The *BMAP/M/s* queue

M.B. Combé

Department of Operations Research, Statistics, and System Theory

Report BS-R9435 December 1994

CWI is the National Research Institute for Mathematics and Computer Science. CWI is part of the Stichting Mathematisch Centrum (SMC), the Dutch foundation for promotion of mathematics and computer science and their applications.
SMC is sponsored by the Netherlands Organization for Scientific Research (NWO). CWI is a member of ERCIM, the European Research Consortium for Informatics and Mathematics.

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

The $BMAP/M/s$ queue

M.B. Combé
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

The paper presents an analysis of the number of customers in the $BMAP/M/s$ queueing model. The analysis supports an overview of current methods for analysing Markov modulated queueing systems. In particular we concentrate on three methods that exploit homogeneity structures in the Markov process that describes the number of customers in the $BMAPM/M/s$ system. With each of the three methods we derive the generating function of the stationary probability vector of this Markov process.

AMS Subject Classification (1991): 60K25, 90B22
Keywords & Phrases: multi-server, batch Markovian arrival process, Markov modulated, queue length
Note: The author was supported by NFI.

1 INTRODUCTION

Over the past decades there has been an increasing interest for the application of Markov modulated processes in queueing theory. Two main reasons can be given. Firstly, the performance measures of Markov modulated queueing systems can often efficiently and rapidly be evaluated, which makes these queueing models of practical use. Secondly, modern queueing characteristics such as dependence structures in the input processes (e.g. bursty arrival processes that lead to correlated interarrival times) can fairly adequately be modelled by relatively simple Markov Processes, whereas the classical $GI/G/1$ queueing models often just lack the degree of freedom for modelling such features. An illustrative example is traffic in a communication network; the arrival process of data packets is typically not a Poisson or renewal process, but the characteristics of such an arrival stream might already be captured by a two- or three-state Markov process; the state of such a Markov process describes the actual characteristics of the arrival process in more detail than the (single-parameter) Poisson process (cf. Grünenfelder & Robert[12] and Bonomi et al.[4]).

Currently there exist a number of methods to analyse Markov modulated queueing models; most of them can be viewed as matrix generalizations of the existing analysis methods for classical queueing models and Markov processes.

In this paper we analyse the queue length process of the $BMAP/M/s$ queue, i.e. the multi-server single-queue system in which customers arrive in batches according to a BMAP and in which the service times of individual customers are exponentially distributed.

A first motivation for studying the $BMAP/M/s$ queue is that it enables the modelling and analysis of a wide range of interesting queueing situations with parallel servers. This is due to the versatility of the BMAP.

A second motivation for studying the $BMAP/M/s$ queue is that it allows us to give an overview and classification of current methods for analysing Markov modulated queueing models.

In the remainder of this section we present a description of the $BMAP/M/s$ queue, a survey of literature related to this queueing model, and an outline of the paper.

Report BS-R9435
ISSN 0924-0659
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Model description

The $BMAP/M/s$ queue is the multi-server single-queue system in which customers arrive in batches according to a BMAP and in which the service times of customers are independent and identically exponentially distributed with parameter μ. Customers of different batches are served in order of the arrival of the batches they belong to, and customers arriving in a same batch are served in random order.

In the BMAP, the interarrival and service times of customers are directed by a continuous time Markov process $\{\mathbf{J}(t), t \geq 0\}$ on a finite state space E. A transition from a state i to a state j may induce the arrival of a *batch customer*, the size of the batch depending on i and j. Let the sojourn time in state $i \in E$ be exponentially distributed with parameter $\lambda_i > 0$, and given that a transition takes place, let p_{ij} be the probability of a transition from i to a state $j \in E\backslash\{i\}$, $\sum_{j \in E\backslash\{i\}} p_{ij} = 1$. Defining $p_{ii} = -1$, then the generator of the Markov process $\{\mathbf{J}(t), t \geq 0\}$ is given by the matrix $D := (p_{ij}\lambda_i)$. Next, conditional on a transition from a state i to a state j, the probability of a batch arrival of size k is denoted by q_{ij}^k, $k = 0, 1, \ldots$, where q_{ij}^0 may be interpreted as the probability of having no arrival or of the arrival of an empty batch. With this notation, $p_{ij}q_{ij}^k\lambda_i$ is the transition rate from state i to state j inducing a batch arrival of size k. The BMAP is completely characterized by the sequence of matrices $D_k := \left(p_{ij}q_{ij}^k\lambda_i\right)$, $k = 0, 1, \ldots$, with $\sum_k D_k = D$.

In this paper we concentrate on the analysis of the stationary distribution of the two-dimensional Markov process $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$, where $\mathbf{X}(t)$ denotes the number of customers in the system at time t, $(\mathbf{X}(t) \in \{0, 1, \ldots\})$. We remark that in general $\mathbf{X}(t)$ is not a Markov process, however the process $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$ is Markovian.

We observe that when $\mathbf{X}(t) \geq s$, i.e. when all servers are busy, the transition dynamics of the process $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$ are identical to the transition dynamics of an associated $BMAP/M/1$ queue, for which the same BMAP describes the arrival process of batches but in which the service times of individual customers are exponentially distributed with parameter $s\mu$. Obviously, as the set $\{(i,j)|0 \leq i < s, j \in E\}$ is finite, the ergodicity condition of $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$ of the $BMAP/M/s$ queue is $\pi \sum_{k=1}^{\infty} kD_k e < s\mu$, which is the stability condition of the associated $BMAP/M/1$ queue (cf. Lucantoni[16]).

Related literature for this model

Over the years many studies on multi-server single-queue models have appeared. The first model we mention here is the $GI/M/s$ queue; this extension of the $GI/M/1$ model is well understood (cf. Wolff[29, p.398-399]). Considering the embedded Markov chain of the number of customers at arrival moments, it is seen that the transition structures in the $GI/M/1$ and $GI/M/s$ are quite similar; they only differ for the subset of boundary states. As a consequence, the tails of the stationary distributions in both embedded Markov processes have a geometrical form (cf. Neuts[20]). When the arrival process is of MAP type the queueing process is generally referred to as a Quasi Birth-and-Death process (QBD). This type of Markov process again possesses the geometric property for the tail of the stationary distribution (cf. Neuts[20]).

The $GI^X/M/s$ queue, the multiple exponential server queue with batches arriving according to a renewal process has been studied in Baily & Neuts[2] and recently in Zhao[30]. In [2] the embedded Markov process describing the queue length at moments of arrivals is analysed with matrix-analytic techniques. In [30] it is shown that the stationary probabilities for the number of customers in the system can be written as a linear combination of s geometric terms.

A special case of the $BMAP/M/s$ queue is the one-dimensional variant, being the $M/M/s$ queue with batch arrivals. This model is included in [2], but here the queue length process itself is Markovian, and the stationary distribution of the number of customers in the system can also be obtained by means of generating function techniques directly applied to the stationary queue length process.

Another special case of the current model is the $N/D/s$ queue, in which the service times of customers

are deterministically distributed and where the N denotes the N-process, which is equivalent to the BMAP. This queueing model has been investigated by Neuts[21, section 5.5B]. However, as pointed out in Neuts[21, p.345] the results appear to be of theoretic use only and are hard to implement numerically.
A last group of models we mention here is related to the $GI/Ph/s$ queue. In the $BMAP/M/s$ queue the individual customers in the batch arrivals are served separately, in the $GI/Ph/s$ queue we have single arrivals but for special cases each customer can be viewed as a batch of customers that will be served as a group by one server. For instance, a single customer with an Erlang distributed service time with k phases can be viewed as a batch of k individual customers (each with an exponentially distributed service time) which all have to be served by the same server. The $GI/Ph/s$ queue and variants, for instance non-homogeneous servers, are studied in the book of Neuts[20]. A particular variant of the $GI/Ph/s$ queue is the $GI/H_m/s$, with H_m indicating a hyper-exponential distribution. This model is analysed by de Smit[25, 26] by means of the Wiener-Hopf factorization technique. This group of models leads to Markov processes with a two-dimensional state space which have the common aspect that the state-space component describing the state of the servers grows exponentially with the number of servers and number of phases of the service time.

Outline of the paper
Section 2 is devoted to a general overview of Markov modulated models in queueing theory. In section 3 we analyse the stationary joint distribution of the number of customers and the underlying Markov process in the $BMAP/M/s$ queue. We present three methods of deriving this joint distribution. A main result is an iterative scheme to obtain stationary probabilities of boundary states.

2 Analysis of Markov modulated queueing systems

A pioneer in the field of Markov modulated queueing systems is Neuts (cf. Neuts[20, 21]). The main contribution of Neuts to this field is the introduction of a probabilistic and algorithmic approach for the analysis of these queueing models. Before the work of Neuts, the analysis remained of a classical nature, following the lines of the complex analysis of Takács (cf. Takács[28], Loynes[15]) and spectral analysis from linear algebra. Due to the efforts of Neuts and others (Ramaswami, Lucantoni, Latouche), Markov modulated queueing models have become an important tool in the analysis of modern queueing systems.
In recent years a revival of classical analysis methods can be detected, cf. Mitrani & Mitra[18], de Smit[25, 26] and Regterschot and de Smit[23], the reason for this being that the increasing computer capacity and the availability of efficient algorithms which have made these methods as practical as the algorithmic methods of Neuts.

In a Markov modulated queueing system the queueing process is directed by a continuous time Markov process, where the state of this Markov process contains all necessary information about the system characteristics. This might be extra information on random variables, e.g. the remaining interarrival time, but could also describe the current configuration of the system, e.g. the number of available servers. For such queueing systems one can often construct a two-dimensional embedded *Markov process* $\{(\mathbf{X}_n, \mathbf{J}_n), n = 1, 2, \ldots\}$, where $\mathbf{J}_n$ denotes the state of the directing Markov process and with $\mathbf{X}_n$ usually being one of the familiar queueing measures, such as the number of customers at a departure moment. $\mathbf{X}_n$ attains non-negative integer values. This typical two-dimensional state space is known as the *semi-infinite strip*.
An important class of analysis methods directly exploits the homogeneity properties in the transition structure of $\{(\mathbf{X}_n, \mathbf{J}_n), n = 1, 2, \ldots\}$. In the one-dimensional case, i.e. in the $M/G/1$ or $GI/M/1$ queue, by the homogeneity property we mean that transitions in the interior of the state-space only depend on the size of the transition and are independent of the actual states that are involved in the transition. To illustrate this: in the Markov process describing the number of customers at departure

epochs in the $M/G/1$ queue, transitions are related to the number of customers that arrive during a service, and obviously the distribution of the latter does not depend on the number of customers in the queue. In the two-dimensional case the homogeneity property means that, next to depending on the state of the directing process $\mathbf{J}_n$, transition probabilities in the process only depend on the *size of the transition* rather than on the *state* of the $\mathbf{X}_n$ component. We remark that the property of homogeneity only has to exist for the interior states, transitions involving boundary states actually may depend on the state of $\mathbf{X}_n$.
Analysis methods that exploit this homogeneity property are the matrix-analytic approach, the transform method, and the spectral expansion method (these are the common names, usually adopted from papers by their main contributors). Below we discuss them in some detail, also pointing at similarities and key differences.

1. The matrix-analytic technique
The matrix-analytic technique can be viewed as the matrix extension of probabilistic analysis for random walks with homogeneous properties; the derivations of the results are of a probabilistic nature and also many elements of the results have probabilistic interpretations. Basically, due to the homogeneity structure, performance measures can explicitly be described by compact expressions for the generating functions with a few boundary state probabilities to be determined. In the matrix-analytic technique the boundary probabilities are obtained with results from Markov renewal theory.
The probabilistic analysis of the matrix-analytic technique was first developed for the matrix-geometric approach. The latter name refers to the matrix generalization of a special feature of stationary probabilities of the $GI/M/1$ queue; let $\pi_0, \pi_1, \ldots$ denote the stationary probabilities of the number of customers in the system, then $\pi_n = \rho(1-\alpha)\alpha^{n-1}$, $n = 1, 2, \ldots$, i.e. the probabilities have a geometrically distributed tail (cf. [20]). Put differently, $\pi_n = \pi_{n-1}\alpha$. In Markov modulated queues of $GI/M/1$ type, where π_n is a vector, there exists a matrix R such that $\pi_n = \pi_{n-1}R$ for n sufficiently large, i.e. for levels in the homogeneous part of the state space. Not all Markov modulated queueing models possess the matrix-*geometric* property, an example being the $BMAP/M/s$ queue (cf. remark 3.3). For this reason the method introduced by Neuts is usually referred to as the matrix-analytic method.
A very appealing aspect of matrix generalizations of $GI/M/1$ and $M/G/1$ queues is that not only stationary probabilities have a form that is analogous to the one-dimensional case, but more generally the expressions for most performance measures can be viewed as matrix-generalizations of one-dimensional results.
Key references for the matrix-analytic technique are the two books of Neuts[20, 21] and papers by Ramaswami[22], Latouche & Ramaswami[14] and Lucantoni[16, 17].

2. The transform method
The transform method, as Gail et al.[9] call it, is the extension of classical generating function and transform analysis. Central in this method are Rouché type arguments. The generating function expressions are obtained in the same way as for the matrix-analytic technique. For continuous valued processes transform expressions follow from recurrence relations at embedded time points, e.g. departure epochs.
With the transform method boundary probabilities are obtained by exploiting the analytic properties of generating functions of a proper distribution. These properties lead to a set of equations from which the boundary state probabilities can be obtained. Key references for this method are Gail et al.[9, 10].
Once the general behaviour of the queueing system has been characterized in terms of generating functions or Laplace transforms, the matrix-analytic method and the transform method can be viewed as different ways of determining the boundary state probabilities.

3. The spectral-expansion method
This technique directly considers state-space probabilities. First a general form of the stationary

probability vector is derived from the balance equations that describe the interior of the state space. Subsequently the stationary probabilities of the boundary states follow from the balance equations. The general form of the stationary probabilities is determined by a multi-dimensional generalization of the spectral expansion techniques from linear algebra that are used for solving linear difference equations. This method is currently advocated in papers by Mitrani and co-authors (cf. [7, 18, 19]). The spectral-expansion technique is restricted to models for which only a finite subset of the state space is non-homogeneous, this means that from a certain level in the $\mathbf{X}$ component of a state $(\mathbf{X}, \mathbf{J})$, the transitions no longer depend on the value of $\mathbf{X}$. As a consequence, the batch size distributions that can be dealt with by this method must have a finite support. Under this restriction, the spectral-expansion technique and the matrix-*geometric* technique can be viewed as different methods of obtaining the above-mentioned matrix R. This observation is discussed in some more detail in remark 3.4.

Generally, it is hard to tell which of the methods is preferable. In general the matrix-analytic method appears to be the most stable method for numerical performance evaluation since it mainly involves iterations, matrix inversions and integrations. On the other hand, in theory the root determination in the transform method and the spectral-expansion method is of a lower complexity (cf. Mitrani & Mitra[18]).

A method of analysing Markov modulated queueing models that does not directly utilize homogeneity properties is Wiener-Hopf factorization. With this method the queueing process at embedded time points is treated from the perspective of random walks; starting point is a multi-dimensional variant of the well known Lindley equation for $GI/G/1$ queues (cf. p.140 of Takács[28]). Moreover, the Wiener-Hopf convolution integral equation that is connected to this Lindley equation also has a multi-dimensional counterpart. In a series of papers by de Smit[25, 26], Regterschot & de Smit[23], and in the thesis of Regterschot[24] the classical one-dimensional Wiener-Hopf factorization is extended to Markov modulated queues.

REMARK 2.1
From the viewpoint of generality, it appears that the Wiener-Hopf factorization technique has more modelling freedom than the three above-mentioned methods because it does not explicitly use the homogeneity structure of the two-dimensional Markov process. For the methods which do use this structure often all the matrices describing the transition probabilities in the embedded Markov process have to be calculated. These matrices might be quite hard to obtain, for instance they might concern the joint probabilities of the number of service completions during an arrival together with the transitions in the underlying Markov process. □

As a broad class of models can be analysed by any of the above-described methods, it might be interesting to determine the connections between various quantities and steps in the analysis. Asmussen[1] attempts to connect the results from the probabilistic analysis of Neuts and others to the Wiener-Hopf factorization method. This is done by deriving a matrix analog of the probabilistic form of the Wiener-Hopf factorization as it exists for one-dimensional random walks.
In the one-dimensional random walk it is known (cf. p.400 of Feller[8]) that the factorization of an LST into two functions, one function being analytic on $\{\omega \in \mathbb{C} | Re\,\omega \leq \delta\}$ for some $\delta > 0$, the other function being analytic in $\{\omega \in \mathbb{C} | Re\,\omega \geq 0\}$, is equivalent to a factorization of the transition probability function of the random walk into two probability functions, one function having its support on the positive real axis, the other function having its support on the non-positive real axis. Asmussen derives a matrix analog of this probabilistic form of the Wiener-Hopf factorization.
This connects the method of Neuts and the Wiener-Hopf factorization of de Smit. However, except for special cases a close connection between the methods is not established. □

REMARK 2.2
From the viewpoint of numerical performance, it appears to be hard to give a fair comparison of the above mentioned methods. This is mainly due to practical reasons:

(i) Implementation of the methods. The methods perform different operations, for which a number of alternative routines may be available.

(ii) The specific parameters of the queueing model that is being evaluated; for each method the choice of parameters might have a different impact on its performance. In one-dimensional single-server queueing models the performance of an implementation often is closely connected to the workload of the system. However, in the multi-dimensional multi-server case other aspects play a role as well. Typical problems for Markov modulated queueing models arise due to the matrix structure; examples are coinciding eigenvalues of matrices or periodicity of the directing Markov chain.

(iii) Demands for the accuracy of the evaluation. The effect of the desired level of accuracy on running time, number of iterations etc. will not be the same for each method.

(iv) Computer specifications. For larger instances even computer specifications can play a role (memory capacity, instruction set of the processor).

Although a fair numerical comparison appears to be a hard problem, it might still be worthwhile to categorize the weak and strong points of each method. However, we feel that this will be quite a considerable task, involving a vast set of test cases. Some efforts have been made for a special multi-server model with breakdown and repair of the servers, Mitrani & Chakka[19] compare the spectral expansion method (which is the main subject of [19]) with the matrix-analytic method. However, this comparison does not lead to strong conclusions. □

REMARK 2.3
There also exist purely numerical techniques for evaluating performance measures of Markovian queueing systems. One such method is the Power Series Algorithm (cfthe tutorial of Blanc[3], which is based on a power series expansion property for stationary probabilities in queueing systems. And finally, much progress has also been made in developing numerical methods that can deal with general (large) Markov chains. Recent developments for such methods and references are presented in the conference proceedings [27]. □

3 ANALYSIS OF THE $BMAP/M/s$ QUEUE

In this section the analysis of the $BMAP/M/s$ queue guides a further discussion of the matrix-geometric, the transform and the spectral-expansion method. We analyse the two-dimensional process of the number of customers in the $BMAP/M/s$ queue as defined in section 1. As indicated in there, the $BMAP/M/s$ queue features homogeneity properties in the interior of the state space; if all servers are busy the transition structure is that of an ordinary $BMAP/M/1$ queue. However, the $BMAP/M/s$ queue has not yet been treated directly, probably because of the technical complications that result from the fact that the batch size distribution not necessarily has a finite support (e.g., geometric batch size distributions are allowed).
One of the consequences is that in general the tail of the stationary distribution of the number of customers in the system is not of the geometric structure as described in the previous section. This will be explained in remark 3.3.

Basic properties
With $D_0, D_1, \ldots$ describing the BMAP arrival process, where D_i is an $M \times M$ matrix, and with I being the $M \times M$ identity matrix, the generator Q of the two-dimensional Markov process $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$ can be represented as

$$Q = \begin{bmatrix}
D_0 & D_1 & D_2 & D_3 & D_4 & D_5 & \dots & & \\
\mu I & D_0{-}\mu I & D_1 & D_2 & D_3 & D_4 & D_5 & \dots & \\
 & 2\mu I & D_0{-}2\mu I & D_1 & D_2 & D_3 & D_4 & D_5 & \dots \\
 & & \dots & \dots & \dots & & & & \\
 & & & \dots & \dots & \dots & & & \\
 & & & & (s{-}1)\mu I & D_0{-}(s{-}1)\mu I & D_1 & D_2 & \dots \\
 & & & & & s\mu I & D_0{-}s\mu I & D_1 & \dots \\
 & & & & & & s\mu I & D_0{-}s\mu I & \dots \\
 & & & & & & & s\mu I & \dots \\
 & & & & & & & & \dots
\end{bmatrix}$$

FIGURE 3.1: GENERATOR OF $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$.

We see that Q is represented in blocks of $M \times M$ matrices. For a state (i, j) of $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$ the component i denotes the $i-th$ row of blocks in Q, the component j specifies the $j-th$ element within the M states of the $i-th$ block. Moreover, the Markov process $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$ is skip-free to the left, i.e. the path from a level i_0 to a level $i_1 < i_0$ visits all levels between i_0 and i_1.

Denote by $\{(\mathbf{X}, \mathbf{J})\}$ the stationary process of $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$, and let $x = \{x_0, x_1, \ldots\}$ be the stationary probability vector of $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$, with $x_i = \{x_{i1}, \ldots, x_{iM}\}$. Then x satisfies

$$xQ = 0, \quad \sum_{i=0}^{\infty} \sum_{j=1}^{M} x_{ij} = 1. \tag{3.1}$$

This results in the set of balance equations

$$\sum_{i=0}^{k} x_i D_{k-i} - k\mu x_k + (k+1)\mu x_{k+1} = 0, \quad k = 0, \ldots, s-1. \tag{3.2}$$

$$\sum_{i=0}^{k} x_i D_{k-i} - s\mu x_k + s\mu x_{k+1} = 0, \quad k = s, s+1, \ldots. \tag{3.3}$$

The set of balance equations given by (3.2) and (3.3) can be written more concisely in terms of generating functions. Define the vector generating function $X(z) := (\sum_{i=0}^{\infty} x_{ij} z^i)$. Then, with $D(z) := \sum_{i=0}^{\infty} D_i z^i$, we obtain

$$X(z)\,[(1-z)s\mu I + zD(z)] = (1-z) \sum_{k=0}^{s-1} (s-k)\mu x_k z^k. \tag{3.4}$$

We observe that the multi-dimensional functional equation has the same form as the one-dimensional case; expression (3.4) is a matrix generalization of the expression for the generating function of the number of customers in the $M/M/s$ queue with batch arrivals.
We next determine the stationary probability vector x with the three different methods that we discussed in the previous section.

The matrix-analytic technique: a probabilistic approach
In (3.4) we observe that $X(z)$ is characterized up to the unknown vectors $x_0, \ldots, x_{s-1}$. Moreover, from (3.2) we find that once x_0 has been determined $x_1, \ldots, x_{s-1}$ follow successively.

From Markov renewal theory (cf. Cinlar[5]) we know that the mean return time, say $\gamma < \infty$, to a state (i,j) of the stationary process $\{(\mathbf{X},\mathbf{J})\}$ and the stationary probability x_{ij} are connected through $x_{ij} = \frac{1}{\gamma}$. Hence we concentrate on the mean return time to level 0, being the set $\{(0,j)|j = 1,\ldots,M\}$. Next we translate the continuous-time Markov process into a discrete-time Markov process by uniformization and determine the mean return times to x_0 in this discrete-time queue.
Let $\theta > \max_i(-D_{ii}) + s\mu$. The one-step transition probability matrix of the discrete-time Markov chain that is the result of the uniformization is $P := Q/\theta + I$.

Define

$$\begin{aligned} A_i &:= \frac{D_{i-1}}{\theta}, \quad i = 2,3,\ldots, \\ A_{1,j} &:= \frac{D_0 - j\mu I}{\theta} + I, \quad j = 0,\ldots,s, \\ A_1 &:= A_{1,s}, \\ A_{0,j} &:= \frac{j\mu I}{\theta}, \quad j = 0,\ldots,s, \\ A_0 &:= A_{0,s}. \end{aligned}$$

Then

$$P = \begin{bmatrix} A_{1,0} & A_2 & A_3 & A_4 & \ldots & & & & & \\ A_{0,1} & A_{1,1} & A_2 & A_3 & A_4 & \ldots & & & & \\ & A_{0,2} & A_{1,2} & A_2 & A_3 & A_4 & \ldots & & & \\ & & \ldots & \ldots & \ldots & & & & & \\ & & & \ldots & \ldots & \ldots & & & & \\ & & & & A_{0,s-1} & A_{1,s-1} & A_2 & A_3 & A_4 & \ldots \\ & & & & & A_0 & A_1 & A_2 & A_3 & \ldots \\ & & & & & & A_0 & A_1 & A_2 & \ldots \\ & & & & & & & A_0 & A_1 & \ldots \\ & & & & & & & & \ldots & \ldots \end{bmatrix}$$

FIGURE 3.2: ONE-STEP TRANSITION MATRIX P.

A next step is the observation that the return time from level i to level 0 is the convolution of the successive first passage times from level l to level $l-1$, for $l = i, i-1, \ldots 1$. As observed in the introduction, and as can be seen from figure 3.1 or equation (3.3), for levels $i \geq s$ the transitions of the Markov process have the characteristics of the $BMAP/M/1$ queue with service times that are exponentially distributed with parameter $s\mu$. Consequently, the first passage times for levels i to $i-1$, $i \geq s$, follow from the results for the $BMAP/M/1$ queue.
Let $G(z)$ define the matrix generating function for the uniformized Markov chain describing the first passage time from level i to level $i-1$, $i \geq s$, together with the transition probabilities in the directing Markov process.
It follows from the $BMAP/G/1$ theory (cf. Lucantoni[17]) that $G(z)$ is a busy period type generating function satisfying the functional equation

$$G(z) = z\sum_{v=0}^{\infty} A_v G^v(z). \tag{3.5}$$

Equation (3.5) is a matrix-generalization of the familiar branching expression for the number of customers served during the busy period in the $M/G/1$ queue (cf. p.32 of Takács[28]).
For the mean return times we need to evaluate the matrices $G(1)$ and $G'(1)$. Starting with some specific initial matrices, $G(1)$ can be obtained by iterating (3.5), and subsequently $G'(1)$ is obtained by iterating the derivative of (3.5). For the case of single arrivals (MAP), an elegant probabilistic

algorithm is proposed in Latouche & Ramaswami[14] as an alternative evaluation of $G(1)$. This algorithm may well be extended for the case of batch size distributions with finite support.

Next define for $i = 1, \ldots, s-1$, the matrix generating function $G_i(z)$ that describes the first passage time from level i to level $i-1$.
Then (cf. Figure 3.2), for $i = 1, \ldots, s-1$,

$$G_{s-i}(z) = zA_{0,s-i} + z\left(A_{1,s-i}G_{s-i}(z) + \sum_{k=2}^{i} A_k \prod_{n=k-1}^{0} G_{s-i+n}(z) + \sum_{k=i+1}^{\infty} A_k G^{k-i}(z) \prod_{n=1}^{i} G_{s-n}(z)\right). \tag{3.6}$$

In (3.6) we use the convention $\prod_{i=a}^{b} G_i(z) := G_a(z) \times \ldots \times G_b(z)$, for $a \geq b$ as well as for $a \leq b$.
Again, for the mean return times we need to evaluate matrices $G_{s-i}(1)$ and $G'_{s-i}(1)$. In (3.6) we observe that $G_{s-i}(z)$ is expressed in terms of $G(z)$ and $G_j(z)$ with $j \geq s-i$. This leads to a recursive scheme for the matrices $G_{s-i}(1)$, $i = 1, \ldots, s-1$. The matrices $G_{s-i}(1)$ can be expressed directly in earlier computed matrices by rewriting (3.6), or they might be computed by iterating (3.6).
After obtaining $G(1), G_1(1), \ldots, G_{s-1}(1)$ and $G'(1), G'_1(1), \ldots, G'_{s-1}(1)$ we subsequently derive the mean return times from the matrix generating function $K(z)$, where $K_{ij}(z)$ is the generating function of the number of transitions until the first return to level 0 and the probability that the underlying Markov chain is in state j at that time, given that at this moment the two-dimensional Markov chain is in state $(0, i)$. By conditioning on the the first transition that is made after a return to level 0 and using (3.6) we obtain

$$K(z) = z\left(A_{1,0} + \sum_{k=2}^{\infty} A_k G^{[0,k-s]^+}(z) \prod_{l=[k-1,s-1]^-}^{1} G_l(z)\right), \tag{3.7}$$

here $[k_0, k_1]^+ := \max\{k_0, k_1\}$ and $[k_0, k_1]^- := \min\{k_0, k_1\}$.
To derive the mean return times to level 0, we first consider $K(1)$ which is a stochastic matrix describing a Markov chain on the state-space of $\mathbf{J}$, this Markov chain being the embedded Markov process at the moments of visits to level 0. Let $\bar{k}$ be the stationary probability (row)vector of this process, defined by $\bar{k}K(1) = \bar{k}$, $\bar{k}e = 1$, where e denotes the M state unit (column)vector.
Then, from Markov renewal theory (cf. Cinlar[5]) we find that x_0 is given by

$$x_0 = \frac{\bar{k}}{\bar{k}K'(1)e}, \tag{3.8}$$

where $K'(1)$ follows from (3.7).
As stated $x_1, \ldots, x_{s-1}$ now successively follow from (3.2). For example

$$x_1 = \frac{-x_0 D_0}{\mu}, \text{ and } x_2 = x_0 \frac{D_0^2/\mu - D_0 - D_1}{2\mu}. \tag{3.9}$$

Finally, we have arrived at:

THEOREM 3.1
The generating function of the stationary distribution of the number of customers in the $BMAP/M/s$ queue is given by (3.4), where the vector x_0 is given by (3.8) and the $s-1$ vectors $x_1, \ldots, x_{s-1}$ follow from (3.2). □

The transform method: an analytic approach

In equation (3.4) the generating function $X(z)$ is characterized up to s unknown boundary vectors. With the matrix-analytic method x_0 is determined in an iterative fashion, after which $x_1, \ldots, x_{s-1}$ successively follow. With the transform method the vectors of the boundary states are obtained by exploring the analytic properties of generating functions of proper distribution functions. After post-multiplying in (3.4) with the inverse of $[(1-z)s\mu I + zD(z)]$ we have isolated $X(z)$. Since the queueing process is ergodic (cf. section 1), $X(z)$ is analytical in the open unit disk $|z| < 1$ and continuous in $|z| \leq 1$. Consequently singularities in the inverse matrix somehow have to be canceled. The principal idea of the transform method is to use this condition for each singularity and obtain an equation for the vectors $x_0, \ldots, x_{s-1}$. The idea of this method is quite transparent, but there are some technical details connected to the central question: does the set of equations that arises contain enough, namely $s \times M$, independent equations to determine the vectors $x_0, \ldots, x_{s-1}$. The two main problems for Markov modulated queueing models are: (i) the number of points on $|z| = 1$ for which $[(1-z)s\mu I + zD(z)]$ is singular, and (ii) the multiplicity of the singularities for $|z| < 1$.

In Gail et al.[9] these technical aspects are addressed in detail, resulting in a well-rounded theory on the extension of the classical Rouché type analysis to the Markov modulated queueing models of $M/G/1$ type. A main result of Gail et al.[9] is that if the Markov process is ergodic in general there can indeed be derived sufficiently many independent equations for the determination of the stationary distribution. In Gail et al.[10] these results are summarized for a similar Markov process which only slightly deviates from the general model, but the extra assumptions that are made reduce the technical complexity to a great extent.

In the following we apply the results from [9] and [10] to the $BMAP/M/s$ queue.

When we conform the transition structure (cf. Figures 3.1 and 3.2) to the notation of [9] and derive the functional equation of the generating function $X(z)$ from the balance equations, we find an equation that is equal to expression (3.4) but with both sides multiplied with a factor z^{s-1} and the uniformization constant θ:

$$X(z)z^{s-1}\theta\,[(1-z)s\mu I + zD(z)] = z^{s-1}(1-z)\theta \sum_{k=0}^{s-1}(s-k)\mu x_k z^k. \tag{3.10}$$

Due to the ergodicity of the queueing process and the polynomial structure of $[(1-z)s\mu I + zD(z)]$ and $\sum_{k=0}^{s-1}(s-k)\mu x_k z^k$ it follows from [9] that the vectors $x_0, \ldots, x_{s-1}$ can be obtained from the set of equations that results from exploring the analytic properties of $X(z)$.

We next concentrate on the determination of the set of equations that provides $x_0, \ldots, x_{s-1}$.

First we make the assumption that $[(1-z)s\mu I + zD(z)]$ has a single singularity for $|z| = 1$, being $z = 1$. This is assumption (A2) from theorem 1 of Gail et al.[10]. While this assumption is not necessary from a theoretical viewpoint, in practice it avoids a number of additional technical difficulties. In [10] it is claimed that this condition is not restrictive. Under this assumption, it follows from theorem 1 of [10] that the determinant of $z^{s-1}\,[(1-z)s\mu I + zD(z)]$ has exactly $sM-1$ zeros in the open unit disk and a simple zero at $z = 1$.

In the set of $sM-1$ zeros inside the unit disk $z = 0$ has multiplicity $(s-1)M$, the remaining $M-1$ zeros of the determinant are connected to the eigenvalues of $D(z)$.

From Lemma 3 of [9] it follows that $z = 0$ results in $M \times (s-1)M$ equations of which exactly $(s-1)M$ are independent. Fortunately, as this set of equations is connected to the derivatives of (3.10), it follows that the independent set consists of the first $s-1$ equations of (3.2). We remark that the independence of this set also follows from the fact that x_0 automatically results in $x_1, \ldots, x_{s-1}$.

Returning to the functional equation (3.4), we can derive a set of M additional equations which together with (3.2) determine $x_0, \ldots, x_{s-1}$. This is done in the following manner. For certain values of z, $|z| \leq 1$, the determinant of $[(1-z)s\mu I + zD(z)]$ equals zero. By multiplying both sides of (3.4)

for these values of z with the right eigenvector belonging to eigenvalue 0, the left-hand side of (3.4) is canceled. From the analytical properties of $X(z)$ it then follows that also the right-hand side must vanish.
Doing so, let $\eta_i,\ i = 1,\ldots,M$ be the values of $z,\ |z| \leq 1$, for which $[(1-z)s\mu I + zD(z)]$ is singular, with $\eta_1 = 1$. Next we make the assumption that all η_i are simple. Again this assumption is non-restrictive; parameters can always be altered such that the assumption holds while the model is hardly affected. We remark that Gail et al.[9] show that again from a theoretical viewpoint this assumption is not necessary.
Next, let r_i be the vector such that $[(1-\eta_i)s\mu I + \eta_i D(\eta_i)]\, r_i = 0$, noting that r_i is an eigenvector of $D(\eta_i)$ and r_1 the unit vector e. From our last assumption it follows that $r_1,\ldots,r_M$ are independent (cf. p.153 of Lancaster & Tismenetsky[13]).
As stated, due to the regularity of $X(z)$ we obtain for $i = 2,\ldots,M$

$$(1-\eta_i)\left(\sum_{k=0}^{s-1}(s-k)\mu x_k \eta_i^k\right) r_i = 0, \tag{3.11}$$

and

$$\sum_{k=0}^{s-1}(s-k)\mu x_k e = \omega_1^*. \tag{3.12}$$

In (3.12) the constant ω_1^* equals $\lim_{z\to 1}\frac{\omega_1(z)}{1-z}$, with $\omega_1(z)$ the eigenvalue belonging to the eigenvector $r_1(z)$ of $[(1-z)s\mu I + zD(z)]$ and $\lim_{z\to 1}\omega_1(z) = \omega_1(\eta_1) = 0$.
In the next theorem we summarize the above:

THEOREM 3.2
The generating function of the stationary distribution of the number of customers in the $BMAP/M/s$ queue is given by (3.4), where the boundary vectors $x_0,\ldots,x_{s-1}$ follow from equations (3.11),(3.12) and the first $s-1$ equations of (3.2). □

REMARK 3.1
In both the matrix-analytic method and the transform method an important role in the analysis is played by the functional $[(1-z)s\mu I + zD(z)]$. In the matrix-analytic approach the aim is to determine the generating function $G(z)$ as the solution of the functional equation (3.5), in the transform method the aim is to find zeros of the determinant of $[(1-z)s\mu I + zD(z)]$, but after rewriting this functional the latter problem is equivalent to determining the zeros of the determinant of $[zI - A(z)]$, with $A(z) := \sum_{v=0}^{\infty} A_v z^v$. By writing $G(z)$ in the spectral decomposition form (cf. Lancaster & Tismenetsky[13] p.314) it is readily verified that the right eigenvector r_i is also a right eigenvector of $G(\zeta_i)$ for some $\zeta_i \leq 1,\ i = 1,\ldots,M$. Incidentally, the same property not necessarily holds for the left eigenvectors. □

REMARK 3.2
When comparing the matrix-analytic method with the transform method we see that with the matrix-analytic method the operations mainly involve iterations of matrices of size $M \times M$, whereas in the transform method a set of $s \times M$ linear equations with as many variables has to be solved. Although the vectors $x_1,\ldots,x_{s-1}$ can be expressed in terms of x_0 via the balance equations (3.2) (cf. (3.9)), still a set of equations has to be solved. □

Next we derive an explicit expression for $\mathrm{E}\tilde{\mathbf{N}} := X'(1)e$, the mean number of customers in the system.

For convenience we first define for $|z| \leq 1$

$$C(z) := [(1-z)s\mu I + zD(z)], \tag{3.13}$$

and

$$v(z) := \sum_{k=0}^{s-1}(s-k)\mu x_k z^k. \tag{3.14}$$

Then we can rewrite (3.4) into

$$X(z)C(z) = (1-z)v(z), \quad |z| \leq 1. \tag{3.15}$$

Let $l(z)$ and $r(z)$ be the left and right eigenvector of $C(z)$ respectively belonging to the eigenvalue $\omega_1(z)$ that has the property $\omega_1(1) = 0$. Moreover, let $l(z)$ and $r(z)$ be defined such that for $|z| \leq 1$

$$\begin{aligned} l(z)r(z) &= 1, \\ l(1) &= \pi,\ r(1) = e, \\ l(z)e &= 1. \end{aligned}$$

Then by exploiting the analytical properties of $l(\cdot), r(\cdot)$ and $\omega_1(\cdot)$ (cf. [9]), and the identities

$$[C(z) - \omega_1(z)I]r(z) = 0, \quad l(z)[C(z) - \omega_1(z)I] = 0,$$

we find

$$\begin{aligned} \omega_1'(1) &= \pi C'(1)e, \\ r'(1) &= [e\pi + C(1)]^{-1}[\omega_1'(1)I - C'(1)]e, \\ \omega_1''(1) &= \pi C''(1)e + 2\pi C'(1)r'(1). \end{aligned} \tag{3.16}$$

By multiplying both sides of (3.15) to the right with $r(z)$ and taking derivatives we obtain for $z = 1$ after some rewriting:

THEOREM 3.3
$\mathrm{E}\tilde{\mathbf{N}}$, the mean number of customers in the $BMAP/M/s$ queue, is given by

$$\mathrm{E}\tilde{\mathbf{N}} = X'(1)e = \frac{\omega_1''(1)}{2\left(\omega_1'(1)\right)^2} - \frac{1}{\omega_1(1)'}\left(v'(1)e + v(1)r'(1)\right), \tag{3.17}$$

with $\omega_1'(1)$, $\omega_1''(1)$ and $r'(1)$ given by expressions (3.16), and with $v(1)$ and $v'(1)$ following from (3.14).
□

From (3.4) and (3.17) we can readily obtain the *vector* $X'(1)$, with $X_i'(1)$ denoting the mean number of customers in the system together with the probability that the underlying Markov chain is in state i. This is done by directly taking derivatives in (3.4) and taking the limit $z \to 1$. This results in an expression for $X'(1)D$. Multiplying on the right with π in (3.17) gives an expression for $X'(1)e\pi$. Observing that $[e\pi + D]$ is non-singular leads to $X'(1)$. We remark that performance measures such as mean waiting times and number of customers at moments of arrival follow from $X(z)$ and $X'(1)$ via applications of Little's law and the PASTA property.

Spectral expansion: an algebraic approach
This method concentrates on determining the stationary probability vector x directly from the set of balance equations given by (3.2) and (3.3). First a general form for the states in the homogeneous part of the state-space is derived from (3.3), subsequently the vector x is obtained from (3.2) and the normalization constant $\sum_{i=0}^{\infty} x_i e = 1$. The method extends an algebraic technique for solving linear difference equations to the multi-dimensional case. A general outline of the method has been presented in Mitrani & Mitra[18]. The method has been applied to queueing models in several papers; Chakka & Mitrani[19] and Ettl & Mitrani[7] study multi-server models with breakdown and repair of the servers; Elwalid et al.[6] study a communication model with multiplexing of sources by means of a Markov modulated queueing system.
To start the investigation of the $BMAP/M/s$ queue, suppose the batch size distribution has finite support. This means that $\exists k_0$ such that $D_k = 0$ for $k > k_0$.
Then for $k \geq \max\{s, k_0\}$ we can rewrite (3.3) as

$$\sum_{j=0}^{k_0} x_{k-j} D_j - s\mu x_k + s\mu x_{k+1} = 0. \tag{3.18}$$

Define matrices $H_i := D_{k_0-i}$ for $i = 0, \ldots, k_0 - 1$, $H_{k_0} := D_0 - s\mu I$, and $H_{k_0+1} := s\mu I$. Then for $k \geq \max\{s - k_0, 0\}$ equation (3.18) can be written as

$$x_k H_0 + x_{k+1} H_1 + \ldots + x_{k+k_0} H_{k_0} + x_{k+k_0+1} H_{k_0+1} = 0. \tag{3.19}$$

Equation (3.19) is a vector difference equation of order $k_0 + 1$. Its general solution can be derived from the associated characteristic matrix polynomial $H(\lambda)$, which is defined as

$$H(\lambda) := \sum_{i=0}^{k_0+1} H_i \lambda^i. \tag{3.20}$$

It is known (cf. Gohberg et al.[11]) that equation (3.19) allows a "spectral decomposition": the vectors x_i, $i = s - 1, s, \ldots$ can be written in terms of the zeros λ_k and corresponding left-eigenvectors l_k of (3.20):

$$x_i = \sum_{k=1}^{M} \zeta_k l_k \lambda_k^i, \quad i = s-1, s, \ldots, \tag{3.21}$$

where the constants ζ_k remain to be determined from the boundary conditions (3.2) and the normalization equation $\sum_{i=0}^{\infty} x_i e = 1$.
For this method one has to verify that indeed there exist M zeros of $H(\lambda)$ inside the unit disk, and also that the set of equations is sufficient for determining all ζ_k.
The first condition follows from the ergodicity of the Markov process; if the radius of one of the eigenvalues λ_k would be larger than 1, the expression (3.21) would not converge for $i \to \infty$. The second condition follows from a probabilistic argument: the set of equations that arises has at least one solution, being the stationary solution; would this set provide more than one solution, we would also have obtained more than one stationary vector of the Markov process, which is in contradiction with the uniqueness of the stationary vector.
So we have obtained the stationary probability vector with a third method; however we started with the assumption that the batch size distribution has finite support. When batch sizes can be arbitrarily large, for example in the case of geometric distributions, the tails of the distribution have to be cut off in order to be able to apply the method. However, similar problems arise for the matrix-analytic

method as well as for the transform method; in the first case infinite sums occur in expressions (3.5) and (3.6), in the second case one has to determine the zeros of the determinant of a matrix whose elements contain the batch size generating functions, the latter functions need not be finite polynomials since the batch size distributions not necessarily have finite support (e.g. geometric distributions).

REMARK 3.3
An important observation for the model for which batch size distributions have finite support is that from a certain level not only the rate of transitions going *out* of a state is independent of the number of customers in the system, but also the rate of transitions *into* that state is independent of the level. To illuminate this observation, consider the ordinary $M/M/s$ queue: when the number of customers is $0 \leq i < s$ then the rate out of state i is $i\mu+\lambda$, the rate into state i is $(i+1)\mu+\lambda$. When the number of customers in the system is $i > s$, then the rate out of i as well as the rate into i are independent of i. On the other hand, in the $M/M/s$ queue with batch arrivals, when batch sizes can be arbitrarily large, state i can always be entered by a batch size arrival of size i from the boundary state 0, but state $i-1$ can only be entered with a batch arrival of size $i-1$. With all other transition rates into and out of the states i and $i-1$ being equal, we see that for all i the transition rate for state i depends on i. In cases where the tail of the stationary distribution has the matrix-geometric property $\pi_n = \pi_{n-1}R$ for large n, one of the necessary conditions for this matrix geometric form is that the distribution of the upward jumps has finite support (cf. chapter 1 of Neuts[20]). As a consequence, the stationary probability vector of the queueing process in the general $BMAP/M/s$ queue, with batch sizes of infinite support, is not of matrix-geometric type. □

REMARK 3.4
A second observation is that bounding the batch sizes is not merely a technical issue; for bounded batch sizes the structure of the $BMAP/M/s$ queue is also of $GI/M/1$ type. As a consequence the stationary distribution of the number of customers has the geometric properties we discussed in section 2, this is reflected by the form of expression (3.21). We find that the analysis of the spectral expansion technique is closely related to the matrix-analytic method for Quasi Birth-and-Death models with bounded jumps as is discussed by Neuts[20]. As the transform method concentrates on the solution of a functional equation connected to the matrix G, the spectral expansion method and the matrix-analytic method are connected via the vector difference equation (3.18). Moreover, from the structure of (3.21) it follows that the matrix R in the geometric tail of x_n (as mentioned in our discussion of the matrix-analytic technique in section 2), is determined by the eigenvalues ζ_k and their respective eigenvectors l_k. □

REFERENCES

[1] Asmussen, S. (1989). Aspects of matrix Wiener-Hopf factorization in applied probability. *Math. Scientist* **14**, 101-116.

[2] Baily, D.E., Neuts, M.F. (1981). Algorithmic methods for multi-server queues with group arrivals and exponential services. *Eur. J. Oper. Res.* **8**, 184-196.

[3] Blanc, J.P.C. (1993). Performance analysis and optimization with the power-series algorithm. In: *Performance Evaluation of Computer and Communication Systems,* Eds. L. Donatiello and R.D. Nelson (North-Holland, Amsterdam), 53-80.

[4] Bonomi, F., Meyer, J., Montagna, S., Paglino, R. (1994). Minimal on/off source models for ATM traffic. In: *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*, Eds. J. Labetoulle, J.W. Roberts (Elsevier, Amsterdam), 387-400.

[5] Cinlar, E. (1975). Markov renewal theory: A survey. *Management Science* **21**, 727-752.

[6] Elwalid, A.I., Mitra, D., Stern, T.E. (1991). Statistical multiplexing of Markov modulated sources: theory and computational algorithms. In: *Teletraffic and Datatraffic*, Eds. A. Jensen, V.B. Iversen, 495-500.

[7] Ettl, M., Mitrani, I. (1994). Applying spectral expansions in evaluating the performance of multiprocessor systems. In: *Performance Evaluation of Parallel and Distributed Systems — Solution Methods*, Eds: O.J. Boxma, G.M. Koole (CWI Tract 105 & 106, Amsterdam), 45-58.

[8] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Vol. II* (Wiley, New York, 2nd ed.).

[9] Gail, H.R., Hantler, S.L., Taylor, B.A. (1992). Spectral analysis of M/G/1 type Markov chains. IBM Res. Rep., RC 17765.

[10] Gail, H.R., Hantler, S.L., Konheim, A., Taylor, B.A. (1992). The transform method for $M/G/1$ type Markov chains. IBM Res. Rep., RC 17891.

[11] Gohberg, I., Lancaster, P., Rodman, L. (1982). *Matrix Polynomials* (Academic Press, New York).

[12] Grünenfelder, R., Robert, S. (1994). Which arrival law parameters are decisive for queueing system performance. In: *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*, Eds. J. Labetoulle, J.W. Roberts (Elsevier, Amsterdam), 377-386.

[13] Lancaster, P., Tismenetsky, M. (1985). *The Theory of Matrices* (Academic Press, Orlando).

[14] Latouche, G., Ramaswami, V. (1993). A logarithmic reduction algorithm for quasi-birth-death processes. *J. Appl. Prob.* **30**, 650-674.

[15] Loynes, R.M. (1962). The stability of a queue with non-independent interarrival and service times. *Proc. Cambridge Phil. Soc.* **58**, 497-520.

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

[17] Lucantoni, D.M. (1993). The $BMAP/G/1$ queue: A tutorial. In: *Models and Techniques for Performance Evaluation of Computer and Communication Systems,* Eds. L. Donatiello, R. Nelson (Springer, Berlin), 330-358.

[18] Mitrani, I., Mitra, D. (1992). A spectral expansion method for random walks on semi-infinite strips. In: *Iterative Methods in Linear Algebra,* Eds. R. Bauwens, P. de Groen (North-Holland, Amsterdam), 141-149.

[19] Mitrani, I., Chakka, R. (1993). Application and evaluation of the spectral expansion solution method. In: *QMIPS Workshop on Formalisms, Principles and State-of-the-Art*, Eds. N. Götz, U. Herzog, M. Rettelbach (Erlangen), 253-267.

[20] Neuts, M.F. (1981). *Matrix-Geometric Solutions in Stochastic Models* (The Johns Hopkins University Press, Baltimore).

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

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

[23] Regterschot, G.J.K., de Smit, J.H.A. (1986). The queue $M/G/1$ with Markov modulated arrivals and services. *Math. of Oper. Res.* **11**, 465-483.

[24] Regterschot, G.J.K. (1987). *Wiener-Hopf Factorization Techniques in Queueing Models* (Ph.D Thesis, University of Twente, Dept. of Applied Mathematics).

[25] De Smit, J.H.A. (1983). The queue $GI/M/s$ with customers of different types or the queue $GI/H_m/s$. *Adv. Appl. Prob.* **15**, 392-419.

[26] De Smit, J.H.A. (1985). The queue $GI/H_m/s$ in continuous time. *J. Appl. Prob.* **22**, 214-222.

[27] Stewart, W.J. (editor) (1991). *Numerical Solution of Markov Chains* (Marcel Dekker, New York).

[28] Takács, L. (1962). *Introduction to the Theory of Queues* (Oxford University Press, New York).

[29] Wolff, R.W. (1989). *Stochastic Modeling and the Theory of Queues* (Prentice Hall, Englewood Cliffs, NJ).

[30] Zhao, Y. (1994). Analysis of the $GI^X/M/c$ model. *Queueing Systems* **15**, 347-364.