Journal of Applied Mathematics and Stochastic Analysis
7, Number 2, 1994, 161-178

A FINITE CAPACITY QUEUE WITH MARKOVIAN ARRIVALS AND TWO SERVERS WITH GROUP SERVICES[1]

S. CHAKRAVARTHY
GMI Engineering & Management Institute
Department of Science and Mathematics
Flint, MI 48504-4898 USA

ATTAHIRU SULE ALFA
University of Manitoba
Department of Mechanical and Industrial Engineering
Winnipeg, CANADA R3T 2N2

(Received February, 1994; revised May, 1994)

ABSTRACT

In this paper we consider a finite capacity queuing system in which arrivals are governed by a Markovian arrival process. The system is attended by two exponential servers, who offer services in groups of varying sizes. The service rates may depend on the number of customers in service. Using Markov theory, we study this finite capacity queuing model in detail by obtaining numerically stable expressions for (*a*) the steady-state queue length densities at arrivals and at arbitrary time points; (*b*) the Laplace-Stieltjes transform of the stationary waiting time distribution of an admitted customer at points of arrivals. The stationary waiting time distribution is shown to be of phase type when the interarrival times are of phase type. Efficient algorithmic procedures for computing the steady-state queue length densities and other system performance measures are discussed. A conjecture on the nature of the mean waiting time is proposed. Some illustrative numerical examples are presented.

Key words: Queues, Renewal Process, Finite Capacity, Markovian Arrival Process, Laplace-Stieltjes Transform, Algorithmic Probability.

AMS (MOS) subject classifications: Primary: 60K20, 60K25, 90B22; Secondary: 60J27, 60K05, 60K15.

1. Model Description

We consider a finite capacity queuing system in which arrivals occur singly according to a Markovian arrival process (MAP). Any arrival finding the buffer (or waiting room) of size K full is lost. The system is attended by two servers who offer services to customers (or jobs) in groups of varying sizes. The service times of both servers are assumed to be independent and exponentially distributed with (possibly) different parameters that may depend on the number of

[1]This research was supported in part by Grant No. DDM-9313283 from the National Science Foundation to S. Chakravarthy and Natural Sciences and Engineering Council of Canada Grant No. OGP0006584 to A.S. Alfa.

customers served. The customers are served in groups of size at least L, where the preassigned number $L \geq 1$, is called the *threshold.* The service discipline of the system is as follows. If upon the completion of a service, L or more customers are present, the freed server initiates a new service to **all** the customers present. However, if fewer than L customers are present the server waits until the queue length reaches L. When both servers are free server 1 will initiate a new service with probability p, and server 2 will initiate a new service with probability $q = 1 - p$, $0 \leq p \leq 1$. The service times of server i are exponentially distributed with parameter $\mu_r^{(i)}$, that may depend on the number of customers (r) in the group, for $i = 1, 2$, and $L \leq r \leq K$.

These types of models in the context of GI/PH/1 and MAP/G/1 queuing models with finite capacity were first studied by Chakravarthy [2, 3]. The potential applications were outlined in those paper and for the sake of completeness we will mention a few applications here.

(1) In manufacturing process, jobs that require processing such as flushing with the same coolant, coating bricks with noble metals (done by dipping the bricks in liquid concentrate), sand blasting and heat treatment, can all be done in groups.

(2) In the treatment of hazardous petrochemical and petroleum wastes, certain wastes may all need a specific treatment method such as thermal treatment method requiring the use of high temperatures (900°F-3000°F) to break down the hazardous chemicals into simpler, less toxic forms. These could be handled in groups.

(3) In machine vision systems, the jobs that arrive for processing may all have the same characteristics and hence all the jobs can be placed in a common tray or belt for the camera to photograph the images and send the information to the central computer for analysis.

(4) In order to process the packages (received at a package delivery company) more efficiently at the initial stages of routing, the company may classify the arrivals of packages into two categories: one containing (usually small) packages that require individual attention such as marking codes, posting special care, if any, and routing information; the other may involve packages that are destined to go in only one route or in one vehicle. These packages can be processed in groups of one or more.

(5) In computer communications, suppose the incoming jobs are grouped into two types: one type requiring access to a common data base and the other type needing the use of common input/output device such as a laser printer or color plotter. The central processor can process all the jobs of each type in groups. Another example is in load balancing using probing in distributed processing. When jobs arrive into the dispatcher, it probes the distributed system for the type of load (heavy, moderate or light) and accordingly the jobs are distributed to balance the load among various processors.

In all the above applications, we see that the jobs that require processing of general type can be processed in groups of varying sizes, which motivates the need for the type of service mechanism considered in this paper. For economic reasons, it is better to have a minimum number of jobs to form a batch before they are processed. The maximum number of jobs that can be processed at a time is the size of the buffer, which is given by K.

A MAP with single arrivals is defined as the point process generated by the transitions epochs of an m-state Markov renewal process with transition probability matrix given by

$$F(x) = \left\{ \int_0^x e^{C_0 t} dt \right\} C_1, \quad \text{for } x \geq 0, \tag{1}$$

where C_0 and C_1 are square matrices, each of order m whose sum$Q = C_0 + C_1$ is an **irreducible**

infinitesimal generator. The matrix C_0, with negative diagonal elements and nonnegative off-diagonal elements, governs transitions that correspond to no arrivals, and C_1, a nonnegative matrix, governs transitions that correspond to (single) arrivals. We also assume that $Q \neq C_0$ and hence C_0, being a stable matrix, will be nonsingular.

Let $\boldsymbol{\pi}$ be the stationary probability vector of the Markov process with generator Q. That is, $\boldsymbol{\pi}$ is the unique probability vector satisfying

$$\boldsymbol{\pi}Q = \mathbf{0} \quad \text{and} \quad \boldsymbol{\pi}\mathbf{e} = 1, \tag{2}$$

where $\mathbf{e}$ is a column vector of dimension m, consisting of 1's. Note that π_j gives the probability that at an arbitrary time the MAP will be in phase j, $1 \leq j \leq m$. The constant $\lambda = \boldsymbol{\pi}C_1\mathbf{e}$, referred to as the **fundamental rate**, gives the expected number of arrivals per unit of time in the stationary mode of the MAP.

The MAP (and the extension allowing for group arrivals) has been shown to be equivalent to Neuts' versatile point process [6]. This is a rich class of point processes containing many familiar arrival processes such as Poisson, PH-renewal process, Markov-modulated Poisson process, alternating PH-renewal processes, arrival process obtained as a sequence of PH-interarrival times selected via a Markov chain, and superposition of PH-renewal processes, as very special cases. There is an extensive literature on queuing and communications models in which such point processes are used to model both arrival and service mechanisms. We refer to Lucantoni [4] and Neuts [6, 8] for full details on these point processes and their usefulness in stochastic modeling.

In some manufacturing processes jobs may arrive from various sources such as vendors, shifts, and assembly plants to a common processing area. In these cases the arrival processes can no longer be assumed to form renewal processes. Hence, modeling the arrival processes with MAPs seems to be a natural choice. It should be noted that by an appropriate choice of parameters of the MAP the underlying arrival process can be made a renewal process.

A number of optimization problems, useful in the design of such queuing systems, can be studied in terms of choosing a value for L, by fixing the parameters of the arrival and service processes. One such example would be to choose a value of L for which the jobs do not have to wait for a long time before entering service. Small values of L will result in frequent services with smaller groups and large values of L will result in longer waits for the jobs. It is, therefore, of interest to see how the system performance measures are influenced by the choice of L. Our algorithmic procedures can clearly handle problems of this nature.

The objective of this paper is two-fold. First, we perform a steady-state analysis of the model by deriving expressions for

- (*i*) the densities of the number of customers in the queue;
- (*ii*) the conditional density of the number of customers served during a service by server i, given that server i, $i = 1, 2$, is busy;
- (*iii*) the Laplace-Stieltjes transforms of the steady-state waiting time distributions; and
- (*iv*) various system performance measure useful in the qualitative interpretation of the model studied.

Secondly, we develop implementable algorithms for computing several performance measures such as the throughput, proportion of time both servers are idle, proportion of time server i is busy, the mean and the standard deviation of number of customers in the queue, and the mean waiting time of an admitted customer into the system. A conjecture, based on our computational experience, on the nature of the mean waiting time at an arrival epoch is proposed.

2. The Steady-State Probability Vectors

The queuing model described in Section 1 can be studied as a continuous-time Markov process on the state space $\Omega = \Omega_1 \cup \Omega_2 \cup \Omega_3 \cup \{\boldsymbol{i}: 0 \leq i \leq K\}$, where

$$\Omega_1 = \{(i,k): 0 \leq i \leq L-1, 1 \leq k \leq m\},$$

$$\Omega_2 = \{(i,j_1,k): 0 \leq i \leq L-1, L \leq j_1 \leq K, 1 \leq k \leq m\},$$

$$\Omega_3 = \{(i,j_2,k): 0 \leq i \leq L-1, L \leq j_2 \leq K, 1 \leq k \leq m\},$$

$$i = \{(i,j_1,j_2,k): L \leq j_1, j_2 \leq K, 1 \leq k \leq m\}, 0 \leq i \leq K.$$

The states and their description are given in Table 1 below. Although the number of states in this Markov process is large, the generator of the Markov process is highly sparse. By exploiting this special structure, we will develop efficient algorithms for computing the steady-state probability vector.

Table 1: States and their description

State	Description
(i,k)	i customers in the queue, the arrival process is in state k and both servers are idle.
(i,j_1,k)	i customers are in the queue, server 1 is busy with j_1 customers, server 2 is idle and the arrival process is in state k.
(i,j_2,k)	i customers are in the queue, server 2 is busy with j_2 customers, server 1 is idle and the arrival process is in state k.
(i,j_1,j_2,k)	i customers are in the queue, servers 1 and 2 are busy, respectively, with j_1 and j_2 customers and the state of the MAP is k.

In the sequel the notation $\boldsymbol{e}_i$ denotes a unit column vector with 1 in the i-th position and 0 elsewhere, $\boldsymbol{e}$ a column vector of 1's, and I an identity matrix of appropriate dimensions. The symbol $'$ will be used to denote the transpose and the symbol $\otimes$ will denote the Kronecker product of two matrices. Specifically, $A \otimes B$ stands for the matrix made up of blocks $A_{ij}B$. For more details on the Kronecker products, we refer the reader to Bellman [1] or Marcus and Minc [5]. Denote by $\boldsymbol{\mu}^{(k)}$ a column vector whose r-th component is given by $\mu_r^{(k)}$, $k = 1,2$, and $L \leq r \leq K$ and $\Delta(\boldsymbol{\mu}^{(k)})$ is a diagonal matrix with diagonal elements given by the components of $\boldsymbol{\mu}^{(k)}$, for $k = 1,2$.

The generator Q^* of the Markov process is given by

$$Q^* = \begin{bmatrix}
B_0 & pB_1 & qB_1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\
I \otimes \boldsymbol{\mu}^{(1)} \otimes I & B_2 & 0 & B_3 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\
I \otimes \boldsymbol{\mu}^{(2)} \otimes I & 0 & B_4 & B_5 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\
0 & E_1 & F_1 & A_0 & I \otimes C_1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\
0 & E_2 & F_2 & 0 & A_0 & I \otimes C_1 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\
0 & E_L & F_L & 0 & 0 & 0 & \dots & A_0 & I \otimes C_1 & 0 & \dots & 0 & 0 \\
0 & 0 & 0 & D_1 & 0 & 0 & \dots & 0 & A_0 & I \otimes C_1 & \dots & 0 & 0 \\
0 & 0 & 0 & D_2 & 0 & 0 & \dots & 0 & 0 & A_0 & \dots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\
0 & 0 & 0 & D_{K-L} & 0 & 0 & \dots & 0 & 0 & 0 & \dots & A_0 & I \otimes C_1 \\
0 & 0 & 0 & D_{K-L+1} & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & A_1
\end{bmatrix},$$

where the matrices appearing in Q^* are defined as follows.

$$B_0 = \begin{bmatrix}
C_0 & C_1 & 0 & \dots & 0 & 0 \\
0 & C_0 & C_1 & \dots & 0 & 0 \\
\vdots & \vdots & \vdots & \dots & \vdots & \vdots \\
0 & 0 & 0 & \dots & C_0 & C_1 \\
0 & 0 & 0 & \dots & 0 & C_0
\end{bmatrix}, \quad B_1 = \boldsymbol{e}_L \otimes \boldsymbol{e}_1' \otimes \boldsymbol{e}_1' \otimes C_1,$$

$$B_2 = \begin{bmatrix}
I \otimes C_0 & I \otimes C_1 & 0 & \dots & 0 & 0 \\
0 & I \otimes C_0 & I \otimes C_1 & \dots & 0 & 0 \\
\vdots & \vdots & \vdots & \dots & \vdots & \vdots \\
0 & 0 & 0 & \dots & I \otimes C_0 & I \otimes C_1 \\
0 & 0 & 0 & \dots & 0 & I \otimes C_0
\end{bmatrix} - [I \otimes \Delta(\boldsymbol{\mu}^{(1)}) \otimes I],$$

$$B_3 = \boldsymbol{e}_L \otimes I \otimes \boldsymbol{e}_1' \otimes C_1, \quad B_5 = \boldsymbol{e}_L \otimes \boldsymbol{e}_1' \otimes I \otimes C_1,$$

$$B_4 = \begin{bmatrix}
I \otimes C_0 & I \otimes C_1 & 0 & \dots & 0 & 0 \\
0 & I \otimes C_0 & I \otimes C_1 & \dots & 0 & 0 \\
\vdots & \vdots & \vdots & \dots & \vdots & \vdots \\
0 & 0 & 0 & \dots & I \otimes C_0 & I \otimes C_1 \\
0 & 0 & 0 & \dots & 0 & I \otimes C_0
\end{bmatrix} - [I \otimes \Delta(\boldsymbol{\mu}^{(2)}) \otimes I], \tag{4}$$

$$E_i = e_i' \otimes I \otimes \mu^{(2)} \otimes I, \quad F_i = e_i' \otimes \mu^{(1)} \otimes I, \text{ for } 1 \le i \le L,$$

$$D_i = [e_i' \otimes \mu^{(1)} \otimes I] + [I \otimes e_i' \otimes \mu^{(2)} \otimes I], \text{ for } 1 \le i \le K - L + 1.$$

$$A_0 = [I \otimes C_0] - [\Delta(\mu^{(1)}) \otimes I \otimes I] - [I \otimes \Delta(\mu^{(2)}) \otimes I], \quad A_1 = A_0 + I \otimes C_1.$$

2.1 Steady-State Probability Vector at an Arbitrary Time

The steady-state probability vector x of Q^* is the (unique) solution to

$$xQ^* = 0, \quad xe = 1. \tag{5}$$

We first partition x into vectors of smaller dimensions as $x = (u, v, w, y(0), y(1), \ldots, y(K))$. Note that the vector u is of order mL; the vectors v and w are of order $(K - L + 1)Lm$; and the vectors $y(i)$, for $0 \le i \le K$, are of order $(K - L + 1)^2 m$. We further partition the vectors u, v, w, and $y(i)$ into vectors of smaller dimension as follows: $u = (u(0), \ldots, u(L-1))$; $v = (v(0), \ldots, v(L-1))$ with $v(i) = (v_L(i), \ldots, v_K(i))$, $0 \le i \le L-1$, $w = (w(0), \ldots, w(L-1))$ with $w(i) = (w_L(i), \ldots, w_K(i))$, $0 \le i \le L-1$, $y(i) = (y_{L,L}(i), \ldots, y_{L,K}(i), \ldots, y_{K,L}(i), \ldots, y_{K,K}(i))$, $0 \le i \le K$. By exploiting the sparsity of Q^*, the steady-state equations in (5) can be efficiently solved in terms of smaller matrices of order m. The required equations are given in the Appendix.

The following two lemmas, which are intuitively obvious, can be used as accuracy checks in numerical computation of x.

Lemma 1: *We have*

$$\sum_{k=L}^{K} [v_k(L-1) + w_k(L-1)]C_1 e = \sum_{i=0}^{L-1} \sum_{r=L}^{K} \sum_{k=L}^{K} [\mu_r^{(1)} y_{r,k}(i) + \mu_2^{(2)} y_{k,r}(i)]e, \tag{6}$$

$$p\delta(k-L)u(L-1)C_1 e + \sum_{i=0}^{L-1} \sum_{r=L}^{K} \mu_r^{(2)} y_{k,r}(i)e$$

$$= v_k(L-1)C_1 e + \sum_{i=0}^{L-1} \mu_k^{(1)} v_k(i)e, \quad L \le k \le K, \tag{7}$$

$$q\delta(k-L)u(L-1)C_1 e + \sum_{i=0}^{L-1} \sum_{r=L}^{K} \mu_r^{(1)} y_{r,k}(i)e$$

$$= w_k(L-1)C_1 e + \sum_{i=0}^{L-1} \mu_k^{(2)} w_k(i)e, \quad L \le k \le K, \tag{8}$$

$$u(L-1)C_1 e = \sum_{i=0}^{L-1} \sum_{j=L}^{K} [\mu_j^{(1)} v_j(i) + \mu_j^{(2)} w_j(i)]e, \tag{9}$$

$$y_{j,k}(0)C_1 e = \sum_{i=1}^{K} [\mu_j^{(1)} + \mu_k^{(2)}] y_{j,k}(i)e, \quad L \le j, \ k \le K. \tag{10}$$

Proof: The stated equations follow immediately from equations (A1-A11). For example, postmultiplying each one of the equations (A1-A8) by $\boldsymbol{e}$ and adding we get the state equation in (6).

Remark: The results of Lemma 1 can be obtained intuitively. For example, Equation (6) states that in equilibrium, the rate at which the system enters the state in which both servers are busy should be equal to the rate at which the system will leave that state, as it should be.

Lemma 2: *We have*

$$\sum_{i=0}^{L-1} \boldsymbol{u}(i) + \sum_{i=0}^{L-1} \sum_{k=L}^{K} [\boldsymbol{v}_k(i) + \boldsymbol{w}_k(i)] + \sum_{i=0}^{K} \sum_{r=L}^{K} \sum_{k=L}^{K} \boldsymbol{y}_{r,k}(i) = \boldsymbol{\pi}, \tag{11}$$

where $\boldsymbol{\pi}$ is as given in (2).

Proof: By adding the equations (A1-A11) over appropriate index values, and using the uniqueness of $\boldsymbol{\pi}$, we get the stated equation.

Let p_i denote the stationary probability that server i is busy, for $i = 1,2$. Then it is easy to see that

$$P_1 = (\boldsymbol{ve} + \boldsymbol{ye}) \quad \text{and} \quad P_2 = (\boldsymbol{we} + \boldsymbol{ye}). \tag{12}$$

Let $N^{(i)}$ denote the number of customers served during service by server i, $i = 1,2$. The following lemma, whose proof follows immediately from the definition of $N^{(i)}$, gives expressions for the conditional probability densities of $N^{(i)}$.

Lemma 3: *The conditional probability densities $f_{B_i}^{(i)}$ of the random variables $N^{(i)}$ given $B_i = \{$server i busy$\}$, $i = 1,2$, are given by*

$$\begin{aligned} f_{B_1}^{(1)}(n) &= \frac{1}{P_1}\left[\sum_{i=0}^{L-1} \boldsymbol{v}_n(i)\boldsymbol{e} + \sum_{i=0}^{K} \sum_{k=L}^{K} \boldsymbol{y}_{n,k}(i)\boldsymbol{e}\right], \quad L \le n \le K, \\ f_{B_2}^{(2)}(n) &= \frac{1}{P_2}\left[\sum_{i=0}^{L-1} \boldsymbol{w}_n(i)\boldsymbol{e} + \sum_{i=0}^{K} \sum_{k=L}^{K} \boldsymbol{y}_{k,n}(i)\boldsymbol{e}\right], \quad L \le n \le K, \end{aligned} \tag{13}$$

where P_1 and P_2 are as given in (12).

The **throughput**, defined as the rate at which customers leave the system, can be expressed as follows.

$$\gamma = P_1 \sum_{i=L}^{K} i\mu_i^{(1)} f_{B_1}^{(1)}(i) + P_2 \sum_{i=L}^{K} i\mu_i^{(2)} f_{B_2}^{(2)}(i). \tag{14}$$

2.2 The Stationary Queue Length at Arrivals

The joint stationary density of the number of customers in the queue, the number of customers in service and the phase of the arrival process at arrival epochs is obtained in this section. Suppose we denote by $z_0(i)$, $0 \le i \le L-1$, $z_1(i,j)$, for $0 \le i \le L-1$, $L \le j \le K$, $z_2(i,j,k)$, for $0 \le i \le K$, $L \le j$, $k <$ K, vectors of order m with components given by $z_0(i,r)$, $z_1(i,j,r)$, and $z_2(i,j,k,r)$, respectively. The component $z_0(i,r)$ gives the steady-state probability that an arrival finds both servers idle with i customers in the queue and that the arrival process is in phase r; $z_1(i,j,r)$ is the steady-state probability that an arrival finds exactly one server busy

with j customers, and i customers are in the queue and at that instant the arrival process is in phase r. $z_2(i,j,k,r)$ is the steady-state probability that an arrival finds both servers busy with j and k customers, and i customers are in the queue and at that instant the arrival process in in phase r.

Lemma 4: *The vectors* $z_0(i)$, $z_1(i,j)$, $0 \le i \le L-1$, $L \le j \le K$ *and* $z_2(i,j,k)$, $0 \le i \le K$, $L \le j$, $k \le K$ *are given by*

$$\begin{aligned} z_0(i) &= \tfrac{1}{\lambda}\boldsymbol{u}(i)C_1, \quad 0 \le i \le L-1, \\ z_1(i,j) &= \tfrac{1}{\lambda}(\boldsymbol{v}_j(i) + \boldsymbol{w}_j(i))C_1, \quad 0 \le i \le L-1,\ L \le j \le K, \\ z_2(i,j,k) &= \tfrac{1}{\lambda}\boldsymbol{y}_{j,k}(i)C_1, \quad L-1 \le i \le K,\ L \le j,\ k \le K, \end{aligned} \tag{15}$$

where λ *is the fundamental rate of the arrival process.*

Proof: follows from the definition of $\boldsymbol{x}$.

3. Stationary Waiting Time Distribution

Suppose that $W(\cdot)$ denotes the stationary waiting time distribution of an admitted customer at an arrival. Before we derive an expression for the Laplace-Stieltjes transform $W^*(s)$ of $W(\cdot)$, we need some additional notation.

Let $\theta_{i,j}(t)$, for $0 \le i \le L-2$, $i \le j \le m$, be the probability that, given that an arriving customer finds i customers in the queue with at least one server being idle and the arrival process being in phase j, $(L-1-i)$ arrivals occur at or before time t. Let $\boldsymbol{\theta}_i^*(s)$ denote the (column) vector of order m, whose j-th element gives the LST of $\theta_{i,j}(\cdot)$. Let $\boldsymbol{\delta}^*(i,j,k,s)$, for $0 \le i \le L-2$, denote (column) vector LST whose r-th component gives transform of the maximum of an exponential random variable with parameter $\mu_j^{(1)} + \mu_k^{(2)}$ and the time taken to see $L-1-i$ arrivals in the arrival process which started in phase r.

Theorem 1: *The LST* $W^*(s)$ *is given by*

$$\begin{aligned} W^*(s) = c^*\Bigg\{ &\sum_{i=0}^{L-2} \xi(i)\boldsymbol{\theta}_i^*(s) + \sum_{i=0}^{L-2}\sum_{j=L}^{K}\sum_{k=L}^{K} z_2(i,j,k)\boldsymbol{\delta}^*(i,j,k,s) \\ &+ \sum_{i=L-1}^{K-1}\sum_{j=L}^{K}\sum_{k=L}^{K} \frac{\mu_j^{(1)} + \mu_k^{(2)}}{s + \mu_j^{(1)} + \mu_k^{(2)}} z_2(i,j,k)\boldsymbol{e}\Bigg\}, \quad \text{for } Re(s) \ge 0, \end{aligned} \tag{16}$$

where

$$\begin{aligned} \xi(i) &= z_0(i) + \sum_{j=L}^{K} z_1(i,j), \quad 0 \le i \le L-2 \\ c^* &= [1 - \sum_{j=L}^{K}\sum_{k=L}^{K} z_2(K,j,k)\boldsymbol{e}]^{-1}. \end{aligned} \tag{17}$$

Proof: Immediate from the law of total probability. The probability that the waiting time is zero is given by

$$c^*[z_0(L-1)\boldsymbol{e} + \sum_{j=L}^{K} z_1(L-1,j)\boldsymbol{e}].$$

4. The Case of Phase Type Arrivals

When $C_0 = T$ and $C_1 = \boldsymbol{T}^0\boldsymbol{\alpha}$ the arrival process is described by a PH-renewal whose interarrival times follow a PH-distribution with representation given by $(\boldsymbol{\alpha}, T)$ of order m. The mean interarrival times can be verified to be $\frac{1}{\lambda} = -\boldsymbol{\alpha}T^{-1}\boldsymbol{e}$.

Theorem 2: *When interarrival times follow a PH-distribution with irreducible representation given by $(\boldsymbol{\alpha}, T)$ of order m, the stationary waiting time distribution $W(\cdot)$ of an admitted customer at an arrival also follows a PH-distribution whose LST is given by*

$$W^*(s) = d^*\left\{\sum_{i=0}^{L-2}\boldsymbol{\delta}(i)\boldsymbol{T}^0[\boldsymbol{\alpha}(sI-T)^{-1}\boldsymbol{T}^0]^{L-1-i}\right.$$
$$+\sum_{i=0}^{L-2}\sum_{j=L}^{K}\sum_{k=L}^{K}\boldsymbol{y}_{j,k}(i)\boldsymbol{T}^0(\boldsymbol{\zeta}(sI-M)^{-1}\boldsymbol{M}^0) \tag{18}$$
$$\left.+\sum_{i=L-1}^{K-1}\sum_{j=L}^{K}\sum_{k=L}^{K}\frac{\mu_j^{(1)}+\mu_k^{(2)}}{s+\mu_j^{(1)}+\mu_k^{(2)}}\boldsymbol{y}_{j,k}(i)\boldsymbol{T}^0\right\}, \quad \textit{for } Re(s)\geq 0,$$

where $\boldsymbol{\zeta} = (\boldsymbol{\alpha}, \mathbf{0}, \ldots, \mathbf{0}, \mathbf{0}, \ldots, \mathbf{0}, 0)$,

$$M = \begin{bmatrix}
T-\mu I & \boldsymbol{T}^0\boldsymbol{\alpha} & 0 & \ldots & 0 & \mu I & 0 & 0 & \ldots & 0 & \mathbf{0} \\
0 & T-\mu I & \boldsymbol{T}^0\boldsymbol{\alpha} & \ldots & 0 & 0 & \mu I & 0 & \ldots & 0 & \mathbf{0} \\
\vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & T-\mu I & 0 & 0 & 0 & \ldots & \mu I & \boldsymbol{T}^0 \\
0 & 0 & 0 & \ldots & 0 & T & \boldsymbol{T}^0\boldsymbol{\alpha} & 0 & \ldots & 0 & \mathbf{0} \\
0 & 0 & 0 & \ldots & 0 & 0 & T & \boldsymbol{T}^0\boldsymbol{\alpha} & \ldots & 0 & \mathbf{0} \\
\vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & \ldots & T & \mathbf{0} \\
0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & \ldots & 0 & -\mu
\end{bmatrix},$$

and

$$\mu = \mu_j^{(1)} + \mu_k^{(2)},$$

$$\boldsymbol{\delta}(i) = \frac{1}{\lambda}\left\{\boldsymbol{u}(i) + \sum_{j=L}^{K}[\boldsymbol{v}_j(i) + \boldsymbol{w}_j(i)]\right\}, \quad 0 \leq i \leq L-2, \tag{19}$$
$$d^* = [\lambda - \sum_{j=L}^{K}\sum_{k=L}^{K}\boldsymbol{y}_{j,k}(K)\boldsymbol{T}^0]^{-1}.$$

The column vector $\boldsymbol{M}^0$ is such that $M\boldsymbol{e} + \boldsymbol{M}^0 = \mathbf{0}$.

Proof: Immediately follows from the following observations (see Neuts [7]):

(*a*) the convolution of $L-1-i$ interarrival time distributions is a PH-distribution;

(*b*) the maximum of two independent phase type random variables is also of phase type; and

(*c*) a finite mixture of PH-distributions is also a PH-distribution.

Corollary: *The mean* (μ'_w) *waiting time in the* PH/M/2/K *model is given by*

$$\mu'_w = d^*\left\{\sum_{i=0}^{L-2}[\delta(i)T^0 + \sum_{j=L}^{K}\sum_{k=L}^{K} y_{j,k}(i)T^0]\frac{L-1-i}{\lambda}\right.$$

$$+\sum_{i=0}^{L-2}\sum_{j=L}^{K}\sum_{k=L}^{K} y_{j,k}(i)T^0\frac{(\alpha[(\mu_j^{(1)}+\mu_k^{(2)})I-T]^{-1}T^0)^{L-1-i}}{\mu_j^{(1)}+\mu_k^{(2)}} \tag{20}$$

$$\left.+\sum_{i=L-1}^{K-1}\sum_{j=L}^{K}\sum_{k=L}^{K}\frac{1}{\mu_j^{(1)}+\mu_k^{(2)}}y_{j,k}(i)T^0\right\},$$

where $\delta(i)$ *and* d^* *are as given in* (19).

5. Numerical Examples

Here we discuss three representative numerical examples of the model discussed in this paper. We also propose a conjecture on the nature of the mean waiting time.

Example 1: Our interest in the first example is to study the effect which the parameter p has on various performance measures. The data for this example is as follows: $K = 15$, $L = 5$, the service rates for servers 1 and 2 are geometrically decreasing with $\mu_5^{(1)} = 3$, $\mu_{15}^{(1)} = 2$, $\mu_5^{(2)} = 2$, and $\mu_{15}^{(1)} = 1$. The MAP governing the arrivals has the representation given by

$$C_0 = \begin{bmatrix} -6.5 & 1 \\ 3.5 & -5.5 \end{bmatrix} \quad \text{and} \quad C_1 = \begin{bmatrix} 3.5 & 2 \\ 0.5 & 1.5 \end{bmatrix}.$$

It can be verified that $\lambda = 4$. The parameter p was varied from 0.0 to 1.0 and the results are shown in Table 2 below.

An examination of Table 2 reveals the following information. The impact of this parameter on the mean queue length, the standard deviation of the queue length and the throughput is very negligible. However, as p increases the probability of both servers being busy decreased. This is intuitive since increasing p implies that the server 1 has a higher chance of initiating service when both servers are idle; and since the first server serves at a faster rate, this outcome is not surprising. It is interesting to note that the probability of both servers being idle increases as p increases. Again, a similar intuitive reasoning can be given. An as one expects, as p increases, the probability of server 1 being busy increases and server 2 being busy decreases.

Table 2

p	γ	P_{12}	P_1	P_2	I_{12}	EQL	$SDQL^*$
0.0	4.00	0.0219	0.0347	0.3479	0.6393	2.0014	1.0974
0.1	4.00	0.0209	0.0548	0.3177	0.6484	2.0014	1.0973
0.2	4.00	0.0198	0.0753	0.2870	0.6576	2.0013	1.0973
0.3	4.00	0.0188	0.0961	0.2558	0.6667	2.0012	1.0971
0.4	4.00	0.0177	0.1171	0.2242	0.6763	2.0011	1.0970
0.5	4.00	0.0166	0.1385	0.1921	0.6860	2.0011	1.0969
0.6	4.00	0.0155	0.1603	0.1595	0.6957	2.0010	1.0968
0.7	4.00	0.0144	0.1824	0.1264	0.7056	2.0009	1.0967
0.8	4.00	0.0132	0.2048	0.0927	0.7157	2.0009	1.0966
0.9	4.00	0.0121	0.2276	0.0586	0.7259	2.0008	1.0965
1.0	4.00	0.0109	0.2507	0.0238	0.7363	2.0007	1.0964

$*p = P$ (server 1 will initiate service); $\gamma =$ Throughput, $P_{12} = P$ (both servers are busy); $P_1 = P$ (server 1 is busy); $P_2 = P$ (server 2 is busy); $I_{12} = P$ (both servers are idle); $EQL =$ mean queue length; $SDQL =$ Standard deviation of queue length.

Example 2: Our interest here is to study the effect of different arrival processes on the performance measures as the threshold level, L, is varied. We consider four arrival processes, viz: Erlang, Exponential, general MAP, and Hyperexponential; all with the same arrival rate of 10.0. The matrices C_0 and C_1 for these four arrival processes are given by

I. Erlang: $$C_0 = \begin{bmatrix} -20 & 20 \\ 0 & -20 \end{bmatrix} \text{ and } C_1 = \begin{bmatrix} 0 & 0 \\ 20 & 0 \end{bmatrix}$$

II. Exponential: $$C_0 = \begin{bmatrix} -10 \end{bmatrix} \text{ and } C_1 = \begin{bmatrix} 10 \end{bmatrix}$$

III. MAP: $$C_0 = \begin{bmatrix} -7 & 1 \\ 1 & -17 \end{bmatrix} \text{ and } C_1 = \begin{bmatrix} 5 & 1 \\ 2 & 14 \end{bmatrix}$$

IV. Hyperexponential: $$C_0 = \begin{bmatrix} -25 & 0 \\ 0 & -\frac{100}{604} \end{bmatrix} \text{ and } C_1 = \begin{bmatrix} 24.75 & 0.25 \\ \frac{99}{604} & \frac{1}{604} \end{bmatrix}.$$

The other parameters of the model are taken to be $K = 15$, $p = 0.5$ and the service rates of server 1 vary geometrically from 2.5 to 0.25 and those of server 2 vary geometrically from 1.25 to 0.215.

The performance measures considered are throughput, mean queue length, standard deviation of the queue length, and the mean number of customers served by server 1 and server 2. These five measures, as functions of L, are plotted in Figures 1-5.

Figure 1

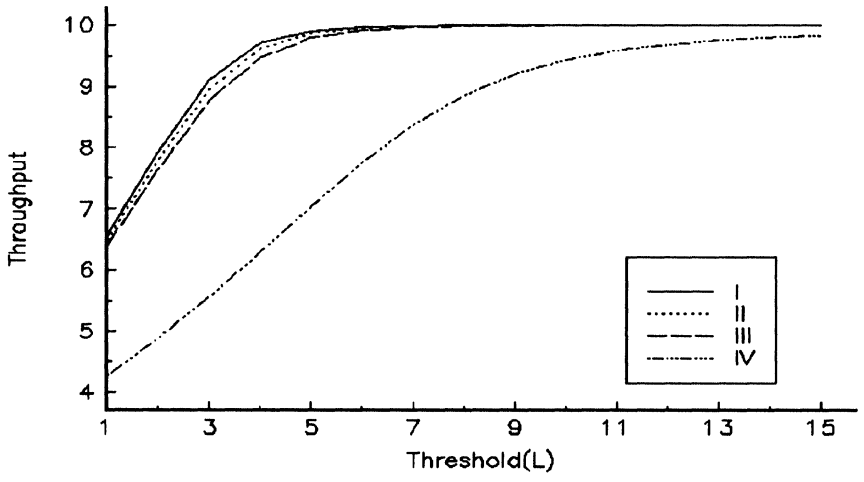
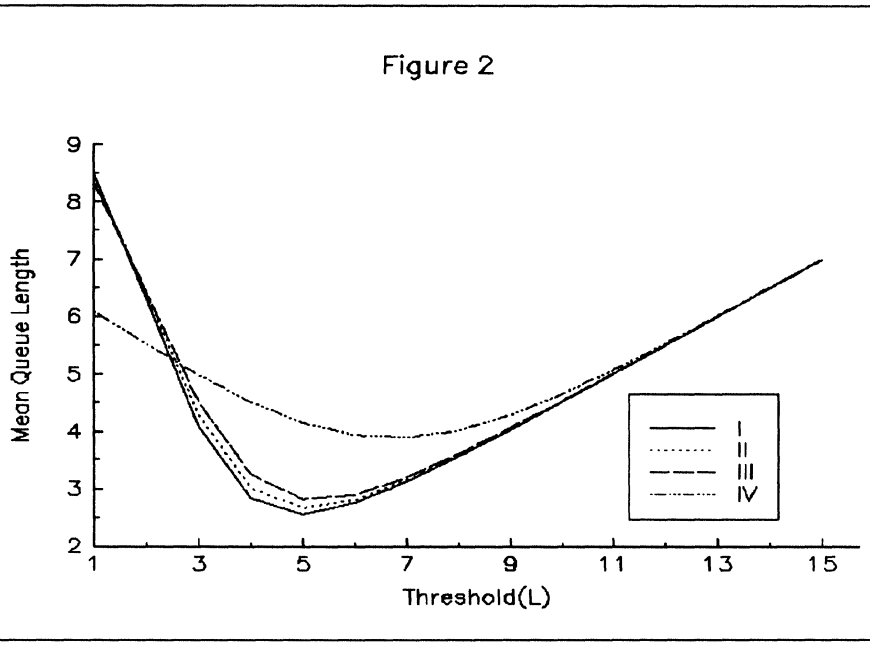


Figure 2



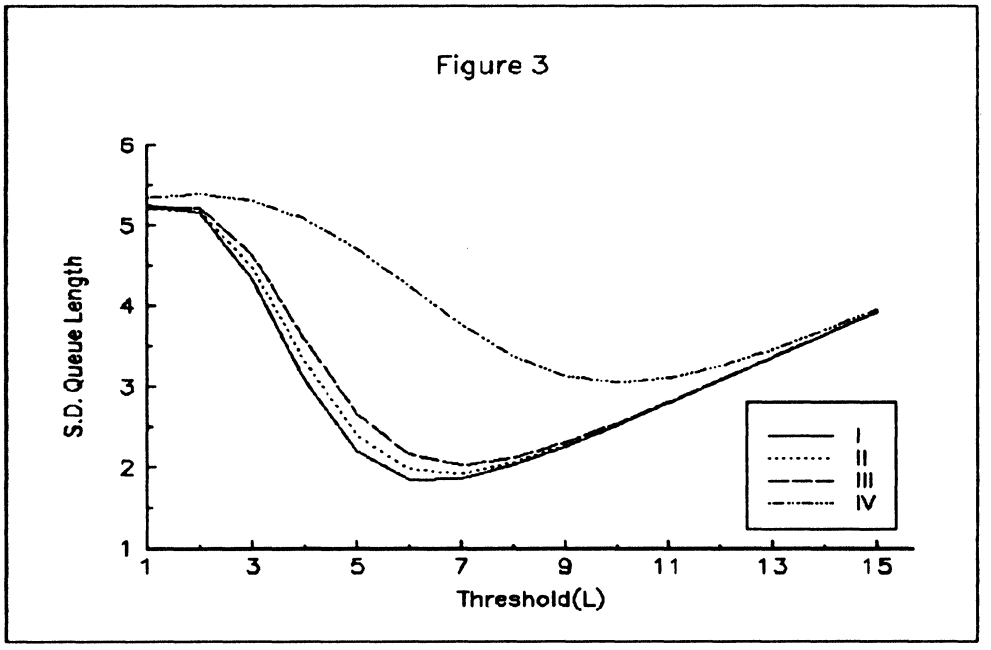


Figure 3
S.D. Queue Length
6
5
4
3
2
1
1
3
5
7
9
11
13
15
Threshold(L)
I
II
III
IV

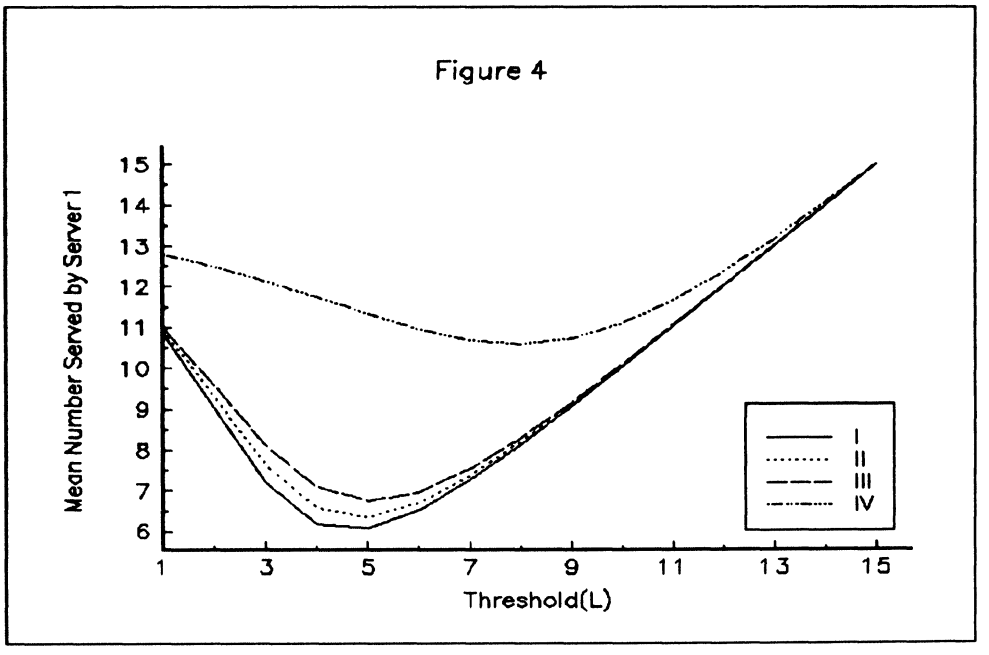
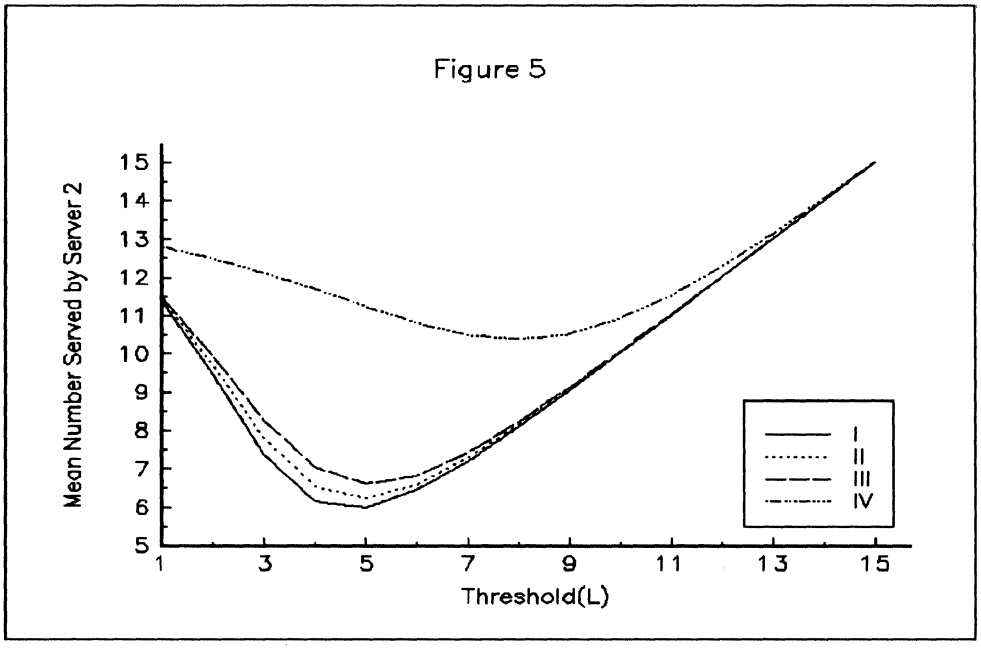


Figure 4
Mean Number Served by Server 1
15
14
13
12
11
10
9
8
7
6
1
3
5
7
9
11
13
15
Threshold(L)
I
II
III
IV



An examination of Figures 1-5 yield the following observations. The throughput, as expected, increases in all cases as L increases and then levels off at the value of 10.0 (the arrival rate). however, for the hyperexponential distribution the throughput started at a much lower value compared to the rest. The standard deviation of the queue length for the hyperexponential is also much higher than that of the other three arrival processes. However, the mean queue length was lower for low values of L and higher for higher values of L for the hyperexponential than for the other three. The coefficient of variation of the queue length was also considered as the performance measure of interest. Its value was higher for the hyperexponential than for the the other three. This behavior is understandable due to hyperexponential's high variance compared to the other three distributions. The mean numbers served by servers 1 and 2 are consistently higher for the hyperexponential than for the other three distributions.

Suppose that $\mu'_w(L)$ denotes the mean waiting time of an admitted customer ta points of arrival as a function of L. The purpose of the following example is to see the effect of the threshold on the mean waiting time distribution, by fixing all other parameters.

Example 3: We consider the following three distributions for the interarrival times:

A: Erlang of order 5 with parameter 50.

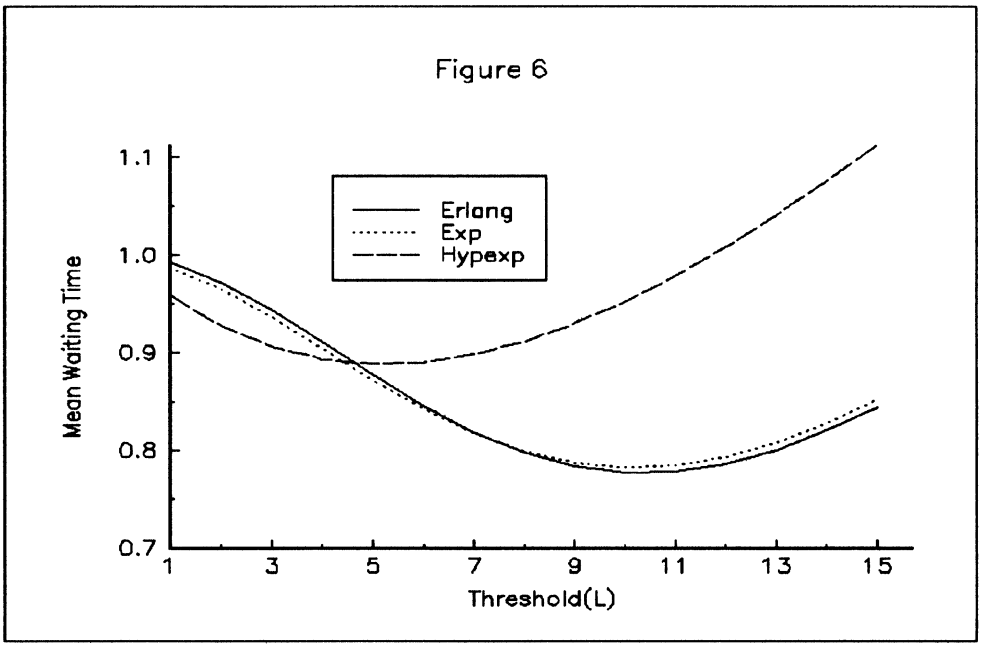
B: Exponential with parameter 10.

C: Hyperexponential with the mixing probabilities 0.85, 0.14 and 0.01. The respective exponentials have parameters 100, 10 and 4/31.

It can be verified that all these have the same mean 0.1. The service rates of both servers are assume to be 0.5 and do not depend on the number of customers served. The mean stationary waiting times of an admitted customer at points of arrivals for these three systems are plotted in Figure 6. Suppose L^* denotes the optimum value of L for which $\mu'_w(L)$ is minimum. An examination of this figure reveals the following observations.

(i) The L^* values are 11, 10 and 5 respectively.

(ii) The value of L^* appears to decrease with increasing variance of the arrival times, which we have also seen in many other types of examples we have run so far.



Our computational experience suggests the following conjecture, which is similar to the one proposed in Chakravarthy [3].

Conjecture: $\mu'_w(L)$ *is nonincreasing on* $\left[\ 1,\ \ L^*\ \right]$ *and is nondecreasing on* $\left[L^*,\ \ K\ \right]$.

If the above conjecture is valid, then finding L^* for a given system is very simple. We start computing the mean waiting time with $L = 1$ and continue until the mean waiting time starts to increase.

Acknowledgement

Thanks are due to a referee and the editor for their helpful suggestions that improved the presentation of the paper.

References

[1] Bellman, R.E., *Introduction to Matrix Analysis*, McGraw Hill, New York, NY 1960.

[2] Chakravarthy, S., Analysis of a finite MAP/G/1 queue with group services, *Queuing Systems: Theory and Applications* **13** (1993), 385-407.

[3] Chakravarthy, S., A finite capacity GI/PH/1 queue with group services, *Naval Research Logistics* **39** (1992), 345-357.

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

[5] Marcus, M. and Minc, H., *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, MA 1964.

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

[7] Neuts, M.F., *Matrix-geometric Solutions in Stochastic Models - An Algorithmic Approach*, Johns Hopkins University Press, Baltimore, MD 1981.

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

Appendix

$$u(0)C_0 + \sum_{j=L}^{K} [v_j(0)\mu_j^{(1)} + w_j(0)\mu_j^{(2)}] = \mathbf{0}, \tag{A1}$$

$$u(i)C_0 + u(i-1)C_1 + \sum_{j=L}^{K} [v_j(i)\mu_j^{(1)} + w_j(i)\mu_j^{(2)}] = \mathbf{0}, \quad 1 \le i \le L-1, \tag{A2}$$

$$pu(L-1)C_1 + v_L(0)[C_0 - \mu_L^{(1)}I] + \sum_{r=L}^{K} \mu_r^{(2)} y_{L,r}(0) = \mathbf{0}, \tag{A3}$$

$$v_k(0)[C_0 - \mu_k^{(1)}I] + \sum_{r=L}^{K} \mu_r^{(2)} y_{k,r}(0) = \mathbf{0}, \quad L+1 \le k \le K, \tag{A4}$$

$$v_k(i)[C_0 - \mu_k^{(1)}I] + v_k(i-1)C_1 + \sum_{r=L}^{K} \mu_r^{(2)} y_{k,r}(i) = \mathbf{0}, \quad L \le k \le K, \ 1 \le i \le L-1, \tag{A5}$$

$$qu(L-1)C_1 + w_L(0)[C_0 - \mu_L^{(2)}I] + \sum_{r=L}^{K} \mu_r^{(1)} y_{r,L}(0) = \mathbf{0}, \tag{A6}$$

$$w_k(0)[C_0 - \mu_k^{(2)}I] + \sum_{r=L}^{K} \mu_r^{(1)} y_{r,k}(0) = \mathbf{0}, L+1 \le k \le K, \tag{A7}$$

$$w_k(i)[C_0 - \mu_k^{(2)}I] + w_k(i-1)C_1 + \sum_{r=L}^{K} \mu_r^{(1)} y_{r,k}(i) = \mathbf{0}, L \le k \le K, 1 \le i \le L-1, \tag{A8}$$

$$\begin{aligned}&[\delta(k-L)v_j(L-1) + \delta(j-L)w_k(L-1)]C_1 - y_{j,k}(0)[\mu_j^{(1)} + \mu_k^{(2)}] \\ &+ y_{j,k}(0)C_0 + \sum_{r=L}^{K} [\mu_r^{(1)} y_{r,k}(j) + \mu_r^{(2)} y_{j,r}(k)] = \mathbf{0}, L \le j,k \le K,\end{aligned} \tag{A9}$$

where $\delta(\cdot)$ is the Kronecker delta defined as $\delta(x) = 1$ for $x = 0$; $\delta(x) = 0$, for $x \neq 0$, and for $L \le j$, $k \le K$, we have

$$y_{j,k}(i)[C_0 - \mu_j^{(1)}I - \mu_k^{(2)}I] + y_{j,k}(i-1)C_1 = \mathbf{0}, \quad 1 \le i \le K-1, \tag{A10}$$

and

$$\boldsymbol{y}_{j,k}(K)[C_0 - \mu_j^{(1)} I - \mu_k^{(2)} I] + [\boldsymbol{y}_{j,k}(K-1) + \boldsymbol{y}_{j,k}(K)]C_1 = \mathbf{0}. \tag{A11}$$

Equations (*A*1-*A*11) can be rewritten in a form that is well-suited for computation by (block) Gauss-Seidlel iterative procedure. Defining

$$\mu_{max}^{(r)} = max\,\{\mu_j^{(r)}: L \le j \le K\},\ M_r = [\mu_{max}^{(r)} I - C_0]^{-1},\ r = 1,2,$$

$$\mu_{max} = \mu_{max}^{(1)} + \mu_{max}^{(2)},\ \ M_3 = [\mu_{max} I - C_0]^{-1},$$

we can rewrite (*A*1-*A*11) as follows.

$$\boldsymbol{u}(0) = \left[\sum_{j=L}^{K} [\boldsymbol{v}_j(0)\mu_j^{(1)} + \boldsymbol{w}_j(0)\mu_j^{(2)}]\right](-C_0)^{-1},$$

$$\boldsymbol{u}(i) = \left[\boldsymbol{u}(i-1)C_1 + \sum_{j=L}^{K} [\boldsymbol{v}_j(i)\mu_j^{(1)} + \boldsymbol{w}_j(i)\mu_j^{(2)}]\right](-C_0)^{-1},\ 1 \le i \le L-1,$$

$$\boldsymbol{v}_L(0) = \left[p\boldsymbol{u}(L-1)C_1 + (\mu_{max}^{(1)} - \mu_L^{(1)})\boldsymbol{v}_L(0) + \sum_{r=L}^{K} \mu_r^{(2)} \boldsymbol{y}_{L,r}(0)\right] M_1,$$

$$\boldsymbol{v}_k(0) = \left[(\mu_{max}^{(1)} - \mu_k^{(1)})\boldsymbol{v}_k(0) + \sum_{r=L}^{K} \mu_r^{(2)} \boldsymbol{y}_{k,r}(0)\right] M_1,\ \ L+1 \le k \le K,$$

$$\boldsymbol{v}_k(i) = \left[(\mu_{max}^{(1)} - \mu_k^{(1)})\boldsymbol{v}_k(i) + \boldsymbol{v}_k(i-1)C_1 + \sum_{r=L}^{K} \mu_r^{(2)} \boldsymbol{y}_{k,r}(i)\right] M_1,\ L \le k \le K,\ 1 \le i \le L-1,$$

$$\boldsymbol{w}_L(0) = \left[q\boldsymbol{u}(L-1)C_1 + (\mu_{max}^{(2)} - \mu_L^{(2)})\boldsymbol{w}_L(0) + \sum_{r=L}^{K} \mu_r^{(1)} \boldsymbol{y}_{r,L}(0)\right] M_2,$$

$$\boldsymbol{w}_k(0) = \left[(\mu_{max}^{(2)} - \mu_k^{(2)})\boldsymbol{w}_k(0) + \sum_{r=L}^{K} \mu_r^{(1)} \boldsymbol{y}_{r,k}(0)\right] M_2,\ \ L+1 \le k \le K,$$

$$\boldsymbol{w}_k(i) = \left[(\mu_{max}^{(2)} - \mu_k^{(2)})\boldsymbol{w}_k(i) + \boldsymbol{w}_k(i-1)C_1 + \sum_{r=L}^{K} \mu_r^{(1)} \boldsymbol{y}_{r,k}(i)\right] M_2,\ L \le k \le K, 1 \le i \le L-1,$$

$$\boldsymbol{y}_{j,k}(0) = \left[[\delta(k-L)\boldsymbol{v}_j(L-1) + \delta(j-L)\boldsymbol{w}_k(L-1)]C_1 + \sum_{r=L}^{K} \mu_r^{(1)} \boldsymbol{y}_{r,k}(j)\right.$$

$$\left. + (\mu_{max} - \mu_j^{(1)} - \mu_k^{(2)})\boldsymbol{y}_{j,k}(0) + \sum_{r=L}^{K} \mu_r^{(2)} \boldsymbol{y}_{j,r}(k)\right] M_3,\ \ L \le j,\ k \le K,$$

and for $L \le j,\ k \le K$,

$$\boldsymbol{y}_{j,k}(i) = \left[(\mu_{max} - \mu_j^{(1)} - \mu_k^{(2)})\boldsymbol{y}_{j,k}(i) + \boldsymbol{y}_{j,k}(i-1)C_1\right] M_3, \quad 1 \leq i \leq K-1,$$

and

$$\boldsymbol{y}_{j,k}(K) = \left[(\mu_{max} - \mu_j^{(1)} - \mu_k^{(2)})\boldsymbol{y}_{j,k}(K) + [\boldsymbol{y}_{j,k}(K-1) + \boldsymbol{y}_{j,k}(K)]C_1\right] M_3.$$