Journal of Applied Mathematics and Stochastic Analysis, **16:**4 (2003), 327-347.
Printed in the USA ©2003 by North Atlantic Science Publishing Company

MULTI-THRESHOLD CONTROL OF THE $BMAP/SM/1/K$ QUEUE WITH GROUP SERVICES

ALEXANDER N. DUDIN
Belarussian State University
Department of Applied Mathematics and Computer Science
Minsk Belarus
E-mail: dudin@bsu.by

and

SRINIVAS R. CHAKRAVARTHY
Kettering University
Department of Industrial and Manufacturing Engineering and Business
Flint, MI USA
E-mail: schakrav@ketting.edu

(Received March 2002; Revised March 2003)

We consider a finite capacity queue in which arrivals occur according to a batch Markovian arrival process ($BMAP$). The customers are served in groups of varying sizes. The services are governed by a controlled semi-Markovian process according to a multi-threshold strategy. We perform the steady-state analysis of this model by computing (a) the queue length distributions at departure and arbitrary epochs, (b) the Laplace-Stieltjes transform of the sojourn time distribution of an admitted customer, and (c) some selected system performance measures. An optimization problem of interest is presented and some numerical examples are illustrated.

Key words: Batch Markovian Arrival Process, Semi-Markovian Service, Algorithmic Probability, Optimal Control.

AMS (MOS) subject classification: 90B22, 60K25

1 Introduction

One of the most popular and effective ways to model the flows of messages in modern communication networks or jobs in a production and manufacturing process is to use the batch Markovian arrival process ($BMAP$). The $BMAP$ is a rich class of point processes that includes many well-known processes such as Poisson, Switched Poisson, Markov-modulated Poisson, and PH-renewal processes. In recent years, there has been a constant and growing interest in the investigation of queues with $BMAP$ input [1, 6, 17, 41].

The behavior of such queues is described usually in terms of continuous or discrete-time Markov chains. These fall under the so-called $M/G/1$ type Markov chains [44] or the quasi-Toeplitz Markov chains [23]. These chains have two important properties. The first one is the Toeplitz-like structure for the (block) transition probability matrix governing the queueing system under study. That is, the $(i,l)^{th}$ block matrix $P_{i,l}$ of the one-step transition probability matrix of the denumerable Markov chain corresponding to the transition from state i to state l depends only on $l-i$, but not on i and l separately for $i > i_0$ where $i_0 \geq 0$ is some integer. The second property is the so-called non-skip free to the left property. That is, $P_{i,l} = 0$, for $0 \leq l < i-1$.

The presence of these two properties enables one to study the stationary distribution of the chain using matrix analytic methods. We refer the reader to [1, 6, 17, 23, 28, 32, 34, 36, 37, 44] for latest developments in this area.

Under this paradigm, the queueing systems $N/G/1$ [45], $M/SM/1$ [44], $BMAP/G/1$ [42], and $BMAP/SM/1$ [43] have been completely investigated. Several modifications of the $BMAP/G/1$ and $BMAP/SM/1$ queues are considered in the book [32].

The "non-skip free to the left" property is an important one in the development of the theory of Markov chains of the $M/G/1$ paradigm. If we do not assume this property (e.g., queues where negative arrivals take place or where services are in groups) the investigation of the Markov chain governing the corresponding queueing system becomes far more difficult (see, for example, [2, 3, 4, 35]). However, there are at least two interesting cases where the analytical investigation of the chain is still possible when the skip free to the left property is not satisfied.

The first case arises in the investigation of queues with disasters. Disaster is a special case of negative arrivals of customers. A disaster causes the removal of all customers (including those in service) from the system instantaneously. For more information on such queueing systems we refer the reader to the survey paper [2]. A Markov chain that describes the queue with disasters has non-zero blocks $P_{i,0}$ for any $i \geq 0$. Hence, such chains do not belong to the class of quasi-Toeplitz type Markov chains. However, the stationary distribution of the queue length can be obtained in a rather nice analytical form (see, for example, [27, 29, 33]).

The second case is where the customers are offered services in groups of varying sizes. In the context of a finite capacity queueing system with finite buffer of size K, a different type of service scheme in which services are offered to groups of varying size, ranging from a predetermined threshold L to the maximum buffer size K, was introduced in [7]. The pre-assigned number $L \geq 1$, called the *threshold*, operates as follows. An idle server finding fewer than L customers in the queue remains idle. However, when i, $L \leq i \leq K$, customers are present, the idle server initiates a service for the entire group. If a service has to be initiated through an arrival of a batch (for example, i jobs, $i < L$, are in the queue with the server being idle, and the arriving batch has at least $L-i$ jobs) only a group of size no larger than K enter into service, and any remaining customers in the batch are considered lost. Service schemes of this type in the context of finite capacity $GI/PH/1$ and $MAP/G/1$ with single arrivals, and $BMAP/M/c$ models were investigated in the papers [7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. Some potential applications of this type of service mechanism in manufacturing processes were outlined in those papers.

Optimization problems in queueing theory play an important role in practice. They can be broadly divided into two groups. One group deals with static optimization, where some system parameters must be tuned to provide the desired quality of service to customers. For example, A.K. Erlang investigated the $M/M/N/0$ model and when the arrival and service intensities are fixed, he calculated the minimal number of channels needed to guarantee that the loss probability does not exceed a prespecified value.

The other group of optimization problems that falls under dynamic optimization involves the control of some parameters of the queueing system such as the arrival rate, the service rate, and the number of channels so as to optimize a given objective function. This class of problems is more complicated and challenging as the objective function is a (nonlinear) function of the steady-state probabilities and for most interesting practical queueing models these probabilities will not be known explicitly. The traditional way of solving such optimal control problems in queues is to use Markov decision processes. However, this approach is not very powerful even in the case of classical queueing systems. Furthermore, in the case of Markov chains describing queues with $BMAP$ input and SM-services there is the *curse of dimensionality* problem. Thus, it appears that when parametric strategies of control such as threshold, hysteretic, and randomization are used in queueing models, the problem of finding an optimal control is better solved by a direct approach. This approach involves an iterative process consisting of three stages:

(1) Computation of the stationary queue length distribution for a given set of values for controlled parameters,

(2) Evaluation of the objective function, and

(3) A search for optimal values for the controlled parameters using an efficient heuristic approach.

Such an approach was used in analyzing the $BMAP/G/1$ system with a multi-threshold service rate control with N service modes in [26], for the $BMAP/PH/1/K$ queue with hysteretic control in [5], for the $BMAP/G/1$ queue with hysteretic control in [30], and for the $BMAP/SM/1/K$ queue with hysteretic control in [31].

In this paper, we extend this approach to the $BMAP/SM/1/K$ system with group services. This generalizes the results of the paper [5] in two aspects. The first one is the assumption of availability of $J, J \geq 2$, service modes (in [5] $J = 2$). Secondly, we assume SM-type service time distributions while in [5] PH-type services were used. Furthermore, here we make a more realistic assumption about the service switching mechanism in that we allow the switching only at the beginning of a service epoch. In [5] the service rates can be changed at any epoch using a hysteretic type control.

The paper is organized as follows. In Section 2 the mathematical model and the service control mechanism are described. The stationary queue length distribution embedded at service completion epochs is investigated in Section 3 and the stationary queue length distribution at an arbitrary time point is obtained in Section 4. In Section 5 some key system performance measures describing the queueing model are presented along with their formulas. The stationary sojourn time distribution in the system of an admitted customer is derived in Section 6. Section 7 contains a brief description of an optimization problem and some illustrative numerical examples are presented in Section 8.

2 The Mathematical Model

We consider a single server queue with a limited buffer of size K. Arrivals of customers occur according to a batch Markovian arrival process ($BMAP$), a tractable class of Markov renewal process. Customers in a batch are admitted to the extent of buffer availability and all others are considered lost. That is, we allow "partial admission" of a batch. We assume that the behavior of the $BMAP$ is governed by the directing process $\{\nu_t, t \geq 0\}$. This process is an irreducible continuous-time Markov chain with the finite state space $\{1, \ldots, W\}$. Suppose that the matrix D_0 governs the transitions corresponding to no arrivals, and the matrices $D_k, k \geq 1$ govern the transitions corresponding to arrivals of batch size $k, k \geq 1$. By assuming D_0 to be a nonsingular matrix, the interarrival times will be finite with probability one and the arrival process does not terminate. Thus, we see that D_0 is a stable matrix. Let N_t denote the number of arrivals in (0, t] of the $\{\nu_t\}$ process. In the sequel, we need the matrices $P(n,t), n \geq 0, t \geq 0$ where the $(j,k)^{th}$ entry of the matrix $P(n,t)$ is defined as $P_{i,j}(n,t)$ $= P(N_t = n, \nu_t = j | N_0 = 0, \nu_0, = i)$. Let

$$\sum_{n=0}^{\infty} P(n,t) z^n = e^{D(z)t}, \tag{2.1}$$

where

$$D(z) = \sum_{k=0}^{\infty} D_k z^k, \ |z| < 1.$$

For use in the sequel, let $\mathbf{0}$ and $\mathbf{1}$ denote, respectively, the row and column vectors of 0's and 1's with appropriate size. The matrix $D(1)$ is the infinitesimal generator of the chain $\{\nu_t, t \geq 0\}$. Let $\mathbf{u}$ denote the stationary vector of this generator. That is, $\mathbf{u}D(1) = \mathbf{0}$, $\mathbf{u1} = \mathbf{1}$. The (group) arrival rate, λ_g, and the fundamental rate (or the average intensity), λ, of the $BMAP$ are defined as:

$$\lambda_g = \mathbf{u}(D(1) - D_0)\mathbf{1},$$

and

$$\lambda = \mathbf{u}D'(1)\mathbf{1}.$$

For full details on $BMAP$ we refer the reader to [32, 42, 44] and for a review and recent work on $BMAP$ we refer the reader to [17].

We assume that the service is offered to groups of varying size $i, L \leq i \leq K$. Here L is some pre-assigned number, $L \geq 1$. If at the completion of a service fewer than L customers are present, the server waits until the queue length reaches at least L and then initiates a service to all those customers present in the system.

In order to attract and serve customers efficiently it may be necessary to control the service rate (or the service time distribution) according to the number of customers in the queue. We assume that the (group) service times of the customers follow an irreducible finite state semi-Markov process $\{m_t, t \geq 0\}$ with state space $\{1, 2, \ldots, M\}$. The sojourn times of the services are given by the entries of the kernel $B(t) = ||B_{m,m'}(t)||$ of dimension M. In this paper we categorize the services into J, $J \geq 2$, modes of operation. During the j^{th} mode, the service times are given by the entries of the kernel $B^{(j)}(t) = ||B^{(j)}_{m,m'}(t)||$ of dimension M. In other words, while the services of all groups are directed by the common process $\{m_t, t \geq 0\}$, the transitions of this process are governed by different kernels depending on the type of strategy control used. We further assume that the semi-Markov services times have finite second moments. Note that independent and identically distributed service time is a special case of the SM-service. In this case, the service time of the group is characterized by the distribution function $B^{(j)}(t), 1 \leq j \leq J$.

We consider a multi-threshold strategy in this paper. This is defined by the integer-valued thresholds $I_1, \ldots, I_{J-1}$. First assume that $L - 1 = I_0 \leq I_1 \leq \ldots \leq I_{J-1} \leq I_J = K$. If the number i of customers in the system at a service completion satisfies the inequality $I_{j-1} < i \leq I_j$, then the entire group of i customers is served in the j^{th} mode, $1 \leq j \leq J$. If the number of customers i is less than L, no service is offered until the queue builds up to L or more. In the latter case, we consider two possible variants for selecting the service mode for the next group. In **Variant 1** the service for the next group will always be in mode 1 and in **Variant 2** the service mode for the next group is determined at the beginning of the service according to the multi-threshold strategy. Note that in the case of MAP (which corresponds to single arrivals) these two variants coincide. Other types of variants for choosing the service mode can easily be incorporated without any additional complexity and the details are omitted. In the case when the thresholds are equal, the number of service modes is reduced appropriately. Thus, if r thresholds are equal then the number of service modes will be reduced by $r - 1$ to $J - r + 1$ and the service modes will be relabelled as 1, 2,..., $J - r + 1$.

In situations where there are various costs associated with the waiting of customers as well as for providing faster modes of services, one of the most popular classes of the parametric strategies is the class of the multi-threshold strategies. The optimality of such multi-threshold strategies in the class of all homogeneous Markovian strategies is proven only in some particular cases (see, e.g., [18, 46]). However, such strategies are practical as well as reasonable for numerical implementations. Various queueing systems, such as $M/M/N$ [19], $GI|M|1$ [22, 38], $M|G|1$ [20, 21, 39], $BMAP|G|1$ [26], $BMAP|SM|1$ (with retrials) [25], and $BMAP|SM|1|N$ [24], with multi-threshold strategies of control have been analyzed in the literature. For the type of group services introduced in [7], to our knowledge this is the first time such a multi-threshold policy is considered.

3 Embedded Queue Length Distribution

Let t_n denote the epoch of the n^{th} service completion; i_n the number of customers in the queue at $t_n + 0, i_n \geq 0; \nu_n$ the state of the $BMAP$ process, $\{\nu_t\}$, at $t_n, 1 \leq \nu_n \leq W$; and m_n be the state of the semi-Markovian process, $\{m_t\}$, that governs the service process at $t_n + 0$, for $1 \leq m_n \leq M, n \geq 1$. When the server becomes idle immediately after a service completion due to not having enough customers in the queue, we assume that the phase at $t_n + 0$ will be the phase of the next service to be initiated. That is, the phase of the service process will be frozen at the instant when the server becomes idle. We can modify this scheme to allow the phase to be chosen according to some initial probability vector, but this will not be addressed in this paper.

Let the thresholds $I_1, \ldots, I_{J-1}$ be fixed and predetermined. It is easy to verify that the process $\{(i_n, \nu_n, m_n), n \geq 1\}$ is a three-dimensional Markov chain. The assumptions of our model imply that this Markov chain has a stationary state distribution.

Define

$$\pi(i,\nu,m) = \lim_{n\to\infty} P\{i_n = i, \nu_n = \nu, m_n = m$$

$$i_0, \nu_0, m_0\}, \quad i \geq 0, 1 \leq \nu \leq W, 1 \leq m \leq M.$$

Enumerating the states of the Markov chain (i_n, ν_n, m_n) in lexicographic order, we form the vectors $\vec{\pi}(i)$, of dimension WM, for $0 \leq i \leq K$, of stationary probabilities.

We now define a number of auxiliary quantities that we need in the sequel. Let $\Omega_i^{(j)}$ denote the matrix of probabilities that during the service of a group of customers in the j^{th} service mode exactly i customers arrive. That is,

$$\Omega_i^{(j)} = \int_0^\infty P(i,t) \otimes dB^{(j)}(t), i \geq 0, 1 \leq j \leq J, \tag{3.1}$$

and

$$\tilde{\Omega}_K^{(j)} = \sum_{r=K}^{\infty} \Omega_r^{(j)} = \int_0^\infty e^{D(1)t} \otimes dB^{(j)}(t) - \sum_{r=0}^{K-1} \Omega_r^{(j)}, 1 \leq j \leq J. \tag{3.2}$$

Let $X^{(i)}$ denote the *first passage* probabilities of going from level 0 to level i or higher. That is, the components of $X^{(i)}$ give the probabilities that the process $\{N_t, \nu_t, m_t\}$ reaches level i or higher for the first time from level 0. It is easy to verify that $X^{(i)}$ is given by

$$X^{(i)} = \sum_{l=0}^{i-1} \int_0^\infty P(l,t)dt(D(1) - \sum_{r=0}^{i-l-1} D_r) \otimes I_M, i \geq 1. \tag{3.3}$$

Define $Y_m^{(l)}$, for $L \leq l \leq K, 0 \leq m \leq L-1$, to be the matrices of probabilities that the process $\{N_t, \nu_t, m_t\}$ reaches level l starting from level m. These matrices are given by

$$Y_m^{(l)} = \sum_{i=0}^{L-m-1} \int_0^\infty P(i,t) D_{l-m-i} dt \otimes I_M, \; L \leq l < K-1, 0 \leq m \leq L-1, \tag{3.4}$$

and

$$Y_m^{(K)} = \sum_{i=0}^{L-m-1} \int_0^\infty P(i,t)(D(1) - \sum_{r=0}^{K-m-i-1} D_r) \otimes I_M dt, \; 0 \leq m \leq L-1. \tag{3.5}$$

Lemma 1: *The vectors $\vec{\pi}(i)$, $i \geq 0$ satisfy the following equations:*

$$\vec{\pi}(i) = \sum_{r=0}^{L-1} \vec{\pi}(r) X^{(L-r)} \Omega_i^{(1)} + \sum_{j=1}^{J} \sum_{r=I_{j-1}+1}^{I_j} \vec{\pi}(r) \Omega_i^{(j)}, 0 \leq i \leq K-1, \tag{3.6}$$

$$\vec{\pi}(K) = \sum_{r=0}^{L-1} \vec{\pi}(r) X^{(L-r)} \tilde{\Omega}_K^{(1)} + \sum_{j=1}^{J} \sum_{r=I_{j-1}+1}^{I_j} \vec{\pi}(r) \tilde{\Omega}_K^{(j)}, \tag{3.7}$$

for Variant 1 (where service is always in mode 1 once the group size builds to L or more); and for Variant 2 (where the mode of service is determined at the beginning of a service after the group size builds to L or more) the equations are

$$\vec{\pi}(i) = \sum_{r=0}^{L-1} \vec{\pi}(r) \sum_{l=L}^{K} Y_r^{(l)} \Omega_i^{(\chi(l))} + \sum_{j=1}^{J} \sum_{r=I_{j-1}+1}^{I_j} \vec{\pi}(r) \Omega_i^{(j)}, 0 \le i \le K-1, \tag{3.8}$$

$$\vec{\pi}(K) = \sum_{r=0}^{L-1} \vec{\pi}(r) \sum_{l=L}^{K} Y_r^{(l)} \tilde{\Omega}_K^{(\chi(l))} + \sum_{j=1}^{J} \sum_{r=I_{j-1}+1}^{I_j} \vec{\pi}(r) \tilde{\Omega}_K^{(j)}. \tag{3.9}$$

The quantity $\chi(l)$ is equal to j if $I_{j-1}+1 \le l \le I_j$, $L \le l \le K$, $1 \le j \le J$, and $\chi(l)$ is equal to J for $l > K$.

Proof: The proof follows directly from the law of total probability.

Theorem 1: *The stationary probability vectors $\vec{\pi}(i)$ are calculated as follows:*

$$\vec{\pi}(i) = \sum_{j=1}^{J} \mathbf{v}_j \Omega_i^{(j)}, 0 \le i \le K-1, \tag{3.10}$$

$$\vec{\pi}(K) = \sum_{j=1}^{J} \mathbf{v}_j \tilde{\Omega}_K^{(j)}, \tag{3.11}$$

where the vector $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_J)$ is the unique solution to the system

$$\mathbf{v} = \mathbf{v}A, \tag{3.12}$$

$$\mathbf{v}\mathbf{1} = 1. \tag{3.13}$$

The matrix A is defined as follows. In the case of Variant 1, the entries A_{rj} are given by:

$$A_{rj} = \begin{cases} \Omega^{(r)}(I_0+1, I_1) + \sum\limits_{i=0}^{L-1} \Omega_i^{(r)} X^{(L-i)}, \quad 1 \le r \le J, \quad j = 1, \\ \\ \Omega^{(r)}(I_{j-1}+1, I_j), \quad 1 \le r \le J, \quad 2 \le j \le J, \end{cases} \tag{3.14}$$

with

$$\Omega^{(r)}(I_{j-1}+1, I_j) = \begin{cases} \sum\limits_{i=I_{j-1}+1}^{I_j} \Omega_i^{(r)}, \ 1 \le j \le J-1, \\ \\ \sum\limits_{i=I_{j-1}+1}^{\infty} \Omega_i^{(r)}, \ j = J. \end{cases} \tag{3.15}$$

For Variant 2, the entries A_{rj} are of the form:

$$A_{rj} = \sum_{k=0}^{L-1} \Omega_k^{(r)} \sum_{l=I_{j-1}+1}^{I_j} Y_k^{(l)} + \Omega^{(r)}(I_{j-1}+1, I_j), 1 \le r, j \le J. \tag{3.16}$$

Proof: We will prove for Variant 2 as the proof is similar for the other case. Defining

$$\mathbf{v}_j = \sum_{k=0}^{L-1} \vec{\pi}(k) \sum_{l=I_{j-1}+1}^{I_j} Y_k^{(l)} + \sum_{k=I_{j-1}+1}^{I_j} \vec{\pi}(k), \quad 1 \le j \le J, \tag{3.17}$$

and substituting (3.17) into (3.8) and (3.9) we get expressions (3.10) and (3.11). Thus, to complete the proof we have to show that the vectors $\mathbf{v}_j$, $1 \le j \le J$ satisfy system of equations given in (3.12) and (3.13) for Variant 2. To this end, we first multiply equation (3.10) by $\sum\limits_{l=I_{j-1}+1}^{I_j} Y_i^{(l)}$ and add over $0 \le i \le L-1$. When this is added to the sum of equations (11) over $I_{j-1} \le i \le I_j$, we get

$$\mathbf{v}_j = \sum_{r=1}^{J} \mathbf{v}_r A_{rj}, \quad 1 \le j \le J, \tag{3.18}$$

where the matrices A_{rj} are defined by formula (3.16). The stated result follows immediately.

Remarks:

(1) Note that the vectors $\vec{\pi}(i)$, $0 \le i \le K$ can be calculated directly by solving the system of linear equations in (3.6) and (3.7) (or (3.8) and (3.9)). However, when $J < K$, use of Theorem 1 will reduce the computational efforts.

(2) The entries of the matrices A_{rj} of dimension WM have a very nice probabilistic interpretation. Partitioning the matrix A_{rj} into W blocks of M by M matrices, the $(l, l')^{th}$ entry of the $(k, k')^{th}$ block matrix gives the transition probability that the next service will start in mode j with phase l' and at that instant the arrival process will be in phase k' given that the current service is in mode r with phase l and the arrival process is in phase k.

4 Stationary Queue Length Distribution at Arbitrary Time Points

Let $\mathbf{z}(0, i)$, $0 \le i \le L-1$, be the row vector of size WM whose entries $z(0, i, j, k)$ give the steady-state probabilities that at an arbitrary time there are i customers in the queue with the arrival process in phase j and the server became idle from service phase k. Similarly, let $\mathbf{z}(n, i)$, for $L \le n \le K, 0 \le i \le K$, $1 \le j \le W, 1 \le m \le M$, be the vector whose entries $z(n, i, j, k)$ define the steady-state probability that at an arbitrary time there are i customers in the queue with the arrival process in phase j and the current service for a batch of n customers is in phase k. To derive an expression relating the steady state probabilities at an arbitrary time and at embedded epochs, first we need the following lemma dealing with the mean interdeparture time.

Lemma 2: *Suppose that τ denotes the mean interdeparture time. It is calculated as*

$$\tau = \left[\sum_{i=0}^{L-1} \vec{\pi}(i) \hat{G}^{(L-i)} + \sum_{j=1}^{J} \mathbf{v}_j \left(I_W \otimes \int_0^\infty t dB^{(j)}(t) \right) \right] \mathbf{1}, \tag{4.1}$$

where

$$\hat{G}^{(L-i)} = \sum_{l=0}^{L-i-1} \int_0^\infty tP(l,t)(D(1) - \sum_{r=0}^{L-i-l-1} D_r)dt \otimes I_M. \tag{4.2}$$

Proof: The proof follows immediately from the law of total probability on noting that the entries of the matrix $\hat{G}^{(L-i)}$ give the (conditional) mean time until the beginning of the next service for a group of L or more customers given that that the previous service left behind i, $0 \le i \le L-1$, customers in the queue.

Theorem 2: *The stationary-state probability vectors* $\mathbf{z}(n,i)$ *in Variant 1 are obtained as follows:*

$$\mathbf{z}(0,i) = \tau^{-1} \sum_{l=0}^{i} \vec{\pi}(l) \int_0^\infty P(i-l,t)dt \otimes I_M, 0 \le i \le L-1, \tag{4.3}$$

$$\mathbf{z}(n,i) = \tau^{-1}\Big[\sum_{l=0}^{L-1} \vec{\pi}(l) \sum_{k=0}^{L-l-1} \int_0^\infty P(k,t)D_{n-l-k}dt \otimes I_M \int_0^\infty \tilde{P}(i,t) \otimes (I_M - \nabla_B^{(1)}(t))dt$$

$$+\vec{\pi}(n) \int_0^\infty \tilde{P}(i,t) \otimes (I_M - \nabla_B^{(\chi(n))}(t))dt\Big], \quad L \le n \le K-1, \ 0 \le i \le K, \tag{4.4}$$

$$\mathbf{z}(K,i) = \tau^{-1}\Big[\sum_{l=0}^{L-1} \vec{\pi}(l) \sum_{k=0}^{L-l-1} \int_0^\infty P(k,t) \sum_{r=K-l-k}^{\infty} D_r dt \otimes I_M \int_0^\infty \tilde{P}(i,t) \otimes (I_M - \nabla_B^{(1)}(t))dt$$

$$+\vec{\pi}(K) \int_0^\infty \tilde{P}(i,t) \otimes (I_M - \nabla_B^{(\chi(K))}(t))dt\Big], \quad 0 \le i \le K, \tag{4.5}$$

where

$$\tilde{P}(i,t) = \begin{cases} P(i,t), & \text{if} \quad i < K, \\ \\ \sum\limits_{l=K}^{\infty} P(l,t), & \text{if} \quad i = K, \end{cases}$$

and $\nabla_B^{(j)}(t)$ *is the diagonal matrix with the diagonal entries defined by the vector* $B^{(j)}(t)\mathbf{1}$, $1 \le j \le J$, *and the indicator function* $\chi(n)$ *is as defined in Lemma* 1.

In Variant 2, *the formulas are corrected by means of replacing* $\nabla_B^{(1)}(t)$ *with* $\nabla_B^{(\chi(n))}(t)$ *in* (4.4) *and* $\nabla_B^{(\chi(K))}(t)$ *in* (4.5).

Proof: Proof follows from a classical argument based on the key renewal theorem (see e.g. [40]).

Remark: Note that the arbitrary time stationary probabilities depend implicitly on the type of variant considered for initiating a service while waiting for the queue size to build up to L or more, through the vectors $\vec{\pi}(l)$.

Corollary 1: *Let $\vec{\eta}_i$ be the vector such that the $(j,k)^{th}$ component gives the probability that at an arbitrary time there are i customers in the system and that the arrival and service processes are respectively in phases j and k. Then we have*

$$\vec{\eta}_i = \mathbf{z}(0,i) + \sum_{l=L}^{min\{i,K\}} \mathbf{z}(l, i-l), \quad 0 \leq i \leq 2K. \tag{4.6}$$

Note that $\mathbf{z}(0,i) = \mathbf{0}$ for $i \geq L$ and in the above summation when the lower bound is greater than the upper bound the value is set to zero.

Corollary 2: *The probability vector, $\mathbf{q}(n)$, $n \geq L$, of seeing n customers at the beginning of a service is obtained as*

$$\mathbf{q}(n) = \tau^{-1}\Big[\sum_{l=0}^{L-1} \vec{\pi}(l) \sum_{k=0}^{L-l-1} \int_0^\infty P(k,t) D_{n-l-k} dt \otimes I_M + \vec{\pi}(n)\Big], L \leq n \leq K. \tag{4.7}$$

5 Selected System Performance Measures

In this section we give a number of system performance measures and their respective formulas useful in qualitative interpretation of the model.

(1) The probability that an arriving customer will be lost is given by

$$P_{reject} = 1 - \lambda^{-1}\Big[\sum_{i=0}^{L-1} \mathbf{z}(0,i) \sum_{k=0}^{K-i} (k+i-K) D_k \otimes I_M$$
$$+ \sum_{n=L}^{K} \sum_{i=0}^{K} \mathbf{z}(n,i) \sum_{k=0}^{K-i} (k+i-K) D_k \otimes I_M\Big]\mathbf{1}. \tag{5.1}$$

(2) Let ζ_j, for $j \geq 0$, denote the probability that exactly j customers are lost at an arrival epoch. Then we have

$$\zeta_j = {\lambda_g}^{-1}\Big[\sum_{i=0}^{L-1} \mathbf{z}(0,i)(D_{K-i+j}\mathbf{1} \otimes \mathbf{1}) + \sum_{n=L}^{K} \sum_{i=0}^{K} \mathbf{z}(n,i)(D_{K-i+j}\mathbf{1} \otimes \mathbf{1})\Big], j \geq 0, \tag{5.2}$$

where λ_g is the (group) arrival rate of the $BMAP$.

(3) The mean number of lost customers at an arrival epoch, μ_{NL}, is calculated as

$$\mu_{NL} = \sum_{j=1}^{\infty} j\zeta_j. \tag{5.3}$$

(4) The *throughput* of the system is given by $\lambda(1 - P_{reject})$.

(5) Denoting by θ_0 the fraction of time the server is idle and by θ_r, $1 \leq r \leq J$, the fraction of time the server is busy serving customers in r^{th} mode, we have

$$\theta_0 = \tau^{-1} \sum_{i=0}^{L-1} \vec{\pi}(i) \hat{G}^{(L-i)} \mathbf{1}, \tag{5.4}$$

$$\theta_r = \tau^{-1}\mathbf{v}_r(I_W \otimes \int_0^\infty tdB^{(r)}(t))\mathbf{1}, \ \ 2 \le r \le J, \tag{5.5}$$

$$\theta_1 = \tau^{-1}\sum_{i=0}^{I_1}\vec{\pi}(i)(I_W \otimes \int_0^\infty tdB^{(1)}(t))\mathbf{1}, \tag{5.6}$$

in the case of Variant 1 and

$$\theta_1 = \tau^{-1}\mathbf{v}_1(I_W \otimes \int_0^\infty tdB^{(1)}(t))\mathbf{1} \tag{5.7}$$

in the case of Variant 2.

(6) The mean number of customers in the queue, μ_{QL}, is given by

$$\mu_{QL} = \sum_{i=0}^{K} i\Big[\mathbf{z}(0,i)\mathbf{1} + \sum_{n=L}^{K}\mathbf{z}(n,i)\mathbf{1}\Big]. \tag{5.8}$$

(7) Suppose that S_{rj}, for $r \ne j$, $1 \le r, j \le J$, denotes the average number of service switches per unit of time. Then from the probabilistic interpretation of the quantities $\mathbf{v}_r$ and A_{rj}, we have

$$S_{rj} = \tau^{-1}\mathbf{v}_r A_{rj}\mathbf{1}, r \ne j, 1 \le r, j \le J, \tag{5.9}$$

where A_{rj} is as defined in (3.14) for Variant 1 and (3.16) for Variant 2.

6 Sojourn Time Distribution

Let $\mathbf{V}(x)$ be the vector distribution function of the sojourn time in the system of an admitted customer at an arrival epoch. We will call this admitted customer a *tagged* customer. Partitioning $\mathbf{V}(x)$ into $\mathbf{V}(x) = (v_{1,1}(x), ..., v_{1,M}(x), ..., v_{W,1}(x), ..., v_{W,M}(x))$, the entry of $v_{j_1,j_2}(x)$ gives the distribution function of the sojourn time of the tagged customer given that the arrival process is in state j_1 and the current service is in phase j_2. Recall that if the server is idle then the new service will start in phase j_2. Let $\mathbf{v}(s) = \int_0^\infty e^{-sx}d\mathbf{V}(x)$, $Re\ s > 0$, be the Laplace-Stieltjes transform of $\mathbf{V}(x)$.

Before we derive an expression for $\mathbf{v}(s)$ for the tagged customer, we need the following result. Suppose $\xi_{l,r}, 0 \le l \le L-1, L \le r \le K$, denotes the sojourn time from the epoch when the server is idle with l customers waiting in the system to the epoch when the service starts for a group of r customers.

Let $F_{l,r}(x)$ be the matrix whose $(j,j')^{th}$ entry is the conditional distribution function:

$$P(\xi_{l,r} \le x, \nu_t = j'|\nu_0 = j).$$

Let $\phi_{l,r}(s) = \int_0^\infty e^{-sx}dF_{l,r}(x)$ and $\Phi(s) = ||\phi_{l,r}(s)||_{l=\overline{0,L-1},\ r=\overline{L,K}}$.

Lemma 3: *The (matrix) Laplace-Stieltjes transform* $\Phi(s)$ *is given by*

$$\Phi(s) = (I - \Delta_2(s))^{-1}\Delta_1(s), \tag{6.1}$$

where $\Delta_r(s)= (I_L \otimes (-D_0 + sI)^{-1})\ \tilde{\Delta}_r$ *with*

$$\tilde{\Delta}_1 = \begin{pmatrix} D_L & D_{L+1} & \cdots & D_{K-1} & \sum\limits_{m=K}^{\infty} D_m \\ D_{L-1} & D_L & \cdots & D_{K-2} & \sum\limits_{m=K-1}^{\infty} D_m \\ \vdots & \vdots & & \vdots & \vdots \\ D_1 & D_2 & \cdots & D_{K-L} & \sum\limits_{m=K-L+1}^{\infty} D_m \end{pmatrix},$$

and

$$\tilde{\Delta}_2 = \begin{pmatrix} 0 & D_1 & D_2 & \cdots & D_{L-1} \\ 0 & 0 & D_1 & \cdots & D_{L-2} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & D_1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Proof: The proof follows immediately on noting that

$$dF_{l,r}(x) = e^{D_0 x}D_{r-l}dx + \int\limits_0^x e^{D_0 y}\sum_{m=1}^{L-l-1} D_m dy dF_{l+m,r}(x-y), \quad L \le r \le K-1,$$

$$dF_{l,K}(x) = e^{D_0 x}\sum_{m=K-l}^{\infty} D_m dx + \int\limits_0^x e^{D_0 y}\sum_{m=1}^{L-l-1} D_m dy dF_{l+m,K}(x-y), \quad 0 \le l \le L-1.$$

In order to derive an expression for $\mathbf{v}(s)$ for the tagged customer, we need to consider two cases.

Case 1: Suppose that the tagged customer is part of a batch of k customers at the epoch when the server is idle and when i, $0 \le i \le L-1$, customers are waiting in the queue. In this case we need to consider the following three scenarios depending on the size of k:

Case 1A: Suppose that k is such that $L \le i+k \le K$. In this case the tagged customer and the others enter service immediately and the service time distribution function is given by $B^{(\chi(k+i))}(t)$.

Case 1B: Suppose that k is such that $1 \le i+k \le L-1$. In this case the tagged customer waits for a random duration, $\xi_{k+i,r}$ until the number of customers in the system reaches r, $L \le r \le K$, before entering into service. The service time of the tagged customer will then be given by $B^{(\chi(r))}(t)$.

Case 1C: Suppose that k is such that $i+k > K$. In this case the tagged customer along with $K-i-1$ customers will enter into service immediately and their service time is governed by $B^{(J)}(t)$.

Case 2: Suppose that the tagged customer arrives as part of a batch of k customers at the epoch when the server is busy with m customers and that there are i customers waiting in the queue. In this case, we need to consider the following two scenarios.

Case 2A: Suppose that k is such that $i + k \leq K$. Suppose that l customers arrive during the residual service time of the current service. Note that the current service is governed by $B^{(\chi(m))}(t)$. If $i + k + l \geq L$, then the service time of the tagged customer is given by $B^{(\chi(i+k+l))}(t)$. If $i + k + l < L$, then the tagged customer waits for a duration $\xi_{i+k+l,r}$, $L \leq r \leq K$ before entering service which is governed by $B^{(\chi(r))}(t)$.

Case 2B: Suppose that k is such that $i + k > K$. In this case the tagged customer along with $K-i-1$ customers will enter into service immediately and their service time is governed by $B^{(J)}(t)$.

Combining all of the above scenarios, the following relation holds good for $x > 0$:

$$\begin{aligned}
d\vec{V}(x) = \lambda^{-1}\Bigg\{ &\sum_{i=0}^{L-1} \mathbf{z}(0,i)\Bigg[\sum_{k=1}^{L-i-1} kD_k \otimes I \sum_{r=L}^{K} \int_0^x d_x F_{i+k,r}(x-y) \otimes dB^{(\chi(r))}(y) \\
&+ \sum_{k=L-i}^{K-i} kD_k \otimes dB^{(\chi(k+i))}(x) + \sum_{k=K-i+1}^{\infty} (K-i) D_k \otimes dB^{(J)}(x)\Bigg] \\
&+ \sum_{m=L}^{K} \vec{q}(m)\Bigg[\sum_{i=0}^{L-1} \int_0^\infty P(i,u) \sum_{k=1}^{L-i-1} kD_k du \otimes I \sum_{l=0}^{L-i-k-1} \int_0^x P(l,y) \otimes d_y B^{(\chi(m))}(u+y) \cdot \\
&\quad \cdot \sum_{r=L}^{K} \int_0^{x-y} d_x F_{i+k+l,r}(x-y-v) \otimes dB^{(\chi(r))}(v) \\
&+ \sum_{i=0}^{K} \int_0^\infty P(i,u) \sum_{k=1}^{K-i} kD_k du \otimes I \sum_{l=L-l-k}^{\infty} \int_0^x P(l,y) \otimes d_y B^{(\chi(m))}(y+u) \cdot \\
&\quad \cdot I \otimes d_x B^{(\chi(i+k+l))}(x-y) + \sum_{i=L}^{K} \int_0^\infty P(i,u) \sum_{k=K-i+1}^{\infty} (K-i) D_k du \otimes I \cdot \\
&\quad \cdot \int_0^x e^{D(1)y} \otimes d_y B^{(\chi(m))}(y+u) I \otimes d_x B^{(J)}(x-y)\Bigg]\Bigg\}.
\end{aligned} \tag{6.2}$$

Defining

$$T(i,k,l,j,s) = \int_0^\infty P(i,u) D_k du \int_0^\infty P(l,y) e^{-sy} \otimes d_y B^{(j)}(y+u), \tag{6.3}$$

$$i \geq 0,\ k \geq 1,\ l \geq 0,\ 1 \leq j \leq J,\ Res > 0,$$

the following theorem is easily verified.

Theorem 3: *The vector Laplace-Stieltjes transform* $\mathbf{v}(s)$ *of the sojourn time of an admitted customer in the case of Variant 2 is given by*

$$
\begin{aligned}
\mathbf{v}(s) = \frac{1}{\lambda(1-P_{reject})}\Bigg\{\sum_{i=0}^{L-1}\mathbf{z}(0,i)\Bigg[&\sum_{k=1}^{L-i-1} kD_k\otimes I\sum_{r=L}^{K}\phi_{i+k,r}(s)\beta^{(\chi(r))}(s)+\\
&+\sum_{k=L-i}^{K-i} kD_k\otimes\beta^{(\chi(k+i))}(s)+\sum_{k=K-i+1}^{\infty}(K-i)D_k\otimes\beta^{(J)}(s)\Bigg]+\\
&+\sum_{m=L}^{K}\vec{q}(m)\Bigg[\sum_{i=0}^{L-1}\sum_{k=1}^{L-i-1}\sum_{l=0}^{L-i-1-k} kT(i,k,l,\chi(m),s)\cdot \qquad (6.4)\\
&\cdot\sum_{r=L}^{K}\phi_{i+k+l,r}(s)\otimes\beta^{(\chi(r))}(s)+\\
&+\sum_{i=0}^{K}\sum_{k=1}^{K-i}\sum_{l=L-i-k}^{\infty} kT(i,k,l,\chi(m),s)I\otimes\beta^{(\chi(i+k+l))}(s)+\\
&+\sum_{i=L}^{K}\sum_{k=K-i+1}^{\infty}\sum_{l=0}^{\infty}(K-i)T(i,k,l,\chi(m),s)I\otimes\beta^{(J)}(s)\Bigg]\Bigg\},
\end{aligned}
$$

where $\beta^{(j)}(s)=\int\limits_0^\infty e^{-sx}dB^{(j)}(x), 1\le j\le J.$

Note: The expression in (6.4) can be used to calculate the mean sojourn time of an admitted customer in the case of Variant 2.

Corollary 3: *The (marginal) Laplace-Stieltjes transform* $\hat{v}(s)$ *of the tagged customer is given by*

$$\hat{v}(s)=\mathbf{v}(s)\mathbf{1}. \qquad (6.5)$$

Remark: Theorem 3 can easily be modified for Variant 1 and the details are omitted.

7 An Optimization Problem

In this section we consider an optimization problem. Let c_j, $0\le j\le J$, denote the cost per unit of time of service in mode j. Note that when j=0, the server is considered to be idle. Let d_1 denote the holding cost per customer per unit of time of waiting in the queue and d_2 the cost per lost customer per unit of time. Then the optimization problem of interest is given by

$$\min_{I_1,\ldots,I_{J-1}}\{\sum_{j=0}^{J}c_j\theta_j+d_1\mu_{QL}+d_2\lambda P_{reject}\}, \qquad (7.1)$$

where P_{reject}, θ_j, and μ_{QL} are as given in (5.1),(5.4)–(5.8).

Finding an optimal solution in the set of all multi-threshold policies is very complex. Furthermore, the solution of this problem is complicated due to the fact that the objective function is known only implicitly in terms of the steady state measures. Hence, developing a very clever and efficient heuristic algorithm and the numerical implementation of this algorithm require a very careful and detailed analysis. In the next section we provide some interesting numerical examples.

8 Numerical Examples

In this section we will discuss some illustrating numerical examples. We consider three arrival processes with the following $BMAP$ representations $\{D_k\}$.

1. Erlang:

$$D_0 = \begin{pmatrix} -5.65218 & 5.65218 \\ 0 & -5.65218 \end{pmatrix}, \quad D_1 = D_2 = \begin{pmatrix} 0 & 0 \\ 2.82609 & 0 \end{pmatrix}$$

$$D_k = 0, k \geq 3.$$

2. Hyperexponential:

$$D_0 = \begin{pmatrix} -6 & 0 \\ 0 & -1.2648646 \end{pmatrix}, \quad D_1 = D_2 = \begin{pmatrix} 2.1 & 0.9 \\ 0.4427026 & 0.1897297 \end{pmatrix},$$

$$D_k = 0, k \geq 3.$$

3. BMAP with Positive Correlation:

$$D_0 = \begin{pmatrix} -1.45 & 0.2 & 0.15 & 0.1 \\ 0.2 & -2.6 & 0.1 & 0.3 \\ 0.2 & 0.1 & -3.7 & 0.4 \\ 0.1 & 0.05 & 0.15 & -4.3 \end{pmatrix}, \quad D_1 = D_2 = \begin{pmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1.5 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix},$$

$$D_k = 0, k \geq 3.$$

For all these $BMAPs$ it is easy to verify that $\lambda_g = 2.8261$ and $\lambda = 4.2391$. While the first two arrival processes correspond to renewal processes, the third arrival process corresponds to a correlated process with the correlation between the successive interarrival epochs given by 0.12795. The standard deviation of these three arrival processes are, respectively, 0.25021, 0.53731, and 0.43652.

We take $J = 3$. That is, the system has at most three operation modes. The service times are of semi-Markov type. For all three operation modes, the transition matrix P of the embedded Markov chain for the directing process $\{m_t, t \geq 0\}$ and the semi-Markovian kernels $B^{(j)}(t), 1 \leq j \leq 3$, are taken to be

$$P = \begin{pmatrix} 0.6 & 0.4 \\ 0.35 & 0.65 \end{pmatrix}, \quad B^{(j)}(t) = \begin{pmatrix} B_1^{(j)}(t) & 0 \\ 0 & B_2^{(j)}(t) \end{pmatrix} P,$$

where the distribution functions $B_r^{(j)}(t)$ correspond to degenerate random variables $T_r^{(j)}$ with

$$T_1^{(1)} = 0.8, T_2^{(1)} = 0.9, T_1^{(2)} = 0.7, T_2^{(2)} = 0.65, T_1^{(3)} = 0.5, T_2^{(3)} = 0.4.$$

The mean service time, $b_1^{(j)}, 1 \leq j \leq 3$, when the system operates in the j^{th} mode is calculated as

$$b_1^{(1)} = 0.85333, b_1^{(2)} = 0.67333, b_1^{(3)} = 0.44667.$$

We fix $K = 10$, $L = 2$, and take

$$c_0 = 2, c_1 = 4, c_2 = 6.3, c_3 = 9.8, d_1 = 3, \text{and } d_2 = 10.$$

In the following we consider Variant 1 for the service mechanism. That is, the service initiated through an arrival of a customer will always be in mode 1.

In Tables 1 through 3 that follow we list the values of $\theta_j, 0 \leq j \leq 3, P = P_{reject}$, and μ_{QL} for various combinations of the thresholds I_1 and I_2, for the three arrival processes. Note that $I_3 = K = 10$. Also, when $I_1 = I_2$ we have only two service modes, namely mode 1 and mode 3, and when $I_1 = I_2 = I_3$, there is only one service mode which is mode 1.

Table 1: Cost function and its components for various values of I_1 and I_2 for Erlang arrivals

I_1, I_2	Cost	μ_{QL}	$P \times 10^5$	$\theta_0 \times 10^3$	$\theta_1 \times 10^3$	$\theta_2 \times 10^3$	$\theta_3 \times 10^3$
2,2	9.37472	1.322	24	119	600	0	281
2,3	9.38577	1.372	24	102	548	162	188
3,3	9.28403	1.441	31	95	708	0	197
2,4	9.39593	1.417	23	88	501	306	105
3,4	9.29263	1.488	31	81	666	139	114
4,4	9.20693	1.547	37	75	804	0	121
2,5	9.40151	1.443	23	79	473	394	54
3,5	9.29762	1.517	31	71	640	228	61
4,5	9.21053	1.579	37	65	780	89	66
5,5	9.15483	1.619	41	61	870	0	69
2,6	9.40475	1.460	23	73	456	448	23
3,6	9.30060	1.536	31	65	623	285	27
4,6	9.21267	1.600	37	58	765	147	30
5,6	9.15589	1.642	41	54	856	58	32
6,6	9.11905	1.669	44	51	916	0	33
2,7	9.40605	1.467	23	71	449	471	9
3,7	9.30184	1.544	31	62	616	311	11
4,7	9.21356	1.610	37	55	757	175	13
5,7	9.15628	1.653	41	50	849	87	14
6,7	9.11910	1.681	44	47	910	29	14
7,7	9.10052	1.695	45	45	940	0	15
2,8	9.40662	1.470	23	70	445	482	3
3,8	9.30240	1.548	31	61	612	323	4
4,8	9.21395	1.615	37	53	754	188	5
5,8	9.15644	1.658	41	48	846	101	5
6,8	9.11909	1.687	44	45	907	43	5
7,8	9.10044	1.701	45	43	937	14	6
8,8	9.09137	1.708	46	42	952	0	6
2,9	9.40681	1.471	23	69	444	486	1
3,9	9.30259	1.550	31	60	611	327	2
4,9	9.21409	1.617	37	53	753	193	1
5,9	9.15650	1.660	41	48	845	105	2
6,9	9.11909	1.689	44	44	906	48	2
7,9	9.10040	1.703	45	43	936	19	2
8,9	9.09133	1.710	46	42	951	5	2
9,9	9.08785	1.713	46	41	957	0	2
2,10	9.40691	1.471	23	69	444	487	0
3,10	9.30270	1.551	31	60	610	330	0
4,10	9.21417	1.618	37	52	752	196	0
5,10	9.15653	1.661	41	47	844	109	0
6,10	9.11908	1.690	44	44	905	51	0
7,10	9.10038	1.705	45	42	935	23	0
8,10	9.09130	1.712	46	41	950	9	0
9,10	9.08782	1.714	46	41	956	3	0
10,10	9.08590	1.716	46	41	959	0	0

Table 2: Cost function and its components for various values of I_1 and I_2 for hyperexponential arrivals

I_1, I_2	Cost	μ_{QL}	$P \times 10^5$	$\theta_0 \times 10^3$	$\theta_1 \times 10^3$	$\theta_2 \times 10^3$	$\theta_3 \times 10^3$
2,2	9.64944	1.409	1341	296	455	0	249
2,3	9.64231	1.435	1360	284	436	84	196
3,3	9.64090	1.474	1479	276	526	0	198
2,4	9.63622	1.459	1378	273	418	163	146
3,4	9.63346	1.497	1495	265	508	79	148
4,4	9.63239	1.533	1606	258	592	0	150
2,5	9.63316	1.478	1393	265	404	224	107
3,5	9.62928	1.515	1507	257	493	141	109
4,5	9.62740	1.551	1618	250	578	61	111
5,5	9.62871	1.579	1706	244	643	0	113
2,6	9.63150	1.495	1406	258	392	276	74
3,6	9.62667	1.531	1518	250	480	193	77
4,6	9.62411	1.567	1628	242	565	114	79
5,6	9.62513	1.595	1716	237	631	52	80
6,6	9.62750	1.619	1791	232	686	0	82
2,7	9.63112	1.507	1417	253	383	314	50
3,7	9.62559	1.544	1527	245	471	232	52
4,7	9.62252	1.579	1636	237	555	154	54
5,7	9.62332	1.608	1725	231	622	91	56
6,7	9.62560	1.632	1801	227	677	39	57
7,7	9.62899	1.651	1860	223	719	0	58
2,8	9.63130	1.517	1425	249	377	342	32
3,8	9.62527	1.553	1534	241	463	262	34
4,8	9.62183	1.588	1643	233	548	184	35
5,8	9.62249	1.617	1732	227	614	122	37
6,8	9.62472	1.641	1808	222	670	70	38
7,8	9.62811	1.661	1868	219	713	30	38
8,8	9.63173	1.675	1914	216	745	0	39
2,9	9.63174	1.523	1430	246	372	362	20
3,9	9.62539	1.559	1539	238	459	282	21
4,9	9.62171	1.595	1647	231	543	204	22
5,9	9.62227	1.624	1737	225	609	143	23
6,9	9.62448	1.649	1813	220	665	91	24
7,9	9.62789	1.668	1874	216	708	51	25
8,9	9.63153	1.683	1920	213	740	22	25
9,9	9.63494	1.694	1954	211	763	0	26
2,10	9.63332	1.534	1439	242	365	393	0
3,10	9.622651	1.571	1547	234	450	316	0
4,10	9.62251	1.606	1656	226	533	241	0
5,10	9.62299	1.636	1746	220	600	180	0
6,10	9.62522	1.661	1824	215	656	129	0
7,10	9.62872	1.681	1885	211	699	90	0
8,10	9.63246	1.697	1932	208	732	60	0
9,10	9.63594	1.708	1967	206	755	39	0
10,10	9.64515	1.729	2033	202	798	0	0

Table 3: Cost function and its components for various values of I_1 and I_2 for positively correlated arrivals

I_1, I_2	Cost	μ_{QL}	$P \times 10^5$	$\theta_0 \times 10^3$	$\theta_1 \times 10^3$	$\theta_2 \times 10^3$	$\theta_3 \times 10^3$
2,2	9.40340	1.356	746	233	511	0	256
2,3	9.39871	1.389	756	221	484	102	193
3,3	9.38464	1.439	863	214	591	0	195
2,4	9.39523	1.419	765	210	459	196	135
3,4	9.38026	1.468	870	203	567	92	138
4,4	9.36943	1.511	967	197	662	0	141
2,5	9.39398	1.442	772	203	441	263	93
3,5	9.37840	1.490	876	195	548	160	97
4,5	9.36718	1.533	973	189	644	67	100
5,5	9.36364	1.565	1047	185	713	0	102
2,6	9.39352	1.460	778	197	427	316	60
3,6	9.37750	1.507	881	189	533	214	64
4,6	9.36602	1.551	974	183	630	121	66
5,6	9.36246	1.584	1053	179	699	53	69
6,6	9.36190	1.610	1115	175	754	0	71
2,7	9.39364	1.472	782	193	417	352	38
3,7	9.37734	1.520	884	185	523	251	41
4,7	9.36572	1.564	981	179	619	159	43
5,7	9.36218	1.597	1057	174	689	92	45
6,7	9.36168	1.624	1120	171	745	38	46
7,7	9.36349	1.644	1168	169	784	0	47
2,8	9.39389	1.481	786	190	410	377	23
3,8	9.37743	1.529	887	183	515	278	24
4,8	9.36574	1.573	984	176	611	187	26
5,8	9.36223	1.607	1061	171	682	119	28
6,8	9.36180	1.634	1124	168	738	66	28
7,8	9.36369	1.654	1172	166	777	27	30
8,8	9.36619	1.669	1208	164	806	0	30
2,9	9.39416	1.486	788	189	405	393	13
3,9	9.37761	1.534	889	181	510	294	15
4,9	9.36589	1.579	986	174	606	204	16
5,9	9.36243	1.613	1063	170	677	137	16
6,9	9.362063	1.641	1127	166	733	84	17
7,9	9.36401	1.661	1176	163	773	46	18
8,9	9.36656	1.676	1212	162	802	18	18
9,9	9.36906	1.686	1237	161	821	0	18
2,10	9.39473	1.494	791	187	399	414	0
3,10	9.37810	1.542	891	179	503	318	0
4,10	9.36640	1.588	988	172	599	229	0
5,10	9.36306	1.623	1066	167	670	163	0
6,10	9.36283	1.651	1131	164	725	111	0
7,10	9.36491	1.672	1181	161	766	74	0
8,10	9.36756	1.688	1217	159	795	46	0
9,10	9.37013	1.698	1243	158	814	28	0
10,10	9.37611	1.714	1285	156	844	0	0

From these tables we notice the following observations.

- For all I_2 and for the three arrival processes considered, the mean queue length (μ_{QL}), the loss probability (P_{reject}), and the fraction of time the server is busy in service mode 1 (θ_1) appear to increase as I_1 increases. The rate of increase decreases as I_1 increases. It is interesting to note that as I_1 increases, the mean queue length for Erlang arrivals appears to dominate the other two arrival processes for all values of I_2.

- For all I_2 and for the three arrival processes considered, the fraction of time the server is idle (θ_0), the fraction of time the server is busy serving in mode 2 (θ_2), and the fraction of time the server is busy in service mode 3 (θ_3) appear to decrease as I_1 increases. The rate of decrease appears to decrease as I_1 increases.

- With respect to the measure, θ_2, there appears to be a cut-off point, say, $I_2^*(I_1)$, such that for $I_1 \leq I_2 < I_2^*$, Erlang has the largest value for this measure and hyperexponential has the least value. For $I_2 \geq I_2^*$, hyperexponential has the largest value and Erlang has the least. For example, when $I_1 = 2$, we have $I_2^* = 2$ and when $I_1 = 4$, $I_2^* = 6$. The same type of phenomenon appears to hold true for the measure θ_3.

- While for all values of I_1 and I_2, the measure θ_2 appears to decrease with increasing variance of the arrival times, the measures θ_0 and P_{reject} appear to increase with increasing variance of the arrival times. However, more experimentation is needed to see how correlation plays a role.

- We see that the optimal value 9.085900 for the cost criterion in (41) occurs at $I_1 = 10$ and $I_2 = 10$ for Erlang arrivals; the optimal value of 9.621708 occurs at $I_1 = 4$ and $I_2 = 9$ for the hyperexponential case; and for $BMAP$ with positive correlation the optimal value of 9.36168 occurs at $I_1 = 6$ and $I_2 = 7$. It is interesting to note that for this choice of parameters, there is no control mechanism required for the Erlang arrivals; however, for the other two arrival processes, control strategy yields a better solution. This indicates that whenever the arrival processes have a larger variation or are correlated the situation will be quite different and needs a careful analysis.

References

[1] Alfa, A.S. and Chakravarthy, S.R., *Advances in Matrix Analytic Methods for Stochastic Models*, Notable Publications, New Jersey 1998.

[2] Artalejo, J., G-networks: A versatile approach for work removal in queueing networks, *European J. of Ops. Res.* **126** (2000), 233–249.

[3] Bailey, N.T.J., On queueing processes with bulk services, *J. of the Royal Stat. Soc. Series B* **16** (1954), 80–87.

[4] Bayer, N. and Boxma, O., Wiener-Hopf analysis of an $M/G/1$ queue with negative customers and of a related class of random walks, *Queueing Sys.* **23**(1996), 301–316.

[5] Bini, D.A., et al., Control of the $BMAP/PH/1/K$ queue with group services, In: *Adv. in Alg. Methods for Stoch. Models* (ed. by G. Latouche and P. Taylor), Notable Publications, Inc., New Jersey (2000), 57–72.

[6] Chakravarthy, S.R. and Alfa, A.S., *Matrix-Analytic Methods in Stochastic Models*, Marcel Dekker, New York 1997.

[7] Chakravarthy, S., A finite capacity $GI/PH/1$ queue with group services, *Naval Res. Log. Qtrly***39** (1992), 345–357.

[8] Chakravarthy, S., Analysis of a finite $MAP/G/1$ queue with group services, *Queueing Sys.: Theory and Appl.* **13**(1993), 385–407.

[9] Chakravarthy, S., A finite capacity queueing network with single and multiple processing nodes, In: *Queueing Network with Finite Capacity* (ed. by R.F. Onvural and I.F. Akyildiz), North-Holland, Netherlands (1993), 197–211.

[10] Chakravarthy, S., Two finite queues in series with nonrenewal input and group services, In: *Proc. of the Seventh Intern. Symp. on Appl. Stoch. Models and Data Anal.* (1995), 78–87.

[11] Chakravarthy, S., Analysis of the $MAP/PH/1/K$ queue with service control, *Appl. Stoch. Models and Data Anal.* **12** (1996), 179–191.

[12] Chakravarthy, S., Analysis of a priority polling system with group services, *Stoch. Models* **14** (1998), 25–49.

[13] Chakravarthy, S., Analysis of a multi-server queue with batch Markovian arrivals and group services, *Eng. Simul.* **18** (2000), 51–66.

[14] Chakravarthy, S. and Alfa, A.S., A finite capacity queue with Markovian arrivals and two servers with group services, *J.Appl. Math. Stoch. Anal.* **7** (1994), 161–178.

[15] Chakravarthy, S. and Bin, L., A finite capacity queue with nonrenewal input and exponential dynamic group services, *INFORMS J. on Comp.***9** (1997), 276–287.

[16] Chakravarthy, S. and Lee, S.Y., An optimization problem in a finite capacity $PH/PH/1$ queue with group services, In: *Stoch. Models, Optim. Tech. and Comp. Appl.* (ed. by G.V. Krishna Reddy, et. al) (1994), 3–13.

[17] Chakravarthy, S.R., The batch Markovian arrival process: A review and future work, *Adv. in Prob. Theory and Stoch. Proc.* (ed. by A. Krishnamoorthy), Notable Publications Inc. New Jersey (2001), 21–49.

[18] Crabill, T.B., Optimal control of a service facility with variable exponential service times and constant arrival rate, *Mgmt. Sci.* **18** (1972), 560–566.

[19] Dudin, A.N., Optimal service rate assignment in the multi-server queueing system, *Reports of Belarussian Academy of Sciences* (Russian. English Summary) Dokl. Akad.Nauk BSSR **24**:9 (1980), 780–783, 859.

[20] Dudin, A.N., Optimal control by multi-rate queueing system. *Eng. Cyber.* **19** (1981), 109–115.

[21] Dudin, A.N. and Ikhsan, H., Optimal control by the current CS/PS border under the adaptive commutation, *Proc. of the 13th All-Union Sem. on Comp. Net. Moscow* **2**(1988), 82–86 (in Russian).

[22] Dudin, A.N. and Klimenok, V.I., Optimal control by multi-mode queue of $GI/M/1$ type, In: *Math. Modeling of Manuf. Proc. Petrozavodsk* (1990), 14–21 (in Russian).

[23] Dudin, A.N. and Klimenok, V.I., Characteristics calculation for the single server queueing system, which operates in the synchronized Markov random environment, *Auto. and Remote Contr.* **58** (1997), 74–84.

[24] Dudin, A.N., Characteristics calculation and optimization of the $BMAP/SM/1/N$ queue with controlled operation modes, In: *Proc. of the Intern. Conf. "Stat. and Appl. Anal. of Time Series" Brest* (1997), 165–171. (in Russian).

[25] Dudin, A.N. and Klimenok, V.I. Klimenok, Optimal multi-threshold control for a $BMAP_N/SM_N/1$ retrial queue, In: *Abstracts of the First Intern. Workshop on Retrial Queues Madrid* (1998), 13–14.

[26] Dudin, A.N., Optimal multi-threshold control for a $BMAP/G/1$ queue with N service modes, *Queueing Sys.: Theory and Appl.* **30** (1998), 273–287.

[27] Dudin, A.N. and Nishimura, S., Embedded stationary distribution for the $BMAP/SM/1/N$ queue with disasters, *Queues: Flows, Sys., Networks* **14** (1998), 92–97.

[28] Dudin, A.N. and Klimenok, V.I., Multi-dimensional quasitoeplitz Markov chains, *J.Appl. Math. Stoch. Anal.* **12**(1999), 393–415.

[29] Dudin, A.N. and Nishimura, S., A $BMAP/SM/1$ queueing system with Markovian arrival input of disasters, *J. of Appl. Prob.* **36** (1999), 868–881.

[30] Dudin, A.N. and Nishimura, S., Optimal control for a $BMAP/G/1$ queue with two service modes, *Math. Prob. in Eng.* **5** (1999), 255–273.

[31] Dudin, A.N. and Nishimura, S. Nishimura, Optimal hysteretic control for a $BMAP/SM/1/N$ queue with two service modes, *Math. Prob. in Eng.* **5** (2000), 397–420.

[32] Dudin, A.N. and Klimenok, V.I., *Queues with correlated input flow*, Belarussian State Univ. Publ., Minsk 2000 (in Russian).

[33] Dudin, A.N. and Karolik, A.V., $BMAP/SM/1$ queue with Markovian input of disasters and non-instantaneous recovery, *Perf. Eval.* **45**(2001), 19–32.

[34] Dudin, A.N., et al., Characteristics calculation for a single-server queue with the $BMAP$-input, SM-service and a finite buffer, *Auto. and Remote Contr.* (2001)-submitted.

[35] Dukhovny, A., Matrix-geometric solutions for bulk $GI/M/1$ systems with unbounded arrival groups, *Stoch. Models* **15** (1999), 547–560.

[36] Gail, H.R., et al., Linear independence of root equations for $M/G/1$ type Markov chains, *Queueing Sys.* **20** (1995), 321–339.

[37] Gail, H.R., Hantler, S.L. and Taylor, B.A., Spectral analysis of $M/G/1$ and $GI/M/1$ type Markov chains, *Adv. in Appl. Prob.* **28** (1996), 114–165.

[38] Klimenok, V.I., Optimization of the traffic restriction strategy for the node of computer network, *Auto. Contr. and Comp. Sci.* **27** (1993), 49–56.

[39] Klimenoki, V.I., About the properties of the optimal strategy of control by multi-mode $M/G/1$ type queue, *Queues: Flows, System, Networks* **11**(1995), 66–67 (in Russian).

[40] Klimenok, V.I., Characteristics calculation for the multi-server queueing system with losses and bursty traffic, *Auto. Contr. and Comp. Sci.* **33**(1999), 43–53.

[41] Latouche, G. and Taylor, P.,ed., *Advances in Algorithmic Methods for Stochastic Models*, Notable Publications, Inc., New Jersey 2000.

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

[43] Lucantoni, D.M. and Neuts, M.F., Some steady-state distributions for the $MAP/SM/1$ queue, *Stoch. Models* **10**(1994), 575–598.

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

[45] Ramaswami, V., The $N/G/1$ queue and its detailed analysis, *Adv. in Appl. Prob.* **12**(1980), 222–261.

[46] Tijms, H., On the optimality of a switch-over policy for controlling the queue size in an $M/G/1$ queue with variable service rate, *Lecture Notes in Computer Sciences* **40** (1976), 736–742.