The $MMAP[K]/PH[K]/1$ queues with a last-come-first-served preemptive service discipline

Qi-Ming HE [a] and Attahiru Sule Alfa [b]

[a] *Department of Industrial Engineering, DalTech, Dalhousie University, Halifax, Nova Scotia, Canada B3J 2X4*
E-mail: Qi-Ming.He@dal.ca

[b] *Department of Mechanical and Industrial Engineering, University of Manitoba, Canada*
E-mail: alfa@cc.umanitoba.ca

Received 27 August 1997; revised 26 February 1998

This paper studies two queueing systems with a Markov arrival process with marked arrivals and PH-distribution service times for each type of customer. Customers (regardless of their types) are served on a last-come-first-served preemptive resume and repeat basis, respectively. The focus is on the stationary distribution of queue strings in the system and busy periods. Efficient algorithms are developed for computing the stationary distribution of queue strings, the mean numbers of customers served in a busy period, and the mean length of a busy period. Comparison is conducted numerically between performance measures of queueing systems with preemptive resume and preemptive repeat service disciplines. A counter-intuitive observation is that for a class of service time distributions, the repeat discipline performs better than the resume one.

Keywords: queueing theory, matrix analytic methods, tree structure, LCFS, quasi-birth-and-death Markov process

1. Introduction

This paper studies two types of queueing systems in which there are several classes of customers whose arrivals are correlated and each has a different type of service time requirement. The arrivals are described by a Markov arrival process with marked arrivals and service times are phase type for each customer. The two types of systems studied are last-come-first-serve (LCFS) preemptive resume and preemptive repeat. These types of queueing systems with LCFS discipline are encountered in computer communications systems. For example, in a multi-access communications system the LCFS discipline is sometimes used in the splitting algorithm with tree structure. It has been considered as an alternative to the first-come-first-serve (FCFS) discipline. The FCFS splitting algorithm has a weakness in that all nodes are required to monitor the channel feedback at all times. The LCFS approach does not have this requirement, instead it allows nodes to monitor the feedback only after receiving a packet.

Another potential application of these types of queueing systems is in the area of studying the Advanced Intelligent Networks (AIN). Subscribers to AIN are usually different types whose arrivals or requests for service are usually correlated within and across classes and each class usually has different service time requirements. A subscriber in an AIN usually submits a request for service through a service switching point which then goes via a signal transfer point into the service control point where the database is usually stored. After a service request is received, information is requested from the subscriber. While the subscriber is preparing the information, it is usually preempted in order to attend to another service request. The server eventually returns to the preempted subscriber in order to accept its information for processing. This processing may also involve several further requests for more information, which result in preemption. Most of the preempted services are resumed without loss and some others are repeated for several reasons. Knowing that users are usually impatient if it takes longer than expected for them to receive service it is possible that some of the customers may hang up thereby decreasing the throughput of the system. The idea of last-come-first-serve may be more efficient than the standard first-come-first-serve approach as discovered in the overload systems for telephone systems. We therefore study this system as last-come-first-served preemptive resume and also repeat. We obtain results for the stationary distribution of the queue strings in the system and the busy periods. These results will be useful for analysts interested in the performance of AIN.

The system considered is modelled using the results for the classical quasi-birth-and-death (QBD) Markov processes with a tree structure presented by Yeung and Alfa [20], and HE [4], which are all based on the results of Yeung and Sengupta [19] and Takine et al. [18]. The work by Yeung and Alfa [20] focused on the Markov chain of the QBD with a tree structure and the work of HE [4] focused on the nonpreemptive LCFS queueing system. For more research work related to this paper, readers are referred to Kelly [6], Lyons [8,9], Takagi [15,16], and references therein. The current work deals with a queueing system with several types of customers and LCFS preemptive service disciplines. The current paper also gives results on the busy period of the system.

The main contribution of this paper is the formulation of queueing systems with multiple types of customers and LCFS preemptive service disciplines into QBD Markov processes with a tree structure. Such a formulation allows the development of efficient algorithms for computing important performance measures such as the stationary distributions of queue length and busy period. Since customers from different (but correlated) sources are distinguished, results obtained in this paper make it possible to analyze the composition of the queue or what occurs in a busy period at the levels of individual types of customers.

This paper deals with the $MMAP[K]/PH[K]/1/$LCFS queue with preemptive resume or preemptive repeat service discipline. By studying the two cases together, it is possible to compare the queueing processes when the two service disciplines are applied respectively. A few examples shall be presented in this paper for such purposes. Another reason the two cases are put in one paper is that the two cases can be solved using the same general approach, though details are different.

The rest of the paper is organized as follows. First, the $MMAP[K]/PH[K]/1/$ LCFS preemptive resume (repeat) queue is introduced in section 2. In section 3, the queueing process of the queueing system with preemptive resume service discipline is formulated into a QBD Markov process with a tree structure. An algorithm for computing the stationary distribution of the queue string is developed in section 4. Sections 5 and 6 deal with the preemptive repeat queue. In section 7, the fundamental periods of the preemptive repeat queue are studied and an algorithm is developed for computing the mean number of customers served in a busy period and mean length of a busy period. Numerical examples are presented in section 8 with brief discussions. Finally, in section 9, the results obtained in this paper are summarized.

2. The $MMAP[K]/PH[K]/1/\mathrm{LCFS}$ preemptive resume or repeat queue

This section defines 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 a "last-come-first-served" (LCFS) preemptive resume or repeat basis. To define the queueing systems of interest explicitly, the input process $MMAP[K]$ is introduced first and then the service time distributions are specified.

A Markov arrival process with marked transitions is defined by a set of matrices $\{D_k,\ 0 \leqslant k \leqslant K\}$. 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 nonsingular. Let

$$D = D_0 + \sum_{k=1}^{K} D_k. \tag{2.1}$$

Then matrix D is the infinitesimal generator of the underlying Markov process. Let $I(t)$ be the phase of the underlying Markov process at time t. An arrival is called a type k arrival if the arrival is marked by k. The (matrix) marking rate of type k arrivals is D_k. Let $\boldsymbol{\theta}$ be the stationary probability vector of the matrix D. The stationary arrival rate of type k arrivals is given by $\lambda_k = \boldsymbol{\theta} D_k \boldsymbol{e}$, $1 \leqslant k \leqslant K$, where $\boldsymbol{e}$ is the column vector with all elements one.

The advantage of the $MMAP[K]$ is that it can capture correlation between interarrival times within a class of customers and across different classes of customers with yet a computationally tractable queueing model. See Asmussen and Koole [1], HE [3,4], or HE and Neuts [5] for more about the $MMAP[K]$. See Neuts [11,13], and Lucantoni [10] for more about Markov arrival processes. A simple example of $MMAP[K]$ is the superposition process of K independent Poisson processes, where if 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 $D_0 = -(\lambda_1 + \cdots + \lambda_K)$, $D_1 = \lambda_1, \ldots, D_K = \lambda_K$. More complicated and interesting examples can be constructed as well. For instance, consider an $MMAP[K]$ with $K = 2$, $m = 2$,

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

For this point process, the arrival pattern is that every type 2 customer is followed by a type 1 customer.

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-dimensional vector ($\boldsymbol{\alpha}_k = (\alpha_{k,1}, \alpha_{k,2}, \ldots, \alpha_{k,m_k})$) and T_k is an $m_k \times m_k$ matrix. Let $\boldsymbol{T}_k^0 = -T_k\boldsymbol{e}$. The mean service time is given by $1/\mu_k = -\boldsymbol{\alpha}_k T_k^{-1}\boldsymbol{e}$. Then μ_k is the average service rate of type k customers. For $(\boldsymbol{\alpha}_k, T_k)$, its associated Markov process has m_k phases. For more details about PH-distribution, see Neuts [12, chapter 2].

The preemptive resume queue. When there is a customer in service and a new customer arrives, the new customer enters the server immediately and the old customer moves back into the queue with its service phase (of its PH-distribution) recorded. Customers in queue will resume their services when the server becomes available on a last-come-first-served basis. *When a customer resumes its service, the PH-distribution starts in the phase recorded when the customer was pushed out of the server.*

The preemptive repeat queue. When there is a customer in service and a new customer arrives, the new customer enters the server immediately and the old customer moves back into the queue. Customers in queue will restart their services on a last-come-first-served basis when the server becomes available. *When a customer restarts its service, the PH-distribution starts with its original initial distribution.*

The traffic intensity of the queueing system is defined as $\rho = \lambda_1/\mu_1 + \cdots + \lambda_K/\mu_K$. It shall be proved that for the preemptive resume case the queueing system is ergodic only if $\rho < 1$. For the preemptive repeat case, $\rho < 1$ does not guarantee the ergodicity of the queueing system. A condition for the ergodicity is yet to be found. In this paper, the two queueing systems are analyzed given that they can reach their steady state. Sections 3 and 4 deal with the preemptive resume queue. Sections 5 and 6 present results of the preemptive repeat queue. Section 7 introduces an approach to study the busy period and busy cycle, which is useful to both the preemptive resume and repeat queues.

3. The QBD Markov process with a tree structure of the resume queue

Sections 3 and 4 focus on the preemptive resume case. The queueing process of the $MMAP[K]/PH[K]/1/$LCFS preemptive resume queue can be modelled into a QBD Markov process with a tree structure. The stationary distribution associated with the queue length can be computed using an algorithm developed for the QBD Markov process with a tree structure. Readers are referred to Yeung and Sengupta [19], Yeung and Alfa [20], and HE [4] for more details about the QBD Markov process with a tree structure.

To study the queue length of the queueing system of interest, a key observation is that the types of customers in queue and their service phases (at the time they are pushed out of service) must be recorded. This is where a tree structure comes into play. To model the queueing process into a QBD Markov process with a tree structure, the following integer string sets are introduced. Let

$$\begin{aligned}
\aleph &= \bigl\{\mathrm{J}\colon\ \mathrm{J} = k_1k_2\ldots k_n,\ 1 \leqslant k_i \leqslant K,\ 1 \leqslant i \leqslant n,\ n \geqslant 1\bigr\} \cup \{0\},\\
\Gamma(\mathrm{J}) &= \bigl\{(\mathrm{J},\mathrm{S})\colon\ \mathrm{S} = s_1s_2\ldots s_n,\ 1 \leqslant s_i \leqslant m_{k\_i},\ 1 \leqslant i \leqslant n\bigr\},\\
&\qquad \mathrm{J} = k_1k_2\ldots k_n \in \aleph \quad \text{and} \quad \mathrm{J} \neq 0,\\
\Gamma(0) &= \bigl\{(0,0)\bigr\}, \qquad \Gamma = \bigcup_{\mathrm{J}\in\aleph} \Gamma(\mathrm{J}).
\end{aligned}$$

Notice that "k_i" represents "k_i" for typographical reasons. Any string vector (J,S) in Γ is a node. Two operations associated with integer strings in $\aleph$ are defined as:

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

The length of string $\mathrm{J} = k_1\ldots k_n$ is denoted as $|\mathrm{J}|$ $(= n)$. Similar operations are defined for string S.

The queueing system of interest can be represented by the following three dimensional stochastic process:

$q(t)$: the string of types of customers in queue (including the one in server), $q(t) \in \aleph$,

$s(t)$: the string of service phases of customers in queue, $(q(t), s(t)) \in \Gamma(q(t))$, $q(t) \in \aleph$,

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

When there is no customer in the system at time t, $q(t) = 0$ and $s(t) = 0$ since there is no service. When there are customers in the system at time t, $q(t)$ is a nonzero string in $\aleph$. For example, for $K = 2$, $(q(t) = 122,\ s(t) = 435)$ implies that there are 3 customers in the system at time t: the customer who arrived first is of type 1 and its service halt in phase 4; the customer who arrived second is of type 2 and its service halt in phase 3; and the customer who arrived last (in service) is of type 2 and its current service phase is 5. When a new customer of type k arrives, $q(t)$ becomes $122k$ and $s(t)$ becomes $435s$, where s is the initial phase of the PH-distribution of the newly arrived customer of type k, $1 \leqslant s \leqslant m_k$, When the current service is completed (before the next arrival), $(q(t), s(t))$ returns to (122, 435), i.e., the type 2 customer who arrived before the type k customer resumes its service in phase 5.

In general, from a node (J,S), the stochastic process $(q(t), s(t))$ may move, in one transition (see figure 1), to

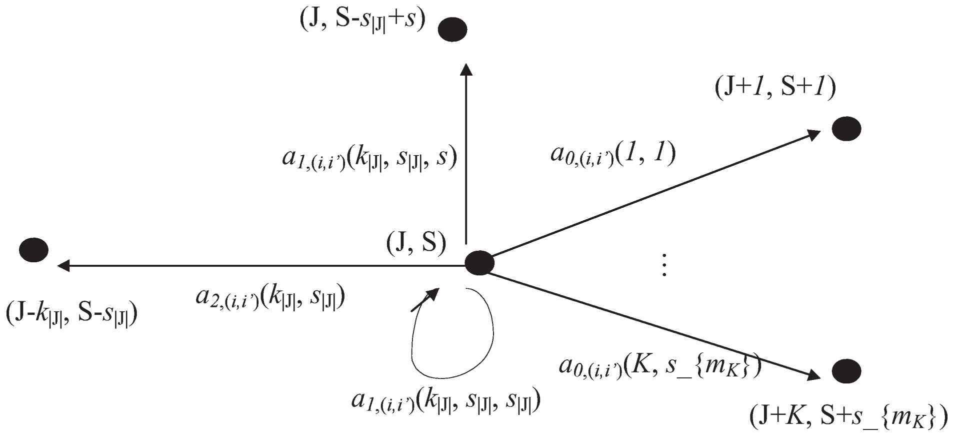

Figure 1. Possible one step transitions when J > 0.

(1) one of its $m_1+m_2+\cdots+m_K$ children $\{(\mathrm{J}+k, \mathrm{S}+s),\ 1 \leqslant k \leqslant K,\ 1 \leqslant s_k \leqslant m_k\}$ when there is an arrival of type k with a transition rate $a_{0,(i,i')}(k,s)$;

(2) the node itself when the service phase changes to itself or the underlying Markov process changes state with a transition rate $a_{1,(i,j')}(k_{|\mathrm{J}|}, s_{|\mathrm{J}|}, s_{|\mathrm{J}|})$ or $b_{1,(i,i')}$;

(3) one of its sibling nodes $\{(\mathrm{J}, \mathrm{S} - s_{|\mathrm{J}|} + s),\ 1 \leqslant s \leqslant m\_\{k_{|\mathrm{J}|}\},\ s \neq s_{|\mathrm{J}|}\}$ when the service phase changes without a service completion with a transition rate $a_{1,(i,i)}(k_{|\mathrm{J}|}, s_{|\mathrm{J}|}, s)$;

(4) its parent node $(\mathrm{J}-k_{|\mathrm{J}|}, \mathrm{S}-s_{|\mathrm{J}|})$ when the current service completes with a transition rate $a_{2,(i,i)}(k_{|\mathrm{J}|}, s_{|\mathrm{J}|})$.

It is then easy to see that $(q(t), s(t), I(t))$ is a Markov process with a state space: $\Gamma \times \{1, 2, \ldots, m\}$. This is a QBD Markov process with a tree structure when $I(t)$ is defined as the auxiliary random variable taking m values. In matrix form, the transition rates of the QBD Markov process are written as $\{A_0(k,s)$, $A_1(k_{|\mathrm{J}|}, s_{|\mathrm{J}|}, s)$, $A_1(k_{|\mathrm{J}|}, s_{|\mathrm{J}|}, s_{|\mathrm{J}|})$, $A_2(k_{|\mathrm{J}|}, s_{|\mathrm{J}|})\}$ and B_1, respectively. Notice that for $A_0(k,s)$, "(k,s)" is the child the Markov process transits to; for $A_1(k,s,s')$, "(k,s')" is the sibling node the Markov process transits to; and for $A_2(k,s)$, "(k,s)" is the last element of the string vector of the current node. According to the law of total probability, the matrix blocks satisfy the following equalities:

$$\left[\sum_{l=1}^{K}\sum_{s'=1}^{m\_l} A_0(l,s') + \sum_{s'=1}^{m\_l} A_1(k,s,s') + A_2(k,s)\right]\boldsymbol{e} = \mathbf{0}, \quad 1 \leqslant k \leqslant K,\ 1 \leqslant s \leqslant m_k,$$

$$\left[\sum_{l=1}^{K}\sum_{s'=1}^{m\_l} A_0(l,s') + B_1\right]\boldsymbol{e} = \mathbf{0}, \tag{3.1}$$

where $\boldsymbol{e}$ is the vector with all components one. The transitions of the QBD Markov process are illustrated in figure 2 for $K = 2$.

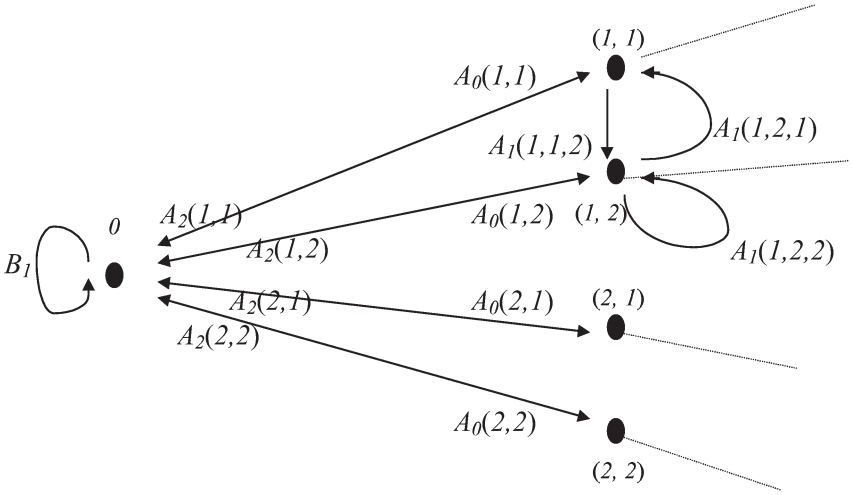

Figure 2. The QBD Markov process with a tree structure.

The infinitesimal generator of the QBD Markov process $(q(t), s(t), I(t))$ is defined by the following transition blocks. For J > 0, $1 \leqslant k \leqslant K$ and $1 \leqslant s, s' \leqslant m_k$,

- $A_0(k, s) = D_k\alpha_{k,s}$ (a type k customer arrives and service starts in phase s),
- $A_1(k, s, s) = D_0 + (T_k)_{s,s}\boldsymbol{I}$ (no service completed, no arrival, and no service phase change),
- $A_1(k, s, s') = (T_k)_{s,s'}\boldsymbol{I}$ (service phase changes from s to s'), $s \neq s'$,
- $A_2(k, s) = (T_k^0)_s\boldsymbol{I}$ (a service completes),

where $\boldsymbol{I}$ is the identity matrix. For J $= 0$, $B_1 = D_0$. All these transition blocks are $m \times m$ matrices.

Note 3.1. HE [4] formulated the queueing process of the $MMAP[K]/PH[K]/1/$ LCFS nonpreemptive queue into a QBD Markov process with a tree structure. Compared to the nonpreemptive case, the QBD Markov process defined in this section has more nodes. Each node has $m_1 + m_2 + \cdots + m_K$ children while there are only K children for the nonpreemptive case. However, the dimensions of the matrix blocks are m, instead of $m(m_1 + m_2 + \cdots + m_K)$ for the nonpreemptive case. Another difference is that there is a “soil” node connected only to the root node (node 0) in the tree defined in HE [4] which is not necessary here.

4. The stationary distribution of the resume queue

Having defined the quasi-birth-and-death Markov process which describes the queueing system of interest explicitly, some results about its stationary distribution are presented next. Let

$$\pi(\mathrm{J},\mathrm{S},i)=\lim_{t\to\infty}\boldsymbol{P}\bigl\{\bigl(q(t),s(t),I(t)\bigr)=(\mathrm{J},\mathrm{S},i)\bigr\} \tag{4.1}$$

and

$$\boldsymbol{\pi}(\mathrm{J},\mathrm{S})=\bigl(\pi(\mathrm{J},\mathrm{S},1),\ldots,\pi(\mathrm{J},\mathrm{S},m)\bigr). \tag{4.2}$$

When the underlying Markov chain D of the arrival process and all the PH-distributions are irreducible, $(q(t),s(t),I(t))$ is irreducible. When the queueing system is ergodic, its stationary distribution vectors $\{\boldsymbol{\pi}(\mathrm{J},\mathrm{S})\colon (\mathrm{J},\mathrm{S})\in\Gamma\}$ satisfy the following equation:

$$\begin{aligned}
\mathbf{0}&=\boldsymbol{\pi}(0,0)B_1+\sum_{l=1}^{K}\sum_{s'=1}^{m\_l}\boldsymbol{\pi}\bigl(l,s'\bigr)A_2\bigl(l,s'\bigr),\\
\mathbf{0}&=\boldsymbol{\pi}(\mathrm{J},\mathrm{S})A_0(k,s)+\sum_{s'=1}^{m\_k}\boldsymbol{\pi}\bigl(\mathrm{J}+k,\mathrm{S}+s'\bigr)A_1\bigl(k,s',s\bigr)\\
&\quad+\sum_{l=1}^{K}\sum_{s'=1}^{m\_l}\boldsymbol{\pi}\bigl(\mathrm{J}+k+l,\mathrm{S}+s+s'\bigr)A_2\bigl(l,s'\bigr),\quad \text{for } 1\leqslant k\leqslant K,\ 1\leqslant s\leqslant m_k,
\end{aligned} \tag{4.3}$$

which is useful in helping understand the following solution intuitively. Furthermore, an ergodic condition of such queueing systems and some elementary performance measures obtained from the stationary distribution are given in the following theorem.

Theorem 4.1. 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

$$\sum_{(\mathrm{J},\mathrm{S})}\boldsymbol{\pi}(\mathrm{J},\mathrm{S})D_k\boldsymbol{e}=\lambda_k;$$

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

$$\sum_{\mathrm{J}\geqslant 1,\ k\_\{|\mathrm{J}|\}=k}\boldsymbol{\pi}(\mathrm{J},\mathrm{S})\boldsymbol{e}=\frac{\lambda_k}{\mu_k};$$

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

$$\rho=\sum_{\mathrm{J}\geqslant 1,\ (\mathrm{J},\mathrm{S})\in\Gamma}\boldsymbol{\pi}(\mathrm{J},\mathrm{S})\boldsymbol{e}=\sum_{k=1}^{K}\frac{\lambda_k}{\mu_k};$$

(d) the probability that the queueing system is empty is $\boldsymbol{\pi}(0,0)\boldsymbol{e}=1-\rho$.

Thus, the queueing system of interest is ergodic only if $\rho<1$.

Proof. In the queueing system of interest with LCFS and preemptive resume, the server starts to serve a type k customer when a type k customer arrives. This proves part (a).

To prove part (b), first notice that the stationary distribution of the Markov process $(q(t), s(t), I(t))$ satisfies the following equation, for $1 \leqslant k \leqslant K$ and $1 \leqslant s \leqslant m_k$,

$$\begin{aligned}\mathbf{0} = {} & \boldsymbol{\pi}(\mathrm{J}, \mathrm{S}) D_k \alpha_{k,s} + \sum_{s'=1:s'\neq s}^{m\_k} \boldsymbol{\pi}\big(\mathrm{J}+k, \mathrm{S}+s'\big)(T_k)_{s',s} \\ & + \boldsymbol{\pi}(\mathrm{J}+k, \mathrm{S}+s)\big[D_0 + (T_k)_{s,s}\boldsymbol{I}\big] \\ & + \sum_{l=1}^{K} \sum_{s'=1}^{m\_l} \boldsymbol{\pi}\big(\mathrm{J}+k+l, \mathrm{S}+s+s'\big)\big(\boldsymbol{T}_l^0\big)_s. \end{aligned} \tag{4.4}$$

Let

$$\begin{aligned} \boldsymbol{X} &= \sum_{(\mathrm{J},\mathrm{S})} \boldsymbol{\pi}(\mathrm{J}, \mathrm{S}), \\ \boldsymbol{X}(k, s) &= \sum_{(\mathrm{J},\mathrm{S})} \boldsymbol{\pi}(\mathrm{J}+k, \mathrm{S}+s) \end{aligned}$$

and

$$\boldsymbol{X}\big(k+l, s+s'\big) = \sum_{(\mathrm{J},\mathrm{S})} \boldsymbol{\pi}\big(\mathrm{J}+k+l, \mathrm{S}+s+s'\big).$$

Apparently, $\boldsymbol{X} = \boldsymbol{\theta}$, i.e., the stationary distribution of the underlying Markov process. Taking summation over all (J, S), equation (4.4) yields

$$\begin{aligned}\mathbf{0} = {} & \boldsymbol{X} D_k \alpha_{k,s} + \sum_{s'=1:s'\neq s}^{m\_k} \boldsymbol{X}\big(k, s'\big)(T_k)_{s',s} + \boldsymbol{X}(k, s)\big[D_0 + (T_k)_{s,s}\boldsymbol{I}\big] \\ & + \sum_{l=1}^{K} \sum_{s'=1}^{m\_l} \boldsymbol{X}\big(k+l, s+s'\big)\big(\boldsymbol{T}_l^0\big)_s. \end{aligned} \tag{4.5}$$

Postmultiplying $\boldsymbol{e}$ on both sides of (4.5), yields

$$\begin{aligned} 0 = {} & \lambda_k \alpha_{k,s} + \sum_{s'=1}^{m\_k} \boldsymbol{X}\big(k, s'\big)\boldsymbol{e}(T_k)_{s',s} + \boldsymbol{X}(k, s) D_0 \boldsymbol{e} \\ & + \sum_{l=1}^{K} \sum_{s'=1}^{m\_l} \boldsymbol{X}\big(k+l, s+s'\big)\boldsymbol{e}\big(\boldsymbol{T}_l^0\big)_{s'}. \end{aligned} \tag{4.6}$$

In steady state, the rate of leaving the set $\{(\mathrm{J}+k, \mathrm{S}+s),\ (\mathrm{J}, \mathrm{S}) \text{ in } \Gamma\}$ equals the rate of entering that set, which leads to

$$-\boldsymbol{X}(k,s)D_0\boldsymbol{e}-\sum_{l=1}^{K}\boldsymbol{X}(k+l,s+s)\boldsymbol{e}\big(\boldsymbol{T}_l^0\big)_s$$

$$=\boldsymbol{X}(k,s)(D_1+\cdots+D_K)\boldsymbol{e}-\sum_{l=1}^{K}\boldsymbol{X}(k+l,s+s)\boldsymbol{e}\big(\boldsymbol{T}_l^0\big)_s$$

$$=\sum_{l=1}^{K}\sum_{s'=1}^{m\_l}\boldsymbol{X}\big(k+l,s+s'\big)\boldsymbol{e}\big(\boldsymbol{T}_l^0\big)_{s'}-\sum_{l=1}^{K}\boldsymbol{X}(k+l,s+s)\boldsymbol{e}\big(\boldsymbol{T}_l^0\big)_s. \quad (4.7)$$

Combining equations (4.6) and (4.7) leads to

$$0=\lambda_k\alpha_{k,s}+\sum_{s'=1}^{m\_k}\boldsymbol{X}\big(k,s'\big)\boldsymbol{e}(\boldsymbol{T}_k)_{s',s},$$

or in vector form:

$$\mathbf{0}=\lambda_k\boldsymbol{\alpha}_k+\big(\boldsymbol{X}(k,1)\boldsymbol{e},\ldots,\boldsymbol{X}(k,m_k)\boldsymbol{e}\big)T_k. \quad (4.8)$$

This leads to

$$\sum_{s=1}^{m\_k}\boldsymbol{X}(k,s)\boldsymbol{e}=\lambda_k\boldsymbol{\alpha}_k\big(-T_k^{-1}\big)\boldsymbol{e}=\frac{\lambda_k}{\mu_k}. \quad (4.9)$$

This completes the proof of part (b).

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).

Apparently,

$$\boldsymbol{\pi}(0,0)\boldsymbol{e}=1-\rho>0$$

when the queueing system is ergodic, i.e., $\rho<1$. This completes the proof. □

A matrix geometric solution can be found for the stationary distribution of $(q(t),s(t),I(t))$ similar to Neuts [12] for classical QBD Markov processes. For general Markov processes with a tree structure and skip-free to the left, a matrix geometric solution is given in Yeung and Sengupta [19]. Therefore, a matrix geometric solution shall be given next without a proof.

Theorem 4.2. When the queueing system is ergodic (i.e., $\rho<1$), the stationary distribution of $(q(t),s(t),I(t))$ is given by

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

$$\boldsymbol{\pi}(0,0)\left[B_1+\sum_{k=1}^{K}\sum_{s=1}^{m\_k}R(k,s)A_2(k,s)\right]=\mathbf{0}, \quad (4.10)$$

$$\boldsymbol{\pi}(0,0)(\boldsymbol{I}-R)^{-1}\boldsymbol{e}=1,$$

where $R = R(1,1) + \cdots + R(K, m_K)$, and $\{R(k,s),\ 1 \leqslant k \leqslant K,\ 1 \leqslant s \leqslant m_k\}$ are the minimal non-negative solutions to

$$0 = A_0(k,s) + \sum_{s'=1}^{m\_k} R\bigl(k,s'\bigr)A_1\bigl(k,s',s\bigr) + R(k,s)\sum_{l=1}^{K}\sum_{s'=1}^{m\_l} R\bigl(l,s'\bigr)A_2\bigl(l,s'\bigr). \quad (4.11)$$

Proof. See Yeung and Sengupta [19] or HE [4]. □

The computation of $\{R(k),\ 1 \leqslant k \leqslant K\}$ can be carried out using the following algorithm which is a generalization of the classical algorithm given in Neuts [12]. Let

$$R(k,s)[0] = 0, \quad 1 \leqslant k \leqslant K,\ 1 \leqslant s \leqslant m_k,$$

and

$$R(k,s)[n+1] = \Biggl[A_0(k,s) + \sum_{s'=1,\ s'\neq s}^{m\_l} R\bigl(k,s'\bigr)[n]A_1\bigl(k,s',s\bigr) + R(k,s)[n] \times \sum_{l=1}^{K}\sum_{s'=1}^{m\_l} R\bigl(l,s'\bigr)[n]A_2\bigl(l,s'\bigr)\Biggr]\bigl[-A_1(k,s,s)\bigr]^{-1}. \quad (4.12)$$

It can be proved that $\{R(k,s)[n],\ n \geqslant 0\}$ is a monotone sequence which converges to $R(k,s)$ from below, for $1 \leqslant k \leqslant K$ and $1 \leqslant s \leqslant m_k$. This algorithm is simple and easy to implement, but may not be the most efficient one. (See Latouche and Ramaswami [7] and Yeung and Alfa [20] for more advanced algorithms.)

The mean number of customers waiting is $\boldsymbol{\pi}(0,0)R(\boldsymbol{I}-R)^{-2}\boldsymbol{e}$. It has been proved in Yeung and Sengupta [19] that the spectrum (or the eigenvalue with the largest real part) of matrix R is less than one when the QBD Markov process is ergodic. The mean number of customers in the system (mean queue length) can be computed by using the formula

$$\overline{L} = \boldsymbol{\pi}(0,0)R(\boldsymbol{I}-R)^{-2}\boldsymbol{e} + \boldsymbol{\pi}(0,0)(\boldsymbol{I}-R)^{-1}\boldsymbol{e} = \boldsymbol{\pi}(0,0)(\boldsymbol{I}-R)^{-2}\boldsymbol{e}. \quad (4.13)$$

To end this section, an algorithm is given for computing the stationary distribution.

Algorithm I.

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.

Step 3: Compute matrices $\{R(k,s),\ 1 \leqslant k \leqslant K,\ 1 \leqslant s \leqslant m_k\}$.

Step 4: Compute vectors $\boldsymbol{\pi}(0,0)$.

Step 5: Compute string distribution $\{\boldsymbol{\pi}(\mathrm{J},\mathrm{S}),\ (\mathrm{J},\mathrm{S}) \in \Gamma\}$.

5. The QBD Markov process with a tree structure of the repeat queue

In this section and the next two sections, the $MMAP[K]/PH[K]/1/$LCFS preemptive repeat queue is studied. In many aspects, the modelling and solution processes of this queueing model are similar to that of the preemptive resume (as well as the nonpreemptive) case. However, there are subtle differences which invite a somehow detailed treatment of this case. First, the QBD Markov process with a tree structure of interest is defined in this section.

Unlike the preemption resume case, the service phase at the pushing out epoch of a customer does not have to be recorded since when a waiting customer enters the server again, its service time starts like new. But it is now necessary to know the type of the customer to be served next when the current service completes; otherwise, the server does not know the initial phase of the next service. In order to do so, the following integer string set are introduced. Let

$$\Omega = \big\{\mathrm{J}\text{: } \mathrm{J} = (0, k_1)(k_1, k_2)(k_2, k_3)\ldots(k_{n-1}, k_n),\ 1 \leqslant k_i \leqslant K,\ 1 \leqslant i \leqslant n\big\} \cup \{0\}.$$

Any string J in Ω is a node. Two operations associated with integer strings are defined as:

1. Addition operation: for $\mathrm{J} = (0, k_1)(k_1, k_2)\ldots(k_{n-1}, k_n) \in \Omega$ and $1 \leqslant k \leqslant K$, $\mathrm{J} + k = (0, k_1)(k_1, k_2)\ldots(k_{n-1}, k_n)(k_n, k) \in \Omega$.
2. Subtraction operation: for $\mathrm{J} = (0, k_1)(k_1, k_2)\ldots(k_{n-1}, k_n) \in \Omega$, $\mathrm{J} - (k_{n-1}, k_n) = (0, k_1)(k_1, k_2)\ldots(k_{n-2}, k_{n-1}) \in \Omega$.

The queueing system of interest can be represented by the following three dimensional stochastic process:

- $q(t)$: the string of the types of customers in queue (including the one in server), $q(t) \in \Omega$,
- $I(t)$: the state of the underlying Markov process D, $1 \leqslant I(t) \leqslant m$,
- $I_1(t)$: the phase of the current service, $1 \leqslant I_1(t) \leqslant m_{k\_n}$ when

$$q(t) = (0, k_1)(k_1, k_2)\ldots(k_{n-1}, k_n).$$

Node transitions of $q(t) = \mathrm{J} = (0, k_1)(k_1, k_2)\ldots(k_{n-1}, k_n)$ can be one of the following three:

(1) to one of its K children $\{\mathrm{J} + k,\ 1 \leqslant k \leqslant K\}$ when there is an arrival of type k,

(2) to its parent node $\mathrm{J} - (k_{n-1}, k_n)$ when the current service completes,

(3) to itself when the state of the underlying Markov process D changes without an arrival or when the service phase changes without a service completion.

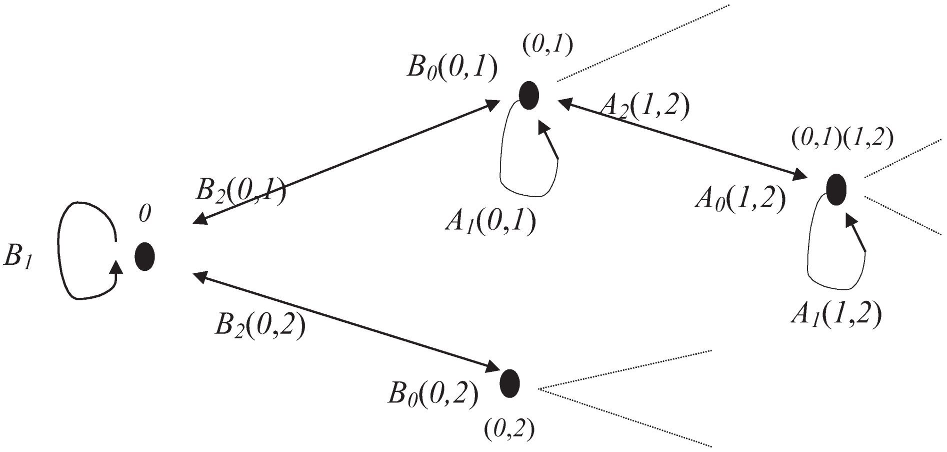

Figure 3. Transitions of the QBD Markov process.

It is then easy to see that $(q(t), I(t), I_1(t))$ is a Markov process with a state space

$$\{\{0\} \times \{1, 2, \ldots, m\}\} \cup \bigcup_{\mathrm{J}=\Omega\setminus\{0\}} \{\mathrm{J} \times \{1, 2, \ldots, m\} \times \{1, 2, \ldots, m_{k\_n}\}\}.$$

This is a QBD Markov process with a tree structure when $(I(t), I_1(t))$ is defined as the auxiliary random variable. In matrix form, the transition rates of the QBD Markov process are written as $\{A_0(k_1, k_2), A_1(k_1, k_2), A_2(k_1, k_2)\}$ and $\{B_1(0), B_0(0, k), B_2(0, k)\}$. This QBD Markov process has no transitions among sibling nodes. The transition of the QBD Markov process is illustrated in figure 3 for $K = 2$.

The transition rate matrices of the QBD Markov process $(q(t), I(t), I_1(t))$ are given as follows. Let $\boldsymbol{e}_k$ be an $m_k \times 1$ vector of one, $1 \leqslant k \leqslant K$. For $\mathrm{J} = (0, k_1)(k_1, k_2) \ldots (k_n, k) \neq 0$,

- $A_0(k_n, k) = D_k \otimes (\boldsymbol{e}_{k\_n}\boldsymbol{\alpha}_k)$ (a type k customer arrives),
- $A_1(k_n, k) = D_0 \otimes \boldsymbol{I} + \boldsymbol{I} \otimes T_k$ (no service completion and no arrival),
- $A_2(k_n, k) = \boldsymbol{I} \otimes (\boldsymbol{T}_k^0 \boldsymbol{\alpha}_{k\_n})$ (a service completion),
- $B_2(0, k) = \boldsymbol{I} \otimes \boldsymbol{T}_k^0$ (a service completion at node $(0, k)$).

For $\mathrm{J} = 0$, $B_1(0) = D_0$ and $B_0(0, k) = D_k \otimes \boldsymbol{\alpha}_k$, $1 \leqslant k \leqslant K$. $\otimes$ represents Kronecker product (see Gantmacher [2]). $A_0(k_n, k)$ is an $mm_{k\_n} \times mm_k$ matrix, $A_1(k_n, k)$ is an $mm_k \times mm_k$ matrix, $A_2(k_n, k)$ is an $mm_k \times mm_{k\_n}$ matrix, $B_0(0, k)$ is an $m \times mm_k$ matrix, B_1 is an $m \times m$ matrix, $B_2(k, 0)$ is an $mm_k \times m$ matrix.

Note. The construction of $A_2(k_n, k)$ shows the necessity to use the representation $\mathrm{J} = (0, k_1)(k_1, k_2) \ldots (k_{n-1}, k_n)$ in this case.

For QBD Markov process $(q(t), I(t), I_1(t))$, the number of states associated with each node can be different since the type of the customer in service can be different. Furthermore, for node J $= (k_1, k_2)(k_2, k_3) \ldots (k_{n-1}, k_n)$, it has K children: $\{\mathrm{J}(k_n, k),\ 1 \leqslant k \leqslant K\}$, which are associated with J through k_n. In the theory of tree structure stochastic processes, such a tree is called asymmetrical. These differentiate $(q(t), I(t), I_1(t))$ from the symmetrical QBD Markov processes studied in sections 3 and 4, HE [4], and Yeung and Alfa [20]. Because of the difference, the solution process is slightly different from that of sections 3 and 4. Therefore, some details about the stationary distribution of $(q(t), I(t), I_1(t))$ are given in the next two sections.

6. The stationary distribution of the repeat queue

In this section, an algorithm for computing the stationary distribution of $(q(t), I(t), I_1(t))$ is presented. Since explicit conditions have not been found for the existence of the stationary distribution, it is assumed in this section that the QBD Markov process $(q(t), I(t), I_1(t))$ can reach its steady state. The following results are parallel to that of theorem 4.2. For J $= (0, k_1)(k_1, k_2) \ldots (k_{n-1}, k_n) \in \Omega$, let

$$\pi(\mathrm{J}, i, i_1) = \lim_{t\to\infty} \mathbf{P}\bigl\{\bigl(q(t), I(t), I_1(t)\bigr) = (\mathrm{J}, i, i_1)\bigr\} \tag{6.1}$$

and

$$\begin{aligned}&\boldsymbol{\pi}(\mathrm{J}, i) = \bigl(\pi(\mathrm{J}, i, 1), \ldots, \pi(\mathrm{J}, i, m_{k\_n})\bigr),\\ &\boldsymbol{\pi}(\mathrm{J}) = \bigl(\boldsymbol{\pi}(\mathrm{J}, 1), \ldots, \boldsymbol{\pi}(\mathrm{J}, m)\bigr), \quad \mathrm{J} \neq 0, \qquad \boldsymbol{\pi}(0) = \bigl(\pi(0, 1), \ldots, \pi(0, m)\bigr).\end{aligned} \tag{6.2}$$

When the underlying Markov chain D of the arrival process and all the PH-distributions are irreducible, $(q(t), I(t), I_1(t))$ is irreducible. When the queueing system can reach its steady state, its stationary distribution vectors $\{\boldsymbol{\pi}(\mathrm{J})\colon \mathrm{J} \in \Omega\}$ satisfy the following equation:

$$\begin{aligned}\mathbf{0} &= \boldsymbol{\pi}(0)B_1 + \sum_{l=1}^{K} \boldsymbol{\pi}(0, l)B_2(0, l),\\ \mathbf{0} &= \boldsymbol{\pi}(0)B_0(0, k) + \boldsymbol{\pi}(0, k)A_1(0, k) + \sum_{l=1}^{K} \boldsymbol{\pi}\bigl((0, k) + l\bigr)A_2(k, l), \quad 1 \leqslant k \leqslant K,\\ \mathbf{0} &= \boldsymbol{\pi}(\mathrm{J})A_0(k_n, k) + \boldsymbol{\pi}(\mathrm{J} + k)A_1(k_n, k) + \sum_{l=1}^{K} \boldsymbol{\pi}(\mathrm{J} + k + l)A_2(k, l), \quad 1 \leqslant k_n, k \leqslant K.\end{aligned} \tag{6.3}$$

A matrix geometric solution is given as follows.

Theorem 6.1. When the queueing system can reach its steady state, the stationary distribution of $(q(t), I(t), I_1(t))$ is given by

$$
\begin{aligned}
&\boldsymbol{\pi}(\mathrm{J}+k) = \boldsymbol{\pi}(\mathrm{J})R(k_n,k), \quad \mathrm{J}=(0,k_1)(k_1,k_2)\ldots(k_{n-1},k_n)\in\Omega,\ \mathrm{J}\neq 0,\\
&\boldsymbol{\pi}(0)B_1 + \sum_{k=1}^{K}\boldsymbol{\pi}(0,k)B_2(0,k) = 0,\\
&\boldsymbol{\pi}(0)B_0(0,k) + \boldsymbol{\pi}(0,k)A_1(0,k) + \boldsymbol{\pi}(0,k)\sum_{l=1}^{K}R(k,l)A_2(k,l) = 0, \quad 1\leqslant k\leqslant K,\\
&\boldsymbol{\pi}(0)\boldsymbol{e} + \sum_{k=1}^{K}\boldsymbol{\pi}(0,k)\left\{\sum_{n=1}^{\infty}\ \sum_{(k_0,k_1)(k_1,k_2)\ldots(k_{n-1},k_n):\ k_0=k}\ \prod_{j=1}^{n}R(k_{j-1},k_j)\right\}\boldsymbol{e} = 1,
\end{aligned}
\tag{6.4}
$$

where $\{R(k,l),\ 1\leqslant k,l\leqslant K\}$ are the minimal non-negative solutions to matrix equations:

$$
0 = A_0(k,l) + R(k,l)A_1(k,l) + R(k,l)\sum_{s=1}^{K}R(l,s)A_2(l,s). \tag{6.5}
$$

Proof. See Yeung and Sengupta [19] or HE [4]. □

The computation of boundary probabilities $\{\boldsymbol{\pi}(0),\boldsymbol{\pi}(0,k),\ 1\leqslant k\leqslant K\}$ can be reduced to solving the following equation of $\boldsymbol{\pi}(0)$ only:

$$
\boldsymbol{\pi}(0)\left\{B_1 + \sum_{k=1}^{K}B_0(0,k)\left[A_1(0,k)+\sum_{l=1}^{K}R(k,l)A_2(k,l)\right]^{-1}B_2(0,k)\right\} = 0. \tag{6.6}
$$

The computation of $\{R(k,l),\ 1\leqslant k,l\leqslant K\}$ can be carried out using the following algorithm. Let $R(k,l)[0]=0$, $1\leqslant k,l\leqslant K$, and

$$
R(k,l)[n+1] = -\left\{A_0(k,l) + R(k,l)[n]\sum_{s=1}^{K}R(l,s)[n]A_2(l,s)\right\}\bigl(A_1(k,l)\bigr)^{-1}. \tag{6.7}
$$

It can be proved that $\{R(k,l)[n],\ n\geqslant 0\}$ is a monotone sequence which converges to $R(k,l)$ from below, for $1\leqslant k,l\leqslant K$.

Intuitively, when the QBD Markov process $(q(t),I(t),I_1(t))$ is ergodic, then the eigenvalue with the largest real part of $R(k,1)+R(k,2)+\cdots+R(k,K)$ is less than one for $1\leqslant k\leqslant K$. However, it is not easy to check such conditions for ergodicity.

An algorithm for computing the stationary distribution can be developed from equations (6.4)–(6.6). As is shown in equation (6.4), the normalization of the stationary distribution (or equivalently the vectors $\{\boldsymbol{\pi}(0),\boldsymbol{\pi}(0,k),\ 1\leqslant k\leqslant K\}$) becomes difficult. It is almost impossible to evaluate the infinite summation since the number of items to be added increases exponentially with respect to n (see equation (6.4)). To overcome this difficulty, a relationship between system idle probability and the mean lengths of a busy period and a busy cycle are used, i.e., $\boldsymbol{\pi}(0)\boldsymbol{e} = 1 -$ mean length of busy period / mean length of busy cycle. This leads to a detailed discussion of the busy period and busy cycle.

7. The busy period and busy cycle

The busy period and busy cycle of the queueing system of interest is closely related to the concept of fundamental period. This section uses the concept of fundamental period to find the joint distribution of the numbers of customers served in a busy period and the moment of the length of a busy period. Since the analysis of the fundamental period is parallel to that of the classical QBD Markov processes (see Neuts [12]), no proof shall be provided.

This section focusses on the preemptive repeat queue. There are several reasons for choosing the preemptive repeat queue to discuss issues associated with busy period and busy cycle, instead of the nonpreemptive and the preemptive resume cases. First, the busy period is equivalent to the fundamental period of the QBD Markov processes. Thus, it is routine to derive formulas for busy period and busy cycle related performance measures (see Takine et al. [18]) once the corresponding QBD Markov processes with a tree structure are setup. Second, for the nonpreemptive queue (studied by HE in [4]) and the preemptive resume queue, their busy periods are equivalent to that of the $MMAP[K]/G[K]/1/$FCFS queues. The busy periods of the FCFS queue have been studied extensively (see HE [3] and Takine and Hasegawa [17]). Thus, it is not so important to derive explicit formulas for the busy periods of the nonpreemptive case nor the preemptive resume case, even though such results are useful. Third, there are no known results for the busy periods and busy cycle of the preemptive repeat queue. Formulas derived using the QBD Markov process approach for this case are new and lead to an algorithm for computing performance measures related to busy periods and busy cycles. Lastly, the mean lengths of busy periods and busy cycles are needed for computing the stationary distribution of the preemptive repeat queue.

In general, a fundamental period is defined as the first passage time during which the (total) queue length decreases by one. Define $\boldsymbol{N} = \{\boldsymbol{n} = (n_1, \ldots, n_K),\ n_k \geqslant 0, 1 \leqslant k \leqslant K\}$. Similar to the classical QBD case (see Neuts [12]), define, for J in Ω, J $= (0, k_1)(k_1, k_2)\ldots(k_{n-1}, k_n)$, $1 \leqslant k \leqslant K$, $1 \leqslant i, i' \leqslant m$, $1 \leqslant j \leqslant m_k$, $1 \leqslant j' \leqslant m_{k\_n}$, and $\boldsymbol{n} = (n_1, \ldots, n_K)$ in $\boldsymbol{N}$,

$g_{(i,j)(i',j')'}(k_n, k, x, \boldsymbol{n})$: the taboo probability that the Markov process $(q(t), I(t), I_1(t))$ reaches node J for the first time in state (J, i', j') in less than x units of time and there are n_1 type 1, n_2 type 2, $\ldots$ and n_K type K customers served during this time, given that the Markov process started in (J $+ k, i, j$).

Let $G(k_n, k, x, \boldsymbol{n})$ be an $mm_k \times mm_{k\_n}$ matrix with elements $g_{(i,j)(i',j')}(k_n, k, x, \boldsymbol{n})$. Because of the special structure of the QBD Markov process, $G(k_n, k, x, \boldsymbol{n})$ does not depend on the node J, when J > 0. $G(k_n, k, x, \boldsymbol{n})$ is defined for a busy period for $0 \leqslant k_n \leqslant K$, $1 \leqslant k \leqslant K$, and $G(0, x, \boldsymbol{n})$ is defined for a busy cycle. Let $G^*(k_n, k, \omega, \boldsymbol{z})$ be the joint transform of $G(k_n, k, x, \boldsymbol{n})$, i.e.,

$$\begin{aligned} G^*(k_n, k, \omega, \boldsymbol{z}) &= \int_0^\infty \sum_{n=1}^\infty \exp\{-\omega x\}\, \mathrm{d}_x G(k_n, k, x, \boldsymbol{n}) \prod_{l=1}^K z_l^{n_l}, \\ &\omega > 0,\ 0 < z_l < 1, \end{aligned} \tag{7.1}$$

where $\boldsymbol{z} = (z_1, \ldots, z_K)$. Then it can be proved that $\{G^*(k_n, k, \omega, \boldsymbol{z}),\ 1 \leqslant k_n, k \leqslant K\}$ are the minimal non-negative solutions to the equations

$$\begin{aligned} G^*(0, \omega, \boldsymbol{z}) &= [\omega \boldsymbol{I} - B_1]^{-1} \sum_{k=1}^K B_0(0, k) G^*(0, k, \omega, \boldsymbol{z}), \\ G^*(0, k, \omega, \boldsymbol{z}) &= \big[\omega \boldsymbol{I} - A_1(0, k)\big]^{-1} \Bigg[z_k B_2(0, k) \\ &\quad + \sum_{l=1}^K A_0(k, l) G^*(k, l, \omega, \boldsymbol{z}) G^*(0, k, \omega, \boldsymbol{z}) \Bigg], \quad 1 \leqslant k \leqslant K, \\ G^*(k_n, k, \omega, \boldsymbol{z}) &= \big[\omega \boldsymbol{I} - A_1(k_n, k)\big]^{-1} \Bigg[z_k A_2(k_n, k) \\ &\quad + \sum_{l=1}^K A_0(k, l) G^*(k, l, \omega, \boldsymbol{z}) G^*(k_n, k, \omega, \boldsymbol{z}) \Bigg], \quad 1 \leqslant k_n, k \leqslant K. \end{aligned} \tag{7.2}$$

Define $G(k_n, k) = G^*(k_n, k, 0+, \mathbf{1}-)$, $0 \leqslant k_n \leqslant K$, $1 \leqslant k \leqslant K$ and $G(0) = G^*(0, 0+, \mathbf{1}-)$, where $\mathbf{1} = (1, 1, \ldots, 1)$. When the QBD Markov process is ergodic, matrices $G(k_n, k)$ and $G(0)$ are stochastic matrices. Matrices $\{G(k_n, k),\ 0 \leqslant k_n \leqslant K,\ 1 \leqslant k \leqslant K\}$ and $G(0)$ are the minimal non-negative solutions to the matrix equations

$$\begin{aligned} G(0) &= -(B_1)^{-1} \sum_{k=1}^K B_0(0, k) G(0, k), \\ G(0, k) &= -\left[A_1(0, k) + \sum_{l=1}^K A_0(k, l) G(k, l) \right]^{-1} B_2(0, k), \quad 1 \leqslant k \leqslant K, \\ 0 &= A_2(k_n, k) + A_1(k_n, k) G(k_n, k) \\ &\quad + \sum_{l=1}^K A_0(k, l) G(k, l) G(k_n, k), \quad 1 \leqslant k_n, k \leqslant K. \end{aligned} \tag{7.3}$$

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 equation (7.2). For instance, let

$$\boldsymbol{u}(k_n, k) = -\frac{\partial G^*(k_n, k, \omega, \boldsymbol{z})\boldsymbol{e}}{\partial \omega}\bigg|_{\omega=0+,\ \boldsymbol{z}=\boldsymbol{1}-}$$

and

$$\boldsymbol{v}(k_n, k, l) = \frac{\partial G^*(k_n, k, \omega, \boldsymbol{z})\boldsymbol{e}}{\partial z_l}\bigg|_{\omega=0+,\ \boldsymbol{z}=\boldsymbol{1}-}$$

for $0 \leqslant k_n, k, l \leqslant K$, where $\boldsymbol{u}(k_n, k)$ is the mean length of a busy period started in the node (k_n, k), conditioning on the initial state of the underlying Markov process D and the service phase $\boldsymbol{v}(k_n, k, l)$ is the mean number of type l customers served in a busy period started in the node (k_n, k), conditioning on the initial state of the underlying Markov process D and the service phase. Similar interpretations go to $\boldsymbol{u}(0)$ and $\boldsymbol{v}(0, l)$. Simple but lengthy calculations lead to the following expressions, for J = 0:

$$\begin{aligned}
\boldsymbol{u}(0) &= -(B_1)^{-1}\left[\boldsymbol{e} + \sum_{k=1}^{K} B_0(0, k)\boldsymbol{u}(0, k)\right], \\
\boldsymbol{v}(0, l) &= -(B_1)^{-1}\left[\sum_{k=1}^{K} B_0(0, k)\boldsymbol{v}(0, k, l)\right], \quad 1 \leqslant l \leqslant K.
\end{aligned} \tag{7.4}$$

For $1 \leqslant k, l \leqslant K$,

$$\boldsymbol{u}(0, k) = -\left[A_1(0, k) + \sum_{t=1}^{K} A_0(k, t)G(k, t)\right]^{-1}\left[\boldsymbol{e} + \sum_{t=1}^{K} A_0(k, t)\boldsymbol{u}(k, t)\right], \tag{7.5}$$

$$\begin{aligned}
\boldsymbol{v}(0, k, l) = &-\left[A_1(0, k) + \sum_{t=1}^{K} A_0(k, t)G(k, t)\right]^{-1} \\
&\times \left[I\{k = l\}A_2(0, k)\boldsymbol{e} + \sum_{t=1}^{K} A_0(k, t)\boldsymbol{v}(k, t, l)\right].
\end{aligned} \tag{7.6}$$

For $1 \leqslant k_n, k, l \leqslant K$,

$$\boldsymbol{0} = \boldsymbol{e} + \left[A_1(k_n, k) + \sum_{t=1}^{K} A_0(k, t)G(k, t)\right]\boldsymbol{u}(k_n, k) + \sum_{t=1}^{K} A_0(k, t)\boldsymbol{u}(k, t), \tag{7.7}$$

$$\mathbf{0} = I\{k = l\}A_2(k_n, k)\boldsymbol{e} + \left[A_1(k_n, k) + \sum_{t=1}^{K} A_0(k, t)G(k, t)\right]\boldsymbol{v}(k_n, k, l).$$

$$+ \sum_{t=1}^{K} A_0(k, t)\boldsymbol{v}(k, t, l), \tag{7.8}$$

where $I\{k = j\} = 1$ if $k = j$; otherwise $I\{k = j\} = 0$.

Performance measures of interest can be obtained by solving equations (7.4)–(7.8). Although the formulas look formidable, the actual programming is not difficult to implement. In fact, equations (7.7) and (7.8) can be solved as linear equations or using an iteration method. More details are given to the mean lengths of a busy period and a busy cycle since they are used in determining the stationary distribution.

Let $\boldsymbol{g}$ be the left invariant vector of matrix $G(0)$. Then $\boldsymbol{g}$ is the probability distribution of the underlying Markov process D at the end of an arbitrary busy period (or busy cycle). The mean length of an arbitrary busy cycle is thus given by $\boldsymbol{gu}(0)$ and the mean length of an arbitrary busy period is given by $\boldsymbol{gu}(0) + \boldsymbol{g}B_1^{-1}\boldsymbol{e}$. Then it has

$$\boldsymbol{\pi}(0)\boldsymbol{e} = 1 - \frac{\boldsymbol{gu}(0) + \boldsymbol{g}B_1^{-1}\boldsymbol{e}}{\boldsymbol{gu}(0)}. \tag{7.9}$$

Equation (7.9) plus equation (6.6) determines $\boldsymbol{\pi}(0)$ and so the stationary distribution of the QBD Markov process $(q(t), I(t), I_1(t))$.

In summary, an algorithm for computing the stationary distribution is given as follows.

Algorithm II.

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

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

Step 3: Compute $\{R(k_n, k),\ 1 \leqslant k_n, k \leqslant K\}$.

Step 4: Compute $\{G(k_n, k), \boldsymbol{u}(k_n, k),\ 1 \leqslant k_n, k \leqslant K\}$.

Step 5: Compute matrices $\{G(0, k), \boldsymbol{u}(0, k),\ 1 \leqslant k \leqslant K\}$ and $\{G(0), \boldsymbol{u}(0)\}$.

Step 6: Compute the mean length of a busy period and the mean length of a busy cycle.

Step 7: Compute vectors $\boldsymbol{\pi}(0)$ and $\{\boldsymbol{\pi}(0, k),\ 1 \leqslant k \leqslant K\}$.

Step 8: Compute string distribution $\{\boldsymbol{\pi}(\mathrm{J}),\ \mathrm{J} \in \Omega\}$.

8. Numerical examples

Using algorithms I and II given in sections 4 and 7, respectively, three numerical examples are presented in this section with brief discussions.

Table 1
Probabilities of queue strings for example 8.1.

(J, S)	(1, 1)	(1, 1)(1, 1)	(1, 1)(1, 1)(1, 1)	(2, 1)	(2, 1)(2, 1)	(2, 1)(2, 1)(2, 1)
$\boldsymbol{\pi}$(J, S)$\boldsymbol{e}$	0.1764	0.0500	0.01421	0.0269	0.0005	0.00001
(J, S)	–	–	–	(2, 2)	(2, 2)(2, 2)	(2, 2)(2, 2)(2, 2)
$\boldsymbol{\pi}$(J, S)$\boldsymbol{e}$	–	–	–	0.0461	0.0012	0.00003

Table 2
Probabilities of queue strings for example 8.1.

J	(0, 1)	(0, 1)(1, 1)	(0, 1)(1, 1)(1, 1)	(0, 2)	(0, 2)(2, 2)	(0, 2)(2, 2)(2, 2)
$\boldsymbol{\pi}$(J)$\boldsymbol{e}$	0.0981	0.0278	0.0079	0.0412	0.0020	0.00010

Example 8.1. Consider an $MMAP[2]/PH[2]/1/$LCFS queue with $m = 2$, $K = 2$,

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

$$m_1 = 1, \qquad \boldsymbol{\alpha}_1 = 1, \qquad T_1 = -2,$$

$$m_2 = 2, \qquad \boldsymbol{\alpha}_2 = (0.2, 0.8), \qquad T_2 = \begin{pmatrix} -3 & 0 \\ 1 & -3 \end{pmatrix}.$$

For the preemptive resume queue, $\boldsymbol{\pi}(0,0)\boldsymbol{e} = 0.5259$. Other probabilities of queue strings are presented in table 1.

For the preemptive repeat queue, $\boldsymbol{\pi}(0)\boldsymbol{e} = 0.2925$. Other probabilities of queue strings are presented in table 2.

Compared to the preemptive repeat queue, the preemptive resume queue has a much shorter queue length (on the average). This agrees with intuition. For both queueing systems, the probabilities of queue strings associated with type 1 customers are higher than that of type 2 customers. This is normal since the total number of type 1 customers served is twice that of type 2 customers.

Furthermore, the difference between system idle probabilities shows that the preemptive repeat queue is much busier than the preemptive resume queue. Two questions naturally follow this observation. First, is there an $MMAP[K]/PH[K]/1/$LCFS queue which is ergodic when the preemptive resume service discipline is applied and unstable when the preemptive repeat service discipline is applied? Second, is there an $MMAP[K]/PH[K]/1/$LCFS queue which is less busy when the preemptive repeat service discipline is applied? The answers to these questions are given by examples 8.2 and 8.3.

Example 8.2. Consider an $MMAP[2]/PH[2]/1/$LCFS queue with $m = 2$, $K = 2$,

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

$$m_1 = 2, \qquad \boldsymbol{\alpha}_1 = (0.2, 0.8), \qquad T_1 = \begin{pmatrix} -5 & 0 \\ 1 & -2 \end{pmatrix},$$

$$m_2 = 2, \qquad \boldsymbol{\alpha}_2 = (0.2, 0.8), \qquad T_2 = \begin{pmatrix} -3 & 0 \\ 1 & -3 \end{pmatrix}.$$

For the preemptive resume queue, $\boldsymbol{\pi}(0,0)\boldsymbol{e} = 0.0251$, which shows a high traffic intensity queueing system. The corresponding preemptive repeat queue is unstable.

Example 8.3. Consider an $MMAP[2]/PH[2]/1/\mathrm{LCFS}$ queue with $m = 2$, $K = 2$,

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

$$m_1 = 2, \qquad \boldsymbol{\alpha}_1 = (0.2, 0.8), \qquad T_1 = \begin{pmatrix} -1 & 0.5 \\ 1 & -4 \end{pmatrix},$$

$$m_2 = 2, \qquad \boldsymbol{\alpha}_2 = (0.3, 0.7), \qquad T_2 = \begin{pmatrix} -0.5 & 0.5 \\ 0.1 & -1 \end{pmatrix}.$$

For the preemptive resume case, $\boldsymbol{\pi}(0,0)\boldsymbol{e} = 0.042$, compared to $\boldsymbol{\pi}(0)\boldsymbol{e} = 0.1218$ corresponding to the preemptive repeat case. Thus, the resume case is busier than the repeat case. The reason (in this case) is that the service times have a special property which can be approximately stated as decreased failure rate (DFR) property. In fact, for the two PH-distributions, they start, with a higher probability, in phase 2 and the service has a large probability to be ended soon. However, if the (underlying) PH-process changes to phase 1, then the residual service time might be much longer than a new service time probabilistically. Thus, if a customer is being pushed out of service when its PH-distribution is in phase 1, its service time would be longer if the preemptive resume service discipline is applied when it returns to service. That is, this customer may prefer to restart its service process afresh rather than being reinstalled from where it was interrupted. In practice, when the service times have the DFR property, the preemptive repeat service discipline is than preferred over the preemptive resume one.

When looking at the levels of individual types of customers, table 3 shows that differences exist between the two types of customers. Consider the preemptive resume case. For type 1 customers, when they wait in the queue, the probabilities that their recorded phases are phase 1 are significantly larger than that of phase 2. However,

Table 3
Probabilities of queue strings for example 8.3 (resume case).

(J, S)	(1, 1)	(1, 1)(1, 1)	(1, 1)(1, 1)(1, 1)	(2, 1)	(2, 1)(2, 1)	(2, 1)(2, 1)(2, 1)
$\boldsymbol{\pi}$(J, S)$\boldsymbol{e}$	0.0166	0.0059	0.00200	0.0048	0.00047	0.000047
(J, S)	(1, 2)	(1, 2)(1, 2)	(1, 2)(1, 2)(1, 2)	(2, 2)	(2, 2)(2, 2)	(2, 2)(2, 2)(2, 2)
$\boldsymbol{\pi}$(J, S)$\boldsymbol{e}$	0.0093	0.0016	0.00027	0.0065	0.00080	0.000098

the situation is reversed for type 2 customers. Since the number of type 1 customers served is several times of that of type 2 customers, the recorded phases of the majority customers in queue are phase 1. Recall that being in phase 1 implies a longer residual service time than that of a new service time. It is easy to see that the preemptive resume queue is busier than the preemptive repeat queue.

9. Summary

This paper gives an analysis of the $MMAP[K]/PH[K]/1$ queue with a LCFS preemptive resume or repeat service discipline. The queueing processes of interest are formulated into QBD Markov processes with a tree structure. Algorithms are developed for computing the stationary distribution of queue strings, the mean length of a busy period, and the mean numbers of customers served in a busy period. An ergodic condition for the preemptive resume queue has been found.

Numerical results show that there is a big difference between queueing systems using the preemptive resume and preemptive repeat service disciplines. In many cases, the preemptive repeat queue would be busier than the preemptive resume queue. However, this is not generally true. For instance, when service times have the DFR property, the preemptive repeat queue can be less busy.

Several issues are pertinent to future research. First, a simple necessary and sufficient condition for the ergodicity of the $MMAP[K]/PH[K]/1$ queue with a LCFS preemption repeat service discipline will be useful. Second, queueing systems with a FCFS or priority service discipline are worth investigating. Obviously, these models generate a lot of interesting problems since it is difficult to construct an analytically tractable QBD Markov process for queueing systems with a FCFS or priority service discipline.

References

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

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

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

[4] Q.-M. HE, Quasi-birth-and-death processes with a tree structure and a detailed analysis of the $MMAP[K]/PH[K]/1$ queue (1997) submitted for publication.

[5] Q.-M. HE and M.F. Neuts, Markov arrival processes with marked transitions, Stochastic Process. Appl. 74(1) (1998) 37–52.

[6] F.P. Kelly, *Reversibility and Stochastic Networks* (Wiley, New York, 1979).

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

[8] R. Lyons, Random walks and percolation on trees, Ann. Probab. 18 (1990) 931–958.

[9] R. Lyons, Random walks, capacity and percolation on trees, Ann. Probab. 20 (1992) 2043–2088.

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

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

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

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

[14] M.F. Neuts, The joint distribution of arrivals and departures in quasi-birth-and-death processes, in: *Numerical Solutions of Markov Chains*, ed. W.J. Stewart (Marcel Decker, New York, 1991) pp. 147–159.

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

[16] H. Takagi, Queueing analysis of polling models: progress in 1990–1994, in: *Frontiers in Queueing*, ed. J.H. Dshalalow (1996) pp. 119–146.

[17] T. Takine 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 10 (1994) 183–204.

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

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

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