A *MAP/G/1* Queue With Cyclic Service Times

Qi-Ming HE[1], Hui Li[2], and Pemberton Cyrus[1]

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

[2] Department of Mathematics
Mount Saint Vincent University
Halifax, Nova Scotia, Canada B3M 2J6

(December, 1996, January, 1997)

AMS Classification 1991: 60K25

Abstract

This paper studies a queueing system with a Markov arrival process and cyclic service times. When customers are distinguished by the types of their service times, the queueing processes experienced by different types of customers are different. First, a computational approach is developed to find the distributions of the queue length and waiting time of each type of customer. The results are useful in comparing the queueing processes experienced by different types of customers. Second, since a number of servers is utilized cyclically, the issue of how to sequence these servers is addressed. Algorithms are proposed for computing the best and worst sequences of servers in terms of the mean queue length or the mean waiting time. Numerical examples are presented to show in which sequence (of servers) the mean queue length or the mean waiting time is minimized or maximized for queueing systems with bursty and nonbursty input processes.

Key words: Queueing theory, matrix analytic methods, Markov chain, integer programming, matrix theory.

[2] Partially supported by the Natural Sciences and Engineering Research Council of Canada through grant OGP0105828.

1. Introduction

This paper considers a queueing system with *K* servers and a Markov arrival process (*MAP*). Customers form a single queue. At any time only one of the *K* servers is utilized to serve customers while the other *K-1* servers await their turns in an idle server pool. After serving a customer, the server joins the idle server pool (at the end of the line), and one of the idle servers (at the head of the line) then becomes responsible for serving the next customer. Thus, the servers serve customers cyclically. This queueing model is denoted as $MAP/G^{[K]}/1$.

The $MAP/G^{[K]}/1$ queue is equivalent to a queueing system with a single server and *K* types of customers arriving cyclically. The service times of the *K* types of customers are different, but the service times of the same type of customers have a common distribution function. From this point of view, the $MAP/G^{[K]}/1$ queue is a special case of the $MMAP(K)/G(K)/1$ considered in HE [3], where the superposition arrival process of *K* types of customers is modeled as a Markov arrival process with marked transitions ($MMAP(K)$). In this paper, the queueing system of interest will be formulated both as a queueing system with one customer type and *K* servers and as a queueing system with *K* types of customers and a single server. The former is used to analyze the queue lengths at departure epochs and the later is used to study the actual waiting time processes.

The $MAP/G^{[K]}/1$ queue is a generalization of the *M/G/1* queue with cyclic service times studied in Iravani and Posner [4] (1996) where arrivals have a Poisson process. Iravani and Posner [4] focused on the queue length distribution and obtained the mean queue length. A literature review and some applications of such queueing models are provided in their paper as well. It is shown in [4] that the $MAP/G^{[K]}/1$ queue finds applications in manufacturing and telecommunications industries. Like [4], other early papers on cyclic queueing systems (Coffman and Gilbert [2], Morrice, Gajulapalli and Tayur [7], etc.) assumed a Poisson input process and considered the queue length and waiting time processes. Unlike [4], this paper considers a more general input process and brings some new issues into consideration.

The queueing model of interest is also a special case of the queueing systems with service times depending on the arrival state (see Takine and Hasegawa [11] (1994) and HE [3] (1996)) and the queueing systems with a semi-Markovian service process (see Neuts [10] (1989)). Many useful performance measures of such queueing systems have been discussed in the literature. Nonetheless, there are still some things to contribute. First, the special structure of the queueing system should be exploited in order to get more explicit results. Second, the queueing process of each type of customer needs to be studied. Third, sequencing of servers so as to minimize or maximize the mean queue length (the mean waiting time) should be investigated. These issues have not been addressed in the literature before.

This paper addresses the above issues from two directions. First, the queue length and waiting time processes are investigated. The focus is on the queue lengths (the waiting times) observed (experienced) by different types of customers served by different servers. By analyzing the difference between these queueing processes, more insights into such queueing systems are learned. In addition, some of the results about the queue length obtained in Iravani and Posner

[4] are generalized. Second, sequencing of servers is studied. It is easy to see that the queueing process and the waiting process have much to do with the sequence of servers because servers have different service times, but it is difficult to find the best and worst sequences of servers in which the mean queue length and the mean waiting time are minimized and maximized respectively. Algorithms are developed for computing performance measures for each sequence of servers for the $MAP/G^{[K]}/1$ queue. Thus, the best and worst sequences can be found by enumeration. However, the enumeration approach is inefficient. Therefore, a queueing system with a deterministic arrival process and constant service times ($D/D^{[K]}/1$) is analyzed in detail. The worst sequence in the $D/D^{[K]}/1$ queue is found, while the best sequence can be found by solving an integer programming problem. It is expected that the solutions of the $D/D^{[K]}/1$ queue are close to the best and worst sequences of its corresponding $MAP/G^{[K]}/1$ queue. It is worth mentioning that both the best and the worst sequences are discussed to see whether sequencing of servers makes a difference and, if it does, how big the difference could be.

The rest of the paper is organized as follows. In Section 2, the $MAP/G^{[K]}/1$ queue is defined explicitly, along with the Markov arrival process and the Markov arrival process with marked transitions. Sections 3 and 4 present results about the queue lengths and the waiting times respectively. In Section 5, attempts are made to solve optimization problems concerning the sequence of servers with respect to the mean queue length (the mean waiting time). Section 5 is focused on the $D/D^{[K]}/1$ queue. In Section 6, numerical results are presented for several special cases. A number of interesting conclusions about the queues observed by different types of customers and the sequences of servers are drawn from numerical experimentation. Finally, in Section 7, some discussion of the obtained results and future research is presented.

2. The $MAP/G^{[K]}/1$ Queue

The queueing system of interest is defined explicitly in this section. Two approaches are used to formulate the queueing system. First, customers are distinguished at the departure epochs depending on what kind of service they have received. This approach is convenient for analyzing the queue lengths at the departure epochs. Second, customers are distinguished at their arrival epochs. This approach is convenient for analyzing the actual waiting time processes.

In general, the $MAP/G^{[K]}/1$ queue has a Markov arrival process and K servers. The K servers serve customers cyclically (one server at a time). Customers are served on a FCFS basis.

The Markov arrival process (MAP) was introduced by Neuts (see Neuts[8]) as a generalization of a phase-type renewal process (see Neuts [9]). It is defined on a finite Markov process $J(t)$ (called the underlying Markov process) which has m states and an irreducible infinitesimal generator D. In the MAP, the sojourn time in state j is exponentially distributed with parameter $(-D_0)_{j,j}$ $(\geq -(D)_{j,j})$. At the end of the sojourn time in state j, there occurs a transition to another (possibly the same) state and that transition may or may not correspond to an arrival. Let D_0 be the (matrix) rate of transitions in the phase process that does not generate arrivals and D_1 be the rate of arrivals. The matrix D_0 has strictly negative diagonal elements and nonnegative off-diagonal elements, the matrix D_1 is a nonnegative matrix, and $D = D_0 + D_1$.

Let $\underline{\theta}$ be the stationary probability vector of the Markov process with generator D, i.e., $\underline{\theta}$ satisfies $\underline{\theta}D = 0$ and $\underline{\theta}\,\underline{e} = 1$, where $\underline{e}$ is a column vector of 1's. The stationary arrival rate is then given as $\lambda = \underline{\theta}D_1\underline{e}$. Note that, throughout this paper, vectors are underlined.

The service times of the K servers have distribution functions $F_k(x)$, and Laplace Stieltjes transform (LST) $f^*_k(x)$, $1 \le k \le K$. Since the K servers serve a single queue cyclically, the service time of a customer is determined by the server which serves the customer. Let $1/\mu_k$ be the mean service time of server k and denote by $\rho_k = \lambda/\mu_k$, $1 \le k \le K$. The traffic intensity of the queueing system is defined as

$$\rho = \frac{1}{K}\left(\sum_{k=1}^{K} \rho_k\right). \tag{2.1}$$

Let $X(t)$ be the queue length (including the one in service, if any) at time t; $I(t)$ the index of the server in use at time t; and $J(t)$ be the phase of the underlying Markov chain at time t. $X(t) \in \{0, 1, 2, ...\}$, $I(t) \in \{1, 2, ..., K\}$, and $J(t) \in \{1, 2, ..., m\}$. Let τ_n be the departure epoch of the nth customer, $n \ge 0$. It is easy to see that $(X(\tau_n +), I(\tau_n -), J(\tau_n +))$ is a Markov chain, which is called the embedded Markov chain of the process $(X(t), I(t), J(t))$ at departure epochs. To determine the transition matrix of the embedded Markov chain, the counting process of the Markov arrival process is defined. Let

$$p_{i,j}(n,t) = \mathbf{P}\{n \text{ arrivals in } (0,t),\ J(t) = j \mid J(0) = i\}, \quad 1 \le i, j \le m. \tag{2.2}$$

Let $P(n, t)$ be an $m \times m$ matrix with elements $p_{i,j}(n, t)$. Define the matrix generating function

$$P^*(z,t) = \sum_{n=0}^{\infty} P(n,t) z^n. \tag{2.3}$$

It has been proved that (see Lucantoni [5])

$$P^*(z,t) = \exp\{(D_0 + zD_1)t\}. \tag{2.4}$$

The transition matrix of the embedded Markov chain $(X(\tau_n +), I(\tau_n -), J(\tau_n +))$ can be written as

$$P = \begin{pmatrix} \hat{P}_0 & \hat{P}_1 & \hat{P}_2 & \hat{P}_3 & \cdots \\ P_0 & P_1 & P_2 & P_3 & \cdots \\ & P_0 & P_1 & P_2 & \ddots \\ & & P_0 & P_1 & \ddots \\ & & & \ddots & \ddots \end{pmatrix}, \tag{2.5}$$

where

$$\hat{P}_n = \begin{pmatrix} 0 & \hat{A}_{n,2} & & & \\ & 0 & \hat{A}_{n,3} & & \\ & & \ddots & \ddots & \\ & & & 0 & \hat{A}_{n,K} \\ \hat{A}_{n,1} & & & & 0 \end{pmatrix}, \quad P_n = \begin{pmatrix} 0 & A_{n,2} & & & \\ & 0 & A_{n,3} & & \\ & & \ddots & \ddots & \\ & & & 0 & A_{n,K} \\ A_{n,1} & & & & 0 \end{pmatrix}, \tag{2.6}$$

and

$$\begin{aligned} \hat{A}_{n,k} &= (-D_0^{-1}D_1)\int_0^\infty P(n,t)dF_k(t),\ 1 \le k \le K; \\ A_{n,k} &= \int_0^\infty P(n,t)dF_k(t),\ 1 \le k \le K. \end{aligned} \tag{2.7}$$

This is a Markov chain of *M/G/1* type, which has been studied extensively (see Neuts [10]).

Although some useful results can be obtained by analyzing the embedded Markov chain, it is inconvenient to study the waiting time processes. Thus, the following formulation of the queueing system, which distinguishes customers at arrival epochs, is established. Suppose there are *K* types of customers and one server. The *K* types of customers are distinguished by marking the arrivals of the *MAP* cyclically, i.e, *1, 2, ..., K, 1, 2, ...,* etc. All customers are served based on FCFS. The service time is determined by the type of the customers. Type *k* customers have a common service time distribution function $F_k(x)$, $1 \le k \le K$. As was discussed in Section 1, this definition of the system is equivalent to the previous one.

Consider a Markov arrival process with marked transitions (*MMAP*(*K*)) (see HE [3]) with a matrix representation $(\hat{D}_0, \hat{D}_1, \cdots, \hat{D}_K)$, where $\hat{D}_0$ is an $mK \times mK$ matrix with all diagonal blocks D_0 and, $\hat{D}_k$, $1 \le k \le K$, are $mK \times mK$ matrices. Partition $\hat{D}_k$ into $m \times m$ blocks. Then all the blocks of $\hat{D}_k$ are zero matrices execpt that the (*k, k+1*)st block is D_1, i.e.,

$$\hat{D}_k = \begin{pmatrix} 0 & & & \\ & \ddots & D_1 & \\ & & \ddots & \\ & & & 0 \end{pmatrix}, \quad 1 \le k \le K - 1, \ \hat{D}_K = \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & \ddots & \\ D_1 & & & 0 \end{pmatrix}. \tag{2.8}$$

The matrix $\hat{D}_k$ represents the arrival rate of the customers with a service time distribution $F_k(x)$, $1 \le k \le K$. Apparently, this marked *MAP* is characterized by the cyclic nature of its arrivals. The *MAP* ($\hat{D}_0$, $\overline{D}_1$), where $\overline{D}_1 = \hat{D}_1 + \hat{D}_2 + \cdots \hat{D}_K$, is the same as the *MAP* (D_0, D_1). The new underlying Markov process $\overline{D} = \hat{D}_0 + \overline{D}_1$ has a state space $\{(k, j), 1 \le k \le K, 1 \le j \le m\}$. The index "*k*" no longer represents the type of server in use, but the type of incoming customer (the next arrival). The stationary distribution of this underlying Markov process is $\underline{\hat{\theta}} = (\underline{\theta}, \cdots, \underline{\theta}) / K$.

With this setting, results obtained in HE [3] (or Takine and Hasegawa [11]) can be used to study the busy cycle, idle period, and waiting time processes of the cyclic queueing system of interest. Some results about the waiting time processes will be presented in Section 4.

3. The Queue Length

This section focuses on the queue length at departure epochs. The main issues are the computation of the probability that the system is empty at a departure epoch and, the queues observed by different types of customers as well as their differences.

Let, for $n \ge 0$, $1 \le k \le K$, and $1 \le j \le m$,

$$x_{n,k,j} = \lim_{l \to \infty} \mathbf{P}\{X(\tau_n +) = n,\ I(\tau_n -) = k,\ J(\tau_n +) = j \mid X(0),\ I(0),\ J(0)\},$$

$$\underline{x}_{n,k} = (x_{n,k,1}, ..., x_{n,k,m}), \quad \underline{x}_n = (\underline{x}_{n,1}, ..., \underline{x}_{n,K}), \quad \text{and } \underline{x} = (\underline{x}_0, \underline{x}_1, ...).$$

The vector $\underline{x}$ is the stationary distribution of the embedded Markov chain. Therefore, $\underline{x}$ satisfies $\underline{x}P = \underline{x}$ and $\underline{x}\underline{e} = \underline{1}$. Let $\underline{X}^*(z) = \sum_{n=0}^{\infty} \underline{x}_n z^n$, $0 \le z \le 1$. Following routine procedures, it can be shown that

$$\underline{X}^*(z)(zI - P^*(z)) = \underline{x}_0(z\hat{P}^*(z) - P^*(z)), \tag{3.1}$$

where

$$P^*(z) = \sum_{n=0}^{\infty} P_n z^n = \begin{pmatrix} 0 & P_2^*(z) & & & \\ & 0 & P_3^*(z) & & \\ & & \ddots & \ddots & \\ & & & 0 & P_K^*(z) \\ P_1^*(z) & & & & 0 \end{pmatrix}, \tag{3.2}$$

$$\hat{P}^*(z) = \sum_{n=0}^{\infty} \hat{P}_n z^n = \begin{pmatrix} 0 & \hat{P}_2^*(z) & & & \\ & 0 & \hat{P}_3^*(z) & & \\ & & \ddots & \ddots & \\ & & & 0 & \hat{P}_K^*(z) \\ \hat{P}_1^*(z) & & & & 0 \end{pmatrix}, \tag{3.3}$$

and, for $1 \le k \le K$,

$$P_k^*(z) = \int_0^{\infty} dF_k(t) \exp\{(D_0 + zD_1)x)\}, \tag{3.4}$$

$$\hat{P}_k^*(z) = (-D_0^{-1} D_1) \int_0^{\infty} dF_k(t) \exp\{(D_0 + zD_1)x) = (-D_0^{-1} D_1) P_k^*(z)\,.$$

After some routine simplifications, the following basic equation can be established.

$$\underline{X}(z)(zI - P^*(z)) = \underline{x}_0(z \ \mathrm{diag}(-D_0^{-1}D_1, \ \ldots, \ -D_0^{-1}D_1) - I)P^*(z)\,. \tag{3.5}$$

Using existing results and algorithms (see Neuts [10] and Lucantoni [6]), the distribution of the queue length and its moments can be found. For completeness, formulas used in computation are presented in Appendix D for the mean queue length. Next, by exploiting the special structure of the embedded Markov chain P, the computation of the vector $\underline{x}_0$ is made easier and an analysis of the queues observed by different types of customers is conducted.

Let G be the minimal nonnegative solution to the equation

$$G = \sum_{n=0}^{\infty} P_n G^n\,. \tag{3.5}$$

The $((k, i), (k', j))$th element of the matrix G is the probability that the Markov chain P enters level q-1 in the state $(q$-$1, k', j)$, given that the Markov chain was in state (q, k, i) initially, $1 \le k, k' \le K$, $1 \le i, j \le m$. Let $\Lambda = (-\hat{D}_0^{-1}\overline{D}_1)G$. According to Neuts [10], matrix Λ is a stochastic matrix and $\underline{x}_0 \Lambda = \underline{x}_0$. To determine the vector $\underline{x}_0$, the following result is useful.

Lemma 3.1. $\lim_{z \to 1} \mathit{diag}(D(z), \ \ldots, \ D(z))[zI - P^*(z)]^{-1}\underline{e} = [\lambda / (1 - \hat{\rho})]\underline{e}$, where $D(z) = D_0 + zD_1$.

Proof. The proof is based on the special structure of the matrix $P^*(z)$. See Appendix B.

Using Lemma 3.1, the following result is obtained immediately.

Theorem 3.2 $\underline{x}_0(-\hat{D}_0^{-1})\underline{e} = \frac{1-\rho}{\lambda}$, where $\hat{D}_0$ is defined in Section 2.

Proof. The result is obtained by using Lemma 3.1, equation (3.4), and $\underline{X}(1)\underline{e} = 1$.

With Theorem 3.2, an algorithm for computing vector $\underline{x}_0$ can be developed. Once $\underline{x}_0$ is found, some other performance measures can be obtained accordingly (see Neuts [10]).

Remark: Let $\underline{y}_0$ be the probability vector that the queue is empty at an arbitrary time. It can be proved that $\underline{y}_0 = \lambda \underline{x}_0(-\hat{D}_0^{-1})$. Then $\underline{y}_0$ $\underline{e} = 1$- ρ, i.e., the probability that the queueing system is empty at an arbitrary time is 1-ρ. This result is obtained in Iravani and Posner [4] for cyclic queueing systems with a Poisson arrival process.

Remark: Theorem 3.2 was obtained in Lucantoni [5] for a standard *MAP/G/1* queue in which there is only one type of customer (or service).

Next, a result of the queues observed by different types of customers is presented. Define $\underline{X}^*(k,z) = \sum_{n=0}^{\infty} \underline{x}_{n,k} z^n,\ 1 \le k \le K$ and $\underline{X}^*(z) = (\underline{X}^*(1,z),\ \ldots,\ \underline{X}^*(K,z))$. It is easy to prove, and intuitively it is true, that $\underline{X}(k,1)\underline{e} = 1/K$. Therefore, $\underline{X}^*(k, z)K$ is the generation function of the queue length right after the departure of an arbitrary type k customer.

Theorem 3.3. The difference between the mean queue lengths left behind by two types of customers is given by, for $1 \le k \le K$,

$$\begin{aligned} K[\underline{X}'(k,1)\underline{e} - \underline{X}'(k-1,1)\underline{e}] = \rho_k - 1 + K\underline{x}_{0,k-1}\underline{e} + K[\underline{x}_{0,k-1}(-D_0^{-1}D_1) + \underline{X}(k-1,1)] \\ \cdot [P^{*\prime}(k-1,1) - I](D + \underline{e}\underline{\theta})^{-1} D_1 \underline{e}. \end{aligned} \tag{3.6}$$

Proof. See Appendix B. The corresponding constant vectors $\{\underline{X}^*(k, 1), 1 \le k \le K\}$ are also given in Appendix B.

For cases with a Poisson input process, equation (3.6) is simplified to

$$K[X^{*\prime}(k,1) - X^{*\prime}(k-1,1)] = Kx_{0,k-1} + \rho_k - 1\ . \tag{3.7}$$

Equations (3.6) and (3.7) show that the computation of the differences between mean queue lengths is simpler than that of the mean queue lengths themselves (see Appendices B and D).

Remark: Higher moments of the queue length will not be discussed further since little more detail has been obtained than those for the Markov chains of *M/G/1* type presented in Neuts [10]. Therefore, readers are referred to [10] for more detail (see Appendix D for a formula for the mean queue length).

4. Waiting times

When the queueing system of interest is interpreted as a queueing system having *K* types of customers (that arrive at the system cyclically), results in HE [3] (also see Takine and Hasegawa [11]) can be applied to study the (actual) waiting time processes of different types of customers. In this section, the LST of the waiting time of each type of customers is derived. The difference between the mean waiting times is derived as well. In addition, necessary results for developing algorithms are given.

Let $W^*_{a,k,j}(s)$ be the LST of the waiting time of an arbitrary customer, given that the state of the underlying Markov process $\hat{D}$ right after the arrival of the customer is (*k*, *j*). According to definition, the state of the underlying Markov process is in one of the states of $\{(k, j), 1 \le j \le m\}$ right after the arrival of a type *k-1* customer. Since the index "*k*" represents the type of incoming customer, $W^*_{a,k,j}(s)$ is the LST of the waiting time of an arbitrary type *k-1* customer. (Note that $W^*_{a,1,j}(s)$ is for an arbitrary type *K* customer.) Let $\underline{W}^*_{a,k}(s) = (W^*_{a,k,1}(s), \ldots, W^*_{a,k,m}(s))$ and $\underline{W}^*_a(s) = (\underline{W}^*_{a,1}(s), \ldots, \underline{W}^*_{a,K}(s))$. It has been obtained in HE [3] that

$$\underline{W}^*_a(s) = -s\underline{\pi}_0\hat{D}_0^{-1}[sI + \hat{D}_0 + \sum_{k=1}^{K} f_k^*(s)\hat{D}_k]^{-1}\overline{D}_1, \tag{4.1}$$

where $\underline{\pi}_0$ is the (vector) probability that the queueing system is empty at a departure epoch, $\underline{\pi}_0 = (\underline{\pi}_{0,1}, \ldots, \underline{\pi}_{0,K})$, and $\underline{\pi}_{0,k} = (\pi_{0,k,j}, \ldots, \pi_{0,k,m})$. Notice that $\underline{\pi}_0$ and $\underline{x}_0$ are different. For $\underline{x}_0$, its index "*k*" represents the type of the server just used (not the current server). For $\underline{\pi}_0$, its index "*k*" represents the type of the incoming customer when an arbitrary customer (its type is unknown) completes its service. Algorithms for computing $\underline{\pi}_0$ and $\underline{x}_0$ are different as well.

Using the special structure of $\overline{D}_1$, $\hat{D}_1$, $\hat{D}_2$, ..., $\hat{D}_K$ defined in (2.8), $\underline{W}^*_a(s)$ can be simplified as follows:

$$[sI + \hat{D}_0 + \sum_{k=1}^{K} f_k^*(s)\hat{D}_k]^{-1} = \begin{pmatrix} sI + D_0 & f_1^*(s)D_1 & & \\ & \ddots & \ddots & \\ & & \ddots & f_{K-1}^*(s)D_1 \\ f_K^*(s)D_1 & & & sI + D_0 \end{pmatrix}^{-1}$$

$$= \left(I - \begin{pmatrix} 0 & D^*(s) f_1^*(s) & & \\ & \ddots & \ddots & \\ & & \ddots & D^*(s) f_{K-1}^*(s) \\ D^*(s) f_K^*(s) & & & 0 \end{pmatrix} \right)^{-1} diag[(sI + D_0)^{-1}, \cdots, (sI + D_0)^{-1}]$$

$$= (\sum_{i=0}^{K-1} [J(s)]^i)\, diag[\,(I - \Delta)^{-1}, \ldots, (I - \Delta)^{-1}]\, diag((sI + D_0)^{-1}, \cdots, (sI + D_0)^{-1})$$

$$= (\sum_{i=0}^{K-1} [J(s)]^i)\, diag[\,(I - \Delta)^{-1}\, (sI + D_0)^{-1}, \ldots, (I - \Delta)^{-1}\, (sI + D_0)^{-1}], \qquad (4.2)$$

where $D^*(s) = -(sI + D_0)^{-1} D_1$ and

$$J(s) = \begin{pmatrix} 0 & D^*(s) f_1^*(s) & & \\ & \ddots & \ddots & \\ & & \ddots & D^*(s) f_{K-1}^*(s) \\ D^*(s) f_{K-1}^*(s) & & & 0 \end{pmatrix}, \qquad (4.3)$$

and

$$\Delta = \prod_{k=1}^{K} f_k^*(s)\ [-(sI + D_0)^{-1} D_1]^K . \qquad (4.4)$$

Then it follows

$$s \underline{\pi}_0 \hat{D}_0^{-1}\, (sI + \hat{D}_0 + \sum_{k=0}^{K} f_k^*(s) \hat{D}_k)^{-1}$$
$$= s\, \underline{\pi}_0 \hat{D}_0^{-1} \sum_{k=0}^{K-1} [J(s)]^k\ [(I - \Delta)^{-1}\, (sI + D_0)^{-1}, \ldots, (I - \Delta)^{-1}\, (sI + D_0)^{-1}]_{.} \qquad (4.5)$$

This leads to, for $1 \le k \le K$,

$$\underline{W}_{a,k}^*(s) = -s \sum_{j=k,k+1,\ldots,j-1} \underline{\pi}_{0,j} D_0^{-1} \prod_{i=k,\ldots j} f_i^*(s)\ [D^*(s)]^{j-i+1}\, (I - \Delta)^{-1} (sI + D_0)^{-1} D_1 . \qquad (4.6)$$

Note that the LST of the waiting time of type *k-1* customers is given by $KW^*_{a,k}(s)$. For brevity, the moments of the waiting times are not discussed. Only a formula for the mean waiting time is presented in Appendix D for use in Section 6. Next, the difference between the mean waiting times is given.

Theorem 4.1. The difference between the mean waiting times of an arbitrary type *k* and an arbitrary type *k-1* customer is given as, for $1 \le k \le K$,

$$K[(-\underline{W}^{*}_{a,k+1}{}'(0)) - (-\underline{W}^{*}_{a,k}{}'(0))]\underline{e} \;=\; K\underline{\pi}_{0,k}(-D_0^{-1})\underline{e} - \frac{1-\rho_{k-1}}{\lambda}\,. \tag{4.7}$$

Proof. See appendix C.

Similar to the mean queue length case, the computation of the difference between mean waiting times is easier. Numerical examples are given in Section 6 to show when the difference is significant and how large the difference could be.

The computation of the mean waiting time (see Appendix D) and equation (4.7) depends on the vector $\underline{\pi}_0$, which can be found using results obtained in HE [3]. Denote by $\overline{G}$ the minimal nonnegative solution to the equation

$$\overline{G} = \int_0^\infty diag\,\big(dF_1(x)I,\ dF_3(x)I,\ \ldots,\ dF_K(x)I\big)\exp\{(T_0 + T_1\overline{G})x\}, \tag{4.8}$$

where T_0 and T_1 are $mK^2 \times mK^2$ matrices defined as

$$T_0 = \begin{pmatrix} \hat{D}_0 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \hat{D}_0 \end{pmatrix} \quad \text{and} \quad T_1 = \begin{pmatrix} \hat{D}_1 & \hat{D}_2 & \cdots & \hat{D}_K \\ \hat{D}_1 & \hat{D}_2 & \cdots & \hat{D}_K \\ \vdots & \vdots & \cdots & \vdots \\ \hat{D}_1 & \hat{D}_2 & \cdots & \hat{D}_K \end{pmatrix}.$$

Let $\gamma > \max_i\{-(D_0)_{ii}\}$. Using $\exp\{(T_0 + T_1\overline{G})x\} = e^{-\gamma x}\exp\{[I + (1/\gamma)(T_0 + T_1\overline{G})]\gamma x\}$,

$$\overline{G} = \sum_{n=0}^{\infty} diag[\int_0^\infty \frac{e^{-\gamma x}x^n}{n!}dF_1(x)I,\ \ldots\ \int_0^\infty \frac{e^{-\gamma x}x^n}{n!}dF_K(x)I][I + \frac{1}{\gamma}(T_0 + T_1\overline{G})]^n. \tag{4.9}$$

Solve the following equation for an mK^2 - dimension vector $\tilde{\underline{\pi}}_0 = (\tilde{\underline{\pi}}_{0,1},\ \cdots,\ \tilde{\underline{\pi}}_{0,K})$,

$$\tilde{\underline{\pi}}_0(-T_0^{-1}T_1)\overline{G} = \tilde{\underline{\pi}}_0 \quad \text{and} \quad \tilde{\underline{\pi}}_0(-T_0^{-1})\underline{e} = \frac{1-\rho}{\lambda}. \tag{4.10}$$

Then $\underline{\pi}_0$ is obtained as $\underline{\pi}_0 = \tilde{\underline{\pi}}_{0,1} + \cdots + \tilde{\underline{\pi}}_{0,K}$. With equations (4.9) and (4.10), an efficient algorithm can be developed for computing $\underline{\pi}_0$ (see Lucantoni [5]) as well as an algorithm for the mean waiting times.

$$\overline{G} = \int_0^\infty diag\left(dF_1(x)I,\ dF_3(x)I,\ \ldots,\ dF_K(x)I\right)\exp\{(T_0 + T_1\overline{G})x\}, \qquad (4.8)$$

where T_0 and T_1 are $mK^2 \times mK^2$ matrices defined as

$$T_0 = \begin{pmatrix} \hat{D}_0 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \hat{D}_0 \end{pmatrix} \quad \text{and} \quad T_1 = \begin{pmatrix} \hat{D}_1 & \hat{D}_2 & \cdots & \hat{D}_K \\ \hat{D}_1 & \hat{D}_2 & \cdots & \hat{D}_K \\ \vdots & \vdots & \cdots & \vdots \\ \hat{D}_1 & \hat{D}_2 & \cdots & \hat{D}_K \end{pmatrix}.$$

Let $\gamma > \max_i\{-(D_0)_{ii}\}$. Using $\exp\{(T_0 + T_1\overline{G})x\} = e^{-\gamma x}\exp\{[I + (1/\gamma)(T_0 + T_1\overline{G})]\gamma x\}$,

$$\overline{G} = \sum_{n=0}^{\infty} diag[\int_0^\infty \frac{e^{-\gamma x}x^n}{n!}dF_1(x)I,\ \ldots\ \int_0^\infty \frac{e^{-\gamma x}x^n}{n!}dF_K(x)I][I + \frac{1}{\gamma}(T_0 + T_1\overline{G})]^n. \qquad (4.9)$$

Solve the following equation for an mK^2 - dimension vector $\underline{\tilde{\pi}}_0 = (\underline{\tilde{\pi}}_{0,1},\ \cdots,\ \underline{\tilde{\pi}}_{0,K})$,

$$\underline{\tilde{\pi}}_0(-T_0^{-1}T_1)\overline{G} = \underline{\tilde{\pi}}_0 \quad \text{and} \quad \underline{\tilde{\pi}}_0(-T_0^{-1})\underline{e} = \frac{1-\rho}{\lambda}. \qquad (4.10)$$

Then $\underline{\pi}_0$ is obtained as $\underline{\pi}_0 = \underline{\tilde{\pi}}_{0,1} + \cdots + \underline{\tilde{\pi}}_{0,K}$. With equations (4.9) and (4.10), an efficient algorithm can be developed for computing $\underline{\pi}_0$ (see Lucantoni [5]) as well as an algorithm for the mean waiting times.

5. Sequencing of servers and system optimization

In this section, an attempt is made to determine the best and worst sequences of the servers in which the mean waiting time and the mean queue length are minimized and maximized, respectively.

It seems difficult to obtain an explicit relationship between the service sequence and the mean waiting time or the mean queue length because of the complexity of the queueing system of interest. The results obtained in Sections 3 and 4 and Appendix D can be used for computing the mean waiting time for each sequence of servers. The search for the best and worst squences can be done by enumerating all the possible sequences. However, this approach is time consuming and expensive. In order to get some insights into such problems, a $D/D^{[K]}/1$ cyclic queue is constructed and studied. The interarrival time of the $D/D^{[K]}/1$ queue is the mean interarrival time $\{1/\lambda\}$ of the original $MAP/G^{[K]}/1$ cyclic queue. The service times of the $D/D^{[K]}/1$ queue are the mean service times $\{1/\mu_k,\ 1 \le k \le K\}$ of the original $MAP/G^{[K]}/1$ queue. Although the $MAP/G^{[K]}/1$

queue and its corresponding $D/D^{[K]}/1$ queue are dramatically different, they do share some common features. It is expected that their best and worst sequences are the same or close to each other. If their best and worst sequences are close, the solutions of the $D/D^{[K]}/1$ queue can be used as approximations for the $MAP/G^{[K]}/1$ queue. An advantage of this approach is that the solutions of the $D/D^{[K]}/1$ queue are much easier to find. Therefore, this section focuses on the $D/D^{[K]}/1$ queue. Algorithms shall be developed for computing the best and worst sequences for this special case. Numerical results will be presented in Section 6.

Because of Little's law on the relationship between the mean queue length and the mean waiting time, discussions in terms of the mean queue length and the mean waiting time are equivalent. Thus, only the mean waiting time will be discussed.

The $D/D^{[K]}/1$ cyclic queue Let v_k be the service time of server k, $k=1, 2, ..., K$, u the interarrival time between two consecutive customers, and W_n the waiting time of the nth customer, $n\geq 0$. It follows

$$W_0 = 0, \qquad W_n = (W_{n-1} + v_n - u)^+, \quad n \geq 1, \tag{5.1}$$

where $v_{nK+k} = v_k, \quad k = 1,\ 2,\ldots,\ K,\ \ n \geq 1$. To ensure the waiting times are finite, it is assumed that $v_1 + v_2 + \cdots + v_K < Ku$, which is equivalent to saying that the traffic intensity ρ is less one. Let $y_k = v_k - u$, $k=1, 2, ..., K$. Equation (5.1) becomes:

$$W_n = \max(0,\ y_n,\ y_n + y_{n-1},\ y_n + y_{n-1} + y_{n-2},\ \ldots,\ y_n + y_{n-1} + \cdots + y_1), \tag{5.2}$$

where $y_{nK+k} = y_k$, for $k=1, 2, ..., K$, and $n\geq 0$.

The condition $y_1 + y_2 + \cdots + y_K < 0$ and the periodicity of the sequence $\{y_n\}$ imply that there is a positive integer N, such that $W_N = 0$. The value of waiting time W_n from N onward, becomes periodic with a period of K, i.e., $W_{nK+k+N} = W_{k+N}$. Hence, for the optimal solution, it is sufficient to consider a single period. For convenience, let $N=0$. $\{W_1, W_2, ..., W_K,\}$ are used to denote the waiting times in a cycle with $W_K = 0$ and, are given by equation (5.2). Let

$$W = W_1 + W_2 + \cdots + W_K \tag{5.3}$$

Then W/K corresponds to the long term average waiting time. The issue of interest is to find the sequences which minimize and maximize W respectively. First, the following proposition gives the worst squence of servers.

Proposition 5.1 When the servers are sqeuenced such that $y_1 \geq y_2 \geq ... \geq y_K$, then the corresponding total waiting time of one cycle, W, is the largest among all the squences.

Proof: From the decreasing nature of $\{y_k,\ k = 1, 2, ..., K\}$, for the stated sequence of servers, equation (5.2) becomes

$$W_k = \max(0, y_1 + y_2 + \cdots y_k)\,,\ k = 1,2,\ldots\ldots, K\,. \tag{5.4}$$

Let W_1*, W_2*, ..., and W_K*, where $W_K^* = 0$, be the waiting times of one cycle for an arbitrary sequence of servers. Then

$$W_0^* = 0\,,\quad W_k^* = \max\{0,\ W_{k-1}^* + x_k\},\ k = 1,2,\ldots\ldots, K\,, \tag{5.5}$$

where (x_1, x_2, ..., x_K) is a permutation of (y_1, y_2, ..., y_K). Denoted by $W^* = W_1^* + W_2^* + \ldots + W_K^*$. It needs to be shown that $W^* \leq W$.

Since $W_0^* = 0 = W_K^*$, the waiting time sequence W_k^*, $k = 1,\ldots,K$ can be broken into P groups, $1 \leq P \leq K$, such that only the last element of each group is zero. Clearly, a group of size one only contains a zero element. For the value of W^*, it is sufficient to consider only those groups of size two or larger. To be more specific, assume that the Lth group contains $W_{n_L+1}^*,\ W_{n_L+2}^*,\ \ldots,\ W_{n_L+k_L}^*$, where $W_{n_L+k_L}^* = W_{n_L}^* = 0$ and $W_{n_L+k}^* > 0$, $k=1,\ 2,\ \ldots,\ k_L - 1$, $2 \leq k_L \leq K$, $L = 1, 2, \ldots, P$ and $2 \leq k_1 + k_2 + \ldots + k_P \leq K$. It follows from equation (5.5)

$$0 < W_{n_L+k}^* = x_{n_L+1} + \cdots + x_{n_L+k}\ ,\ k = 1,\ldots,k_L - 1, \tag{5.6}$$

which implies

$$\sum_{k=1}^{k_L-1} W_{n_L+k}^* \leq \sum_{k=1}^{k_L-1}\Bigl(W_{n_L+k}^* + \sum_{j=1}^{L-1} W_{n_j+k_j-1}^*\Bigr) = \sum_{k=1}^{k_L-1}\Bigl(\sum_{l=1}^{k} x_{n_L+l} + \sum_{j=1}^{L-1}\sum_{l=1}^{k_j-1} x_{n_j+l}\Bigr). \tag{5.7}$$

By its decreasing nature, the sum of any n elements of $\{y_1,\ y_2,\ \ldots,\ y_K,\ 1 \leq n \leq K\}$ is less or equal to the sum of its first n elements. Hence when $n = k_1 + k_2 + \ldots + k_{L-1} - (L-1) + k$, from equations (5.7) and (5.4),

$$\sum_{k=1}^{k_L-1} W_{n_L+k}^* \leq \sum_{k=1}^{k_L-1} y_1 + y_2 + \ldots + y_{k_1+k_2+\ldots+k_{L-1}-(L-1)+k} = \sum_{k=1}^{k_L-1} W_{k_1+\ldots+k_{L-1}-(L-1)+k}\ . \tag{5.8}$$

Therefore,

$$\begin{aligned} W^* &= \sum_{k=1}^{k_1-1} W_{n_1+k}^* + \sum_{k=1}^{k_2-1} W_{n_2+k}^* + \ldots + \sum_{k=1}^{k_P-1} W_{n_P+k}^*\ . \\ &\leq \sum_{k=1}^{k_1-1} W_k + \sum_{k=1}^{k_2-1} W_{k_1-1+k} + \ldots + \sum_{k=1}^{k_p-1} W_{k_1+\ldots+k_{p-1}-(p-1)+k} \leq W \end{aligned} \tag{5.9}$$

This completes the proof.

Notice: The solution of the worst sequence may not be unique.

While the worst sequence is simple and easy to find, the best sequence is usually complicated. In fact, the search for the best sequence is an NP hard problem except for a few special cases (see Chen [1]). Thus, no polynomial algorithm can be developed for computing the best sequence. Nonetheless, when the number of servers to be sequenced is not large (less than *30*), the search for the best sequence can be done by formulating the problem as an integer program and solving it using existing algorithms and software (such as CPLEX or LINDO). The integer programming problem can be formulated as follows.

$$
\begin{array}{lll}
\min \displaystyle\sum_{k=1}^{K} \omega_k & & \\
s.t. & & \\
(s.1) \quad \displaystyle\sum_{k=1}^{K} x_{k,i} = 1, \quad \sum_{k=1}^{K} x_{i,k} = 1, & \forall \;\; i \in \{1,2,\ldots,K\}; & (5.10) \\
(s.2) \quad \omega_k + y_i - \omega_i \le M(1 - x_{k,i}), & \forall \;\; k,i \in \{1,2,\ldots,K\}; & \\
\qquad\quad x_{k,i} \in \{0,1\}, \;\; \omega_k \ge 0, & \forall \;\; k,i \in \{1,2,\ldots,K\}; &
\end{array}
$$

where M is a constant which can be chosen as M = max{0, y_1}+...+max{0, y_K}+max$_{\{1\le k\le K\}}$ {max{0, y_k}}. In this formulation, constraints (*s.1*) construct subtours; constraints (*s.2*) along with the objective calculate the max{0, y_i + ω_k} when i follows k in the sequence. (See Lawler, Lenstra, Rinnooy Kan, and Shmoys [6] for more details about such integer programming formulations.)

The solution to (5.10) consists of a set of subtours, each subtour representing a cyclic sequencing of a subset of servers. In the solution, $x_{i,k} = 1$ means that y_k follows y_i in a cycle. The cycles can therefore be easily constructed from the {$x_{i,k}$, $1\le i$, $k\le K$} by starting with any $x_{i,k}$,= 1, and finding the succeeding members of the sequence by choosing j such that $x_{k,j}$,= 1, etc.

Sequencing of servers becomes even more complicated when randomness is considered. Based on the results of the $D/D^{[K]}/1$ queue, it seems unrealistic to obtain any analytical results of the best sequence for the $MAP/G^{[K]}/1$ queue immediately. Therefore, further discussion of the $MAP/G^{[K]}/1$ queue will continue in Section 6 with numerical examples.

Remark: The worst sequence of servers is of interest because, when compared to the best sequence, it provides information about whether or not one should pay special attention to the issue of server sequencing. For example, when the difference between the mean waiting times associated with the best and worst sequences is insignificant, sequencing of servers is less important. On the other hand, when the difference is significant, some effort must be made to avoid the worst sequence or nearly worst sequences so as to improve system efficiency.

6. Numerical Examples

In this section, two examples are presented with some detailed discussions. The queue lengths (waiting times) observed by different types of customers of some *MAP/G*$^{[K]}$*/1* queues are studied and compared. Special attention goes to the comparison of the queueing systems with the best and worst sequences as well as their deterministic counterparts. Some conclusions are drawn with regard to when sequencing of servers is important in terms of the mean waiting time.

Example 6.1 <u>The *M/M*$^{[K]}$*/1* cyclic queue</u> In this queueing system, customers arrive according to a Poisson process with parameter λ. The service times of the K servers are assumed to be exponential with service rates $\mu_1, \ldots, \mu_K$ respectively. Let $q(t)$ be the queue length and $I(t)$ be the type of the service at time t. It is clear that $(q(t), I(t))$ is a quasi birth-and-death process with an infinitesimal generator

$$Q = \begin{pmatrix} A_{00} & A_0 & & & \\ A_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots \end{pmatrix} \quad \text{where} \quad A_2 = \begin{pmatrix} 0 & \mu_1 & & & \\ & 0 & \mu_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \mu_{K-1} \\ \mu_K & & & & 0 \end{pmatrix}, \qquad (6.1)$$

$A_{00} = -\lambda \mathrm{I}$, $A_0 = \lambda \mathrm{I}$, and $A_1 = -\lambda \mathrm{I} - diag(\mu_1, \mu_2, \ldots, \mu_K)$. Let $x_{q.k} = \lim_{t\to\infty} P\{q(t) = q, I(t) = k\}$, $\underline{x}_q = (x_{q,1}, \ldots, x_{q,K})$ and $\underline{x} = (\underline{x}_0, \underline{x}_1, \ldots)$, which satisfies $\underline{x}Q = 0$ and $\underline{x}\underline{e} = 1$. Let R be the minimal nonnegative solution of the matrix equation

$$A_0 + RA_1 + R^2 A_2 = 0 \qquad (6.2)$$

By Neuts [9], the stationary distribution $\underline{x}$ is given as $\underline{x}_n = \underline{x}_1 R^{n-1}$, for $n \geq 1$, where $\underline{x}_0$ and $\underline{x}_1$ are the unique solution to the following equation:

$$\begin{aligned} -\lambda \underline{x}_0 + \underline{x}_1 A_2 &= 0; \\ \lambda \underline{x}_0 + \underline{x}_1 (A_1 + RA_2) &= 0. \end{aligned} \qquad (6.3)$$

Expanding $\underline{x}Q = 0$ in terms of $\{\underline{x}_n, n \geq 1\}$ and adding all equations together:

$$\underline{y} A_2 = \underline{y} \ \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_K), \qquad (6.4)$$

where $\underline{y} = (y_1, \cdots, y_K) = \sum_{q=1}^{\infty} x_q = \sum_{q=1}^{\infty} x_1 R^{n-1} = x_1 (I - R)^{-1}$. Solving equation (6.4) for $\underline{y}$ yields

$$y_k = \hat{\rho} \frac{1/\mu_k}{\sum_{j=1}^{K} 1/\mu_j}, \quad \text{where} \quad \hat{\rho} = \frac{\lambda}{K} \left(\frac{1}{\mu_1} + \frac{1}{\mu_2} + \ldots + \frac{1}{\mu_K} \right). \qquad (6.5)$$

Let $\hat{x}_k = \sum_{n=0}^{\infty} x_{n,k}$ and $L_k = \sum_{n=0}^{\infty} n x_{n,k}$, $1 \le k \le K$. Then

$$(\hat{x}_1, \cdots, \hat{x}_K) = \underline{x}_0 + \underline{x}_1 (I - R)^{-1} \quad \text{and,}$$
$$(L_1, \cdots, L_K) = \sum_{n=1}^{\infty} n \underline{x}_n = \sum_{n=1}^{\infty} n \underline{x}_1 R^{n-1} = \underline{x}_1 (I - R)^{-2} = \underline{y}(I - R)^{-1}. \qquad (6.6)$$

The mean queue length at an arbitrary departure epoch (and the arrival epoch) of a type k customer is given by $L_k \,/\, \hat{x}_k$. The mean queue length at an arbitrary departure epoch (an arrival epoch or an arbitrary time) is given by $L = L_1 + ... + L_K$ because of the PASTA property.

Since this system is represented by a quasi birth-and-death process, the solutions are much simpler. As a result, a simple algorithm is developed for computing the mean queue length and the mean waiting time, based on equations from (6.2) to (6.6). Briefly, the following steps are involved:

1. Input system parameters: $\{K, \lambda, \mu_1, ..., \mu_K\}$ and construct the transition blocks in the infinitesimal generator (see equation (6.2)).
2. Compute the matrix R using equation (6.2).
3. Solve equation (6.3) for $\underline{x}_0$ and $\underline{x}_1$, which are useful to find the difference between mean queue lengths of different types of customers.
4. Use equation (6.6) to find the mean queue length. The mean waiting time W is obtained by using Little's law: $L = \rho + \lambda W$.

Table 6.1 gives the difference between the mean queue lengths of different types of customers: $L_k \,/\, \hat{x}_k - L_{k-1} \,/\, \hat{x}_{k-1}$. Notice that $L_0 = L_K$ and $\hat{x}_0 = \hat{x}_K$ Each row of Table 6.1 shows an example. The last column in Table 6.1 gives the (overall) mean queue length.

Table 6.1 The difference between mean queue lengths

K	λ (ρ)	Service rate ($\mu_1, \mu_2, ..., \mu_K$)	$L_k / \hat{x}_k - L_{k-1} / \hat{x}_{k-1}$ k=1, 2, ..., K	L
3	6 (0.97)	(4 6 14)	0.76 0.99 -1.75	45.05
3	5 (0.77)	(6 4 20)	0.88 0.82 -1.7	3.91
3	4 (0.71)	(83 39 2)	-3.8 0.03 3.77	4.15
4	0.3 (0.81)	(0.25 0.45 0.22 10)	4.12 -0.3 0.99 -4.81	5.20
4	0.04 (0.99)	(0.024 0.14 0.022 0.2)	4.34 -3.2 3.42 -4.56	173.4
4	0.1 (0.95)	(0.06 0.3 0.07 0.25)	2.15 -2.3 2.03 -1.87	26.54
5	0.04 (0.96)	(0.033 0.05 0.03 0.028 1)	11.23 -0.28 0.61 0.69 -12.2	29.43
5	0.5 (0.89)	(0.4 0.8 0.6 0.3 5)	3.85 -0.62 0.12 1.32 -4.67	9.93
5	1 (0.69)	(2 0.9 8 0.6 20)	1.38 1.02 -0.98 2.51 -2.94	3.02

It is clear that the difference between the mean queue lengths is significant, especially when the service times of different types of customers are dramatically different. For example, the last row in Table 1 shows that the mean queue length after a customer with a service time *0.05* (service rate *20*) is much shorter than the mean queue length after other types of service. This further implies that the average waiting time of the customers served by other servers is shorter than the average waiting time of the customers served by the fastest server, an interesting observation which will be discussed further in Example 6.2.

Table 6.2 presents the best and worst sequences of servers of the $M/M^{[K]}/1$ queues and their corresponding $D/D^{[K]}/1$ queues. Every two consecutive rows in Table 6.2 show the results of an example. The first row shows the results of the $M/M^{[K]}/1$ queue and the second row shows the results of its corresponding $D/D^{[K]}/1$ queue. Notice that $1/\lambda$ and ($1/\mu_1$,, $1/\mu_2$,, ..., $1/\mu_K$) are the interarrival time and the service times of the $D/D^{[K]}/1$ queue respectively. L_{min} and L_{max} are the mean queue lengths corresponding to the best and worst sequences respectively. For the $M/M^{[K]}/1$ queue, L_{min} and L_{max} are obtained using equation (6.6). For the $D/D^{[K]}/1$ queues, L_{min} and L_{max} are obtained using equation (5.3).

Table 6.2 The best and worst sequences, and their mean queue lengths

K	λ	The best sequence		The worst sequence	
	(ρ)	$(\mu_1, \mu_2, ..., \mu_K)$	L_{min}	$(\mu_1, \mu_2, ..., \mu_K)$	L_{max}
3	6	(4 6 14)	45.05	(4 6 14)	45.06
	(0.97)	(4 6 14)	1.132	(4 6 14)	1.300
3	5	(6 4 20)	3.914	(4 6 20)	3.921
	(0.77)	(6 4 20)	0.85	(4 6 20)	0.88
3	4	(83 39 2)	4.149	(39 83 2)	4.158
	(0.71)	(83 39 2)	1.054	(39 83 2)	1.078
4	0.3	(0.25 0.45 0.22 10)	5.204	(0.22 0.25 0.45 10)	5.245
	(0.81)	(0.25 0.45 0.22 10)	0.951	(0.22 0.25 0.45 10)	1.100
4	0.04	(0.024 0.14 0.022 0.2)	173.4	(0.022 0.024 0.14 0.2)	173.5
	(0.99)	(0.024 0.14 0.022 0.2)	1.358	(0.022 0.024 0.14 0.2)	1.750
4	0.1	(0.06 0.3 0.07 0.25)	26.54	(0.06 0.07 0.25 0.3)	26.59
	(0.95)	(0.06 0.3 0.07 0.25)	1.225	(0.06 0.07 0.25 0.3)	1.499
5	0.04	(0.033 0.05 0.03 0.028 1)	29.43	(0.028 0.03 0.033 0.05 1)	29.48
	(0.96)	(0.033 0.05 0.03 0.028 1)	1.229	(0.028 0.03 0.033 0.05 1)	1.550
5	0.5	(0.4 0.8 0.6 0.3 5)	9.937	**(0.4 0.3 0.6 0.8 5)**	10.00
	(0.89)	(0.4 0.8 0.6 0.3 5)	1.073	(0.3 0.4 0.6 0.8 5)	1.432
5	1	(2 0.9 8 0.6 20)	2.916	(0.6 0.9 2 8 20)	3.019
	(0.69)	(2 0.9 8 0.6 20)	0.845	(0.6 0.9 2 8 20)	1.034

The results in Table 6.2 and other examples conducted show that, in general, the difference between the mean waiting times of the best and worst sequences for the stochastic models ($M/M^{[K]}/1$) are relatively small. Thus, sequencing of servers is not an important factor. The reason is that the input process is a Poisson process where the variation coefficient is *1*, which is not small. Such a variation in the input process reduces the influence of the sequence of servers on the queueing process. On the other hand, the difference between the mean waiting times could be considerably large for the deterministic models. Thus, sequencing of servers is more important when the service times and interarrival times are deterministic.

Second, the mean waiting times of the stochastic queueing systems ($M/M^{[K]}/1$) are dramatically larger than the mean waiting times of their corresponding deterministic ones. Apparently, variations in the input process and the service process play an important role. A conclusion drawn from Table 6.2 is that the higher the variation is, the longer the waiting time would be.

Third, it is shown that the best and worst sequences of the $M/M^{[K]}/1$ queue and it corresponding $D/D^{[K]}/1$ queue are the same except for one case (see the fourth line from the bottom of Table 6.2). In fact, results of an extensive numerical experimentation demonstrate that the solutions are either the same or close. Thus, instead of working on the $M/M^{[K]}/1$ queue, it is good enough to find the best (worst) sequence of its corresponding deterministic model, then use the solutions of the deterministic model as approximations of the $M/M^{[K]}/1$ queue. This approach does not guarantee an optimal solution for the $M/M^{[K]}/1$ queue but it is efficient. Besides, Table 6.2 shows that this approach works well most of the time.

Example 6.2 Consider a $MAP/D^{[K]}/1$ queue in which the service times are constants $\{d_1, d_2, ..., d_K\}$. This example gets special attention since it has potential applications in telecommunication industry, especially ATM systems. The focus of this example is on the influence of the input process on waiting time processes and the sequencing of servers. Bursty, nonbursty, and moderate input processes are considered together. A brief description of the algorithm used in computation is given as follows.

1. Input system parameters: $\{K, \mu_1, ..., \mu_K,, m, D_0, D_1\}$ and compute related constants (see Section 4 for details).
2. Compute the matrix $\overline{G}$ using equation (4.9).
3. Compute the vector π_0 using equation (4.10).
4. Calculate the difference between mean waiting times using equation (4.7).
5. Calculate the mean waiting time using equations (D.9) and (D.11).

For this system, computation of the matrix G becomes simpler since $f_k{}^*(s)=\exp\{-sd_k\}$ and

$$a_{n,k} = \frac{1}{n!}(\gamma d_k)^n \exp\{-\gamma d_k\}, \quad 1 \le k \le K,\ n \ge 0. \qquad (6.7)$$

Tables 6.3 and 6.4 present results for several examples with deterministic service times. Two sets of deterministic service times are considered: Set I = $\{0.25, 0.167, 0.07\}$ and Set II = $\{1, 0.7, 0.6, 0.05\}$. Four input processes are used: a bursty process (BURST), a Poisson process (Poisson), an Erlang process (Erlang(3)), and a deterministic process (D). The input processes are chosen to investigate the influence of the input process on the sequence of servers in terms of the mean waiting time.

In Table 6.3, the difference between the waiting times of different types (*k* and *k-1*) of customers, $K[(-W^*_{k+1}(0))\underline{e}-(-W^*_k(0))\underline{e}]$, is presented. Parameter λ is the arrival rate of the Markov arrival process. The last column of Table 6.3 shows the overall mean waiting time.

Table 6.3 The difference between mean waiting times

λ, ρ, and	Input process	The differences	
Service times		$k = 1, 2, ..., K$	$\mathbf{E}W_a$
$\lambda = 6$, $\rho = 0.95$ {*0.167, 0.25, 0.07*}	BURST	*-0.004 -0.086 0.09*	*8.394*
	Poisson	*-0.0035 -0.0865 0.09*	*2.987*
	Erlang(3)	*-0.0027 -0.0843 0.087*	*0.985*
	D	*-0.081 0. 0.081*	*0.027*
$\lambda = 0.5$, $\rho = 0.29$ {*0.7, 0.6, 1, 0.05*}	BURST	*-0.097 -0.003 -0.235 0.336*	*0.343*
	Poisson	*-0.070 0.003 -0.145 0.217*	*0.148*
	Erlang(3)	*-0.017 0.006 -0.050 0.062*	*0.023*
	D	*0 0 0 0*	*0*

Like Table 6.1, Table 6.3 shows that the queues experienced by difference types of customers are different. In general, the waiting time of a customer is small if its predecessor's service time is small or if its own service time is much longer than others. For example, for Set I and Set II, customers with the longest service time *0.25* get the smallest mean waiting time for all four input processes. On the other hand, customers with the shortest service time receive the longest mean waiting time. Although these observation may not always be true, most numerical examples show that customers with the shortest service time usually have the longest (or close to the longest) mean waiting time. The reason is that customers with the shortest service time contribute the least to the waiting times of other types of customers, while all other types of customers contribute more to their waiting times. Since customers arrive cyclically, the opportunity for each type of customer to contribute to the waiting time process is equal. Therefore, customers with the shortest service time actually experience a longer waiting time. Another observation is that the type of customer following the type of customer with the longest service time experiences a longer waiting time as well. (Notice that the observation does not apply to the sojourn times of customers.)

The above observations can be used to argue for setting up express checkouts in supermarkets. Usually, the numbers of items customers at checkout can be dramatically different. If customers with fewer items join the queues with other customers, the above observation shows that their waiting times are, on average, longer than other customers. Thus, by setting up express checkouts, not only can the flow of customers in the supermarket be improved, but also customers with fewer items are treated more fairly.

Next in Table 6.4, the results of the best and worst sequences are presented for series of queueing systems with the same deterministic service times. The main objective is to show how a queueing system performs when its input process is bursty or nonbursty.

Table 6.4 The best and worst sequences, and their mean waiting times

	Input process	Service times		$\mathbf{E}W_a$
$\lambda = 6$ $\rho = 0.95$ Set I	BURST	Best	*0.25 0.07 0.167*	*8.394*
		Worst	*0.25 0.167 0.07*	*8.396*
	Poisson	Best	*0.25 0.07 0.167*	*2.979*
		Worst	*0.25 0.167 0.07*	*2.982*
	Erlang(3)	Best	*0.25 0.07 0.167*	*0.985*
		Worst	*0.25 0.167 0.07*	*0.992*
	D	Best	*0.25 0.07 0.167*	*0.027*
		Worst	*0.25 0.167 0.07*	*0.055*
$\lambda = 0.5$ $\rho = 0.29$ Set II	BURST	Best	*0.7 0.6 1 0.05*	*0.343*
		Worst	*1 0.7 0.6 0.05*	*0.359*
	Poisson	Best	*0.7 0.6 1 0.05*	*0.148*
		Worst	***0.7 1 0.6 0.05***	*0.156*
	Erlang(3)	Best	*0.7 0.6 1 0.05*	*0.024*
		Worst	*1 0.7 0.6 0.05*	*0.026*
	D	Best	*any sequence*	*0*
		Worst	*any sequence*	*0*

Two trends are easy to see: 1) the more bursty the input process, the smaller the difference between the mean waiting times associated with the best and worst sequences of servers; 2) the more bursty the input process, the larger the mean waiting time of any particular sequence of servers. In fact, numerical experimentation shows that 1) and 2) are true whenever the variation in the input process or the service process increases. Also, Tables 6.3 and 6.4 show that sequencing of servers becomes more important when the service times of different servers are dramatically different. Based on these observations, it can be concluded that sequencing of servers makes little sense when the service times of servers are not dramatically different and the uncertainty in the system is high. In such cases, management has to seek other approaches to decrease the mean waiting time, for instance, reducing uncertainty in the input or service process.

Similar to Table 6.2, Table 6.4 shows that the best and worst sequences are almost the same for four different types of input processes. For the exceptions, numerical results (not presented) show that their resulting mean waiting times are close. Therefore, for any cyclic queueing system, it makes sense to use the best and worst seqeunces of its corresponding $D/D^{[K]}/1$ queue as approximations to its own. By doing so, heavy calculation is avoided.

Numerical examples also show that the sequencing of servers has something to do with the traffic intensity of the system. When the traffic intensity ρ is small, sequencing of servers makes a relatively larger difference in mean waiting times. However, the influence of the traffic intensity is not as strong as the uncertainty in the system nor the difference between servers.

7. Summary and future research

In this paper, a queueing system with cyclic service times is studied. Performance measures such as the mean queue lengths and the mean waiting times are analyzed and sequencing of servers is discussed.

By exploiting the special structure of an embedded Markov chain, an equality for the empty probability vector x_0 is obtained, which is useful in developing algorithms for computing x_0. An expression of the difference between the mean queue lengths (the mean waiting times) of customers of different types is obtained as well. These results are useful in analyzing the queueing processes of different types of customers. Numerical examples show that the difference between the queueing processes of different types of customers is significant in general and it becomes smaller when the variation in the input process or the service process becomes larger. Numerical results also show that customers with the shortest (average) service time have a longer (average) waiting time.

For sequencing of servers for a $D/D^{[K]}/1$ queue, the worst sequence (which maximizes the mean waiting time) is obtained explicitly. Finding the best sequence (which minimizes the mean waiting time) turns out to be an NP hard problem. Consequently, the problem of searching for the best sequence is transformed into an integer programming problem so that existing algorithms and software can be used. For a $MAP/G^{[K]}/1$ queue, the search for its best (worst) sequence is difficult. Then an associated $D/D^{[K]}/1$ queue is constructed. Numerical examples suggest the use of the best (worst) sequence of the $D/D^{[K]}/1$ queue as approximations to the best (worst) sequence of the original $MAP/G^{[K]}/1$ queue.

Two more results are worth mentioning. First, a $MAP/G^{[K]}/1$ queue usually has a longer queue (waiting time) than its corresponding $D/D^{[K]}/1$ queue. Thus, reduction in variation may help to increase system efficiency. Second, sequencing of servers becomes more important when variation in the input process or the service process decreases.

Although this paper answered some questions associated with queues with cyclic servers, it also raises a lot of questions to be explored. Based on our experience, three suggestions are made to conclude this paper. First, further exploiting the special structure in order to obtain more explicit results about the queue length and waiting time distributions will be useful. Second, an efficient algorithm should be developed to search for the best sequence for $MAP/G^{[K]}/1$ queues. At the least, more work should be done to improve the algorithm for computing the best sequence of a $D/D^{[K]}/1$ queue. Lastly, stochastic comparison associated with sequences of servers will be useful since the search for the best sequence is a difficult problem.

Reference

[1] Chen, B., On the NP hardness of a sequencing problem, *private communication*, 1996.

[2] Coffman, E.G. and Gilbert, E.N., Service by a queue and a cart, *Management Science*, **38** (1992), 867-876.

[3] HE, Qi-Ming, Queues with marked customers, *Adv. Appl. Prob*. **28** (1996), 567-587.

[4] Iravani, S.M.R. and Posner, M.J.M., An *M/G/1* queue with cyclic service times, Queueing Systems, **22** (1996), 145-169.

[5] Locantoni, D.M., New results on the single server queue with a batch Markovian arrival process, *Stochastic Models*, **7** (1991), 1-46.

[6] Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G., and Shmoys, D.B., eds., *The Travelling Salesman Problem*, John Wiley & Sons, Chichester, 1985.

[7] Morrice, D.J., Gajulapalli, R.S., and Tayur, S.R., A single server queue with cyclically indexed arrival and service rates, *Queueing Systems*, **15** (1994), 165-198.

[8] Neuts, M.F., A versatile Markovian point process, *J. Appl. Prob*., **16** (1979), 764-779.

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

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

[11] Takine, T. and Hasegawa, T., The worklaod 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.

Appendixes

Appendix A Proof of Lemma 3.1.

First, the special structure of the matrix $P^*(z)$ leads to $(P^*(z)/z)^K = diag(\Delta_1, \Delta_K, \ldots, \Delta_2)$, where $\Delta_k = P_k^* P_{k-1}^* \cdot\cdot P_1^* P_K^* P_{K-1}^* \cdot\cdot P_{k+1}^*, .1 \le k \le K$. This leads to

$$\begin{aligned}(zI - P^*(z))^{-1} &= (1/z)\sum_{n=0}^{\infty} (P^*(z)/z)^n \\ &= \frac{1}{z}[(I-\Delta_1)^{-1}, (I-\Delta_K)^{-1}, \ldots, (I-\Delta_2)^{-1}] \sum_{n=0}^{K-1} (P^*(z)/z)^n \end{aligned} \qquad \text{(A.1)}$$

since

$$(P^*(z)/z)^{Kn+l} = \mathrm{diag}(\Delta_1^n, \Delta_K^n, \ldots, \Delta_2^n) \;\; (P^*(z)/z)^l . \qquad \text{(A.2)}$$

Since $D = D_0 + D_1$ is irreducible, $D(z) = D_0 + zD_1$ is irreducible as well. Let $\chi(z)$ be the eigenvalue of $D(z)$ with the largest real part, then $\chi(z)$ is real. Let $\underline{\theta}(z)$ and $\underline{e}(z)$ be the left and right eigenvectors corresponding to $\chi(z)$, i.e. $\underline{\theta}(z)D(z)=\chi(z)\underline{\theta}(z)$ and $D(z)\underline{e}(z)=\chi(z)\underline{e}(z)$. $\underline{\theta}(z)$ and $\underline{e}(z)$ are normalied by $\underline{\theta}(z)\underline{e}=1$ and $\underline{\theta}(z)\underline{e}(z)=1$. Also, $D(z)$ have the following decomposition

$$Q^{-1}(z)D(z)Q(z)=\begin{pmatrix}\chi(z) & 0\\ 0 & T(z)\end{pmatrix}, \tag{A.3}$$

where $Q(z)$ is an $m\times m$ matrix and $T(z)$ is an $(m\text{-}1)\times(m\text{-}1)$ matrix. The first row of $Q^{-1}(z)$ is $\underline{\theta}(z)$ and the first column of $Q(z)$ is $\underline{e}(z)$. Substituting equation (A.3) into equation (3.4), it yields

$$P_k^*(z)=Q(z)\begin{pmatrix}\int_0^\infty \exp\{\chi(z)x\}dF_k(x) & \underline{0}\\ \underline{0} & \hat{T}_k(z)\end{pmatrix}Q^{-1}(z), \tag{A.4}$$

where $\hat{T}_k(z)=\int_0^\infty \exp\{T(z)x\}dF_k(x)$. Since the first column of the matrix $Q(1)$ is $\underline{e}$ and the first row of matrix $Q^{-1}(1)$ is $\underline{\theta}$, it can be proved that (see Neuts [9]):

$$\chi'(1)=\lambda=\underline{\theta}D_1\underline{e}. \tag{A.5}$$

Also, it can be proved that

$$\begin{aligned}(I-\Delta_i)^{-1}&=z^KQ(z)\left[z^KI-\begin{pmatrix}\prod_{j=1}^{K} f_j^*(\chi(z)) & \underline{0}\\ \underline{0} & \prod_{j=i}^{i-1}\hat{T}_j(z)\end{pmatrix}\right]^{-1}Q^{-1}(z)\\ &\equiv z^KQ(z)[z^KI-R(z)]^{-1}Q^{-1}(z),\end{aligned} \tag{A.6}$$

where

$$[z^KI-R(z)]^{-1}=\begin{pmatrix}[z^K-\prod_{k=1}^{K} f_k^*(\chi(z))]^{-1} & \underline{0}\\ \underline{0} & [z^KI-\prod_{j=k,k+1,\ldots,k-1}\hat{T}_j(z)]^{-1}\end{pmatrix}. \tag{A.7}$$

Since $D(z)$ is irreducible, the eigenvalue $\chi(z)$ is single and all other eigenvalues have negative real parts for $z\le 1$. Therefore, $\lim_{z\to 1}[z^KI-\prod_{j=k,k+1,\ldots,k-1}\hat{T}_j(z)]^{-1}$ exists and is finite. Since $\lim_{z\to 1}[z^K-\prod_{k=1}^{K} f_k^*(\chi(z))]=0$ and $\lim_{z\to 1}\chi(z)=0$, applying the L'Hospital's rule, it has

$$\lim_{z\to 1}\frac{\chi(z)}{[z^K-\prod_{k=1}^{K} f_k^*(\chi(z))]}=\lim_{z\to 1}\frac{\chi'(z)}{[Kz^{K-1}-\prod_{k=1}^{K} f_k^{*}{}'(\chi(z))]} \qquad \text{(A.8)}$$
$$=\frac{\lambda}{K(1-\rho)},$$

since

$$\lim_{z\to 1} f_k^{*}{}'(\chi(z))=\lim_{z\to 1}\chi'(z)\int_0^{\infty}\exp\{x\chi(z)\}xdF_k(x)=\lambda\mu_k=\rho_k .$$

It follows from (A.5), (A.7), and (A.8) that

$$\lim_{z\to 1} D(z)(I-\Delta_k)^{-1}=Q(1)\begin{pmatrix}\dfrac{\lambda}{K(1-\rho)} & \underline{0}\\ \underline{0} & A_k\end{pmatrix}Q^{-1}(1), \qquad \text{(A.9)}$$

where $A_k=\lim_{z\to 1} T(z)[z^K I-\prod_{j=k,k+1,\ldots,k-1}\hat{T}_j(z)]^{-1}$. Since $\underline{e}$ is the first column of $Q(1)$, it has $Q^{-1}(1)\underline{e}$ = $(1, 0, \ldots, 0)^{\mathrm{T}}$ (where "T" means matrix transpose). Postmultiplying $\underline{e}$ on both sides of (A.9) yields

$$\lim_{z\to 1} D(z)(I-\Delta_i)^{-1}\underline{e}=\frac{\lambda}{K(1-\rho)}\underline{e},\quad 1\le i\le K . \qquad \text{(A.10)}$$

Also from $P^*(1)\underline{e}=\underline{e}$, it has $\sum_{n=0}^{K-1}(P^*(1))^n\underline{e}=K\underline{e}$. Combining equations (A.1), (A.10), and the following equations gives the expected result.

$$\lim_{z\to 1} diag(D(z),\ \ldots,\ D(z))(zI-P^*(z))^{-1}\underline{e}$$
$$=\lim_{z\to 1} diag(D(z)(I-\Delta_1)^{-1},\ \ldots,\ D(z)(I-\Delta_2)^{-1})\sum_{n=0}^{K-1}(P^*(z))^n\underline{e}. \qquad \text{(A.11)}$$

This completes the proof.

Appendix B. Proof of Theorem 3.3

Expanding equation (3.4) yields

$$z\underline{X}(1,z)=\{\underline{X}(K,z)+\underline{x}_{0,K}[z(-D_0^{-1}D_1)-I]\}P^*(1,z),$$

$$
\begin{aligned}
&z\underline{X}(2,z) = \{\underline{X}(1,z) + \underline{x}_{0,1}[z(-D_0^{-1}D_1) - I]\}P^*(2,z)\,, \\
&\qquad\vdots \\
&z\underline{X}(K,z) = \{\underline{X}(K-1,z) + \underline{x}_{0,K-1}[z(-D_0^{-1}D_1) - I]\}P^*(K,z)\,.
\end{aligned}
\tag{B.1}
$$

Setting $z = 1$ in (B.1) yields

$$
\begin{aligned}
&\underline{X}(1,1) = \{\underline{X}(K,1) + \underline{x}_{0,K}[(-D_0^{-1}D_1) - I]\}P^*(1,1)\,, \\
&\underline{X}(2,1) = \{\underline{X}(1,1) + \underline{x}_{0,1}[(-D_0^{-1}D_1) - I]\}P^*(2,1), \\
&\qquad\vdots \\
&\underline{X}(K,1) = \{\underline{X}(K-1,1) + \underline{x}_{0,K-1}[(-D_0^{-1}D_1) - I]\}P^*(K,1)\,.
\end{aligned}
\tag{B.2}
$$

It is easy to find out that

$$
\underline{X}(k,1) = \underline{X}(k,1)[P^*(k,1)\cdots P^*(K,1)\cdots P^*(k-1,1) + \sum_{j=k,\ldots,k-1} \underline{x}_{0,j}(-D_0^{-1}D_1)\prod_{i=j,j+1}^{k-1} P^*(k,1)]\,. \tag{B.3}
$$

Notice that the summation and product orders are *k*, *k+1*, ..., *K*, *1*, *2*, ..., *k-1*. Since

$$
\underline{\theta}P^*(k,1)\cdots P^*(K,1)P^*(k-1,1) = \underline{\theta}\,, \tag{B.4}
$$

$$
P^*(k,1)\cdots P^*(K,1)P^*(k-1,1)\underline{e} = \underline{e}\,, \tag{B.5}
$$

Since the matrix $I - P^*(k,1)\cdots P^*(K,1)\cdots P^*(k-1,1) + \underline{e}\underline{\theta}$ is invertible, then, for $1 \le k \le K$,

$$
\underline{X}(k,1) = [\sum_{j=k,k+1,\ldots,k-1} \underline{x}_{0,j}(-D_0^{-1}D_1)\prod_{i=j,j+1,\ldots,k-1} P^*(i,1) + \frac{1}{K}\underline{\theta}][I - \prod_{j=k,k+1,\ldots,k} P^*(j,1) + \underline{e}\underline{\theta}]^{-1}\,. \tag{B.6}
$$

Notice that $\underline{X}(k,1)\underline{e} = 1/K$. Differentiating both sides of (B.1) with respect to *z*, yields

$$
\begin{aligned}
&\underline{X}(1,1) + \underline{X}'(1,1) = \underline{X}'(K,1)P^*(K,1) + \underline{X}(K,1)P^{*\prime}(K,1) \\
&\qquad\qquad + \underline{x}_{0,1}(-D_0^{-1}D_1)[P^{*\prime}(K,1) + P^*(K,1)], \\
&\underline{X}'(k,1) = \{\underline{X}'(k-1,1)P^*(k-1,1) + \underline{X}(k-1,1)P^{*\prime}(k-1,1) \\
&\qquad + \underline{x}_{0,k-1}(-D_0^{-1}D_1)[P^{*\prime}(k-1,1) + P^*(k-1,1)] - \underline{X}(k,1), \quad 2 \le k \le K
\end{aligned}
\,. \tag{B.7}
$$

This leads to

$$\begin{aligned}\underline{X}'(k,1) = & \sum_{i=k-1,k-2,\ldots,k} \{\underline{X}(i,1)P^{*}{}'(i,1) + \underline{x}_{0,i}(-D_0^{-1}D_1)[P^{*}{}'(i,1) + P^{*}(i,1)] \\ & - \underline{X}(i+1,1)\} \prod_{j=k+1}^{i-1} P^{*}(j,1)[I - \prod_{i=k,k+1}^{k-1} P^{*}(i,1) + \underline{e}\underline{\theta}]^{-1} + (\underline{X}'(k,1)\underline{e})\underline{\theta}\text{'}\end{aligned} \tag{B.8}$$

$$\begin{aligned}\underline{X}'(k,1)\underline{e} = & \ \underline{X}'(k-1,1)\underline{e} + \underline{X}(k-1,1)P^{*}{}'(k-1,1)\underline{e} \\ & + \underline{x}_{0,k-1}\underline{e} + \underline{x}_{0,k-1}(-D_0^{-1}D_1)P^{*}{}'(k-1,1)\underline{e} - 1/K,\end{aligned} \tag{B.9}$$

where

$$P^{*}{}'(k,1)\underline{e} = \int_0^\infty dF_k(x) \sum_{n=1}^{\infty} D^{n-1} \frac{x^n D_1 \underline{e}}{n!} = (P^{*}(k,1) - I)(D + \underline{e}\underline{\theta})^{-1} D_1 \underline{e} + \lambda\mu_k \underline{e}. \tag{B.10}$$

This completes the proof.

Appendix C. Proof of Theorem 4.1

Let

$$\underline{V}^{*}(s) = (\underline{V}_1^{*}(s), \cdots, \underline{V}_K^{*}(s)) = -s\underline{x}_0 \hat{D}_0^{-1}(sI + \hat{D}_0 + \sum_{k=1}^{K} f_k^{*}(s)\hat{D}_k)^{-1}. \tag{C.1}$$

(Based on experience, $\lambda\underline{V}^*(s)$ should be the LST of the virtual waiting time, but it needs to be proved.) Notice that $\underline{W}_a^{*} = \underline{V}^{*}(s)\hat{D}_1$ and $\underline{W}_{a,k}^{*}(s) = \underline{V}_{k-1}^{*}(s)D_1$. Rewrite (C.1) as

$$\underline{V}^{*}(s)(sI + \hat{D}_0 + \sum_{k=1}^{K} f_k^{*}(s)\hat{D}_k) = -s\underline{x}_0 \hat{D}_0^{-1}. \tag{C.2}$$

Let $\{\underline{V}^*_k(s), 1 \le k \le K\}$ be m-vectors satisfying $\underline{V}^*(s) = (\underline{V}^*_1(s), ..., \underline{V}^*_K(s))$. Then

$$\underline{V}_k^{*}(s)(sI + D_0) + \underline{V}_{k-1}^{*}(s) f_{k-1}^{*}(s) D_1 = -s\underline{x}_{0,k} D_0^{-1}. \tag{C.3}$$

Taking the first and second derivatives on both sides of equation (C.3) leads to

$$\underline{V}_k^{*}{}'(s)(sI + D_0) + \underline{V}_k^{*}(s) + \underline{V}_{k-1}^{*}{}'(s) f_{k-1}^{*}(s) D_1 + \underline{V}_{k-1}^{*}(s) f_{k-1}^{*}(s) D_k = -x_{0,k} D_0^{-1}. \tag{C.4}$$

Letting $s = 0$ in (C.2) yields $\underline{V}^{*}(0)\hat{D} = 0$, which implies

$$\underline{V}_k^{*}(0) = \underline{\theta}/(\lambda K),\ 1 \le k \le K. \tag{C.5}$$

Substituting (C.7) into (C.4) yields

$$\underline{V}_k^{*}{}'(0)D_0 + \underline{V}_{k-1}^{*}{}'(0)D_1 = -\underline{x}_{0,i}D_0^{-1} + \mu_{k-1}\underline{\theta}D_1/(\lambda K) - \underline{\theta}/(\lambda K). \qquad \text{(C.6)}$$

Combining equations (C.6) and $(D_0+D_1)\underline{e}=0$ leads to the conclusion. This completes the proof.

Appendix D. The mean queue length and the mean waiting time

For completeness of the paper and computational purpose, the mean queue length at the departure of an arbitrary customer and the mean waiting time of an arbitrary customer are given in this appendix.

The mean queue length Let $S = diag(-D_0^{-1}D_1, \ldots, -D_0^{-1}D_1)$. Differentiating both sides of equation (3.4) with respect to z yields

$$\underline{X}'(z)(zI - P^*(z)) + \underline{X}(z)(I - P^{*}{}'(z)) = \underline{x}_0 SP^*(z) + \underline{x}_0(zS - I)P^{*}{}'(z). \qquad \text{(D.1)}$$

It follows

$$\underline{X}'(1)(I - P^*(1)) + \underline{X}(1)(I - P^{*}{}'(1)) = \underline{x}_0 SP^*(1) + \underline{x}_0(S - I)P^{*}{}'(1),$$

which implies

$$\underline{X}'(1) = \underline{x}_0[SP^*(1) + (S-I)P^{*}{}'(1)]H - \underline{X}(1)[I - P^{*}{}'(1)]H + \underline{X}'(1)\underline{e}\underline{\hat{\theta}}, \qquad \text{(D.2)}$$

where $H = [I - P^*(1) + \underline{\hat{e}}\underline{\hat{\theta}}]^{-1}$. Letting $z=1$ in (3.5) leads to

$$\underline{X}(1) = \underline{\hat{\theta}} + \underline{x}_0(S - I)P*(1)H. \qquad \text{(D.3)}$$

Substituting (D.3) into (D.2) gives us

$$\begin{aligned}\underline{X}'(1) = \underline{x}_0\{SP^*(1) + (S-I)[P^{*}{}'(1) - P^*(1)H(I - P^{*}{}'(1))]\}H \\ - \underline{\hat{\theta}}[I - P^{*}{}'(1)]H + \underline{X}'(1)\underline{e}\underline{\hat{\theta}}.\end{aligned} \qquad \text{(D.4)}$$

Differentiating both side of equation (D.1) with respect to z, it yields

$$\begin{aligned}\underline{X}''(z)(zI - P^*(z)) + 2\underline{X}'(z)(I - P^{*}{}'(z)) - \underline{X}(z)P^{*}{}''(z) \\ = \underline{x}_0[2SP^*(z) + (zS - I)P^{*}{}''(z)],\end{aligned} \qquad \text{(D.5)}$$

which implies

$$2\underline{X}'(1)\underline{e} = 2\underline{X}'(1)P^{*}{}'(1)\underline{e} + \underline{X}(1)P^{*}{}''(1)\underline{e} + \underline{x}_{0}[(S-I)P^{*}{}''(1) + 2SP^{*}{}'(1)]\underline{e} \qquad \text{(D.6)}$$

Substituting (D.3) and (D.4) into (D.6), after some algebraic simplifications, mean queue length at the departure epoch of an arbitrary customer is given by

$$X'(1)\underline{e} = \frac{1}{1-\rho}\underline{x}_{0}\{\{SP^{*}(1) + (S-I)[P^{*}{}'(1) - P^{*}(1)H(I - P^{*}{}'(1))]\}HP^{*}{}'(1)$$

$$+ 2SP^{*}{}'(1) + \frac{1}{2}(S-I)[I + P*(1)H]P^{*}{}''(1)\}\underline{e}$$

$$+ \frac{1}{1-\rho}\underline{\hat{\theta}}\Big\{\frac{1}{2}P^{*}{}''(1) - [I - P^{*}{}'(1)]HP^{*}{}'(1)\Big\}\underline{e}.$$

The mean waiting time Differentiating both sides of equation (C.1) with respect to *s*, it yields

$$\underline{V}^{*}{}'(s)\ (sI + \hat{D}_{0} + \sum_{k=1}^{K} f_{k}^{*}(s)\hat{D}_{k}) + \underline{V}^{*}(s)\ (I + \sum_{k=1}^{K} f_{k}^{*}{}'(s)\hat{D}_{k}) = -\underline{x}_{0}\hat{D}_{0}^{-1}. \qquad \text{(D.7)}$$

Substituting $s=0$ into (D.7) gives

$$\underline{V}^{*}{}'(0) = -\underline{V}^{*}(0)(I + \sum_{k=1}^{K} f_{k}^{*}{}'(0)\hat{D}_{k})\ (\hat{D} + \underline{\hat{e}}\underline{\hat{\theta}})^{-1} - \underline{x}_{0}\hat{D}_{0}^{-1}\ (\hat{D} + \underline{\hat{e}}\underline{\hat{\theta}})^{-1} + (\underline{V}^{*}{}'(0)\underline{e})\underline{\hat{\theta}}$$

$$= [-\underline{x}_{0}\hat{D}_{0}^{-1} - \frac{\underline{\hat{\theta}}}{\lambda}(I - \sum_{k=1}^{K}\frac{1}{\mu_{k}}\hat{D}_{k})]\ (D + \underline{\hat{e}}\underline{\hat{\theta}})^{-1} + (\underline{V}^{*}{}'(0)\underline{e})\underline{\hat{\theta}}. \qquad \text{(D.8)}$$

This leads to the mean waiting time:

$$-\underline{W}_{a}^{*}(0)e = -\underline{V}^{*}(0)D_{1}\underline{e} = [\underline{x}_{0}\hat{D}_{0}^{-1} + \underline{\hat{\theta}}(I - \sum_{k=1}^{K}\frac{1}{\mu_{k}}\hat{D}_{k})]\ (\hat{D} + \underline{\hat{e}}\underline{\hat{\theta}})^{-1} D_{1}\underline{e} - \lambda(\underline{V}^{*}{}'(0)\underline{e}) . \text{(D.9)}$$

The term $\underline{V}^{*}$’(*0*)e is obtained as follows. Differentiating (D.7), it yields

$$\underline{V}^{*}{}''(s)\ (sI + \hat{D}_{0} + \sum_{k=1}^{K} f_{k}^{*}(s)\hat{D}_{k}) + 2\underline{V}^{*}{}'(s)\ (I + \sum_{k=1}^{K} f_{k}^{*}{}'(s)\hat{D}_{k})$$

$$+ \underline{V}^{*}(s)\sum_{k=1}^{K} f_{k}^{*}{}''(s)\hat{D}_{k} = 0. \qquad \text{(D.10)}$$

Combining (C.1), (D.8) and (D.10) yields

$$2\underline{V}^{*}{}'(0)\underline{e} = 2\underline{V}^{*}{}'(0)[\sum_{k=1}^{K}\frac{1}{\mu_{k}}D_{ki}]\underline{e} - \frac{\underline{\hat{\theta}}}{\lambda}\sum_{k=1}^{K} f_{k}^{*}{}''(0)\hat{D}_{k}\,\underline{e}$$

$$= 2\{[-\underline{x}_0 \hat{D}_0^{-1} - \frac{\underline{\underline{\hat{\theta}}}}{\lambda}(I - \sum_{k=1}^{K} \frac{1}{\mu_k}\hat{D}_k)](\underline{\hat{D}} + \underline{e}\underline{\hat{\theta}})^{-1} + \underline{V}^{*}{}'(0)\underline{e}\underline{\hat{\theta}}\} \\ \cdot [\sum_{k=0}^{K} \frac{1}{\mu_k}\hat{D}_i]\underline{e} - \frac{\underline{\underline{\hat{\theta}}}}{\lambda}\sum_{k=1}^{K} f_k^{*}{}''(0)\hat{D}_k \underline{e} ,$$

that is,

$$\underline{V}^{*}{}'(0)\underline{e} = \frac{-1}{1-\rho}\{[\underline{x}_0 \hat{D}_0^{-1} + \frac{\underline{\underline{\hat{\theta}}}}{\lambda}(I - \sum_{k=0}^{K} \frac{1}{\mu_k}\hat{D}_k)](\hat{D} + \underline{e}\underline{\hat{\theta}})^{-1} \sum_{k=0}^{K} \frac{1}{\mu_k}\hat{D}_k \underline{e} \\ + \frac{1}{2}\frac{\underline{\underline{\hat{\theta}}}}{\lambda}\sum_{k=1}^{K} f_k^{*}{}''(0)\hat{D}_k \underline{e}\} . \qquad (D.11)$$

This completes the proof.