The BMAP/G/1 QUEUE: A Tutorial

David M. Lucantoni

AT&T Bell Laboratories, Room 3K-601, 101 Crawfords Corner Rd, Holmdel, New Jersey 07733-3030, USA, email:dave@buckaroo.att.com

Abstract. We present an overview of recent results related to the single server queue with general independent and identically distributed service times and a batch Markovian arrival process (*BMAP*). The *BMAP* encompasses a wide range of arrival processes and yet, mathematically, the *BMAP*/*G*/1 model is a relatively simple matrix generalization of the *M*/*G*/1 queue. Stationary and transient distributions for the queue length and waiting time distributions are presented. We discuss numerical algorithms for computing these quantities, which exploit both matrix analytic results and numerical transform inversion. Two-dimensional transform inversion is used for the transient results.

1 Introduction

It is well known that the basic *M*/*G*/1 and *GI*/*M*/1 queueing models can be analyzed via embedded Markov chains. As shown in Neuts [1] [2], a large class of interesting queueing models can be analyzed via matrix generalizations of these embedded Markov chains. These are called *M*/*G*/1-type and *GI*/*M*/1-type Markov chains, respectively.

Within the class of models that can be analyzed by *M*/*G*/1-type Markov chains is the *BMAP*/*G*/1 queue; it is the generalization of the *M*/*G*/1 model in which the Poisson arrival process is replaced by a *batch Markovian arrival process* (*BMAP*). Not only is the embedded Markov chain at departure epochs in a *BMAP*/*G*/1 queue an *M*/*G*/1-type Markov chain, but the entire model tends to be a matrix generalization of the *M*/*G*/1 queue (or, more exactly, its batch arrival generalization).

This paper is a tutorial on the *BMAP* and the *BMAP*/*G*/1 queue. The *BMAP* was first introduced with alternative notation as the versatile Markovian point process in Neuts [3], and the *BMAP*/*G*/1 queue was first analyzed (under the name *N*/*G*/1) by Ramaswami [4]. A major focus here is on exact and efficient numerical algorithms for both the steady-state and transient distributions for the *BMAP*/*G*/1 queue. This paper is largely based on the author's (mostly joint) work [5]–[12] but we give a general history in §4.

The idea of a *BMAP* is to keep the tractability of the Poisson arrival process but significantly generalize it in ways that allow the inclusion of dependent interarrival times, non-exponential interarrival-time distributions, and correlated batch sizes. The *BMAP* includes as special cases both phase type renewal processes (which include the Erlang, E_k, and hyperexponential, H_k, renewal processes) and non-renewal processes such as the Markov modulated Poisson process

(*MMPP*) and many other processes in the applied probability literature. These are reviewed in §3. The class also allows correlated batch size distributions and is closed under superpositions, thinning, etc..

Matrix analytic solutions to the $BMAP/G/1$ queue have been available for some time now (see Ramaswami [4]) but the expressions were not in a form that allowed feasible numerical implementation in their full generality. Recent results (reviewed in §4) have resulted in much more transparent solutions that show that this model is indeed a simple generalization of the ordinary $M/G/1$ queue. In fact, many expressions for the performance measures of interest are natural matrix analogues of the corresponding expressions for the $M/G/1$ queue. The purpose of this review is to demonstrate the simplicity of the results which may occasionally have been obscured by the technicalities governing their derivations. This is accomplished by displaying the relevant results next to the corresponding $M/G/1$ formulas. There are no proofs in this paper; we refer the reader to referenced papers for proofs.

The rest of the paper is organized as follows. In §2 we define the *BMAP*, and in §3 we describe a number of special cases. A brief history of the *BMAP* and $BMAP/G/1$ queue is presented in §4. This section may serve as an annotated bibliography of recent work related to this model. Expressions for the transforms of the stationary and transient queue length and waiting time distributions are presented in §5. Actual algorithmic procedures for obtaining numerical results for the performance measures of interest are discussed in §6. In particular, we discuss some of the standard matrix analytic algorithms as well as new transform inversion algorithms. Several examples are discussed in §7. A small list of current and future extensions to this model is discussed in §8.

2 The Batch Markovian Arrival Process

We motivate the *BMAP* by starting with a constructive definition of the Poisson process. Start with a continuous-time Markov process with one state where the same state is visited successively. Since sojourn times in a Markov process are exponential, a sojourn time in this state expires after an exponentially distributed interval with some rate, say λ, but since there is only one state, it is immediately revisited and another sojourn begins. If we construct a point process by associating an arrival with each transition in the above Markov process then the resulting process is Poisson.

One way to generalize the Poisson process is to relax the assumption that interarrival times are exponentially distributed. In the context of the Markov process representation above, we can do this by adding additional (auxiliary) states to the Markov process and associating arrivals in the point process with certain transitions in the underlying Markov process. Let the underlying Markov process be irreducible and have infinitesimal generator D. The sojourn time in state i is thus exponentially distributed with parameter $\lambda_i \geq -D_{ii}$. At the end of a sojourn time in state i, there occurs a transition to another (or possibly the same) state and that transition may or may not correspond to an arrival epoch.

With probability $p_i(0,j)$, $1 \le j \le m$, $j \ne i$, there will be a transition to state j *without* an arrival. With probability $p_i(k,j)$, $k \ge 1$, $1 \le j \le m$, there will be a transition to state j with a batch arrival of size k. We therefore have, for $1 \le i \le m$,

$$\sum_{\substack{j=1 \\ j \ne i}}^{m} p_i(0,j) + \sum_{k=1}^{\infty} \sum_{j=1}^{m} p_i(k,j) = 1 \ .$$

It is convenient to represent the evolution of the system in terms of a sequence of matrices $\{D_k, k \ge 0\}$, by letting $(D_0)_{ii} = -\lambda_i$, $1 \le i \le m$, $(D_0)_{ij} = \lambda_i p_i(0,j)$, $1 \le i,j \le m$, $j \ne i$, and $(D_k)_{ij} = \lambda_i p_i(k,j)$, $k \ge 1$, $1 \le i,j \le m$. This definition implies that $\sum_{k=0}^{\infty} D_k = D$, the infinitesimal generator of the underlying Markov process. Intuitively, we think of D_0 as governing transitions in the phase process that do not generate arrivals and D_k as the rate of arrivals of size k (with the appropriate phase change).

The matrix D_0 has strictly negative diagonal elements, nonnegative off-diagonal elements, row sums less than or equal to zero and we assume it is nonsingular. In other words D_0 is a stable matrix (i.e., all of its eigenvalues have negative real parts; see e.g., p. 251 of Bellman [13]). This implies that the inter-arrival times are finite with probability one (see Lemma 2.2.1 of Neuts [1]) and that the arrival process does not terminate.

Let $\boldsymbol{\pi}$ be the stationary probability vector of the Markov process with generator D, i.e., $\boldsymbol{\pi}$ satisfies

$$\boldsymbol{\pi} D = \mathbf{0}, \qquad \boldsymbol{\pi}\mathbf{e} = 1 \ , \tag{1}$$

where $\mathbf{e}$ is a column vector of 1's. Then the component π_j is the stationary probability that the arrival process is in state j. The stationary arrival rate of the process is

$$\lambda = \boldsymbol{\pi} \sum_{k=1}^{\infty} k D_k \mathbf{e} = \boldsymbol{\pi}\boldsymbol{\eta} \ , \tag{2}$$

where $\boldsymbol{\eta} \equiv \sum k D_k \mathbf{e}$.

A key quantity for analyzing the $BMAP/G/1$ queue is the matrix generating function

$$\underset{(BMAP)}{D(z) = \sum_{k=0}^{\infty} D_k z^k}, \qquad \underset{(Poisson)}{D(z) = -\lambda + \lambda z}, \qquad \text{for } |z| \le 1 \ . \tag{3}$$

Here, and throughout this paper, we display certain results for the Poisson process or $M/G/1$ queue alongside the corresponding results for the $BMAP$ or $BMAP/G/1$ queue, respectively, to highlight the similarity of these expressions.

Let $N(t)$ count the number of arrivals in $(0,t]$ and $J(t)$ represent the auxiliary state or phase at time $t+$. It is significant that the pair $(N(t), J(t))$ is a

continuous-time Markov chain on the state space $\{(i,j) : i \geq 0, 1 \leq j \leq m\}$ with infinitesimal generator

$$Q = \begin{bmatrix} D_0 & D_1 & D_2 & D_3 & \cdot & \cdot \\ & D_0 & D_1 & D_2 & \cdot & \cdot \\ & & D_0 & D_1 & \cdot & \cdot \\ & & & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot \end{bmatrix} .$$

where the states are listed in lexicographic order.

Let $P_{ij}(n,t) = P(N(t) = n,\ J(t) = j \mid N(0) = 0,\ J(0) = i)$ be the (i,j) element of a matrix $P(n,t)$. That is, $P(n,t)$ represents the probability of n arrivals in $(0,t]$ plus the phase transition. Let the matrix generating function $P^*(z,t)$ be defined by

$$P^*(z,t) = \sum_{n=0}^{\infty} P(n,t) z^n, \quad \text{for } |z| \leq 1,\ t \geq 0 .$$

By a routine argument conditioning on the first transition, one can get

$$\underset{(BMAP)}{P^*(z,t) = e^{D(z)t}}, \quad \underset{(Poisson)}{P^*(z,t) = e^{(-\lambda+\lambda z)t}}, \quad \text{for } |z| \leq 1,\ t \geq 0 , \tag{4}$$

where $e^{D(z)t}$ is an exponential matrix (see e.g., p. 169 of Bellman, [13]). By differentiating with respect to z and setting $z = 1$ in (4) we get the expected number of arrivals in $(0,t]$ given that the phase at time $t = 0$ is i as the i^{th} element of the vector $\lambda t\mathbf{e} + (I - e^{Dt})(\mathbf{e}\boldsymbol{\pi} - D)^{-1}\boldsymbol{\eta}$. See Narayana and Neuts [14] for higher moment formulas, asymptotic expansions and the correlation of the number of arrivals in nonoverlapping intervals.

3 Special Cases

Many familiar arrival processes can be obtained as special cases of the *BMAP*. Here is a selected sample of some of the more useful examples.

3.1 Single Arrivals

(A *BMAP* with all batch sizes equal to one is called a *Markovian arrival process* (*MAP*).)

a) *Poisson process.* As described above; in this case $D_0 = -\lambda$, $D_1 = \lambda$ and $D_k = 0$, for $k \geq 2$.

b) *PH-renewal process.* The phase type (PH) renewal process introduced in Neuts [15] (see Neuts [1], chapter 2) contains Erlang, E_k, and hyperexponential, H_k, as well as common renewal processes with interarrival time distributions distributed as finite mixtures of these. A phase type renewal process with representation $(\boldsymbol{\alpha}, T)$, is a *BMAP* with $D_0 = T$, $D_1 = -T\mathbf{e}\boldsymbol{\alpha}$, and $D_k = 0$, for $k \geq 2$.

c) *Markov-modulated Poisson process (*MMPP*).* The *MMPP* is the doubly stochastic Poisson process whose arrival rate is given by $\widehat{\lambda}[J(t)] \geq 0$, where $J(t)$, $t \geq 0$, is an m-state irreducible Markov process. The arrival rate therefore takes on only m values $\lambda_1, \ldots, \lambda_m$, and is equal to λ_j whenever the Markov process is in the state j. If the underlying Markov process has infinitesimal generator R and if $\Lambda = diag(\lambda_1, \ldots, \lambda_m)$, then we have $D_0 = R - \Lambda$, $D_1 = \Lambda$, and $D_k = 0$, for $k \geq 2$. (See Heffes and Lucantoni [16] for an application of this process to the superposition of packetized voice.)

d) *A sequence of PH interarrival times selected via a Markov chain.* For example, assume we have three PH distributions with representations $(\boldsymbol{\alpha}, T)$, $(\boldsymbol{\beta}, S)$, $(\boldsymbol{\gamma}, L)$ labeled 1, 2, and 3, respectively, and that successive interarrival times are chosen from these according to a Markov chain with transition matrix P. The resulting semi-Markov process is the *MAP* defined by

$$D_0 = \begin{bmatrix} T & 0 & 0 \\ 0 & S & 0 \\ 0 & 0 & L \end{bmatrix}, \quad D_1 = \begin{bmatrix} p_{11}\mathbf{T}^o\boldsymbol{\alpha} & p_{12}\mathbf{T}^o\boldsymbol{\beta} & p_{13}\mathbf{T}^o\boldsymbol{\gamma} \\ p_{21}\mathbf{S}^o\boldsymbol{\alpha} & p_{22}\mathbf{S}^o\boldsymbol{\beta} & p_{23}\mathbf{S}^o\boldsymbol{\gamma} \\ p_{31}\mathbf{L}^o\boldsymbol{\alpha} & p_{32}\mathbf{L}^o\boldsymbol{\beta} & p_{33}\mathbf{L}^o\boldsymbol{\gamma} \end{bmatrix},$$

and $D_k = 0$, for $k \geq 2$, where $\mathbf{T}^o = -T\mathbf{e}$, $\mathbf{S}^o = -S\mathbf{e}$ and $\mathbf{L}^o = -L\mathbf{e}$. This process was originally studied in Latouche [17]. As an interesting special case of this, we could have an arrival stream consisting of a sequence of Erlang interarrival times where the orders of successive Erlang random variables form a Markov chain. Also, a trivial special case is the alternating PH-renewal process.

e) *Output and overflows from finite Markovian networks.* Since the *MAP* is defined as point process where arrivals are associated with transitions of an underlying Markov process, it is clear that the overflows and/or outputs of any finite state Markovian network can be modeled as a *MAP*. From a practical viewpoint the size of the network could, however, become a limiting factor.

3.2 Batch Arrivals

a) *Batch Poisson.* Let the arrival rate of batches be γ and the successive batch sizes have probability mass function $\{p_k,\ k \geq 1\}$ and probability generating function $p(z)$. In this case, $m = 1$, and the sequence $\{D_k\}$ are scalars with $D_0 = -\gamma$, and $D_k = \gamma p_k$, for $k \geq 1$. Note that $D(z) = -\gamma + \gamma p(z)$.

b) *A MAP with i.i.d. batch arrivals.* Consider a *MAP* defined by the pair (D_0, D_1) where each arrival epoch corresponds to a batch arrival. If successive batch sizes are independent and identically distributed (i.i.d.) with

probability mass function $\{p_j,\ j \geq 1\}$ then this process is a *BMAP* with $D_j = p_j D_1,\ j \geq 1$.

c) *A batch Poisson process with correlated batch arrivals.* Consider a batch Poisson process where the batch size distribution of successive batch arrivals is chosen according to a Markov chain. For example, let $\{q_i(k),\ k \geq 1\}$, $1 \leq i \leq m$, be a set of m discrete probability mass functions and let P be the transition probability matrix of an m-state, irreducible Markov chain. Let the rate of the Poisson process be λ and assume that successive batch size distributions are chosen from the set $\{q_i(\cdot),\ 1 \leq i \leq m\}$ according to P. This process is then a *BMAP* with $D_0 = -\lambda I$ and $(D_k)_{ij} = \lambda P_{ij} q_j(k)$. A simple example is the overflow process of a $M^X/M/1/N$ system. This example is easily extended to a *MAP* with correlated batch sizes.

d) *Neuts' versatile Markovian point process.* This process, introduced in Neuts [3], is constructively defined by starting with a PH-renewal process as a substratum. There are three types of arrival epochs which are related to the evolution of the PH-renewal process as follows. There are Poisson arrivals with arbitrary batch size distributions during sojourns in the states of the Markov process governing the renewal process. The arrival rates of the Poisson process and the batch size distributions may depend on the state of the Markov process. The underlying Markov process can change states either with or without a corresponding renewal. Each time the process changes states there is a batch arrival (the batch size may be 0) where the batch size distribution can depend on the states before and after the change and whether a renewal occurred. The construction of this process allows easy incorporation of certain qualitative features into a model of a traffic stream. For example, background arrivals could be inhibited or stimulated by upcoming batch arrivals, etc. It can be shown that this process is equivalent to the *BMAP*. In Lucantoni [7] it was shown how an N-process could be represented as a *BMAP*. To go the other way is less obvious but the idea is to pick a state j for which $(D_n)_{jk}$ is nonzero for some n and k. Whenever a batch arrival of size n occurs at a transition from j to k assume that it happened through an artificial absorbing state. It is then easy to construct the parameters of the N-process in terms of the sequence $\{D_k\}$. An advantage of viewing the process in the framework of the *BMAP* is that the notation is much simplified.

e) *A Markov-compound Poisson arrival process.* The class of processes discussed in Pacheco and Prabhu [18] is equivalent to the *BMAP* except that their general formulation allows the auxiliary phase variable to have an infinite state space and also allows D to be reducible and D_0 to be singular. If D is reducible then the stationary version of the process might not be unique; moreover, when D_0 is singular there are two further possibilities. Either the auxiliary phase will enter an irreducible subset from which point the process is a *BMAP* or the process will terminate (see Lemma 2.2.1 of Neuts [1]). When D is irreducible and D_0 is singular then the process is trivial and has no arrivals, i.e., $P(N(t) = 0) = 1$ for all t (see Lemma 5.4.1 and the following

remark on pg. 289 of Neuts [2]).

f) *A superposition of BMAP's.* The class of *BMAP*'s is closed under superposition. That is, the superposition of n independent *BMAP*'s with representations $\{D_k(i)\}$, $1 \leq i \leq n$, is also a *BMAP* with

$$D_k = D_k(1) \oplus \cdots \oplus D_k(n) , \qquad k \geq 0,$$

where "$\oplus$" denotes the matrix Kronecker sum (see, e.g., Bellman [13] or Graham [19]). This formulation has been useful in establishing certain asymptotic results; see Abate, Choudhury, and Whitt [20] and Choudhury and Whitt [21]. These extend results for the models of the $GI/M/1$ paradigm in Neuts [22]. This superposition property has recently been exploited by Choudhury, Lucantoni and Whitt [23]–[25] to study the effect of multiplexing bursty traffic streams in an ATM (asynchronous transfer mode) network.

Other examples of incorporating qualitative features of a traffic stream into a model by using a *BMAP* are presented in Neuts [26]. Although the *BMAP* is a special case of a semi-Markov process (*SMP*), its relationship to continuous time Markov processes leads to far more tractable expressions than would be afforded by a general *SMP*.

We also note that stationary *BMAP*'s are dense in the family of all stationary point processes; see Asmussen and Koole [27]. This shows that the *BMAP* can represent a wide range of behavior although, from a practical point of view, the dimension of the matrices may be a limiting factor. A negative result for the applicability of *MAP*'s is obtained in Olivier and Walrand [28] where it is shown that the output of an $MMPP/M/1$ queue is not a *MAP* unless the input is Poisson. This has implications for modeling networks of $MAP/M/1$ queues.

Finally, we point out that there is a completely analogous discrete time version of the *BMAP*; see Blondia [29] [30], Blondia and Theimer [31], Blondia and Casals [32], Briem and Theimer [33], Herrmann [34], Ohta, et al., [35], Alfa and Neuts [36], Berger [37], Garcia and Casals [38] [39], Ramaswami and Latouche [40] [41], and Neuts [26] [42]. The discrete model is often referred to as the *DMAP* or *DBMAP* for the single and batch arrival versions, respectively. In this case D_k is nonnegative for $k \geq 0$ and the sum of these matrices, D, is irreducible and stochastic. Many results related to the counting function, and moments of the number of arrivals in $(0,t]$, etc., can be derived in an analogous fashion. These involve matrix geometric expressions instead of the matrix exponential expressions in the continuous case. The results for the corresponding discrete time queue, however, are not as explicit as the ones for the continuous time $BMAP/G/1$ queue.

4 A Brief History of the BMAP and BMAP/G/1 Queue

In this section, we provide a brief history of the $BMAP/G/1$ queue. Since it is impractical to review the numerous papers written on special cases, we restrict attention to papers that are directly applicable to the model in its general

formulation. We also recommend the annotated bibliography on phase-type distributions by Neuts [43]. There are several hundred papers cited there, many of which are related to special cases of the $BMAP/G/1$ queue.

The *BMAP* is subclass of stochastic point processes on the real line; see Daley and Vere-Jones [44]. A distinguishing feature of the *BMAP* is the underlying Markovian structure. An early point process exploiting Markov structure is the Wold process; see Wold [45] and p. 89 of Daley and Vere-Jones [44]. In a Wold process the successive interarrival times form a Markov chain.

A process very similar to the *MAP* was introduced by Rudemo [46] where arrivals occur at a subset of transitions of a finite Markov process. In that model the probability of an arrival at a transition from i to j was either zero or one. A number of interesting quantities for that process were derived but it was not used as the arrival process to a queue. In the current context of matrix analytic solutions to queues, the earliest cases of the *MAP* and *BMAP* were constructed by Marcel Neuts, as follows.

In the mid-seventies, Neuts generalized the Erlang and hyperexponential distributions to the class of *phase type* distributions which are distributions that can be represented as those of the time till absorption in finite state absorbing Markov processes with one absorbing state [15]. Using matrix formalism it was shown that many quantities of interest such as the distribution and density functions, moments, the random modification, etc., could be written in a compact form which bore a striking similarity to the ordinary exponential distribution. Neuts [57] then defined the corresponding renewal process of phase type for which the inter-renewal times have a phase type distribution. This then formed the basis for the construction of Neuts' versatile Markovian point process [3]; see Example (d) in §3.2 above. Although the notation was fairly complex (to distinguish between all of the different types of arrivals) the matrix formalism showed that the process was indeed a natural generalization of the ordinary Poisson process.

During this same time, Neuts was developing a matrix-analytic methodology for analyzing complex queueing models which used purely probabilistic arguments to derive expressions and algorithms for computing the performance measures of interest; see, e.g., Neuts [1] [2]. One of the motivations for this effort was to provide an alternative solution technique which avoided the sometimes difficult problem of numerically searching for the roots of a (usually) transcendental equation. Neuts distinguished between two different paradigms: Markov chains of $GI/M/1$-type and $M/G/1$-type, respectively. Solutions to the models of $GI/M/1$-type tend to have a very elegant *matrix geometric solution*, which is a matrix generalization of the geometric distribution, (see Theorem 1.2.1 in Neuts [1]). The solution to the models of $M/G/1$ type are usually more complicated.

The key ingredient to the matrix analytic solution to models of $M/G/1$-type is the solution, G, of a matrix functional equation. In the general $M/G/1$ paradigm, G is related to the fundamental period of the queue; see §2.2 of Neuts [2]. In the context of the $BMAP/G/1$ queue, G is related to the busy period of the queue and indirectly, to the behavior of the arrival process during successive

idle periods of the queue; see §5.2 below. The relationship between the matrix analytic approach and the traditional approach is that the roots in the traditional analysis are the eigenvalues of the matrix G. It has been shown in some cases that when some roots are close or identical, thus causing problems in the root finding algorithms, the matrix G can still be easily computed.

The matrix analytic methodology was applied by Ramaswami [4] to analyze the single server queue with general service times and Neuts' versatile Markovian point process (or N-process) as the arrival stream. Ramaswami derived expressions for the stationary queue length and waiting time distributions which generalized those for the $M/G/1$ queue. In particular, a matrix generalization of the familiar Pollaczek-Khinchin formula for the waiting time transform was obtained. The resulting algorithms were in principle computable, however, developing a program to solve the model in its full generality was a formidable task. The algorithms at that time required the explicit calculation of the parameters of the transition probability matrix of the Markov chain embedded at departures; see Lucantoni [7] for a full description of the algorithm needed to compute performance measures using the results in [4]. Subsequently, several other queues with this arrival process were analyzed. In particular, the $N/G/\infty$, $N/D/c$, finite $N/G/1$ models and the $N/G/1$ departure process were analyzed by Ramaswami [59], Neuts [60], Blondia [61], and Saito [62], respectively. The special case of a Markov modulated Poisson process (*MMPP*) (see Example (c) in §3.1 above) has been extensively studied; see e.g., Kuczura [47], Neuts [48] [49], Heffes [50], van Hoorn and Seelen [51], Heffes and Lucantoni [16], Burman and Smith [52], Rossiter [53], Ide [54], Baiocchi, et al., [55], Asmussen [56], Zhu and Prabhu [58], etc. It is a doubly stochastic Poisson process or Cox process directed by a Markov chain; see p. 532 of Daley and Vere-Jones [44].

Using uniformization, Lucantoni and Ramaswami [5] developed an algorithm for computing the matrix G without having to compute the parameters of the transition matrix of the embedded Markov chain. This drastically reduced the computational effort required for computing the waiting time distributions and other performance measures and lead to feasible implementations of special cases of the $N/G/1$ queue. In particular, using this result, Heffes and Lucantoni [16] outlined the general procedure for computing the waiting time in the $MMPP/G/1$ queue. Also, a recursive scheme for the queue length distribution eliminating the previous Gauss-Seidel iterations was obtained by Ramaswami [89] for general M/G/1-type models.

The Markovian arrival process (*MAP*) in its current formulation was introduced in Lucantoni, Meier-Hellstern, and Neuts [6] as a generalization of the phase type renewal process and the Markov modulated Poisson process. In studying a single server queue with server vacations, it was convenient to have a simple process which contained both renewal and non-renewal processes as special cases. This process was then easily generalized to the *BMAP* by allowing batch arrivals [7]. It was immediately obvious that this process included many processes described in the applied probability literature and that the streamlined notation used for this process was a very natural generalization of that used for the Pois-

son process. In fact, all of the expressions derived for this process were direct matrix analogues of the simpler expressions for the Poisson process. It was also clear that the N-process was a special case of the $BMAP$ but we had originally conjectured that the class of $BMAP$'s was larger than the class of N-processes. Later we observed that these processes are in fact equivalent (see Example (d) in §3.2 above). Once you realize they are the same, it is not difficult to show this equivalence, although our notes on this were never published. By the time the equivalence was observed, there were already a number of papers referring to the $BMAP$. We therefore decided to keep the new name to distinguish the simplified notation and the results using it from the original more complex notation. It should be noted that all of the results for the models involving the N-process mentioned above could be put into simpler forms using the new notation.

The next major result in the solution to the $BMAP/G/1$ queue came about indirectly. Sengupta [63] showed that the functional equation arising in the solution to the $GI/PH/1$ queue could be written in a matrix exponential form. It was immediately clear that a similar situation occurred in the $PH/G/1$ queue and, consequently, also for the $BMAP/G/1$ queue for the matrix G. Neuts [64] proved the result for the $MMPP/G/1$ queue and the result for the $BMAP/G/1$ queue was derived simultaneously by Lucantoni [7] and Ramaswami [65]. Machihara [66] obtained the exponential form using a last-in-first-out (LIFO) argument and Asmussen [67] later obtained it using ladder heights. Recently, simple proofs of this result along with other key relationships were obtained in Lucantoni and Neuts [10]. Some key quantities were obtained explicitly by exploiting certain commutative properties unobserved earlier. That resulted in major algorithmic simplifications in [7]. Ramaswami [65] also contains detailed results for the $GI/BMAP/1$ queue where the $BMAP$ is used as a model for the service process; that model is a substantial generalization of the $GI/PH/1$ queue.

A goal in the development of the matrix analytic methods was to develop stable numerical algorithms for computing quantities of interest without resorting to numerical transform inversion. However, with the recent availability of improved transform inversion methods, it has become evident that numerical transform inversion can contribute significantly to the numerical solution of these models. Indeed, numerical transform inversion can provide extremely accurate results (see, e.g., Abate and Whitt [68] and Choudhury, Lucantoni and Whitt [12]). Hence, a good strategy for solving the $BMAP/G/1$ queue is to combine the standard matrix-analytic techniques with transform inversion routines. We discuss this more in §6.

Very recent results related to the $BMAP/G/1$ queue are the solutions for the transient queue length and waiting time distributions presented in Lucantoni, Choudhury and Whitt [8]; this involves two-dimensional transform inversion. An algorithm for inverting multi-dimensional transforms is presented in Choudhury, Lucantoni and Whitt [12].

5 The BMAP/G/1 Queue

Consider a single-server queue with a *BMAP* arrival process specified by the sequence of matrices $\{D_k,\ k \geq 0\}$. Let the service times be i.i.d. and independent of the arrival process; let the service time have an arbitrary distribution function H with Laplace-Stieltjes transform (*LST*) h and n^{th} moment α_n. We assume that the mean $\alpha \equiv \alpha_1$ is finite and define the *traffic intensity* to be $\rho \equiv \lambda\alpha$.

5.1 The Embedded Markov Renewal Process at Departures

The embedded Markov renewal process at departure epochs is defined as follows. Define $X(t)$ and $J(t)$ to be the number of customers in the system (including in service, if any) and the phase of the arrival process at time t, respectively. Let τ_k be the epoch of the k^{th} departure from the queue, with $\tau_0 = 0$. (We understand that the sample paths of these processes are right continuous and that there is a departure at $\tau_0 = 0$.) Then $(X(\tau_k),\ J(\tau_k),\ \tau_{k+1} - \tau_k)$, for $k \geq 0$, is a semi-Markov process on the state space $\{(i,j) :\ i \geq 0,\ 1 \leq j \leq m\}$ and is *positive recurrent* when $\rho < 1$; see, e.g., Ramaswami [4].

The matrices of mass functions, $\widetilde{A}_n(x)$, have elements defined by

$[\widetilde{A}_n(x)]_{ij} = P$ (Given a departure at time 0 which left at least one customer in the system and the arrival process in phase i, the next departure occurs no later than time x with the arrival process in phase j, and during that service there were n arrivals).

We introduce the transform matrix

$$A(z,s) = \sum_{n=0}^{\infty} \int_0^{\infty} e^{-sx} d\widetilde{A}_n(x) z^n \ , \tag{5}$$

where $\text{Re}(s) \geq 0$ and $|z| \leq 1$. Then

$$\begin{array}{ll} A(z,s) = \displaystyle\int_0^{\infty} e^{-sx} e^{D(z)x} dH(x) & A(z,s) = \displaystyle\int_0^{\infty} e^{-(s+\lambda-\lambda z)x} dH(x) \\ \equiv h(sI - D(z)), & = h(s + \lambda - \lambda z) \ . \\ (BMAP/G/1) & (M/G/1) \end{array} \tag{6}$$

The definition in (6) above is consistent with the usual definition of a scalar function evaluated at a matrix argument (see Theorem 2, p. 113 of Gantmacher, [69]). In particular, since h is analytic in the right half-plane, the above function is defined by using the matrix argument in the power series expansion of h. This is well defined as long as the spectrum of the matrix argument also lies in the right half plane, which can be shown to hold. Note that from (6) we see that $A(z,s)$ is a power series in $D(z) - sI$. Thus, $A(z,s)$ and $D(z)$ commute. This important property is used repeatedly in the proofs. For later use, we define $A(z) \equiv A(z,0)$.

5.2 The Busy Period

Define $\widetilde{G}^{[r]}_{jj'}(k;x)$, $k \geq 1$, $x \geq 0$, as the probability that the first passage from the state $(i+r,\ j)$ to the state (i,j'), $i \geq 1$, $1 \leq j,\ j' \leq m$, $r \geq 1$, occurs in exactly k transitions and no later than time x, and that (i,j') is the first state visited in level $\mathbf{i}$, where level $\mathbf{i} \equiv \{(i,1),\ldots,(i,m)\}$. The matrix with elements $\widetilde{G}^{[r]}_{jj'}(k;x)$ is $\widetilde{G}^{[r]}(k;x)$.

The joint transform matrix, $G(z,s)$, is defined by

$$G(z,s) = \sum_{k=1}^{\infty} \int_0^{\infty} e^{-sx} d\widetilde{G}^{[1]}(k;x) z^k, \quad \text{for } |z| \leq 1, \quad \text{Re}(s) \geq 0 .$$

In the context of the $BMAP/G/1$ queue, $G(z,s)$ is the two-dimensional transform of the number served during, and the duration of, the busy period (with the appropriate phase change information). It can be shown that the joint transform matrix governing the number served during and the duration of a busy period starting with r customers, is given by $G(z,s)^r$ (see, e.g., Neuts [70] [2]).

It was shown in Lucantoni [7] and Ramaswami [65] that $G(z,s)$ is the solution to

$$\begin{aligned} G(z,s) &= z \int_0^{\infty} e^{-sx} e^{D[G(z,s)]x} dH(x) & G(z,s) &= z \int_0^{\infty} e^{-(s+\lambda-\lambda G(z,s))x} dH(x) \\ &\equiv zh(sI - D[G(z,s)]), & &= zh(s+\lambda-\lambda G(z,s)) , \\ & (BMAP/G/1) & & (M/G/1) \end{aligned} \tag{7}$$

for $|z| \leq 1$, $\text{Re}(s) \geq 0$, where $D[G(z,s)] \equiv \sum_{k=0}^{\infty} D_k G(z,s)^k$. Equation (7) is the matrix analogue of Takács' equation for the busy period in the ordinary $M/G/1$ queue [71]. Equation (7) with $z = 1$ is the matrix analogue of the Kendall functional equation, (see (59) in Kendall [72], and the discussion of I. J. Good on p. 182 there). Note from (7) that $G(z,s)$ commutes with $D[G(z,s)]$. The exponential form of the matrix $G(z,s)$ results in substantial reduction in the computational complexity of the implementation of the matrix analytic solution (see [7]).

Next define the matrices $G(s) \equiv G(1,s)$ and $G \equiv G(0)$ and note that G satisfies

$$G = \int_0^{\infty} e^{D[G]x} dH(x) . \tag{8}$$

The matrix G is stochastic when $\rho \leq 1$ (see, e.g., Theorem 2.3.1 in Neuts [2]) and is the key ingredient in the solution of the stationary version of this system. Efficient algorithms for computing this matrix are discussed in §6. For $\rho \leq 1$, the invariant probability vector $\mathbf{g}$, of the positive stochastic matrix G, satisfies

$$\mathbf{g}G = \mathbf{g}, \quad \mathbf{g}\mathbf{e} = 1 . \tag{9}$$

The matrix $D[G] \equiv D[G(1,0)]$ has a nice probabilistic interpretation which was originally pointed out in Lucantoni, Meier-Hellstern and Neuts [6]. Since G is

strictly positive, it follows that the off-diagonal entries of $D[G]$ are nonnegative. When the queue is stable, G is stochastic so that $D[G]\mathbf{e} = \mathbf{0}$; that is, $D[G]$ is the infinitesimal generator of a finite-state, irreducible Markov process. From the structure of the matrix we see that starting in some state i, there will be an exponential sojourn time with rate $|(D_0)_{ii}|$. Then there will either be a transition to state j, $j \neq i$, with rate $(D_0)_{ij}$ (i.e., without an arrival), or a transition to state j with rate $(\sum_{k=1}^{\infty} D_k G^k)_{ij}$. That is, a batch of size k arrives followed by k busy periods which end in phase j, corresponding to a phase change from i to j in this process. It is clear that this process is the phase of the arrival process observed only during idle periods, i.e., the time during the busy periods are *excised.* In the unstable case, i.e., $\rho > 1$, G is strictly substochastic so that $D[G]$ is a stable matrix. In other words, in this case the total amount of idle time observed before the last busy period (that never ends) is phase type (see, e.g., [1]) with representation $(\mathbf{a},\ D[G])$, where $\mathbf{a}$ is the vector of initial phase probabilities at time 0 Note that Equation (8) implies that $\mathbf{g}$ is also the stationary vector of the matrix $D[G]$ and therefore its j^{th} component is the stationary probability that the arrival process is in state j given that the server is idle. This has major implications in the computational algorithm.

5.3 The Stationary Distributions

For proofs of the results in this section, see Ramaswami [4], Lucantoni [7], and Lucantoni and Neuts [9]. Exponential asymptotics for the steady-state distributions below appear in Abate, Choudhury and Whitt [20], Choudhury and Whitt [21], and Baiocchi [73]; we will not discuss them here.

The Queue Length Distribution at Departures Let

$$x_{ij} = \lim_{k \to \infty} P(X(\tau_k) = i,\ J(\tau_k) = j),$$

(recalling that the sample paths are assumed to be right continuous), and define the vectors $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$, for $i \geq 0$. Then the vector generating function $\mathbf{X}(z) \equiv \sum_{i=0}^{\infty} \mathbf{x}_i z^i$, is given by

$$\mathbf{X}(z) = (1-\rho)\lambda^{-1}\mathbf{g}D(z)A(z)[zI - A(z)]^{-1}, \quad X(z) = \frac{(1-\rho)(z-1)A(z)}{z - A(z)}\ ,$$
$$(BMAP/G/1) \qquad\qquad (M/G/1) \tag{10}$$

for $|z| \leq 1$. Note that $(z-1)$ in the numerator for $M/G/1$ plays the role of $\lambda^{-1}\mathbf{g}D(z)$ for $BMAP/G/1$; see Equation (3). The general form of the generating function $\mathbf{X}(z)$ in (10) was proved in Ramaswami [4], however, the simplicity of the constant vector $(1-\rho)\mathbf{g}$ went unnoticed until Lucantoni [7].

The Queue Length Distribution at an Arbitrary Time We first consider the continuous parameter process $\{[X(t),\ J(t)],\ t \geq 0\}$. The time-dependent joint distribution of the queue length and the arrival phase is given by the conditional probabilities

$$Y_{i_0 i}^{j\,k}(t) = P(X(t) = i,\ J(t) = k \mid X(0) = i_0,\ J(0) = j,\ \tau_0 = 0)\ , \qquad (11)$$

for $i_0,\ i \geq 0,\ 1 \leq j, k \leq m,\ t \geq 0$. We can show that the limits

$$y_{ik} \equiv \lim_{t \to \infty} Y_{i_0 i}^{j\,k}(t)\ ,\ \text{ for } i \geq 0,\ 1 \leq k \leq m\ ,$$

exist and are independent of i_0 and j. For $i \geq 0$ let $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im})$ and define the vector generating function, $\mathbf{Y}(z) = \sum_{i=0}^{\infty} \mathbf{y}_i z^i$. Then

$$\underset{(BMAP/G/1)}{\mathbf{Y}(z) = (1-\rho)\mathbf{g}(z-1)A(z)[zI - A(z)]^{-1}} \quad , \underset{(M/G/1)}{Y(z) = \frac{(1-\rho)(z-1)A(z)}{z - A(z)}}\ , \qquad (12)$$

for $|z| \leq 1$. Note that $\mathbf{y}_0 = (1-\rho)\mathbf{g}$ which is consistent with the probabilistic interpretation of $\mathbf{g}$ as the stationary phase probabilities given that the system is empty (see §5.2). Although (12) shows the similarity between solutions to the $BMAP/G/1$ and $M/G/1$ queues, a more convenient representation of $\mathbf{Y}(z)$ in the $BMAP/G/1$ queue is

$$\mathbf{Y}(z)D(z) = \lambda(z-1)\mathbf{X}(z).$$

Therefore, numerically inverting $\mathbf{X}(z)$ and $\mathbf{Y}(z)$ can be done simultaneously. Alternatively, a recursion for $\mathbf{y}_i$ in terms of $\mathbf{x}_i$ can be derived by equating coefficients of z^i (see (38) in Lucantoni [7]; note that this recursion is much simpler than (3.3.16) in Ramaswami [4]).

The Queue Length Distribution at an Arrival Let $L(z)$ be the generating function of the number of customers in the system at an arbitrary arrival (not including the customers in the arriving batch). Then

$$\begin{aligned} L(z) &= (1-\rho)(\boldsymbol{\pi} D_0 \mathbf{e})^{-1}\mathbf{g}(z-1)A(z)[zI - A(z)]^{-1}D_0\mathbf{e}, \quad (BMAP/G/1) \\ L(z) &= \frac{(1-\rho)(z-1)A(z)}{z - A(z)}\ , \quad (M/G/1) \end{aligned} \qquad (13)$$

for $|z| \leq 1$.

The Virtual Waiting Time Distribution In this section, we state results for the virtual waiting time or workload distribution. First, we define the following quantities $\widetilde{\mathbf{W}}_V(x) = (\widetilde{W}_{V,1}(x), \ldots, \widetilde{W}_{V,m}(x))$, where $\widetilde{W}_{V,j}(x)$ is the joint probability that at an arbitrary time the arrival process is in phase j and that a *virtual* customer who arrives at that time waits at most a time x before entering service. The Laplace-Stieltjes transform of $\widetilde{\mathbf{W}}_V(x)$ is $\mathbf{W}_V(s) = \int_0^\infty e^{-sx} d\widetilde{\mathbf{W}}_V(x)$. Then

$$\underset{(BMAP/G/1)}{\mathbf{W}_V(s) = s(1-\rho)\mathbf{g}[sI + D(h(s))]^{-1},\ \mathbf{W}_V(0) = \boldsymbol{\pi},} \quad \underset{(M/G/1)}{W_V(s) = \frac{s(1-\rho)}{s - \lambda + \lambda h(s)},} \tag{14}$$

for $\mathrm{Re}(s) \geq 0$. The waiting time transform in (14) is a matrix generalization of the Pollaczek-Khinchin formula for the $M/G/1$ queue and was obtained using Markov renewal arguments in Ramaswami [4] along with the associated system of Volterra integral equations. An alternative derivation was obtained by Ide [54]. The simplified expression for the constant vector $(1-\rho)\mathbf{g}$ is due to Lucantoni [7].

The Waiting Time Distribution of the First Customer in a Batch Let $W_B(s)$ be the LST of the delay of the first customer in a batch. Then

$$\underset{(BMAP/G/1)}{W_B(s) = s(1-\rho)(\boldsymbol{\pi} D_0 \mathbf{e})^{-1}\mathbf{g}[sI + D(h(s))]^{-1} D_0 \mathbf{e},} \quad \underset{(M^X/G/1)}{W_B(s) = \frac{s(1-\rho)}{s - \lambda + \lambda h(s)},} \tag{15}$$

for $\mathrm{Re}(s) \geq 0$.

The Waiting Time Distribution of an Arbitrary Customer Let

$$\widetilde{\mathbf{W}}_A(x) = (\widetilde{W}_{A,1}(x), \ldots, \widetilde{W}_{A,m}(x)),$$

where $\widetilde{W}_{A,j}(x)$ is the joint probability that the arrival process is in phase j and the delay of an arbitrary arrival is at most x, and let $\mathbf{W}_A(s)$ be the LST of $\widetilde{\mathbf{W}}_A(x)$. Note that this delay includes the delay due to customers in the arriving batch who are ahead of the arbitrarily chosen customer. Then

$$\begin{aligned} \underset{(BMAP/G/1)}{\mathbf{W}_A(s)} &= \frac{1}{\lambda(1-h(s))}\mathbf{W}(s)[D - D(h(s))] \\ &= \frac{1}{\lambda(h(s)-1)}[\mathbf{W}(s)(sI + D) - s(1-\rho)\mathbf{g}], \end{aligned} \quad \begin{aligned} \underset{(M^X/G/1)}{W_A(s)} &= W(s)\frac{1 - p(h(s))}{\overline{p}(1-h(s))} \\ &= \frac{s(W(s) - (1-\rho))}{\lambda(h(s)-1)}, \end{aligned} \tag{16}$$

for $\mathrm{Re}(s) \geq 0$, where, for the $M^X/G/1$ queue, $p(z)$ is the probability generating function of the batch size distribution and $\overline{p}$ is the mean batch size. Note that the

first expression (for the $M^X/G/1$ queue) in (16) is Equation (2) in Burke [74]. This shows that the total delay is factored into the delay of the first customer in a batch plus the delay due to customers in the batch ahead of the tagged customer. That number of customers is distributed as the forward recurrence time in a discrete-time renewal process where the inter-renewal times are distributed as the batch size distribution. The second expressions in (16) are more convenient for deriving moment formulas since they do not involve the batch size distribution; see Lucantoni and Neuts [9].

5.4 The Transient Distributions

For proofs of the results in this section see Lucantoni, Choudhury and Whitt [8]. The $M/G/1$ counterparts appear in Takács [71].

The Emptiness Functions In this section we characterize the probability that the system is empty at time t. The key role of this function for general systems was demonstrated by Beneš [75]. Let $V(t)$ be the amount of work in the system at time t; for $1 \leq i, j \leq m$, $x \geq 0$ and $t \geq 0$, let

$$P_{x0}^{ij}(t) = P(V(t) = 0,\ J(t) = j \mid V(0) = x\ ,\ J(0) = i\ ), \tag{17}$$

and let the $m \times m$ matrix $P_{x0}(t)$ have (i,j)-entry $P_{x0}^{ij}(t)$. The unconditional emptiness function, starting with initial workload distributed according to cdf F, defined by

$$P_0(t) \equiv \int_0^\infty P_{x0}(t)dF(x), \qquad \text{for } t \geq 0\ , \tag{18}$$

has Laplace transform $p_0(s) \equiv \int_0^\infty e^{-st}P_0(t)dt$ given by

$$\underset{(BMAP/G/1)}{p_0(s) = f(sI - D[G(s)])(sI - D[G(s)])^{-1}}, \quad \underset{(M/G/1)}{p_0(s) = \frac{f(s+\lambda-\lambda G(s))}{s+\lambda-\lambda G(s)}}\ , \tag{19}$$

for $\mathrm{Re}(s) > 0$, where f is the LST of F. The expression for $M/G/1$ in (19) is Equation (9) on p. 52 of Takács [71]. Since the components of the vector $G(s)\mathbf{e}$ are Laplace-Stieltjes transforms and $|G(s)\mathbf{e}| < 1$, for $\mathrm{Re}(s) > 0$, the eigenvalues of $D[G(s)]$ are in the left half-plane. Therefore, for $\mathrm{Re}(s) > 0$, the eigenvalues of $sI - D[G(s)]$ are in the right half-plane and the inverse appearing in (19) is well defined.

We can show from (19) that

$$\lim_{t\to\infty} P_0(t) = \begin{cases} (1-\rho)\mathbf{eg} & \text{for } \rho \leq 1\ , \\ 0 & \text{for } \rho > 1\ , \end{cases} \tag{20}$$

as expected.

The Transient Workload In this section we present the transform of the workload (work in the system in uncompleted service time) at time t. Let $W(t,x)$ be the matrix whose $(i,j)^{\text{th}}$ element is the probability that the work in the system is less than or equal to x and the phase is j at time t, given that at time 0 the phase was i and the initial workload (including the customer in service, if any) was distributed according to F; let the Laplace-Stieltjes transform of F be f. Let $w(t,s)$ and $\widetilde{w}(\xi,s)$ be the transforms

$$w(t,s) = \int_0^\infty e^{-sx} d_x W(t,x) \quad \text{and} \quad \widetilde{w}(\xi,s) = \int_0^\infty e^{-\xi t} w(t,s) dt \ .$$

Then the matrix $\widetilde{w}(\xi,s)$ is given by

$$\widetilde{w}(\xi,s) = (f(s)I - sp_0(\xi))[\xi I - sI - D(h(s))]^{-1}, \quad \widetilde{w}(\xi,s) = \frac{f(s) - sp_0(\xi)}{\xi - s + \lambda - \lambda h(s)} \ , \tag{21}$$

$(BMAP/G/1)$ $\qquad$ $(M/G/1)$

and the matrix $w(t,s)$ is given by

$$w(t,s) = \left(f(s)I - s\int_0^t P_0(u) e^{-[sI + D(h(s))]u} du \right) e^{[sI + D(h(s))]t}, \tag{22}$$

$(BMAP/G/1)$

$$w(t,s) = \left(f(s) - s\int_0^t P_0(u) e^{-(s - \lambda + \lambda h(s))u} du \right) e^{(s - \lambda + \lambda h(s))t}, \tag{23}$$

$(M/G/1)$

for $\mathrm{Re}(s) \geq 0$, $\mathrm{Re}(\xi) > 0$, where $P_0(u)$ and $p_0(\xi)$ are given in (18) and (19), respectively. The equations in (21) are formulas (28) in Lucantoni, Choudhury and Whitt [8] and (15) on p. 53 of Takács [71], respectively. Equation (23) for the $M/G/1$ queue is given on pg. 53 of Takács [71].

Although we are able to express the transform of the delay explicitly in terms of t in (22), we note that this expression is not trivial to evaluate numerically. It involves numerically inverting a Laplace transform where the evaluation of the transform at a value of s requires the numerical integration of the emptiness function times an exponential matrix. The values of the emptiness function are themselves obtained by inverting a Laplace transform. The corresponding expression in (23) for the ordinary $M/G/1$ queue also suffers from the same difficulty. This may partly explain why the known formulas for that case have not been widely used for practical computations.

In contrast, however, the transform expressions in (21) are relatively simple to evaluate, so that with an inversion algorithm for 2-dimensional Laplace transforms, we have a practical method for obtaining numerical results. An efficient and accurate multi-dimensional transform inversion algorithm is presented in Choudhury, Lucantoni and Whitt [12] and is briefly discussed in §6.

It can be shown using Rouché's theorem that for each s, $\mathrm{Re}(s) \geq 0$, the determinant of the matrix $\xi I - sI - D(h(s))$ appearing in the inverse in (21)

has exactly m roots in the region $\mathrm{Re}(\xi) > 0$. (For similar arguments see Çinlar [76] and Neuts [77] [78].) Since $\widetilde{w}$ is a Laplace-Stieltjes transform and is therefore analytic in the interior of the above region, these pairs of (ξ, s) must also be zeros of the first matrix on the right in (21). That is, they are removable singularities. The classical approach to this type of problem would then assume that the roots are distinct to obtain m independent linear equations for each row of the unknown matrix $p_0(\xi)$. In practice, the roots may not be distinct, or if they are close, there may be numerical difficulties in locating these roots. These technical problems are circumvented in the present case since we have an explicit expression for $p_0(\xi)$ (see Equation 19).

We see from (21) that the transform of the limiting distribution of the workload is given by

$$w(s) \equiv \lim_{\xi \to 0} \xi \widetilde{w}(\xi, s) = \begin{cases} s(1-\rho)\mathbf{eg}[sI + D(h(s))]^{-1}, & \text{for } \rho \leq 1 \ , \\ 0, & \text{for } \rho > 1 \ , \end{cases}$$

which agrees with (14).

The Transient Queue Length In this section, we present the transient queue length distribution at time t given an initial number of customers present immediately after a departure at time $t = 0$. Let $Y_{i_0 i}^{jk}(t)$ be as defined in (11) and let $Y_{i_0 i}(t)$ have (j,k)-entry $Y_{i_0 i}^{jk}(t)$. Recall that $\tau_0 = 0$ means that there is a departure at time 0. Let $y_{i_0 i}(s)$ be the Laplace transform of $Y_{i_0 i}(t)$. Then $y_{i_0 0}(s) = G(s)^{i_0}(sI - D[G(s)])^{-1}$ and the probability generating function of the queue length at time t, defined by $\widetilde{y}_{i_0}(z,s) \equiv \sum_{i=0}^{\infty} y_{i_0 i}(s) z^i$, is given by

$$\begin{aligned} \widetilde{y}_{i_0}(z,s) = [&z^{i_0+1}(I - A(z,s))(sI - D(z))^{-1} \\ &+ (z-1)G(s)^{i_0}(sI - D[G(s)])^{-1} A(z,s)][zI - A(z,s)]^{-1} \ , \end{aligned}$$

$$(BMAP/G/1) \qquad (24)$$

$$\widetilde{y}_{i_0}(z,s) = \frac{1}{z - A(z,s)} \left(\frac{z^{i_0+1}(1 - A(z,s))}{s + \lambda - \lambda z} + \frac{(z-1)G(s)^{i_0} A(z,s)}{s + \lambda - \lambda G(s)} \right) \ ,$$

$$(M/G/1) \qquad (25)$$

for $\mathrm{Re}(s) > 0$, $|z| < 1$ and $A(z,s)$ is given in (6).

Let the Laplace transform of the complementary queue length distribution be defined by

$$y_{i_0 i}^{*}(s) = \int_0^{\infty} e^{-st} \sum_{n=i+1}^{\infty} Y_{i_0 n}(t) dt \ ,$$

with the corresponding generating function $\widetilde{y}_{i_0}^{*}(z,s) \equiv \sum_{i=0}^{\infty} y_{i_0 i}^{*}(s) z^i$. Then since $\widetilde{y}_{i_0}(1,s) = (sI - D)^{-1}$, the transform of the complementary queue length distribution, $\widetilde{y}_{i_0}^{*}(z,s)$, is given by

$$\widetilde{y}_{i_0}^{*}(z,s) = \frac{1}{1-z}\left[(sI - D)^{-1} - \widetilde{y}_{i_0}(z,s)\right] \ , \quad \widetilde{y}_{i_0}^{*}(z,s) = \frac{1}{1-z}\left[\frac{1}{s} - \widetilde{y}_{i_0}(z,s)\right] \ ,$$

$$(BMAP/G/1) \qquad (M/G/1) \qquad (26)$$

for $\mathrm{Re}(s) > 0$ and $|z| < 1$.

Transient Results at Arrivals and Departures Further transient results for the $BMAP/G/1$ queue have recently been derived and will be reported in Lucantoni [79]. In particular, we derived explicit expressions for the transform of the queue length at the n-th departure, assuming a departure at time $t = 0$, and the workload at the n-th arrival (keeping track of the appropriate phase changes). The departure process is characterized by the double transform of the probability that the n-th departure occurs at time less than or equal to time x (similar to that derived by Saito [62]). This leads to an explicit expression for the LST of the expected number of departures up to time t. All of these expressions are direct matrix analogues of the corresponding $M/G/1$ results in Takács [71].

6 Numerical Algorithms

It is evident from the results in §5 that the transforms for the stationary queue length and delay distributions are completely specified in terms of the vector $\mathbf{g}$, the stationary probability vector of the matrix G. We discuss below several algorithms for computing this matrix. The transient distributions require evaluation of the matrix $G(s)$ for complex s. Computing the queue length distributions themselves can be done either by transform inversion or by other probabilistic algorithms. In the former case, typically the matrix $A(z)$ needs to be evaluated for complex z. In the latter case, the matrices $A_n \equiv \widetilde{A}_n(\infty)$, $n \geq 0$, need to be evaluated. Both of these require additional overhead. Detailed comparisons of these methods have not yet been performed, however, it is our opinion that a full package for analyzing the $BMAP/G/1$ queue should include several alternatives for computing the desired results. Every numerical algorithm has limitations either in accuracy or in processing requirements and storage so it is useful to compute the results several different ways in order to check for consistency.

Due to space limitations, we do not present the explicit algorithms for computing the quantities of interest but instead, refer to appropriate references for the details and clear demonstrations of implementability.

6.1 Computing the Matrices G, G(s), and A(z)

The matrix G is the unique solution, in the class of substochastic matrices, to the matrix equation (8). It has been shown (see Neuts [2]) that Equation (8) can be solved by successive substitutions. Although early proofs of this result required an initial iterate of $G_0 = 0$, this is not a good starting solution for computations. It has been observed that convergence is slower with this initial solution as the traffic intensity, ρ, increases. By starting with G_0 equal to a stochastic matrix, each iterate will itself be stochastic and we have observed that in many cases the speed of convergence is insensitive to ρ. An efficient technique for evaluating the right hand side of Equation (8) based on uniformization is presented in

Lucantoni [7]. This requires the computation of a scalar sequence $\{\gamma_n,\ n \geq 0\}$ where $\gamma_n = \int_0^\infty (e^{-\theta x}(\theta x)^n/n!)dH(x)$, and $\theta = \max_i(-(D_0)_{ii})$. If H is phase type then the γ_n's can be computed recursively without any numerical integrations; see Theorem 2.2.8 in Neuts [1]. If H is arbitrary then the γ_n's can be computed by numerically inverting the probability generating function $h(\theta - \theta z)$; we have had much success with the transform inversion algorithm presented in Abate and Whitt [80] (see Choudhury, Lucantoni and Whitt [81]).

If H is Coxian (i.e., H has a rational Laplace-Stieltjes transform) then there is an efficient algorithm for evaluating the integral on the right hand side of (8) which is presented below (see Lucantoni, Choudhury and Whitt [8]). This algorithm is also directly applicable to computing $A(z)$ and $G(s)$, for complex z and s, from Equations (6) and (7), respectively.

Let

$$M' = \int_0^\infty e^{-Mx} dH(x)\ , \tag{27}$$

where M and M' are, in general, complex matrices. If H has a Coxian distribution with the Laplace-Stieltjes transform

$$h(s) = \int_0^\infty e^{-sx} dH(x) = \frac{\sum_{k=0}^i a_k s^k}{\sum_{k=0}^j b_k s^k}, \qquad i \leq j, \tag{28}$$

then (27) may be evaluated as

$$M' = \left(\sum_{k=0}^{j} b_k M^k\right)^{-1} \left(\sum_{k=0}^{i} a_k M^k\right)\ . \tag{29}$$

Note that (29) only requires the computation of two matrix polynomials and one matrix inversion. Often the matrix polynomials may be computed with just a few matrix multiplications. As an example, note that when H is an Erlang distribution of order n, i.e., E_n, with mean μ, then $h(s) = (1 + \mu s/n)^{-n}$ and (27) is evaluated as

$$M' = \left(I + \frac{\mu}{n} M\right)^{-n}\ . \tag{30}$$

If $n = 2^m$, for an integer m, then only m matrix multiplications are needed. For example, evaluating (27) when H is E_{1024} requires one matrix inverse and ten matrix multiplications. This example also shows that we can approach very close to the practically important, non-Coxian, deterministic distribution with relatively few matrix operations. (Of course, the results in this paper apply directly to general service-time distributions, but computing integrals like (27) even in the deterministic case is more computationally intensive than in the E_n case, for large n. Moreover, it is harder to get high accuracy in the transform inversion algorithms for deterministic service due to discontinuities in the derivative of the waiting time cdf in that case).

Application of the inversion formulas in Choudhury, Lucantoni and Whitt [12] to the transient workload (21) and the transient queue length (24) respectively, requires numerous evaluations of the matrix $G(s)$ for complex s. It was first

proved in Lucantoni [82] that $G(s)$ can be computed by successively iterating in (7) starting with $G_0(s) \equiv 0$. That proof exploited the fact (which apparently was not well known even for the *M/G/1* queue) that the successive iterates were in fact Laplace-Stieltjes transforms of first passage times along restricted sets of sample paths. (Similar arguments were later used in Latouche [83] to study various iterative algorithms in the general matrix paradigms.) Neuts [2] later proved convergence of the iterations using arguments from functional analysis. It was also shown in Choudhury, Lucantoni, and Whitt [11] that starting the iterations with $G_0(s) \equiv 0$ and $G_0(s) \equiv G$, respectively, produce the sequences of iterates, $\underline{G}_k(s)$ and $\overline{G}_k(s)$ where the matrices $\underline{G}_k(s)$ and $\overline{G}_k(s)$ are themselves Laplace-Stieltjes transforms of distribution functions $\underline{F}_k(x)$ and $\overline{F}_k(x)$. These distributions bound the true distribution in the sense that

$$\underline{F}_k(x) \leq \widehat{G}(x) \leq \overline{F}_k(x) ,$$

for all x, where $\widehat{G}(x)$ is the distribution function with *LST* $G(s)$. (This extends results for the $M/G/1$ queue in Abate and Whitt [84].) Thus, stopping the iteration at any point will give useful bounds on the true distributions. We usually carry out the computations until the bounds are within 10^{-12}. It is clear that the numerous evaluations of $G(s)$ may constitute a significant component of the total computational cost of evaluating the transient distributions; see §6.6. In the examples we have run so far, in order to get successive iterates to within 10^{-13}, the number of iterations has ranged from several tens to a few hundreds.

6.2 Computing Moments and Asymptotics of the Distributions

Explicit formulas for the moments of the distributions discussed in this paper can be derived from the corresponding transform expressions. The first two moments for the number served during and the duration of the busy period in the general $M/G/1$ paradigm were first derived in Neuts [70] [85]. Simpler expressions in the case of the $BMAP/G/1$ queue were derived in Choudhury, Lucantoni and Whitt [11]. The explosion in complexity of the expressions precludes explicit formulas for higher moments of the busy period.

A recursive algorithm for computing the moments of the queue length distribution is given in 3.3.11–13 of Neuts [2]. These require the evaluation of successive derivatives of $A(z)$ at $z = 1$. Recursive expressions for the moments of the virtual and actual delay distributions are given in Lucantoni and Neuts [9]. (We note that these correct several typos in the first two moments given in Equations (47) and (48) of Lucantoni [7]). These recursive expressions are efficient for the low order moments but lose precision if higher moments are required.

A recent algorithm by Choudhury and Lucantoni [86] has been found to be very effective in computing a large number of moments from a probability transform. Also, the high order moments can be used to compute the exact asymptotic parameters of the distributions; see [86] and Choudhury, Lucantoni and Whitt [87]. The asymptotic parameters can be used in approximating high percentiles or as a stopping criterion for the numerical inversion of the distributions; see

[87] for more details. We also note that an alternative technique for computing the asymptotic parameters for the delay and queue length distributions is given in Abate, Choudhury and Whitt [20].

6.3 Numerical Transform Inversion

The distributions presented in this paper can be computed by numerically inverting the corresponding transforms. We recommend first computing the asymptotic parameters as discussed above to be used as a stopping criteria for an inversion algorithm such as the one presented in Abate and Whitt [68] for the inversion; see [8], [11], [12], [81] and [88] for applications and discussions of this.

6.4 Computing the Stationary Queue Length Densities

We discuss two procedures for computing the queue length density. The first is a recursive scheme due to Ramaswami [89] and is a generalization of a device by P. J. Burke that eliminates the loss of precision in computing the queue length density in the $M/G/1$ queue; see p. 186 of Neuts [90]. This algorithm requires the explicit computation and storage of the matrices A_n, $n \geq 0$. An efficient algorithm for computing these is given in Lucantoni [7]. We have implemented Ramaswami's algorithm and generally find it very effective. However, cases where it loses precision for small tail probabilities have been reported by Wang and Silvester [91] [92]. One possibility for this loss of significance is that for very bursty arrival processes it may not be possible to compute a sufficient number of A_n matrices (with enough precision) in order to compute the queue length density to a high degree of accuracy. These problems were avoided by first computing the asymptotic tail by the methods mentioned in §6.2.

The second technique is to invert the transform directly using the algorithms mentioned in §6.3. These do not require the computation of the sequence $\{A_n\}$ but instead need the evaluation of $A(z)$ for several complex values of z, as discussed in §6.1. The inversion is most effectively done by first computing the asymptotic parameters discussed in §6.2 and using these as a stopping rule for the direct inversion of the queue length.

6.5 Computing Waiting Time Distributions

The waiting time distributions given by the transforms (14)–(16) can be computed by numerically solving the associated Volterra integral equations (see, e.g., Neuts [93] [2]) or by numerically inverting the transforms directly (see e.g., Abate and Whitt [68] and Choudhury, Lucantoni and Whitt [81]).

6.6 Computing Transient Distributions

We compute the transient queue length and waiting time distributions by numerically inverting the double transforms in (24) and (21), respectively, using

new multi-dimensional transform inversion algorithms presented in Choudhury, Lucantoni and Whitt [12]. Note that $G(s)$ needs to be computed a large number of times (particularly if the inversion is needed for several values of queue length/waiting time and for several time points). However, upon closer examination of the inversion algorithms it is apparent that not all of the evaluations of $G(s)$ are at distinct s values. For a particular example in [8], $G(s)$ is needed 250,000 times but only at 500 distinct s values. By precomputing and storing these matrices for later use, the computational burden is greatly reduced. In fact, the computational effort for evaluating $G(\cdot)$ becomes an insignificant fraction of the overall computation. This allows feasible numerical computations for more complicated models such as the $BMAP_t/G_t/1$ discussed in §8 below.

7 Numerical Results

The computability of the results in this paper has been demonstrated recently in Choudhury, Lucantoni and Whitt [23] [8] and already put to practical uses. For example, in [23] we study the effectiveness of recently proposed "effective bandwidth" measures to be used for call admission algorithms in ATM networks. In particular, we compute the stationary delay and queue length distributions for a superposition of up to sixty identical bursty sources (modeled as *MMPP*'s). We found that effective bandwidth could lead to extremely conservative or extremely optimistic predictions depending on whether the individual streams were more or less bursty then Poisson, respectively. See [23] for more discussion of this and for references on effective bandwidth.

The transient delay and queue length distributions are computed in [8]. There we consider a *BMAP* corresponding to the superposition of four i.i.d. *MMPP*'s, each having two environment states, and a gamma service-time distribution. As should be anticipated, these examples show, that the transient distributions can be very different from the steady-state distributions. We also compute the transient distributions when $\rho > 1$. This has applications to studying overloads in situations that cannot be studied by stationary models.

8 Current and Future Work

The $MAP/G/1$ queue with server vacations was analyzed in Lucantoni, Meier-Hellstern and Neuts [6], where it was shown that known factorization theorems in the $M/G/1$ queue with vacations carry over to the *MAP* case. Several new factorizations were also derived there. It is easy to generalize the expressions in [6] to the $BMAP/G/1$ queue with server vacations. The $BMAP/G/1$ queue with a more general vacation discipline than in [6] is analyzed in Ferrandiz [95].

All of the results for the $BMAP/G/1$ queue generalize to the $BMAP/SM/1$ queue; that is a single server queue with a *BMAP* arrival process where successive service times form a semi-Markov process; see Lucantoni and Neuts [95] for the analogous stationary results.

Recently, results have been obtained for the $MMPP/G/1$ queue where the service time depends on the phase at arrivals (Asmussen [56], Zhu and Prabhu [58] [100]) and also for the corresponding $BMAP/G/1$ queue in He [96] and Takine and Hasegawa [97]. These papers allow a superposition of traffic streams each with its own service time distribution. Note that the latter paper also contains the transform expression for the transient workload and an application to priority queues with $BMAP$ arrivals.

Application of the transient results in [8] to the $M_t/G_t/1$ and $BMAP_t/G_t/1$ queues where the arrival processes and service time distributions are fixed on nonoverlapping intervals is currently in progress; see Choudhury, Lucantoni and Whitt [98] [99]. Here a novel use of the double transform inversion routines reduces the enormous amount of information that needs to be carried along at successive intervals.

From this tutorial it should be clear that successful algorithms for the $BMAP/G/1$ queue have benefitted from several different approaches. The combination of matrix analytic methods, transform inversion, embedded Markov chain analysis, techniques based on workload processes, etc., has yielded the possibility of carrying the algorithmic approach to these classes of queues to much greater levels than imagined possible.

Acknowledgements: The author thanks Gagan Choudhury, V. Ramaswami and Ward Whitt for many useful comments on this manuscript, Pat Wirth for her encouragement and support of this work and Susan Pope for converting the original TROFF manuscript to LaTeX.

References

1. Neuts, M. F., *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* Baltimore: The Johns Hopkins University Press, 1981.
2. Neuts, M. F., *Structured stochastic matrices of M/G/1 type and their applications.* New York: Marcel Dekker, 1989.
3. Neuts, M. F., A versatile Markovian point process. *J. Appl. Prob.,* **16**, (1979) 764–79.
4. Ramaswami, V., The $N/G/1$ queue and its detailed analysis, *Adv. Appl. Prob.*, **12**, (1980) 222–61.
5. Lucantoni, D. M. and Ramaswami, V., Efficient algorithms for solving the non-linear matrix equations arising in phase type queues, *Stoch. Models*, **1**, (1985) 29–52.
6. Lucantoni, D. M., Meier-Hellstern, K. S, Neuts, M. F., A single server queue with server vacations and a class of non-renewal arrival processes, *Adv. Appl. Prob.*, **22**, (1990) 676–705.
7. Lucantoni, D. M., New results for the single server queue with a batch Markovian arrival process, *Stoch. Mod.*, **7**, (1991) 1–46.
8. Lucantoni, D. M., Choudhury, G. L., and Whitt, W., The transient $BMAP/G/1$ queue, submitted to *Stoch. Models*, 1993.
9. Lucantoni, D. M., and Neuts, M. F., The customer delay in the single server queue with a batch Markovian arrival process, submitted for publication, 1993.

10. Lucantoni, D. M., and Neuts, M. F., Simpler proofs of some properties of the fundamental period of the $MAP/G/1$ queue, *J. Appl. Prob.*, to appear in April, 1994.
11. Choudhury, G. L., Lucantoni, D. M., Whitt, W., The distribution of the duration and number served during a busy period in the $BMAP/G/1$ queue, in preparation.
12. Choudhury, G. L., Lucantoni, D. M., Whitt, W., Multi-dimensional transform inversion with applications to the transient $M/G/1$ queue, submitted for publication, 1993.
13. Bellman, R., *Introduction to Matrix Analysis*, New York: McGraw Hill, 1960.
14. Narayana, S. and Neuts, M. F., The First Two Moment Matrices of the Counts for the Markovian Arrival Process, *Stoch. Models*, **8**, (1992) 459–477.
15. Neuts, M. F., Probability distributions of phase type. In *Liber Amicorum Prof. Emeritus H. Florin*, Department of Mathematics. Belgium: University of Louvain, (1975) 173–206.
16. Heffes, H. and Lucantoni, D. M., A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE J. on Selected Areas in Communication, Special Issue on Network Performance Evaluation*, **6**, (1986) 856–868.
17. Latouche, G., A phase-type semi-Markov point process. *SIAM J. Alg. Disc. Meth.*, **3**, (1982) 77–90.
18. Pacheco, A. and Prabhu, N. U., Markov-compound Poisson arrival processes, Tech. Rept. No. 1059, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853, 1993.
19. Graham, A., *Kronecker Products and Matrix Calculus With Applications.* Chichester: Ellis Horwood Limited, 1981.
20. Abate, J., Choudhury, G. L. and Whitt, W., Asymptotics for steady-state tail probabilities in structured Markov queueing models, *Stoch. Models*, to appear 1994.
21. Choudhury, G. L. and Whitt, W., Heavy-traffic approximations for the asymptotic decay rates in $BMAP/GI/1$ queues, submitted to *Stoch. Models*, 1992.
22. Neuts, M. F. The caudal characteristic curve of queues. *Adv. Appl. Prob.*, **18**, (1986) 221-54.
23. Choudhury, G. L., Lucantoni, D. M., and Whitt, W., Squeezing the most out of ATM, submitted for publication, 1993.
24. Choudhury, G. L., Lucantoni, D. M., and Whitt, W., Tail probabilities in queues with many independent sources, in preparation.
25. Choudhury, G. L., Lucantoni, D. M., and Whitt, W., On the effectiveness of effective bandwidths for admission control in ATM networks, submitted to the *14th International Teletrtaffic Congress*, 1994.
26. Neuts, M. F., Models based on the Markovian arrival process, *IEICE Trans. Commun.* Vol. E75-B, No. 12, (1992) 1255–65.
27. Asmussen, S., and Koole, G., Marked point processes as limits of Markovian arrival streams, *J. Appl. Prob.*, to appear.
28. Olivier, C. and Walrand, J., On the existence of finite dimensional filters for Markov modulated traffic, *J. Appl. Prob.*, to appear.
29. Blondia, C., A discrete-time Markovian arrival process, *RACE Document* PRLB-123-0015-CD-CC, August 1989.
30. Blondia, C., A discrete time batch Markovian arrival process as B-ISDN traffic model, to appear in *JORBEL*, 1993.

31. Blondia, C. and Theimer, T., A discrete-time model for ATM traffic, *RACE Document* PRLB-123-0018-CD-CC/UST-123-0022-CD-CC, October 1989.
32. Blondia, C. and Casals, O., Statistical multiplexing of VBR sources: a matrix analytic approach, *Perf. Eval.*, **16**, (1992) 5–20.
33. Briem, U., Theimer, T. H. and Kröner, H., A general discrete-time queueing model: analysis and applications, *TELETRAFFIC AND DATATRAFFIC in a period of change*, ITC-13, A.Jensen and V. B. Iversen (Eds.), Elsevier Science Publishers B.V. (North-Holland) 1991.
34. Herrmann, C., Analysis of the discrete-time $SMP/D/1/s$ finite buffer queue with applications in ATM, *Proc. of IEEE Infocom '93*, San Francisco, March 28–29, (1993) 160–7.
35. Ohta, C., Tode, H., Yamamoto, M., Okada, H. and Tezuka, Y., Peak rate regulation scheme for ATM networks and its performance, *Proc. of IEEE Infocom '93*, San Francisco, March 28-29, (1993) 680–9.
36. Alfa, A. S. and Neuts, M. F., Modelling vehicular traffic using the discrete time Markovian arrival process, submitted for publication, 1991.
37. Berger, A., Performance analysis of a rate-control throttle where tokens and jobs queue, *J. of Selected Areas in Communications*, **9**, (1991) 165–170.
38. Garcia J. and Casals O., Priorities in ATM Networks, *High-capacity local and metropolitan area networks. Architecture and performance issues* G. Pujolle (Ed.) NATO ASI Series, Springer-Verlag, (1990) 527–1536.
39. Garcia J. and Casals O., Space priority mechanisms with bursty traffic, *Proc. Internat. Conf. of Distributed Systems and Integrated Communication Networks*, Kyoto, (1991) 363–82.
40. Ramaswami, V. and Latouche. G., Modeling packet arrivals from asynchronous input lines, In *"Teletraffic Science for New Cost-Effective Systems, Networks and Services,"* ITC-12, M. Bonatti Ed., North-Holland, Amsterdam, (1989) 721–27.
41. Latouche, G. and Ramaswami, V., A unified stochastic model for the packet stream from periodic sources, *Perf. Eval.*, **14**, (1992) 103-21.
42. Neuts, M. F., Modelling data traffic streams, *TELETRAFFIC AND DATATRAFFIC in a period of change*, ITC-13, A.Jensen and V. B. Iversen (Eds.), Elsevier Science Publishers B.V. (North-Holland) 1991.
43. Neuts, M. F., Phase-type distributions: a bibliography, (1989) Working Paper.
44. Daley, D. J. and Vere-Jones, D., *An Introduction to the Theory of Point Processes*, Springer-Verlag, 1988.
45. Wold, H., On stationary point processes and Markov chains. *Skand. Aktuar.*, **31**, (1948) 229–240.
46. Rudemo, M, Point processes generated by transitions of Markov chains, *Adv. Appl. Prob.*, **5**, (1973) 262–286.
47. Kuczura, A., The interrupted Poisson process as an overflow process, *Bell. Syst. Tech. J.*, **52**, (1973) 437–48.
48. Neuts, M. F., The $M/M/1$ queue with randomly varying arrival and service rates, *Opsearch*, **15**, (1978) 139-57.
49. Neuts, M. F., Further results on the $M/M/1$ queue with randomly varying rates, *Opsearch*, **15**, (1978) 158-68.
50. Heffes, H. A class of data traffic processes — Covariance function characterization and related queueing results, *Bell Syst. Tech. J.*, **59**, (1980) 897–929.
51. van Hoorn, M. H., and Seelen, L. P., The $SPP/G/1$ queue: a single server queue with a switched Poisson process as input process. *O. R. Spektrum*, **5**, (1983) 207–218.

52. Burman, D. Y. and Smith, D. R., An asymptotic analysis of a queueing system with Markov-modulated arrivals, *Oper. Res.*, **34**, (1986) 105–119.
53. Rossiter, M., The switched Poisson process and the $SPP/G/1$ queue. *Proceedings of the 12th International Teletraffic Congress*, 3.1B.3.1-3.1B.3.7, Torino, 1988.
54. Ide, I., Superposition of interrupted Poisson processes and its application to packetized voice multiplexers, *Proceedings of the 12th International Teletraffic Congress*, 3.1B.2.1-3.1B.2.7, Torino, 1988. Torino, 1988.
55. Baiocchi, A., Blefari Melazzi, N., Listanti, M., Roveri, A. and Winkler, R., Loss performance analysis of an ATM multiplexer, *IEEE J. on Selected Areas in Communication, Special Issue on Teletraffic Analysis of ATM Systems*, **9**, (1991) 388–93.
56. Asmussen, S., Ladder heights and the Markov-modulated $M/G/1$ queue, submitted for publication, 1993,
57. Neuts, M. F., Renewal processes of phase type, *Nav. Res. Logist. Quart.*, **25**, (1978) 445–54.
58. Zhu, Y. and Prabhu, N. U., Markov-modulated queueing systems, *Queueing Systems*, **5**, (1989) 215–46.
59. Ramaswami, V., The $N/G/\infty$ queue, Tech. Rept., Dept. of Math., Drexel Univ., Phila., Pa., October, 1978.
60. Neuts, M. F., The c-server queue with constant service times and a versatile Markovian arrival process. In *"Applied probability - Computer science: The Interface,"* Proc. of the Conference at Boca Raton, Florida, Jan. 1981, R. L. Disney and T. J. Ott, eds, Boston: Birkhüser, Vol I, 31–70, 1982.
61. Blondia, C. The $N/G/1$ finite capacity queue, *Stoch. Models*, **5**, (1989) 273–94.
62. Saito, H., The departure process of an $N/G/1$ queue, *Perf. Eval.*, **11**, (1990) 241–251.
63. Sengupta, B., Markov processes whose steady state distribution is matrix-exponential with an application to the $GI/PH/1$ queue. *Adv. Appl. Prob.*, **21**, (1989) 159–80.
64. Neuts, M. F., The fundamental period of the queue with Markov-modulated arrivals, In *Probability, Statistics and Mathematics, in Honor of Professor Samuel Karlin*, J. W. Anderson, K. B. Athreya and D. L. Iglehart (eds.) Academic Press, NY 1989, pp. 187–200.
65. Ramaswami, V., From the matrix-geometric to the matrix-exponential, *Queueing Systems*, **6**, (1990) 229–60.
66. Machihara, F., A new approach to the fundamental period of a queue with phase-type Markov renewal arrivals, *Stoch. Models*, **6**, (1990) 551–60.
67. Asmussen, S., Phase-type representations in random walk and queueing problems, *The Annals of Probability*, **20**, (1992) 772–789.
68. Abate, J. and Whitt, W., The Fourier-series method for inverting transforms of probability distributions, *Queueing Systems*, **10**, (1992) 5–88.
69. Gantmacher, F. R., *The Theory of Matrices, Vol. 1*, New York: Chelsea, 1977.
70. Neuts, M. F., Moment formulas for the Markov renewal branching process. *Adv. Appl. Prob.*, **8**, (1976) 690–711.
71. Takács, L., *Introduction to the Theory of Queues*, New York: Oxford University Press, 1962.
72. Kendall, D. G., Some problems in the theory of queues, *J. Roy. Statist. Soc.*, Ser. B **13**, (1951) 151–185.
73. Baiocchi, A., Asymptotic behaviour of the loss probability of the $MAP/G/1/K$ queue, Part I: Theory, submitted for publication, 1992.

74. Burke, P. J., Delays in single-server queues with batch input, *Oper. Res.*, **23**, (1975) 830–833.
75. Beneš, V., *General Stochastic Processes in the Theory of Queues*, Reading, MA: Addison-Wesley, 1963.
76. Çinlar, E., The time dependence of queues with semi-Markovian service times. *J. Appl. Prob.*, **4**, (1967) 356–64.
77. Neuts, M. F., The single server queue with Poisson input and semi-Markov service times, *J. Appl. Prob.*, **3**, (1996) 202–230.
78. Neuts, M. F., Two queues in series with a finite, intermediate waitingroom, *J. Appl. Prob.*, **5**, (1968) 123–42.
79. Lucantoni, D. M., Further transient analysis of the $BMAP/G/1$ queue, in preparation.
80. Abate, J. and W. Whitt, Numerical Inversion Of Probability Generating Functions. *OR letters.* 12/4, (1992) 2450–251.
81. Choudhury, G. L., Lucantoni, D. M., and Whitt, W., An algorithm for a large class of $G/G/1$ queues, in preparation.
82. Lucantoni, D. M., *An Algorithmic Analysis of a Communication Model with Retransmission of Flawed Messages.* London: Pitman, 1983.
83. Latouche, G., Algorithms for infinite Markov chains with repeating columns, *IMA Workshop on Linear Algebra, Markov Chains and Queueing Models*, January, 1992.
84. Abate, J., and Whitt, W., Solving probability transform functional equations for numerical inversion, *OR Letters*, **12**, (1992) 275–281.
85. Neuts, M. F., The second moments of the absorption times in the Markov renewal branching process, *J. Appl. Prob.*, **15**, (1978) 707–714.
86. Choudhury, G. L. and Lucantoni, D. M., Numerical computation of the moments of a probability distribution from its transform, *Oper. Res.*, to appear, 1994.
87. Choudhury, G. L., Lucantoni, D. M. and Whitt, W., Asymptotic analysis based on the numerical computation of a large number of moments of a probability distribution, in preparation.
88. Choudhury, G. L., Lucantoni, D. M. and Whitt, W., Numerical transform inversion to analyze teletraffic models, submitted to the *14th International Teletrtaffic Congress*, 1994.
89. Ramaswami, V., Stable recursion for the steady state vector for Markov chains of $M/G/1$ type, *Stoch. Models*, **4**, (1988) 183–88.
90. Neuts, M. F., Algorithms for the waiting time distributions under various queue disciplines in the $M/G/1$ queue with service time distribution of phase type, in *"Algorithmic Methods in Probability,"* TIMS Studies in Management Sciences, No. 7, London: North Holland Publishing Co. (1977) 177–97.
91. Wang S. S. and Silvester, J. A., A discrete-time performance model for integrated service ATM multiplexers, submitted for publication, 1993.
92. Wang S. S. and Silvester, J. A., A fast performance model for real-time multimedia communications, submitted for publication, 1993.
93. Neuts, M. F., Generalizations of the Pollaczek-Khinchin integral equation in the theory of queues, *Adv. Appl. Prob.*, **18**, (1986) 952–90.
94. Ferrandiz, J. M., The $BMAP/G/1$ queue with server set-up times and server vacations, *Adv. Appl. Prob.*, **25**, (1993) 235–54.
95. Lucantoni, D. M. and Neuts, M. F., Some steady-state distributions for the $MAP/SM/1$ queue, submitted for publication, 1993.
96. He, Q-M., Queues with marked customers, submitted for publication, 1993.

97. Takine, T., and Hasegawa, T., The workload process of the $MAP/G/1$ queue with state-dependent service time distributions and its application to a preemptive resume priority queue, submitted for publication, 1993.
98. Choudhury, G. L., Lucantoni, D. M. and Whitt, W., Numerical solution of the $M_t/G_t/1$ queue, submitted for publication.
99. Choudhury, G. L., Lucantoni, D. M. and Whitt, W., Numerical solution of the $BMAP_t/G_t/1$ queue, in preparation.
100. Zhu, Y. and Prabhu, N. U., Markov-modulated $PH/G/1$ queueing systems, *Queueing Systems*, **9**, (1991) 313–22.