that a type k customer is in service is*

$$\sum_{J\geqslant 0}\sum_{i=1}^{m}\boldsymbol{\pi}(J,i,k)\mathbf{e} = \lambda_k/\mu_k;$$

(c) *the probability that the queueing system is busy is*

$$\rho = \sum_{J\geqslant 0}\boldsymbol{\pi}(J)\mathbf{e} = \sum_{k=1}^{K}\lambda_k/\mu_k;$$

(d) *the probability that the queueing system is empty is* $\boldsymbol{\pi}(-1)\mathbf{e} = 1-\rho$.

Furthermore, the queueing system of interest is positive recurrent if and only if $\rho < 1$.

Proof. In the queueing system of interest with LCFS, the server starts to serve a type k customer when it completes a service and the customer who arrived last is of type k, or at the beginning of a busy period, the customer who just arrived is of type k. Then Part (a) of Theorem 3.1 is obtained since the arrival rate of type k customer equals the start-to-service rate of type k customer at an arbitrary time. Part (a) can be rewritten into

$$\boldsymbol{\pi}(-1)D_k\mathbf{e} + \sum_{J\geqslant 0}\sum_{l=1}^{K}\sum_{i=1}^{m}\boldsymbol{\pi}(J+k,i,l)\mathbf{T}_l^0 = \lambda_k.$$

To prove part (b), first notice that the stationary distribution of the Markov process ($q(t)$, $I(t)$, $I_{1,1}(t)$, $I_{1,2}(t)$) satisfies the following equation:

for $J > 0$,

$$0 = \boldsymbol{\pi}(J) D_k \otimes \mathbf{I} + \boldsymbol{\pi}(J+k)\left[D_0 \otimes \mathbf{I} + \mathbf{I} \otimes \begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_K \end{pmatrix}\right] + \sum_{l=1}^{K} \boldsymbol{\pi}(J+k+l) \mathbf{I} \otimes \left[\begin{pmatrix} \mathbf{T}_1^0 \boldsymbol{\alpha}_k \\ \vdots \\ \mathbf{T}_K^0 \boldsymbol{\alpha}_k \end{pmatrix} (0,\ldots,0,\mathbf{I},0,\ldots,0)\right]; \qquad (16)$$

for $J = 0$,

$$0 = \boldsymbol{\pi}(-1)(D_k \otimes \boldsymbol{\alpha}_1 \cdots D_k \otimes \boldsymbol{\alpha}_K) + \boldsymbol{\pi}(0)\left[D_0 \otimes \mathbf{I} + \mathbf{I} \otimes \begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_K \end{pmatrix}\right] + \sum_{l=1}^{K} \boldsymbol{\pi}(l) \mathbf{I} \otimes \left[\begin{pmatrix} \mathbf{T}_1^0 \boldsymbol{\alpha}_k \\ \vdots \\ \mathbf{T}_K^0 \boldsymbol{\alpha}_k \end{pmatrix} (0,\ldots,0,\mathbf{I},0,\ldots,0)\right]. \qquad (17)$$

Expanding Eq. (16) in terms of vectors $\{\boldsymbol{\pi}(J,w,k)\}$ for fixed k, l, i $(1 \leqslant k, l \leqslant K,\ 1 \leqslant i \leqslant m)$ yields

$$0 = \sum_{w=1}^{m} \boldsymbol{\pi}(J,w,k)(D_l)_{w,i} + \sum_{w=1}^{m} \boldsymbol{\pi}(J+l,w,k)(D_0)_{w,i} + \boldsymbol{\pi}(J+l,i,k)T_k + \sum_{t=1}^{K} \boldsymbol{\pi}(J+k+l,i,t)\mathbf{T}_t^0 \boldsymbol{\alpha}_k. \qquad (18)$$

Expanding Eq. (17) for fixed k and i $(1 \leqslant k \leqslant K,\ 1 \leqslant i \leqslant m)$ yields

$$0 = \sum_{w=1}^{m} \boldsymbol{\pi}(-1,w)(D_k)_{w,i}\boldsymbol{\alpha}_k + \sum_{w=1}^{m} \boldsymbol{\pi}(0,w,k)(D_0)_{w,i} + \boldsymbol{\pi}(0,i,k)T_k + \sum_{t=1}^{K} \boldsymbol{\pi}(k,i,t)\mathbf{T}_t^0 \boldsymbol{\alpha}_k. \qquad (19)$$

For fixed k $(1 \leqslant k \leqslant K)$, adding Eq. (18) together for all $J \geqslant 0,\ 1 \leqslant l \leqslant K,\ 1 \leqslant i \leqslant m$, yields

$$0 = \sum_{w=1}^{m} \sum_{J \geqslant 0} \sum_{i=}^{m} \boldsymbol{\pi}(J,w,k) \sum_{l=1}^{K} (D_1)_{w,i} + \sum_{w=1}^{m} \sum_{J \geqslant 0} \sum_{i=1}^{m} \sum_{l=1}^{K} \boldsymbol{\pi}(J+l,w,k)(D_0)_{w,i} + \sum_{J \geqslant 0} \sum_{i=1}^{m} \sum_{l=1}^{K} \boldsymbol{\pi}(J+l,i,k)T_k + \sum_{t=1}^{K} \sum_{J \geqslant 0} \sum_{i=1}^{m} \sum_{l=1}^{K} \boldsymbol{\pi}(J+k+l,i,t)\mathbf{T}_l^0 \boldsymbol{\alpha}_k. \qquad (20)$$

This leads to

$$0 = \sum_{w=1}^{m} \sum_{J \geqslant 0} \sum_{i=}^{m} \boldsymbol{\pi}(J,w,k) \sum_{l=0}^{K} (D_\ell)_{w,i} - \sum_{w=1}^{m} \sum_{i=1}^{m} \boldsymbol{\pi}(0,w,k)(D_0)_{w,i} + \sum_{J \geqslant 0} \sum_{i=1}^{m} \boldsymbol{\pi}(J,i,k)T_k - \sum_{i=1}^{m} \boldsymbol{\pi}(0,i,k)T_k + \sum_{t=1}^{K} \sum_{J \geqslant 0} \sum_{i=1}^{m} \boldsymbol{\pi}(J+k,i,t)\mathbf{T}_t^0 \boldsymbol{\alpha}_k - \sum_{t=1}^{K} \sum_{i=1}^{m} \boldsymbol{\pi}(k,i,t)\mathbf{T}_t^0 \boldsymbol{\alpha}_k. \qquad (21)$$

For fixed k $(1 \leqslant k \leqslant K)$, adding Eq. (19) together for $1 \leqslant i \leqslant m$, yields

$$0 = \sum_{w=1}^{m}\sum_{i=1}^{m}\boldsymbol{\pi}(-1,w)(D_k)_{w,i}\boldsymbol{\alpha}_k + \sum_{w=1}^{m}\sum_{i=1}^{m}\boldsymbol{\pi}(0,w,k)(D_0)_{w,i} + \sum_{i=1}^{m}\boldsymbol{\pi}(0,i,k)T_k + \sum_{t=1}^{K}\sum_{i=1}^{m}\boldsymbol{\pi}(k,i,t)\mathbf{T}_t^0\boldsymbol{\alpha}_k. \tag{22}$$

Eqs. (21) and (22), part (a), and $(D_0 + D_1 + \cdots + D_K)\mathbf{e} = 0$ lead to, for fixed k ($1 \leqslant k \leqslant K$):

$$\begin{aligned} 0 &= \left(\sum_{J\geqslant 0}\sum_{i=1}^{m}\boldsymbol{\pi}(J,i,k)\right)T_k + \left(\sum_{t=1}^{K}\sum_{J\geqslant 0}\sum_{i=1}^{m}\boldsymbol{\pi}(J+k,i,t)\mathbf{T}_t^0 + \sum_{w=1}^{m}\sum_{i=1}^{m}\boldsymbol{\pi}(-1,w)(D_k)_{w,i}\right)\boldsymbol{\alpha}_k \\ &= \left(\sum_{J\geqslant 0}\sum_{i=1}^{m}\boldsymbol{\pi}(J,i,k)\right)T_k + \lambda_k\boldsymbol{\alpha}_k. \end{aligned} \tag{23}$$

Part (b) is obtained since $1/\mu_k = -\boldsymbol{\alpha}_k T_k^{-1}\mathbf{e}$.

Part (c) is obtained by taking summation of the results obtained in part (b) with respect to k. Part (d) is obtained from part (c).

Clearly, $\boldsymbol{\pi}(-1)\mathbf{e} = 1 - \rho > 0$ when the queueing system is positive recurrent, i.e., $\rho < 1$. The sufficiency of $\rho < 1$ for positive recurrence is proved in Ref. [3]. This completes the proof.

Note. Intuitively, $\rho < 1$ implies that the system has enough capacity to serve all customers. Consider the interval (0, t). On an average, $\lambda_1 t$ type 1 customers, …, and $\lambda_K t$ type K customers arrive in (0, t). On an average, $\mu_1 t_1$ type 1 customers, …, and $\mu_K t_K$ type K customers are served in (0, t) if t_1 units of time, …, and t_K units of time are used to serve type 1 customers, …, and type K customers in (0, t), respectively. It has $t_1 + \cdots + t_K \leqslant t$. If the system is positive recurrent, there must be a set $\{t, t_1, \ldots, t_K\}$ such that $\lambda_1 t \leqslant \mu_1 t_1, \ldots,$ and $\lambda_K t \leqslant \mu_K t_K$. This leads to $\rho \leqslant (t_1 + t_2 + \cdots + t_K)/t < 1$.

When the queueing system (or the Markov process) is positive recurrent, using formulas presented in Section 2, the following theorem is obtained.

Theorem 3.2. *When the queueing system of interest is positive recurrent, the stationary distribution of* $(q(t), I(t), I_{1,1}(t), I_{1,2}(t))$ *is given by*

$$\begin{aligned} &\boldsymbol{\pi}(J+k) = \boldsymbol{\pi}(J)R(k), \quad J \in \aleph,\ 1 \leqslant k \leqslant K; \\ &(\boldsymbol{\pi}(-1),\boldsymbol{\pi}(0))\begin{pmatrix} D_0 & (D_1 \otimes \boldsymbol{\alpha}, \ldots, D_K \otimes \boldsymbol{\alpha}_K) \\ \mathbf{I} \otimes \begin{pmatrix} \mathbf{T}_1^0 \\ \vdots \\ \mathbf{T}_K^0 \end{pmatrix} & A_1 + \sum_{k=1}^{K} R(k)A_2(k) \end{pmatrix} = 0; \\ &\boldsymbol{\pi}(-1)\mathbf{e} + \boldsymbol{\pi}(0)(\mathbf{I} - R)^{-1}\mathbf{e} = 1, \end{aligned} \tag{24}$$

where

$$A_1 + \sum_{k=1}^{K} R(k)A_2(k) = D_0 \otimes \mathbf{I} + \mathbf{I} \otimes \begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_K \end{pmatrix} + \sum_{k=1}^{K} R(k)\mathbf{I} \otimes \left[\begin{pmatrix} \mathbf{T}_1^0\boldsymbol{\alpha}_k \\ \vdots \\ \mathbf{T}_K^0\boldsymbol{\alpha}_k \end{pmatrix}(0,\ldots,0,\mathbf{I},0,\ldots,0)\right],$$

$R = R(1) + \cdots + R(K)$, *and* $R(k)$, $1 \leqslant k \leqslant K$ *are the minimal non-negative solutions to*

$$0 = D_k \otimes \mathbf{I} + R(k)\left[D_0 \otimes \mathbf{I} + \mathbf{I} \otimes \begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_K \end{pmatrix}\right] + R(k)\sum_{l=1}^{K} R(l)\, \mathbf{I} \otimes \left[\begin{pmatrix} \mathbf{T}_1^0\boldsymbol{\alpha}_k \\ \vdots \\ \mathbf{T}_K^0\boldsymbol{\alpha}_k \end{pmatrix}(0,\ldots,0,\mathbf{I},0,\ldots,0)\right]. \tag{25}$$

The computation of $\{R(k),\ 1 \leqslant k \leqslant K\}$ can be carried out using the algorithm given in Section 2.

Later in Section 4, for some special cases, the computation of $\{R(k),\ 1 \leqslant k \leqslant K\}$ shall be simplified.

Proof. It is obvious from Eqs. (3) and (7). This completes the proof.

Let $L = |q(t)|$, i.e., the length of the queue string at an arbitrary time t, then

$$\mathbf{P}\{L = n\} = \sum_{J:|J|=n} \boldsymbol{\pi}(J)\mathbf{e} = \boldsymbol{\pi}(0)R^n\mathbf{e}, \quad n \geqslant 0. \quad (26)$$

Clearly, Eq. (26) shows the exponential decay of the queue length in the queueing system of interest. Then $\mathbf{E}L = \boldsymbol{\pi}(0)R(\mathbf{I} - R)^{-2}\mathbf{e}$. The mean number of customers in the system (mean queue length) can be computed by using the formula

$$\begin{aligned}\overline{L} &= \boldsymbol{\pi}(0)R(\mathbf{I} - R)^{-2}\mathbf{e} + \boldsymbol{\pi}(0)(\mathbf{I} - R)^{-1}\mathbf{e} \\ &= \boldsymbol{\pi}(0)(\mathbf{I} - R)^{-2}\mathbf{e}. \quad (27)\end{aligned}$$

Algorithm II

Step 1: Input data: $m, K, (D_0, D_1 \ldots, D_K)$, $(m_k, \boldsymbol{\alpha}_k, T_k),\ 1 \leqslant k \leqslant K$;

Step 2: Construct the transition blocks of the corresponding QBD;

Steps 3, 4, and 5 are the same as Steps 2, 3, and 4 of Algorithm I.

4. Some special cases of the *MMAP*[*K*]/*PH*[*K*]/1 queue

In this section, two groups of special *MMAP*[*K*]/*PH*[*K*]/1/*LCFS* non-preemptive queues are discussed. Queueing systems with a common service time distribution for all types of customers are discussed first. Queueing systems with a Markov arrival process marked by a multiple Bernoulli distribution are then studied. More explicit solutions are obtained for these special cases. Special computation approaches are developed for computing matrices $\{R(k),\ 1 \leqslant k \leqslant K\}$.

Example 4.1. The simplest special case of the queueing system of interest is the *M*[*K*]/*M*/1 case, where the arrival process consists of K independent Poisson processes with parameter $\{\lambda_1, \ldots, \lambda_K\}$ and the service times have the same exponential distribution with a parameter μ for all types of customers. The arrival process is equivalent to marking customers of a Poisson process with parameter $\lambda = \lambda_1 + \cdots + \lambda_K$ with probabilities $\{p_1, \cdots, p_K\}$, where $p_k = \lambda_k/(\lambda_1 + \cdots + \lambda_K)$, $1 \leqslant k \leqslant K$. Let $\rho_k = \lambda_k/\mu,\ 1 \leqslant k \leqslant K$. It is assumed that $\rho = \rho_1 + \cdots + \rho_K < 1$ to ensure a positive recurrent queueing system. The queueing system is represented by $q(t)$ (see Section 3 for definition), while $I(t)$, $I_{1,1}(t)$, and $I_{1,2}(t)$ are unnecessary in this case. However, when it is necessary to trace the type of the customer in service, $(q(t), I_{1,1}(t))$ should be considered.

When the service discipline is LCFS, using Eq. (25), it can be verified that $\{R(k) = \rho_k,\ 1 \leqslant k \leqslant K\}$, $\pi(-1) = 1 - \rho$, $\pi(0) = (1 - \rho)\rho$, and $\pi(J) = (1 - \rho)\rho\rho_{k_1} \ldots \rho_{k_n}$, where $J = k_1 \ldots k_n \in \aleph$ represents the types of waiting customers. (Notice that k_i represents k_i.) Thus, there is a product form solution for the stationary distribution in this case. Let x_n be the probability that there are n customers in the system. Then it is easy to see that $x_n = (1 - \rho)\rho^n,\ n \geqslant 0$. Since

$$\rho_k = \rho p_k = \rho\lambda_k/(\lambda_1 + \cdots + \lambda_K),\ 1 \leqslant k \leqslant K,$$

it is intuitive that the product solution can be obtained by marking customers in queue by $p_k,\ 1 \leqslant k \leqslant K$ in an *M*/*M*/1 queue.

Example 4.2. The *MMAP*[*K*]/*PH*/1 case. For this case, since the service time distributions are the same for all types of customers, $I_{1,1}(t)$ can be removed. Thus, $(q(t), I(t), I_{1,2}(t))$ will be used to represent the queueing system.

For the LCFS non-preemptive case, the transition blocks are given by

$$\begin{aligned}&B_0 = (D_1 + \cdots + D_K) \otimes \boldsymbol{\alpha}, \quad B_1 = D_0, \\ &B_2 = \mathbf{I} \otimes \mathbf{T}^0; \\ &A_0(k) = D_k \otimes \mathbf{I}, \\ &A_1 = A_1(k) = D_0 \otimes \mathbf{I} + \mathbf{I} \otimes T, \\ &A_2(k) = \mathbf{I} \otimes (\mathbf{T}^0\boldsymbol{\alpha}). \quad (28)\end{aligned}$$

Matrices $\{R(k),\ 1 \leqslant k \leqslant K\}$ satisfies

$$0 = D_k \otimes \mathbf{I} + R(k)(D_0 \otimes \mathbf{I} + \mathbf{I} \otimes T) + R(k)\sum_{l=1}^{K} R(l)\mathbf{I} \otimes (\mathbf{T}^0\boldsymbol{\alpha}), \quad 1 \leqslant k \leqslant K. \tag{29}$$

Summing up all K equalities in Eq. (29) yields the equation

$$0 = \left(\sum_{k=1}^{K} D_k\right) \otimes \mathbf{I} + R(D_0 \otimes \mathbf{I} + \mathbf{I} \otimes T) + R^2\mathbf{I} \otimes (\mathbf{T}^0\boldsymbol{\alpha}), \tag{30}$$

for matrix R with $[D_0 \otimes \mathbf{I} + R \otimes (\mathbf{T}^0\boldsymbol{\alpha})]\mathbf{e} = 0$. Eq. (29) leads to

$$R(k) = -(D_k \otimes \mathbf{I})[D_0 \otimes \mathbf{I} + \mathbf{I} \otimes T + R(\mathbf{I} \otimes (\mathbf{T}^0\boldsymbol{\alpha}))]^{-1}, \quad 1 \leqslant k \leqslant K. \tag{31}$$

Thus, the computation of matrices $\{R(k),\ 1 \leqslant k \leqslant K\}$ is significantly simplified. It can be proven that $\mathrm{sp}(R) < 1$ when $\rho < 1$.

Example 4.3. The $M[K]/M[K]/1$ case. In this case, assume that the service times of type k customers have a common exponential distribution with parameter $\mu_k,\ 1 \leqslant k \leqslant K$. The system is represented by $(q(t),\ I_{1,1}(t))$ which has an infinitesimal generator Q with transition blocks:

$$B_0 = -(\lambda_1 + \cdots + \lambda_K), \quad B_1 = (\lambda_1, \ldots, \lambda_k),$$
$$A_0(k) = \lambda_k\mathbf{I},$$
$$A_1 = A_1(k) = \left(\sum_{l=0}^{K} \lambda_l\right)\mathbf{I} - \begin{pmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_K \end{pmatrix},$$
$$B_2 = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix},$$
$$A_2(k) = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix}(0, \ldots, 0, 1, 0, \ldots, 0), \quad 1 \leqslant k \leqslant K. \tag{32}$$

From the equation

$$0 = \lambda_k\mathbf{I} + R(k)\left[-\left(\sum_{l=1}^{K}\lambda_l\right)\mathbf{I} - \begin{pmatrix} \mu_1 & & \\ & \vdots & \\ & & \mu_K \end{pmatrix}\right] + R(k)\sum_{l=1}^{K} R(l)\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix}(0, \ldots, 0, 1, 0, \ldots, 0), \tag{33}$$

it is easy to see that $R(k),\ 1 \leqslant k \leqslant K$ are invertible, $\lambda_l R(k) = \lambda_k R(l)$, and $R(k) = \lambda_k R/(\lambda_1 + \cdots + \lambda_K)$, $1 \leqslant k,\ l \leqslant K$. This leads to the following equation for matrix R:

$$0 = \lambda\mathbf{I} + R\left[-\left(\sum_{l=1}^{K}\lambda_l\right)\mathbf{I} - \begin{pmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_K \end{pmatrix}\right] + R^2\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix}\frac{(\lambda_1 \cdots \lambda_K)}{\lambda_1 + \cdots + \lambda_K} \equiv \lambda\mathbf{I} + RA_1 + R^2\Delta. \tag{34}$$

This provides a simpler algorithm for computing $\{R(k), 1 \leqslant k \leqslant K\}$ in this case. Let $A = \lambda\mathbf{I} + A_1 + \Delta$. Matrix A is an infinitesimal generator with a stationary distribution $\boldsymbol{\xi}$ with

$$\xi_k = \frac{\lambda_k/\mu_k}{\sum_{l=1}^{K}\lambda_l/\mu_l}, \quad 1 \leqslant k \leqslant K. \tag{35}$$

According to Neuts [13], the matrix R has a spectrum less than one if and only if $\boldsymbol{\xi}\lambda\mathbf{I}\mathbf{e} < \boldsymbol{\xi}\Delta\mathbf{e}$, which is equivalent to $\rho = \lambda_1/\mu_1 + \cdots + \lambda_K/\mu_K < 1$.

Example 4.4. The $MAP(p_1, \ldots, p_K)/PH[K]/1$ case. In this case, the arrival process is a Markov arrival process with matrix representation

$$\left(D_0, p_1\sum_{k=1}^{K} D_k, \ldots, p_K\sum_{k=1}^{K} D_k\right),$$

where $p_1 + \cdots + p_K = 1$, and the service times of type k customers have a common PH-distribution with parameter $(T_k, \boldsymbol{\alpha}_k)$, and the mean service time

$-\boldsymbol{\alpha}_k T_k^{-1}\mathbf{e} = 1/\mu_k$, $1 \leqslant k \leqslant K$. The system is represented by $(q(t), I(t), I_{1,1}(t), I_{1,2}(t))$.

It can be proved that $p_l R(k) = p_k R(l)$, and $R(k) = p_k R$, $1 \leqslant k$, $l \leqslant K$. This leads to the following equation for matrix R:

$$0 = \sum_{k=1}^{K} D_k \otimes \mathbf{I} + R\left[D_0 \otimes \mathbf{I} + \mathbf{I} \otimes \begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_K \end{pmatrix}\right] + R^2 \sum_{k=1}^{K} p_k \mathbf{I} \otimes \begin{pmatrix} \mathbf{T}_1^0 \boldsymbol{\alpha}_k \\ \vdots \\ \mathbf{T}_K^0 \boldsymbol{\alpha}_k \end{pmatrix} (0, \ldots, 0, \mathbf{I}, 0, \ldots, 0). \tag{36}$$

This provides a simpler algorithm for computing $\{R(k),\ 1 \leqslant k \leqslant K\}$ in this case.

5. Numerical results

This section presents a few numerical examples so as to gain insights into queueing systems with multiple types of customers. The issues of interest are, not limited to, how the dependence between arrival processes of different types of customers influences the queueing processes, how the arrival pattern of different types of customers (such as cyclic and one type after the other) influences the queueing process.

Example 5.1. Consider a queueing system with arrival process $K = 2$, $m = 1$, $D_0 = -1.2$, $D_1 = 0.8$, $D_2 = 0.4$, and service times: $m_1 = m_2 = 1$, $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = 1$, $T_1 = -2$, $T_2 = -1$. For this queueing system, $\lambda_1 = 0.8$, $\lambda_2 = 0.4$, $\mu_1 = 2$, $\mu_2 = 1$, and $\rho = 0.8$. The string probabilities are given in Fig. 3.

First, the probability distribution decreases exponentially with the queue length since it is essentially an $M/M/1$ queueing system. Since the arrival rate of type 1 customers is twice as much of type 1 customers, the probabilities of strings $J = 1$, $J = 121$, $J = 12\,121$, are twice of that of $J = 2$, $J = 212$, $J = 21\,212$. This shows that the product form solution obtained in Example 5.1 for common service time distribution cases does not exist when the service time distributions of different types of customers are different. Further, because of the difference in service times, the queue becomes longer when compared to an $M/M/1$ queue with $\lambda = 1.2$, $\mu = 1.5$, and $\rho = 0.8$.

Example 5.2. Consider a queueing system with an arrival process $K = 2$, $m = 2$,

$$D_0 = \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0.5 & 0.5 \\ 0 & 0 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}$$

and service times: $m_1 = m_2 = 1$, $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = 1$, $T_1 = -2$, $T_2 = -1$.

For this queueing system, $\lambda_1 = 0.8$, $\lambda_2 = 0.4$, $\mu_1 = 2$, $\mu_2 = 1$, and $\rho = 0.8$. Thus, this queueing system has the same arrival rates, service rates, and traffic intensity as that of Example 5.1. However, the arrivals of the two types of customers show a special pattern. That is, any type 2 customer is followed by a type 1 customer. It is interesting to see how the special arrival pattern influences the queueing process.

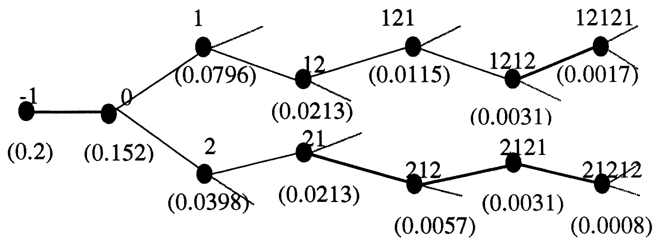

Fig. 3. The queue string distribution.

The string probabilities are given in Fig. 4. It is shown that the probabilities of strings with substring 21 are still small, even when a type 2 arrival is followed by a type 1 arrival. It is also shown that the probabilities of strings "$2\cdots2$" are positive but small. The reason is that the type 1 customer following a type 2 customer may be in service when the next type 2 customer arrives.

Example 5.3. Consider a queueing system with an arrival process $K=2$, $m=2$,

$$D_0=\begin{pmatrix}-0.8 & 0\\ 0 & -1.6\end{pmatrix},\quad D_1=\begin{pmatrix}0 & 0.8\\ 0 & 0\end{pmatrix},$$

$$D_2=\begin{pmatrix}0 & 0\\ 1.6 & 0\end{pmatrix},$$

and service times: $m_1=m_2=1$, $\boldsymbol{\alpha}_1=\boldsymbol{\alpha}_2=1$, $T_1=-2$, $T_2=-1$.

For this queueing system, $\lambda_1=0.533$, $\lambda_2=0.533$, $\mu_1=2$, $\mu_2=1$, and $\rho=0.8$. The string probabilities are given in Fig. 5.

The two types of customers arrive cyclically. It is expected that strings with substring 21 or 12 have larger probabilities. Thus, this example does show that the arrival pattern influences the queue in a particular way. Therefore, when analyzing queueing systems with multiple types of customers, one should take into consideration the arrival pattern of customers. Notice that the probabilities of strings such as 11 and 22 are positive. This is different from the cases where all customers are served on a first-come-first-served (FCFS) basis. For the FCFS case, the probability that the queue string is 11, 111, 22, 222, or 2222 is zero since the two types of customers arrive cyclically.

6. Modelling of the *MMAP*[*K*]/*PH*[*K*]/*N*/*LCFS* non-preemptive queue

This section formulates the *MMAP*[*K*]/*PH*[*K*]/*N* queue with an LCFS non-preemptive service discipline into a QBD Markov process with a tree structure. In this queueing system, there are N identical servers. Service times of each type of customer at all N servers have the same PH-distributions defined in Section 3.

The queueing system of interest can be represented by the following stochastic process $\{q(t),\ I(t),\ (I_{n,1}(t),\ I_{n,2}(t)),\ 1\leqslant n\leqslant N\}$, where

$q(t)$: the string of waiting customers in the queueing system, $q(t)\in \aleph\cup\{-1,\ldots,-N\}$;

$I(t)$: the state of the underlying Markov process D, $1\leqslant I(t)\leqslant m$;

$I_{n,1}(t)$: the type of the customer in server n (if any) $1\leqslant I_{n,1}(t)\leqslant K$ and $1\leqslant n\leqslant N$;

$I_{n,2}(t)$: the phase of the PH-distribution of server n (if working), $1\leqslant I_{n,2}(t)\leqslant m_{I_{n,1}(t)}$ and $1\leqslant n\leqslant N$.

When there is no customer in the system at time t, $q(t)=-N$; when there is one customer in the system at time t, $q(t)=-N+1$; …; when there are N customers in the system, $q(t)=0$. When

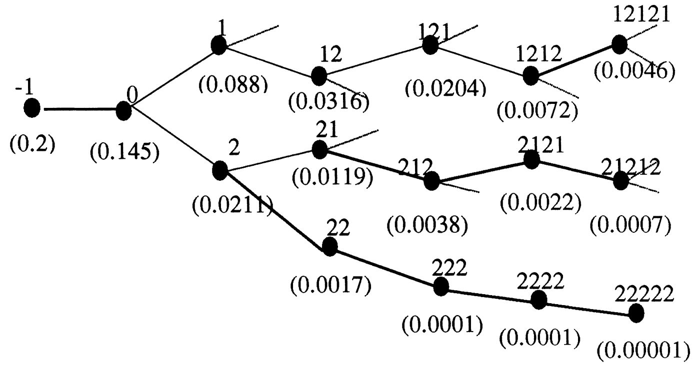

Fig. 4. The queue string distribution.

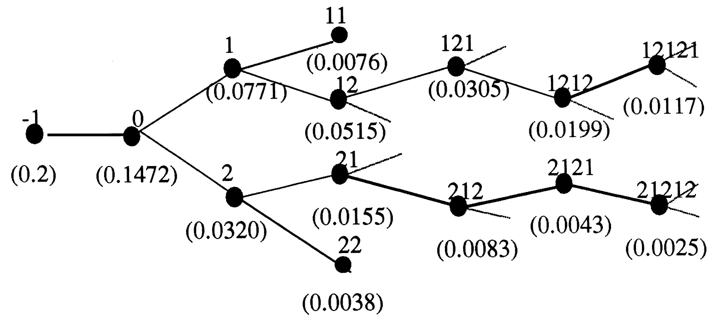

Fig. 5. The queue string distribution.

there are customers waiting at time t, $q(t)$ is a string in $\aleph$. It is easy to see that $\{q(t), I(t), (I_{n,1}(t), I_{n,2}(t)), 1 \leqslant n \leqslant N\}$, is a Markov process with a state space:

$$\bigcup_{q=1}^{N}\{\{-q\} \times \{1,2,\ldots,m\} \bigtimes_{n=1}^{N-q}\{1,2,\ldots,m_k\}\} \quad \text{and}$$
$$\aleph \times \{1,2,\ldots,m\} \bigtimes_{n=1}^{N}\left\{\bigcup_{k=1}^{K}\{1,2,\ldots,m_k\}\right\}.$$

This is a QBD process with a tree structure when $\{I(t), (I_{n,1}(t), I_{n,2}(t)), 1 \leqslant n \leqslant N\}$ is defined as the auxiliary random variable with $m\bar{m}^N$ states (where $\bar{m} = (m_1 + \cdots + m_K)$), except that when $q(t) = -N, \ldots, -q, \ldots, -1$, the auxiliary variable takes values

$$\{1,2,\ldots,m\} \bigtimes_{n=1}^{N-q}\left\{\bigcup_{k=1}^{K}\{1,2,\ldots,,m_k\}\right\}.$$

For instance, when $K=2$, the QBD Markov process $\{q(t), I(t), (I_{n,1}(t), I_{n,2}(t)), 1 \leqslant n \leqslant N\}$ is shown in Fig. 6 as follows.

Compared to the single server case, this QBD Markov process of the multiple server case has more soil nodes (boundary states) and the auxiliary variable has more states. The infinitesimal generator of this QBD Markov process with a tree structure is defined by the following transition blocks.

For $J = k_1 \cdots k_{n-1}k \in \aleph$ and $1 \leqslant k \leqslant K$, $\{A_0(l), 1 \leqslant l \leqslant K, A_1(k), A_2(k)\}$ are $m(\bar{m})^N \times m(\bar{m})^N$ matrices and

$$A_0(l) = D_l \otimes \mathbf{I}_{(\bar{m})^N \times (\bar{m})^N}, \quad 1 \leqslant l \leqslant K;$$

$$A_1(k) = D_0 \otimes \mathbf{I}_{(\bar{m})^N \times (\bar{m})^N} + \sum_{n=0}^{N-1}\mathbf{I}_{m\times m} \otimes \mathbf{I}_{(\bar{m})^n \times (\bar{m})^n} \otimes \begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_K \end{pmatrix} \otimes \mathbf{I}_{(\bar{m})^{N-1-n} \times (\bar{m})^{N-1-n}}; \tag{37}$$

$$A_2(k) = \sum_{n=0}^{N-1}\mathbf{I}_{m\times m} \otimes \mathbf{I}_{(\bar{m})^n \times (\bar{m})^n} \otimes \begin{pmatrix} 0 & \cdots & 0 & \mathbf{T}_1^0\boldsymbol{\alpha}_k & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & \cdots & 0 & \mathbf{T}_K^0\boldsymbol{\alpha}_k & 0 & \cdots & 0 \end{pmatrix} \otimes \mathbf{I}_{(\bar{m})^{N-1-n} \times (\bar{m})^{N-1-n}}.$$

For $J=0$, the transition to $J=k$ is $A_0(k)$, $1 \leqslant k \leqslant K$; the transition to itself is $A_1(0) = A_1(1)$, and the transition to node -1 is

$$B_2(0) = \sum_{n=0}^{N-1}\mathbf{I}_{m\times m} \otimes \mathbf{I}_{(\bar{m})^n \times (\bar{m})^n} \otimes \begin{pmatrix} T_1^0 \\ \vdots \\ T_K^0 \end{pmatrix} \otimes \mathbf{I}_{(\bar{m})^{N-1-n} \times (\bar{m})^{N-1-n}}. \tag{38}$$

For $-N+1 \leqslant J = q \leqslant -1$,

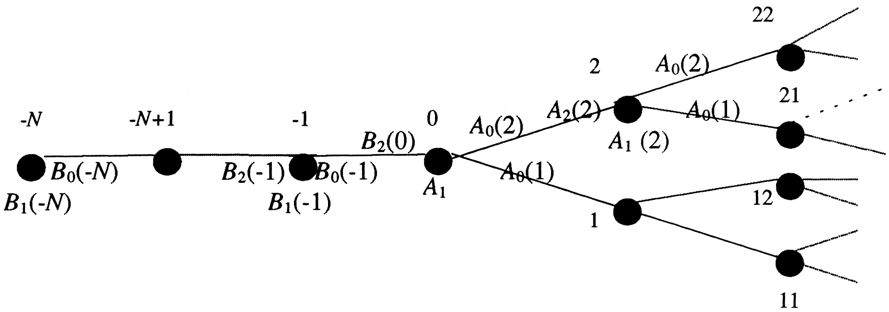

Fig. 6. A 2-ary tree with N soil nodes.

$$B_0(q) = (D_1 \otimes \mathbf{I}_{(\bar m)^{N+q}\times(\bar m)^{N+q}} \otimes \boldsymbol{\alpha}_1 \cdots D_K \otimes \mathbf{I}_{(\bar m)^{N+q}\times(\bar m)^{N+q}} \otimes \boldsymbol{\alpha}_K);$$

$$B_1(q) = D_0 \otimes \mathbf{I}_{(\bar m)^{N+q}\times(\bar m)^{N+q}} + \sum_{n=0}^{N-1+q} \mathbf{I}_{m\times m} \otimes \mathbf{I}_{(\bar m)^n\times(\bar m)^n} \otimes \begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_K \end{pmatrix} \otimes \mathbf{I}_{(\bar m)^{N-1+q-n}\times(\bar m)^{N-1+q-n}}; \tag{39}$$

$$B_2(q) = \sum_{n=0}^{N-1+q} \mathbf{I}_{m\times m} \otimes \mathbf{I}_{(\bar m)^n\times(\bar m)^n} \otimes \begin{pmatrix} T_1^0 \\ \vdots \\ T_K^0 \end{pmatrix} \otimes \mathbf{I}_{(\bar m)^{N-1+q-n}\times(\bar m)^{N-1+q-n}}.$$

For $J = -N$,

$$B_0(-N) = (D_1 \otimes \boldsymbol{\alpha}_1 \cdots D_K \otimes \boldsymbol{\alpha}_K); \quad B_1(-N) = D_0. \tag{40}$$

Once the transition blocks are determined, performance measures can be obtained similar to the single server case. Define the traffic intensity as $\rho = (\lambda_1/\mu_1 + \cdots + \lambda_K/\mu_K)/N$. Similar to Theorem 3.1, it can be proven that when the queueing system is positive recurrent, $\rho < 1$. All details are omitted.

7. Summary

This paper gives a detailed analysis of the *MMAP*[*K*]/*PH*[*K*]/*N* queue with an LCFS non-preemptive service discipline, especially the single server case. Using results of the QBD Markov process with a tree structure, an efficient algorithm is developed for computing the stationary distribution of queue string. All the matrices, variables, parameters involved are explicitly given. While a general algorithm is given, several special algorithms are presented for some special cases, which are much more efficient computationally.

Taking advantages of the problem formulation and efficient algorithms developed, this paper presents some numerical results to show how the interactions among the arrival processes and/or the uncertainty in service times influence the queueing process. The impact of the pattern of the arrival process (cyclic, one type follows the other type, etc.) on the queueing process is discussed as well.

Several issues are pertinent to future research. First, it will be interesting to look at queueing systems with multiple types of customers and an FCFS service discipline. The development of efficient algorithms for computing the distributions of queue strings of such queueing systems is an interesting but difficult problem. Second, there are many differences between the discrete time and continuous time *MMAP*[*K*]/*PH*[*K*]/*N*/*LCFS* queues. A detailed study of the discrete time *MMAP*[*K*]/*PH*[*K*]/*N*/*LCFS* queue can be interesting as well.

Acknowledgements

The author would like to thank two anonymous referees for their valuable comments and suggestions.

References

[1] S. Asmussen, G. Koole, Marked point processes as limits of Markovian arrival streams, Journal of Applied Probability 30 (1993) 365–372.

[2] F.R. Gantmacher, The Theory of Matrices, Chelsea, New York, 1959.

[3] Q.-M. He, Classification of Markov processes of matrix *M/G/*1 type with a tree structure and its applications to the *MMAP*[*K*]/*G*[*K*]/1 queue, Working paper #98-05, Department of Industrial Engineering, DalTech, Dalhousie University, 1998.

[4] Q.-M. He, Queues with marked customers, Advances in Applied Probability 28 (1996) 567–587.

[5] Q.-M. He, A.S. Alfa, The *MMAP*[*K*]/*PH*[*K*]/1 queue with a last-come-first-served preemptive service discipline, Queueing Systems 29 (1998) 269–291.

[6] Q.-M. He, M.F. Neuts, Markov arrival processes with marked transitions, Stochastic Processes and their Applications 74 (1998) 37–52.

[7] F.P. Kelly, Reversibility and Stochastic Networks, Wiley, New York, 1979.

[8] G. Latouche, A note on two matrices occurring in the solution of quasi-birth-and-death processes, Stochastic Models 3 (1987) 251–257.

[9] G. Latouche, V. Ramaswami, A logarithmic reduction algorithm for quasi birth and death processes, Journal of Applied Probability 30 (1993) 650–674.

[10] R. Lyons, Random walks, capacity and percolation on trees, Annals of Probability 20 (1992) 2043–2088.

[11] D.M. Lucantoni, New results on the single server queue with a batch Markovian arrival process, Stochastic Models 7 (1991) 1–46.

[12] M.F. Neuts, A versatile Markovian point process, Journal of Applied Probability 16 (1979) 764–779.

[13] M.F. Neuts, Matrix-Geometric Solutions in Stochastic Models: An algorithmic Approach, Johns Hopkins University Press, Baltimore, MD, 1981.

[14] M.F. Neuts, Structured Stochastic Matrices of M/G/1 Type and Their Applications, Marcel Dekker, New York, 1989.

[15] M.F. Neuts, The joint distribution of arrivals and departures in quasi-birth-and-death processes, 1989.

[16] H. Takagi, Queueing Analysis: A Foundation of Performance Evaluation, Vol. 1: Vacation and Priority Systems, Part 1, Elsevier, Amsterdam, 1990.

[17] H. Takagi, Queueing analysis of polling models: Progress in 1990–1994, in: J.H. Dshalalow (Ed.), Frontiers in Queueing, 1996, pp. 119–146.

[18] T. Takine, 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 10 (1994) 183–204.

[19] T. Takine, B. Sengupta, 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.

[20] V. Ramaswami, P.G. Taylor, Some properties of the rate operators in level-dependent quasi-birth-and-death processes with a countable number of phases, Commun. Statist. Stochastic Models 12 (1) (1996).

[21] R.W. Yeung, A.S. Alfa, The quasi-birth-and-death type Markov chain with a tree structure, submitted.

[22] R.W. Yeung, B. Sengupta, Matrix product-form solutions for Markov chains with a tree structure, Advances in Applied Probability 26 (4) (1994) 965–987.