Computational Aspects of the Workload Distribution in the MMPP/GI/1 Queue

Alain Jean-Marie[1], Zhen Liu[1], Philippe Nain[1] and D. Towsley[2]

[1]INRIA, B.P. 93, 06902, Sophia Antipolis Cedex, France
E-mail: {Alain.Jean-Marie, Zhen.Liu, Philippe.Nain}@sophia.inria.fr

[2]Dpt. of Computer Science, Univ. of Massachusetts, Amherst, MA 01003, USA
E-mail: towsley@cs.umass.edu

March 6, 1998

Abstract

In this paper, we show how the analyses of Markov Modulated Rate Processes can be used to address the problem of computing the distribution of W, the stationary workload in the MMPP/GI/1 queue. Using the results of papers by Anick, Mitra and Sondhi (1982), Mitra (1988), and Elwalid, Mitra and Stern (1991), we present decomposition properties of the Laplace transform of W and efficient computational algorithms for computing its distribution. The techniques are also applied to compute bounds on the distribution of W developed in Liu, Nain and Towsley (1997). Numerical results illustrating the usefulness of the methods are given for the case of a superposition of independent, non-identical sources.

Keywords: MMPP model, workload distribution, computation algorithms, exact method, bound.

1 Introduction

The problem of switch design and admission control in high speed networks (in particular ATM) has spawned a large amount of research on stochastic models that are numerically tractable and, at the same time, capable of a realistic representation of highly variable traffic sources.

To this end, many authors have studied queueing systems with *Markov modulated* input processes. The standard renewal processes are judged as inadequate for representing the complexity of the behavior of sources originating in multimedia (including video and voice) applications.

When modeling packet level behavior, the network is represented by classical queueing systems in which the customers represent the data packets. A commonly used source model is the *Markov Modulated Poisson Process* (MMPP) in which customers arrive according to a Poisson process whose instantaneous rate is a function of $\{X(t), t \geq 0\}$, a finite-state, continuous-time Markov chain. Other families of modulated arrival processes have been studied, such as *Markov Arrival Processes* (MAP) and the "batch" variant (BMAP) (see *e.g.* [14]). The modeling power of these families is larger, but their analysis techniques are quite similar.

The popularity of these families of processes is motivated by the facts that

- they can represent, or approach by density, very wide classes of stationary processes;
- they are closed under superposition, that is: the superposition of several independent Markov modulated Poisson processes (or MAP, or BMAP) still belongs to the same family;
- they contain "on/off" processes, which are simple yet realistic models for several classes of traffic (voice, data, etc.).

A Markov modulated Poisson source is represented by a couple $(\mathbf{Q}, \mathbf{\Lambda})$, the first matrix being the generator of the Markov chain $\{X(t),\ t \geq 0\}$, and the second being the diagonal matrix of the arrival intensities associated with each state.

The results of many years of research on the MMPP/GI/1 queue have been collected in the "MMPP Cookbook" [9]. Regarding the distribution of the workload in the queue, there exist general formulas and algorithms that enable its numerical computation. Unfortunately, a number of technical difficulties limit their applications to systems containing a small number of sources. These difficulties relate to the solution of matrix functional equations and the inversion of Laplace transforms.

There exists a strong similarity between the MMPP/GI/1 queue and a class of *fluid data* models that have been studied in detail in the literature. In the latter systems, the source is modeled by a *Markov Modulated Rate Process* (MMRP, an acronym appearing in [17]), in which the value of the instantaneous arrival rate of information is a function of $\{X(t), t \geq 0\}$, a finite-state, continuous-time Markov chain.

The analysis of MMRP models centers around matrices of the form

$$\mathbf{A}(h) \;=\; \mathbf{Q} \;-\; h\mathbf{\Lambda}\ .$$

For this reason, the literature on MMRP includes a detailed analysis of the spectrum and eigenvectors of such matrices. In [2], the case of a superposition of identical and independent two-state sources (referred to as *binary sources* throughout this paper) is solved. In [17], this analysis is

generalized to the superposition of independent sources. This is made possible through the use of *Kronecker algebra* [4, 10] and related spectral theory.

The algebraic similarity exhibited by queueing systems fed either by MMPP sources or by MMRP sources has been observed and exploited in several papers authored by Elwalid, Mitra and Stern. In [5, 6, 7], the authors consider that the source is a superposition of identical independent MMPP sources and exponential service times. In [5], the authors address the case of the superposition of heterogeneous sources and Markov modulated exponential servers. In all of these studies the distribution of interest is that of the *number of customers* in the system. In complement to the Kronecker product representations, the authors propose an aggregation procedure applicable to homogeneous superpositions of arbitrary MMPP sources.

The principal objective of this paper is to take advantage of the analysis techniques developed for MMRP queues to improve the computation algorithms for the *workload* distribution in the MMPP single server queue with *general* service times. Our analysis starts from the results of [9], which are briefly summarized in Section 2. It then proceeds along the lines of [17] of Stern and Elwalid. The aggregation techniques of [5, 6] are not exploited here: this is a topic left for future research.

The paper is organized as follows. In Section 2 we describe the model and collect and derive some basic results related to the structure of MMPP/GI/1 queues. In Section 3 we consider the exact computation of the workload distribution for *reversible sources*. Both the single source and the superposition of such sources are analyzed. In Section 4 we address the exact computation of the workload distribution when the input process is the superposition of homogeneous/heterogeneous two-state (*binary*) MMPP sources. In Section 5, we revisit and extend computational aspects of the exponential bounds for the waiting distribution reported in [13]. Throughout these three sections (3, 4 and 5), we will discuss both the computation algorithms and the analysis of their computational complexity. In Section 6 we present numerical results obtained with these algorithms. Finally, we conclude in Section 7 with remarks on future research directions.

2 Preliminaries and Summary of the Results

2.1 Basic Model and Known Results

We consider a single-server MMPP/GI/1 queueing system equipped with an infinite buffer. Customers arrive according to an MMPP process and require independent and identically distributed (i.i.d.) service times. Let $H(x)$ denote the probability distribution of the service times and let $H^*(s)$ be its Laplace transform. Let $\overline{m}$ be the average service time and $\overline{m}^{(2)}$ its second moment. An MMPP is a continuous-time irreducible Markov chain $\{X(t), t \geq 0\}$ with finite state space $\{0, 1, 2, \ldots, N\}$. When the Markov chain is in state i, $0 \leq i \leq N$, customers arrive to the queue according to a Poisson process with parameter λ_i. The arrival process is represented by a couple $(\mathbf{Q}, \mathbf{\Lambda})$, the first matrix being the infinitesimal generator of the Markov chain $\{X(t),\ t \geq 0\}$, the second being the diagonal matrix of the arrival intensities associated with each state (usually referred to as the rate matrix). Let $\boldsymbol{\pi}$ denote the row vector of the stationary distribution of the matrix $\mathbf{Q}$. Also, let $\mathbf{1}$ denote the column vector constituted of ones. The average arrival rate of the process is:

$$\overline{\lambda} \;=\; \boldsymbol{\pi}\ \mathbf{\Lambda}\ \mathbf{1}\ .$$

The load factor of the system is then $\rho = \overline{\lambda}\overline{m}$.

A special case, and also one of the most studied cases, is the MMPP with a two-state Markov chain ($N = 1$). Such a source will be referred to as a *binary source*.

The study of the MMPP/GI/1 queue has recently been revived, due to its important applications in the analysis of high speed communication networks. The paper of Fischer and Meier-Hellstern [9] surveys the current knowledge on this system.

It appears that matrices of the form $\mathbf{Q} + a\mathbf{\Lambda} + b\mathbf{I}$, $\mathbf{I}$ being the identity matrix, play a pivotal role in the theory of the MMPP/GI/1 queue. For instance, if $F_{i,j}(x) = \mathbb{P}(X_k = j, \tau_k \leq x \,|\, X_{k-1} = i)$ is the probability transition of the state of the Markov chain $\{X(t), t \geq 0\}$ at arrival instants, joint with inter-arrival times, then [9, eq. (5)]:

$$\mathbf{F}(x) \;:=\; (F_{i,j}(x))_{i,j} \;=\; \left(\mathbf{I} - e^{(\mathbf{Q}-\mathbf{\Lambda})x}\right)(\mathbf{\Lambda} - \mathbf{Q})^{-1}\mathbf{\Lambda}\ . \tag{1}$$

Here matrix $\mathbf{P} := (\mathbf{\Lambda} - \mathbf{Q})^{-1}\mathbf{\Lambda}$ is the transition matrix of the Markov chain embedded at arrival epochs. We will denote by $\mathbf{p}$ its invariant measure.

Likewise, the Laplace transform of the joint distribution of the first n inter-arrival times is given by (see [9, eq. (14)]):

$$\mathbf{F}^*(s_1, \ldots, s_n) \;=\; \prod_{k=1}^{n} \left((s_k\mathbf{I} - \mathbf{Q} + \mathbf{\Lambda})^{-1}\mathbf{\Lambda}\right)\ . \tag{2}$$

In particular,

$$\mathbf{F}^*(s) := (\mathbf{F}^*_{i,j}(s))_{i,j} = \int_0^\infty e^{-sx}\, dF_{i,j}(x) \;=\; (s\mathbf{I} - \mathbf{Q} + \mathbf{\Lambda})^{-1}\mathbf{\Lambda}\ . \tag{3}$$

Finally, it is known that the Laplace transform of the joint distribution of the state of the chain $\{X(t),\, t \geq 0\}$ and the workload in the queue, $\mathbf{W}^*(s)$, satisfies the equation (see [9, eq. (48)] and [16]):

$$\mathbf{W}^*(s) \;=\; s(1-\rho)\ \mathbf{g}\ [s\mathbf{I} + \mathbf{Q} - (1 - H^*(s))\mathbf{\Lambda}]^{-1}\ , \tag{4}$$

for $\Re(s) > 0$, with $\mathbf{W}^*(0) = \boldsymbol{\pi}$. Here, $\mathbf{g}$ is a probability vector that is the solution of $\mathbf{g} = \mathbf{g}\mathbf{G}$, where $\mathbf{G}$ is the matrix of the transition probabilities of the chain $\{X(t),\, t \geq 0\}$ between the beginning and the end of busy periods. $\mathbf{G}$ is in turn the solution of the matrix functional equation:

$$\mathbf{G} \;=\; \int_0^\infty \exp\left((\mathbf{Q} - \mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{G})x\right) \mathrm{d}H(x)\ . \tag{5}$$

According to (1), (2) or (4), it is clear that computing the inverse or the exponential of matrices of the form $\mathbf{Q} + a\mathbf{\Lambda} + b\mathbf{I}$ is a central issue. Also, when computing the workload distribution with (4), the vector $\mathbf{g}$ has to be determined. This is normally carried out by first computing the matrix $\mathbf{G}$ (using an iterative procedure based on (5)), then by solving for $\mathbf{g}$.

The exact computation of the distribution further requires the inversion of the Laplace transform. This can be performed using the EULER algorithm devised by Abate and Whitt [1]. However, this

approach requires numerous computations of $\mathbf{W}^*(s)$ and, therefore, numerous matrix inversions. The size of the sources that can be handled by this method is quite limited, unless structural information can be used to improve the computation of $\mathbf{W}^*(s)$ or otherwise simplify the inversion. For this reason, the use of easily computable bounds on the distribution of W is of interest. Liu, Nain and Towsley have proposed in [13] a methodology for computing such bounds which involve the determination of the principal eigenvalue of a matrix $H^*(s)(s\mathbf{I}+\mathbf{\Lambda}-\mathbf{Q})^{-1}\mathbf{\Lambda}$, and the associated left-eigenvector. It turns out that the approach developed below applies to the computation of such bounds as well.

2.2 Markov modulated Sources and Diagonalization

In this section, we collect the principal algebraic and analytical properties that are commonly used in the spectral analysis of Markov Modulated sources.

The first result relates the diagonalization of a *Kronecker sum* of matrices, in terms of the diagonalization of each components. The importance of this property comes from the fact that Kronecker sum $\oplus$ and Kronecker product $\otimes$ [4, 10] are related to the *superposition* of independent Markov chains.

The second details the diagonalization of a certain class of matrices. These matrices are related to the *aggregation* of the (compound) Markov chain resulting from the superposition of identical two-state chains. Most of the results presented here appear in one form or in another in various papers of the recent literature on fluid sources (MMRP) [2, 15, 17]. Note that in these references, the focus is mostly on one-sided spectral decomposition. This is due to the fact that all problems at hand are "vectorial" in nature, and can be solved by considering one-sided spectral problems (such as (74) below). Note also that Anick, Mitra and Sondhi exhibit in [2] a system of right eigenvectors, without however completing the diagonalization of the matrix.

A complete diagonalization proves useful in cases of "matrix" nature: for instance for transition probabilities, and more generally when the transient behavior is involved. For this reason and for the sake of completeness, we state these results in full detail, in the form of a complete diagonalization result (Lemma 2.3). These results will be proved in the Appendix A, together with some remarks on the analyticity of the decomposition, and on singular cases.

Superpositions The following results provide a way of computing the spectral elements (eigenvalues, left and right eigenvectors) of a matrix based on its structure.

Lemma 2.1 *Assume that*

$$\mathbf{A} \;=\; \mathbf{A}^{(1)} \oplus \;\ldots\; \oplus \mathbf{A}^{(K)} \;,$$

and that for all k, $\mathbf{A}^{(K)}$ is diagonalizable with

$$\mathbf{A}^{(k)} \;=\; \mathbf{R}^{(k)}\mathbf{D}^{(k)}\mathbf{S}^{(k)},$$

where $\mathbf{R}^{(k)}\mathbf{S}^{(k)} = \mathbf{I}^{(k)}$ and $\mathbf{D}^{(K)} = diag(\omega_i^{(k)})$. Then:

$$\mathbf{A} \;=\; \left(\bigotimes_{k=1}^{K}\mathbf{R}^{(k)}\right)\;\left(\bigoplus_{k=1}^{K}\mathbf{D}^{(k)}\right)\;\left(\bigotimes_{k=1}^{K}\mathbf{S}^{(k)}\right)\;.$$

Corollary 2.2 *With the assumptions of Lemma 2.1, we have:*

$$e^{\mathbf{A}x} \;=\; \left(\bigotimes_{k=1}^{K}\mathbf{R}^{(k)}\right)\;\left(e^{\bigoplus_{k=1}^{K}\mathbf{D}^{(k)}x}\right)\;\left(\bigotimes_{k=1}^{K}\mathbf{S}^{(k)}\right)\;.$$

Also, if no diagonal element of $\oplus_{k=1}^{K}\mathbf{D}^{(k)}$ *is zero, that is, if*

$$\sum_{k=1}^{K}\omega_{i_k}^{(k)} \neq 0, \quad \forall(i_1,\ldots,i_K)\in\{0..N_1\}\times\ldots\times\{0..N_K\}, \tag{6}$$

then $\mathbf{A}$ *is invertible, and:*

$$\mathbf{A}^{-1} \;=\; \left(\bigotimes_{k=1}^{K}\mathbf{R}^{(k)}\right)\;\left(\bigoplus_{k=1}^{K}\mathbf{D}^{(k)}\right)^{-1}\;\left(\bigotimes_{k=1}^{K}\mathbf{S}^{(k)}\right)\;.$$

Proof (of Lemma 2.1.) The proof is based on the known property that $(v_1\otimes\ldots\otimes v_K)(A_1\oplus\ldots\oplus A_K)=(v_1A_1)\oplus\ldots\oplus(v_KA_K)$. This allows us to construct the eigenvalues and (left and right) eigenvectors of the Kronecker sum in function of those of the components: eigenvalues are sums of the eigenvalues of the components, and eigenvectors are Kronecker products of the corresponding eigenvectors. ■

Homogeneous superpositions of binary sources A very useful particular case is where matrix $\mathbf{A}$ corresponds to sources that are the superposition of several independent and homogeneous binary sources. The generator and rate matrix have a special structure.

Assume that $\mathbf{A} = \mathbf{M}(N;\lambda,\mu)+a\mathbf{I}(N)+b\mathbf{J}(N)$, where a and b are complex numbers, $\mathbf{I}(N)$ is the $(N+1)\times(N+1)$ identity matrix,

$$\mathbf{J}(N)=\mathrm{diag}(0,1,\ldots,N)\;, \tag{7}$$

and

$$\mathbf{M}(N;\lambda,\mu) \;=\; \begin{pmatrix} -N\lambda & N\lambda & & & \\ \mu & -(\mu+(N-1)\lambda) & (N-1)\lambda & & \\ & \ddots & \ddots & \ddots & \\ & & (N-1)\mu & -((N-1)\mu+\lambda) & \lambda \\ & & & N\mu & -N\mu \end{pmatrix}\;. \tag{8}$$

Define, for $k\in\{0,\ldots,N\}$:

$$\omega_k \;=\; a \;+\; (\frac{N}{2}-k)\sqrt{(b+\lambda-\mu)^2+4\lambda\mu}-\frac{N}{2}\,(-b+\lambda+\mu)\;. \tag{9}$$

Define further:

$$\sigma_{1,2} \;=\; \frac{(b+\lambda-\mu) \pm \sqrt{(b+\lambda-\mu)^2+4\lambda\mu}}{2\lambda} \;. \tag{10}$$

These are the roots of the equation

$$\lambda\sigma^2 \;-\; (b+\lambda-\mu)\sigma \;-\; \mu \;=\; 0. \tag{11}$$

Let $\boldsymbol{\phi}_k = (\phi_k(0), \ldots, \phi_k(N))$ be the row vector with coordinates given by the coefficient of x^i in the polynomial function $(x-\sigma_1)^k(x-\sigma_2)^{N-k}$:

$$\phi_k(i) \;=\; [x^i](x-\sigma_1)^k\,(x-\sigma_2)^{N-k}. \tag{12}$$

Finally, define the $(N+1)\times(N+1)$ matrices out of the row vectors $\boldsymbol{\phi}_i$ and the *column* vectors $\boldsymbol{\psi}_j$:

$$\boldsymbol{\Phi} \;=\; ((\phi_i(j)))_{i,j} \tag{13}$$

$$\boldsymbol{\Psi} \;=\; ((\psi_j(i)))_{i,j} \;:=\; ((\phi_i(j)\ \sigma_2^{i+j-N}))_{i,j}. \tag{14}$$

We then have the following result:

Lemma 2.3 *Assume that the matrix* $\mathbf{A}$ *has the form:*

$$\mathbf{A} \;=\; \mathbf{M}(N;\lambda,\mu) + a\mathbf{I}(N) + b\mathbf{J}(N) \;, \tag{15}$$

where a *and* b *are arbitrary complex numbers. Then,* $\mathbf{A}$ *is diagonalizable if and only if*

$$b \;\neq\; -(\sqrt{\lambda} \pm i\sqrt{\mu})^2 \;. \tag{16}$$

In that case,

$$\mathbf{A} \;=\; (\sigma_2-\sigma_1)^{-N}\ \boldsymbol{\Psi\Omega\Phi} \tag{17}$$

where $\boldsymbol{\Omega} = diag(\omega_0, \ldots, \omega_N)$.

The proof is provided in Appendix A.
As a consequence,

Corollary 2.4 *A right eigenvector of* $\mathbf{A}$ *corresponding to the eigenvalue* ω_k *is the vector* $\boldsymbol{\psi}_k$ *such that* $\psi_k(i) = \psi_{ik}$.

Also, as a consequence of (9) and (17):

Corollary 2.5 *The characteristic polynomial of* $\mathbf{A}$ *is given by:*

$$\det(\mathbf{A} - z\mathbf{I}) \;=\; \prod_{k=0}^{N} \left(a - z - \frac{N}{2}\,(-b+\lambda+\mu) + (\frac{N}{2} - k)\sqrt{(b+\lambda-\mu)^2 + 4\lambda\mu} \right) \;. \qquad (18)$$

When $\omega_k \neq 0$ *for all* k, $\mathbf{A}$ *is invertible, and*

$$\mathbf{A}^{-1} \;=\; (\sigma_2 - \sigma_1)^{-N}\; \mathbf{\Psi \Omega^{-1} \Phi} \;. \qquad (19)$$

In Appendix A, some technical remarks are made regarding analyticity and degenerate cases of the result of Lemma 2.3.

Remark 2.1 (Jordan forms) When condition (16) fails to hold, the matrix $\mathbf{A}$ is not diagonalizable. Its standard Jordan form is then of the form

$$\begin{pmatrix} \omega & 1 & 0 & & \\ 0 & \omega & 1 & \ddots & \\ & \ddots & \ddots & \ddots & \\ & & & 0 & \omega \end{pmatrix} \;,$$

with $\omega = a + (N/2)(b - \lambda - \mu)$. The construction of the change of basis will be omitted here.

Remark 2.2 The following identity holds:

$$\boldsymbol{\phi}_k \cdot \mathbf{1} \;=\; \sum_{i=0}^{N} [x^i](x-\sigma_1)^k (x-\sigma_2)^{N-k} \;=\; (1-\sigma_1)^k (1-\sigma_2)^{N-k} \;. \qquad (20)$$

It is useful for computing the (unconditional) workload distribution (see Section 4.3). It can be used in conjunction with

$$(1-\sigma_1)(1-\sigma_2) = -b/\lambda,$$

which is easily derived from (11).

Remark 2.3 Being the generator of the aggregated chain obtained by superposing independent and identical two-state chains, the matrix $\mathbf{M}(N;\lambda,\mu)$ coupled with the results of Lemma 2.1 and the spectral analysis of the two-state problem, can be used to derive again the results above. See [8] for details.

2.3 Summary of the Main Results

In this paper, we use the results known for MMRP systems to provide a new perspective on the computation of the distribution of W. The principle of the analysis goes in parallel with that of the paper by Stern and Elwalid [17]. In particular, our results

- provide an alternative and more efficient way to compute $\mathbf{g}$;

- allow one to compute formally $[s\mathbf{I} + \mathbf{Q} - (1 - H^*(s))\mathbf{\Lambda}]^{-1}$, and provide an approach to the formal inversion of the Laplace transform $\mathbf{W}^*(s)$;
- allow one to compute more efficiently the distribution of W, and bounds on this distribution, in the case of superposed heterogeneous sources.

The results obtained depend naturally on the degree of generality of the source considered. We shall consider two classes of sources:

- sources whose underlying Markov chain is reversible, and with an arbitrary rate matrix $\mathbf{\Lambda}$. We shall refer to these sources as *reversible sources*;
- sources that are the superposition of several independent, identical two-state (binary) sources. These sources are reversible, but their generator and rate matrices have a special structure. These sources are referred to as *superposition of homogeneous binary sources.*

We shall also consider the superposition of several independent but heterogeneous reversible sources.

Sources with two states arise naturally as simple models of voice or Web traffic. Other types of traffic, such as images and video, are not always adequately modeled by such simple processes. Reversible sources provide a class wide enough to approximate arbitrarily complex traffic. Note that birth-and-death processes are reversible.

We begin with general results on sources with a reversible generator (Section 3). In Section 4, we specialize the results to the case of superposition of homogeneous binary sources. In Section 5, we use the same analysis to compute the bounds.

3 The Workload Distribution for Reversible Sources

3.1 Single MMPP Source Case

Assume that $\mathbf{Q}$ is the generator of a reversible *ergodic* Markov chain with $N+1$ states. It is known that $\mathbf{Q}$ can be *symmetrized.* More precisely, if $\mathbf{E} = \text{diag}(\pi_k^{1/2})$, then $\mathbf{E}^{-1}\mathbf{Q}\mathbf{E}$ is a symmetric, negative and semidefinite matrix ([12], see also [15, 17]). It is therefore diagonalizable. Its eigenvalues, which are the same as those of $\mathbf{Q}$, are real and negative. Consequently, the matrix

$$\mathbf{A}(s) \;=\; \mathbf{Q} \;+\; s\mathbf{I} \;-\; (1 - H^*(s))\mathbf{\Lambda} \tag{21}$$

is also diagonalizable when s is real, and its eigenvalues are then real. Let $\omega_k(s)$, $0 \le k \le N$, denote these eigenvalues, sorted in decreasing order, and let $\mathbf{\Omega}(s) = \text{diag}(\omega_k(s),\, k = 0, 1, \ldots, N)$. For any real s, there exist invertible matrices $\mathbf{R}(s)$ and $\mathbf{S}(s)$, such that $\mathbf{R}(s)\mathbf{S}(s) = \mathbf{I}$ and

$$\mathbf{A}(s) \;=\; \mathbf{R}(s)\; \mathbf{\Omega}(s)\; \mathbf{S}(s). \tag{22}$$

These matrices are respectively composed of *right* and *left* eigenvectors associated with the ω_k's, say $\mathbf{R}_k(s)$ and $\mathbf{S}_k(s)$. Introducing this decomposition in (4) yields the *spectral expansion*

$$\mathbf{W}^*(s) \;=\; s(1-\rho)\; \mathbf{g} \sum_{k=0}^{N} \frac{1}{\omega_k(s)} \mathbf{R}_k(s)\; \mathbf{S}_k(s). \tag{23}$$

We shall now perform a singularity analysis of (23). As the Laplace transform of a positive random variable, $\mathbf{W}^*(s)$ is analytic in the right-hand half-plane, the right-hand side of (23) must have the same property.

It is well known that the $\omega_k(s)$, being eigenvalues of a parametric and continuous matrix $\mathbf{A}(s)$, are continuous functions of the parameter s, as are the vectors $\mathbf{R}_k(s)$ and $\mathbf{S}_k(s)$. For $s = 0$, these are the eigenvalues of $\mathbf{Q}$, so that $\omega_0(0) = 0$ and $\omega_k(0) < 0, k \geq 1$. When s is large enough, the matrix $\mathbf{A}(s)$ is strictly diagonally dominant, so that all $\omega_k(s)$'s are strictly positive [11, p. 349]. Therefore, for each $k > 0$, there is a strictly positive real number s_k such that $\omega_k(s_k) = 0$. Necessarily, we have $\mathbf{g}\mathbf{R}_k(s_k)\mathbf{S}_k(s_k) = \mathbf{0}$. The vector $\mathbf{S}_k(s_k)$ is not null, because $\mathbf{R}(s_k)\mathbf{S}(s_k) = \mathbf{I}$. Therefore,

$$\forall k > 1, \qquad \mathbf{g}\ \mathbf{R}_k(s_k)\ =\ 0\ .$$

The reasoning above does not apply to the eigenvalue $\omega_0(s)$. Fortunately, the *a priori* condition that $\mathbf{g}$ is a probability vector, that is $\mathbf{g1} = 1$, may be adjoined to the N other conditions. Finally, the $N+1$ conditions may be put in matrix form by defining the (square) matrix $\mathbf{\Xi} = (\mathbf{1}, \mathbf{R}_1(s_1), \ldots, \mathbf{R}_N(s_N))$. The unknown vector $\mathbf{g}$ is then solution of

$$\mathbf{g}\ \mathbf{\Xi}\ =\ (1,\ 0,\ \ldots\ , 0\ )\ . \tag{24}$$

Remark 3.1 The condition $\mathbf{g1} = 1$ is also a direct consequence of the requirement that $\mathbf{W}^*(0) = \boldsymbol{\pi}$. Indeed, letting $s \to 0$ in (23), we have:

$$\boldsymbol{\pi}\ =\ (1-\rho)\ \mathbf{g}\ \mathbf{R}_0(0)\ \mathbf{S}_0(0)\ \lim_{s\to 0}\frac{s}{\omega_0(s)}\ =\ (1-\rho)\ \mathbf{g}\ \mathbf{1}\ \boldsymbol{\pi}\ \lim_{s\to 0}\frac{s}{\omega_0(s)}.$$

Using the fact that $\mathbf{\Omega}(s) = \mathbf{S}(s)\mathbf{A}(s)\mathbf{R}(s)$ and that $\mathbf{R}_0(0) = \mathbf{1}, \mathbf{S}_0(0) = \boldsymbol{\pi}$, one obtains

$$\lim_{s\to 0}\frac{\omega_0(s)}{s}\ =\ \omega_0'(0)\ =\ \boldsymbol{\pi}(\mathbf{I} - \overline{m}\mathbf{\Lambda})\mathbf{1}\ =\ 1-\rho\ .$$

The condition $\mathbf{g1} = 1$ follows.

Remark 3.2 There is no known guarantee that the system (24) has a unique solution. This also happens to be the case for MMRP processes, although it has never been reported that it causes any practical problems. Here, however, if the matrix $\mathbf{\Xi}$ should happen to be singular, then the standard computation of $\mathbf{g}$ could still be performed (see [9]).

3.2 Heterogeneous Superpositions

Consider now the case where the input process of the queue is a superposition of sources such as described in Section 3. There are K sources, characterized by generators $\mathbf{Q}^{(k)}$ and rate matrices $\mathbf{\Lambda}^{(k)}$ of dimensions $N_k + 1$. Then, as in the case of Markov Modulated Rate Processes [17, 8], the generator and the rate matrix of the superposed process admits the representation:

$$\begin{aligned}\mathbf{Q}\ &=\ \mathbf{Q}^{(1)}\ \oplus\ \ldots\ \oplus\ \mathbf{Q}^{(K)}\\ \mathbf{\Lambda}\ &=\ \mathbf{\Lambda}^{(1)}\ \oplus\ \ldots\ \oplus\ \mathbf{\Lambda}^{(K)}\ .\end{aligned} \tag{25}$$

Consequently, the matrix $\mathbf{A}(s) = \mathbf{Q} + s\mathbf{I} - (1 - H^*(s))\mathbf{\Lambda}$ admits a similar representation:

$$\mathbf{A}(s) \;=\; s\mathbf{I} \;+\; \left(\mathbf{B}^{(1)} \;\oplus\; \ldots \;\oplus\; \mathbf{B}^{(k)}\right) \;,$$

where

$$\mathbf{B}^{(k)} \;=\; \mathbf{Q}^{(k)} - (1 - H^*(s))\mathbf{\Lambda}^{(k)}.$$

Denote by $\omega_0^{(k)}, \ldots, \omega_{N_k}^{(k)}$ the eigenvalues of the matrix $\mathbf{B}^{(k)}$ and $\boldsymbol{\phi}_0^{(k)}, \ldots, \boldsymbol{\phi}_{N_k}^{(k)}$ the corresponding left eigenvectors. According to Lemma 2.1, the eigenvalues of $\mathbf{A}$ are of the form:

$$\omega \;=\; \omega(i_1, \ldots, i_K) := s \;+\; \sum_{k=1}^{K} \omega_{i_k}^{(k)}, \qquad i_k \in \{0, \ldots, N_k\},\ 1 \le k \le K \;, \tag{26}$$

with the associated eigenvector:

$$\boldsymbol{\phi} \;=\; \boldsymbol{\phi}(i_1, \ldots, i_K) := \boldsymbol{\phi}_{i_1}^{(1)} \;\otimes\; \ldots \;\otimes\; \boldsymbol{\phi}_{i_K}^{(K)}. \tag{27}$$

Likewise, the right eigenvectors of $\mathbf{A}$ are obtained as Kronecker products of the right eigenvectors $\boldsymbol{\psi}^{(k)}$ of the matrices $\mathbf{B}^{(k)}$.

Assume now that each source k is the superposition of N_k homogeneous binary sources. In this case,

$$\mathbf{W}^*(s) \;=\; \prod_{k=1}^{K} (\sigma_2^{(k)} - \sigma_1^{(k)})^{-N_k}\; s(1-\rho)\; \mathbf{g} \sum_{(i_1,\ldots,i_K)} \frac{\boldsymbol{\psi}(i_1, \ldots, i_K) \cdot \boldsymbol{\phi}(i_1, \ldots, i_K)}{\omega(i_1, \ldots, i_K)} \;. \tag{28}$$

3.3 Computational Algorithms

In this paragraph, we recapitulate the algorithmic steps needed for computing the distribution of W. The algorithmic complexity is evaluated as a function of N, with $N + 1$ being the size of the matrices involved.

Single reversible sources. The general algorithm is the following:

1. Find, for all $k \ge 1$, the positive root s_k of the equation $s + \omega_k(s) = 0$. Construct the matrix $\mathbf{\Xi}$. This can be done using an iterative procedure of the Newton-type, taking advantage of the fact that an expression for $d\omega_k(s)/ds$ is known. The reader is referred to [8, 17] for details. Each step n in the approximation requires the spectral analysis of a matrix $\mathbf{A}\left(s_k^{(n)}\right)$;

2. Solve for $\mathbf{g}$ by inverting (24);

3. Compute the inverse of $\mathbf{W}^*(s)$ using for instance the EULER algorithm of Abate and Whitt [1]. This requires the computation of $\mathbf{W}^*(s)$ for a certain number M of values of s.

Complexity of the algorithm:

1. It is reported that due to the fast convergence of Newton's algorithms, the number of steps to perform seldom exceeds 4 or 5. The construction of the matrix Ξ of (24) takes therefore $\mathcal{O}(N^4)$ operations;

2. This step takes $\mathcal{O}(N^3)$ operations;

3. Each computation with (4) involves the inversion of the matrix $\mathbf{A}(s)$ (or the solution of a system $\mathbf{A}(s)\mathbf{x} = \mathbf{g}$). The complexity of this step is therefore $\mathcal{O}(MN^3)$.

The main part of the computation time may be spent on step 1 or on step 3, depending on the value of M that is actually needed. It is likely that M will be larger than N most of the time in practice.

This algorithm appears to be an improvement on the method advocated in [9] (at least measured in computation time) in that it avoids the computation of $\mathbf{G}$ by iterative means.

Superposition of reversible sources. The use of the structure of the matrices allows us to improve the general algorithm. If the size of the component generators are $N_1 + 1, \ldots, N_K + 1$, then, according to Section 3.2, the size of the superposed generator is $N+1 = \prod_k (N_k+1)$. Note that as far as the time complexity analysis is concerned, $\mathcal{O}(N_k + 1) = \mathcal{O}(N_k)$ and $\mathcal{O}(N) = \mathcal{O}(\prod_k N_k)$, so that N_k will be used instead of $N_k + 1$.

1. When finding the quantities $s_j, 1 \leq j \leq N$, by an iterative procedure, one takes advantage of property (26). At each step n in the Newton algorithm, the eigenvalues $\omega^{(k)}(s_j^n)$ of the matrices $\mathbf{A}^{(k)}(s_j^n)$ are computed and added.

   This takes $\mathcal{O}(\sum_k N_k^3)$ operations, resulting in a total complexity of $\mathcal{O}(N \sum_k N_k^3)$;

2. Solving for $\mathbf{g}$ still requires $\mathcal{O}(N^3)$ operations;

3. For computing a particular value of $\mathbf{W}^*(s)$, one may use Corollary 2.2 to construct the inverse of $\mathbf{A}(s)$.

   The construction of the matrices $\mathbf{R}^{(k)}$ requires $\mathcal{O}(\sum_k N_k^3)$, Once they are computed, the matrix $\otimes_k \mathbf{R}^{(k)}$ is constructed in $\mathcal{O}(KN^2)$ operations.

   The cost of multiplication $\mathbf{g}\mathbf{A}^{-1}$ is of $\mathcal{O}(N^2)$. Thus, the total computation cost of this step is $\mathcal{O}(M \sum_k N_k^3) + \mathcal{O}(MKN^2)$.

The advantage of structuring the computation appears both in steps 1 and 3. As observed in [17], in step 1 the value of $N \sum_k N_k^3$ will be much smaller than the original value of $N^4 = \prod_k N_k^4$. Likewise, the principal part of the computation in step 3 will be to fill the (large) matrix $\otimes_k \mathbf{R}^{(k)}$, an improvement on the initial complexity of $\mathcal{O}(N^3)$. Note that the Kronecker structure may be used to save storage capacity when performing the product $\mathbf{g}\mathbf{A}^{-1}$. However, step 2 still requires the inversion of a full $(N + 1) \times (N + 1)$ matrix.

When some of the sources are superpositions of homogeneous binary sources, additional savings can be made, as discussed in Section 4.

Compared with the general case, the complexity of the computation of $\mathbf{g}$ has decreased from $\mathcal{O}(N^4)$ to $\mathcal{O}(N^3)$, and the computational bottleneck has slipped from the computation of the quantities s_k to the solution of (24). The ongoing research aims at eliminating this bottleneck by a proper structuring of the problem. As mentioned above, another direction for research is the use of the aggregation technique in [5].

4 Superpositions of Binary Sources

4.1 The Workload Distribution for Superpositions of Binary Sources

4.1.1 Superpositions of homogeneous binary sources

Consider now the case where the source is the superposition of homogeneous and independent binary sources, each described by the matrices:

$$\mathbf{Q}^{(k)} = \begin{pmatrix} -q_0 & q_0 \\ q_1 & -q_1 \end{pmatrix} \qquad \text{and} \qquad \mathbf{\Lambda}^{(k)} = \begin{pmatrix} \lambda_0 & 0 \\ 0 & \lambda_1 \end{pmatrix} .$$

Due to the homogeneity of the individual sources, the superposition may be described by the number of sources in state 0 which behave as a Markov chain. The superposition can therefore be seen as a MMPP with $N+1$ states, a generator $\mathbf{Q} = \mathbf{M}(N; q_0, q_1)$ and a rate matrix $\mathbf{\Lambda} = N\lambda_0\mathbf{I}+(\lambda_1-\lambda_0)\mathbf{J}(N)$, where $\mathbf{M}$ and $\mathbf{J}$ are defined in (7) and (8). This property is well known in the context of MMRPs [2, 15], where the matrices $\mathbf{Q}$ and $\mathbf{\Lambda}$ have exactly the same form.

The stationary distribution of the Markov chain $\{X(t), t \geq 0\}$ is

$$\boldsymbol{\pi} = \frac{1}{(q_0+q_1)^N} \left(q_1^N, \ldots, \binom{N}{i} q_0^i q_1^{N-i}, \ldots, q_0^N \right) . \tag{29}$$

The overall arrival rate of the process is

$$\overline{\lambda} = N \, \frac{q_1\lambda_0 + q_0\lambda_1}{q_0+q_1} .$$

We can therefore apply Lemma 2.3 with $\lambda = q_0$, $\mu = q_1$, $a = s - (1 - H^*(s))N\lambda_0$, $b = (1 - H^*(s))(\lambda_0 - \lambda_1)$. Accordingly,

$$\begin{aligned} s\mathbf{I} + \mathbf{Q} - (1 - H^*(s))\mathbf{\Lambda} &= (\sigma_2 - \sigma_1)^{-N} \, \mathbf{\Psi\Omega\Phi} \\ \Longrightarrow \quad [s\mathbf{I} + \mathbf{Q} - (1 - H^*(s))\mathbf{\Lambda}]^{-1} &= (\sigma_2 - \sigma_1)^{-N} \mathbf{\Psi\Omega}^{-1}\mathbf{\Phi} . \end{aligned}$$

Here, the values of $\sigma_{1,2}$ and ω_k are given by

$$\sigma_{1,2} = \frac{(1 - H^*(s))(\lambda_0 - \lambda_1) + q_0 - q_1 \pm \sqrt{((1 - H^*(s))(\lambda_0 - \lambda_1) + q_0 - q_1)^2 + 4q_0q_1}}{2q_0} \tag{30}$$

$$\begin{aligned} \omega_k = \; & s + (\frac{N}{2} - k)\sqrt{((1 - H^*(s))(\lambda_0 - \lambda_1) + q_1 - q_0)^2 + 4q_0q_1} \\ & - \frac{N}{2}\left((1 - H^*(s))(\lambda_0 + \lambda_1) + q_0 + q_1\right) . \end{aligned} \tag{31}$$

The matrices $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are computed according to (12), (13) and (14) in Appendix 2.2. The numbers σ_i, ω_k and the matrices $\mathbf{\Psi}$ and $\mathbf{\Phi}$ are functions of s, but we shall often omit the explicit "(s)" in the formulas.

The expansion (23) takes now the form

$$\mathbf{W}^*(s) \quad = \quad (\sigma_2 - \sigma_1)^{-N} \; s(1-\rho) \; \mathbf{g} \; \sum_{k=0}^{N} \frac{1}{\omega_k(s)} \; \boldsymbol{\psi}_k \cdot \boldsymbol{\phi}_k \; . \tag{32}$$

The singularity analysis of Section 3 applies in this case, $\mathbf{Q}$ being the matrix of a birth and death process (hence reversible), which is ergodic as long as q_0 and q_1 are not zero. Here, we can take advantage of the precise knowledge of the eigenvalues and of the eigenvectors to provide greater justification. Indeed, observe that $\sigma_1, \sigma_2, \omega_k, \boldsymbol{\phi}_k$ and $\boldsymbol{\psi}_k$ are all functions of s. As discussed in Remark A.3, each of these functions is analytic only in domains where condition (16) holds. However, it turns out that their product is regular. This may be seen by carefully combining the terms corresponding to ω_k and ω_{N-k} in (32): square roots either cancel or factor out, according to whether N is odd or even. This is similar to a cancelation of imaginary parts.

The only non-removable singularities of (32) may lie therefore at points s where, for some k, $s + \omega_k(s) = 0$. These are the numbers s_k defined above. In the present case, these are easily computed as the roots of

$$\begin{aligned} 0 \quad = \quad & s + (\frac{N}{2} - k)\sqrt{(1 - H^*(s))(\lambda_0 - \lambda_1) + q_0 - q_1)^2 + 4q_0 q_1} \\ & - \frac{N}{2} \left((1 - H^*(s))(\lambda_0 + \lambda_1) + q_0 + q_1\right) \; . \end{aligned} \tag{33}$$

4.1.2 Superposition of heterogeneous binary sources

The analysis of the superposition of heterogeneous binary sources can be performed by first partitioning the sources into classes of homogeneous binary sources and analyzing as above the superposition of homogeneous sources within each class. Then, one can use the results of Section 3.2 to analyze the resulting heterogeneous superposition of sources.

4.2 Computational Algorithms

Superposition of homogeneous superpositions. The general algorithm is simplified, taking advantage of the numerous available closed form solutions.

1. Compute the $\{s_k\}$ and construct $\mathbf{\Xi}$. Computing $\{s_k\}$ only requires solving equation (33) for $1 \leq k \leq N$. The construction of $\mathbf{\Xi}$ requires the computation of N coefficients ϕ_{ij} from formula (12). Various tricks may be used to speed up this computation, which may easily be completed in $\mathcal{O}(N^2)$ for each line vector, and $\mathcal{O}(N^3)$ for the matrix. The total complexity of this step becomes $\mathcal{O}(N^3)$;

2. Solving for $\mathbf{g}$ stills takes $\mathcal{O}(N^3)$;

3. Computing a particular value of $\mathbf{W}^*(s)$ using equation (32) necessitates the construction of the matrix $\mathbf{\Psi}$ each time. Using (20), it is not necessary to build matrix $\mathbf{\Phi}$, if only the unconditional waiting time is needed. Constructing $\mathbf{\Psi}$ takes $\mathcal{O}(N^3)$, as in step 1.

   Note however that directly inverting $\mathbf{A}(s)$, or rather, solving the system $\mathbf{A}(s)\mathbf{x} = \mathbf{g}$, is a faster option, because $\mathbf{A}(s)$ is a sparse matrix. The complexity of this step by this method is then $\mathcal{O}(N^3)$.

The main improvement of the algorithm, with respect to that of the general case, is therefore the simplification of step 1. However, one might expect that making use of the structure of the solution should bring a gain in step 3 as well. Minimally, not inverting matrices may result in a gain in numerical stability.

Note also that the availability of a closed form (32) for $\mathbf{W}^*(s)$ gives the possibility of a formal inversion of the transform, at least in the case where $H^*(s)$ is rational. This property is exploited in Section 4.4 for exponential services. More investigation is needed in this direction.

4.3 The Expected Workload

In this section, we apply the above results to derive expressions for the expected workload in the MMPP/GI/1 queue. In this section, we shall concentrate on the scalar distribution of the workload, W, which is denoted W_v in [9]. We have: $W^*(s) = \mathbf{W}^*(s)\mathbf{1}$.

According to [9, eq. (52)], the expected workload $w = \mathbb{E}W$ is given by:

$$w = \frac{1}{2(1-\rho)}\left[2\rho + \overline{\lambda}\overline{m}^{(2)} - 2\overline{m}((1-\rho)\mathbf{g} + \overline{m}\boldsymbol{\pi}\mathbf{\Lambda})(\mathbf{Q} + \mathbf{1}\boldsymbol{\pi})^{-1}\boldsymbol{\lambda}\right].$$

We shall prove the alternate formulas (34) and (39), respectively for superpositions of homogeneous and heterogeneous binary sources.

4.3.1 Superpositions of homogeneous binary sources

Proposition 4.1 *When the source is the superposition of homogeneous binary sources, the expected virtual waiting time in the MMPP/GI/1 is:*

$$w = \frac{1}{1-\rho}\left(\rho\frac{\overline{m}^{(2)}}{2\overline{m}} + N\overline{m}^2\frac{(\lambda_0-\lambda_1)^2}{(q_0+q_1)^3}q_0q_1\right) + N\overline{m}\ q_0\ \frac{\lambda_0-\lambda_1}{(q_0+q_1)^2} - \overline{m}\frac{\lambda_0-\lambda_1}{q_0+q_1}\ \mathbf{g}.\mathbf{j}\ , \qquad (34)$$

where $\mathbf{j} = (0, 1, 2, \ldots, N)^T$.

Proof We shall use the notation $\nu = 1 + q_1/q_0$. First of all, the following special values are easily computed from the definitions of the $\sigma_i(s)$, $\boldsymbol{\psi}_k(s)$ and $\boldsymbol{\phi}_k(s)$:

$$\sigma_1(0) = 1, \quad \sigma_2(0) = -\frac{q_1}{q_0}, \quad \sigma_2(0) - \sigma_1(0) = -\nu$$

$$\sigma_2'(0) - \sigma_1'(0) = -\frac{\overline{m}}{q_0}(\lambda_0 - \lambda_1)\frac{q_0 - q_1}{q_0 + q_1}$$

$$\boldsymbol{\phi}_0(0) = \nu^N\boldsymbol{\pi}, \quad \boldsymbol{\psi}_0(0) = (-1)^N\mathbf{1}.$$

Also, recall that $\omega_0'(0) = 1 - \rho$ so that

$$\lim_{s\to 0} \frac{s}{\omega_0(s)} \;=\; \frac{1}{1-\rho}\;.$$

We start from equation (32) and notice that when $k \neq 0$, $\omega_k(s) \neq 0$. Therefore, we have:

$$\begin{aligned}
W^{*\prime}(0) \;&=\; \frac{d}{ds}\left((1-\rho)(\sigma_2(s)-\sigma_1(s))^{-N}\mathbf{g}\boldsymbol{\psi}_0(s)\boldsymbol{\phi}_0(s)\frac{s}{\omega_0(s)}\right)\Big|_{s=0}\;\mathbf{1} \\
&\quad+\;(1-\rho)(\sigma_2(0)-\sigma_1(0))^{-N}\sum_{k=1}^{N}\mathbf{g}\boldsymbol{\psi}_k(0)\boldsymbol{\phi}_k(0)\mathbf{1}\;\frac{1}{\omega_k(0)} \\
&=\;(1-\rho)\;\mathbf{g}\mathbf{1}\boldsymbol{\pi}\mathbf{1}\;\frac{d}{ds}\;\frac{s}{\omega_0(s)}\Big|_{s=0} \qquad (35)\\
&\quad+\;(1-\rho)\frac{N(\sigma_2'(0)-\sigma_1'(0))}{(\sigma_2(0)-\sigma_1(0))^{N+1}}(-1)^N\nu^N\mathbf{g}\mathbf{1}\boldsymbol{\pi}\mathbf{1}\frac{1}{1-\rho} \\
&\quad+\;(1-\rho)(-1)^N\nu^{-N}\mathbf{g}\left(\boldsymbol{\psi}_0'(0)\boldsymbol{\phi}_0(0)+\boldsymbol{\psi}_0(0)\boldsymbol{\phi}_0'(0)\right)\mathbf{1}\frac{1}{1-\rho} \qquad (36)\\
&\quad+\;(1-\rho)(-1)^N\nu^{-N}\sum_{k=1}^{N}\mathbf{g}\boldsymbol{\psi}_k(0)\boldsymbol{\phi}_k(0)\frac{1}{\omega_k(0)}\;\mathbf{1} \qquad (37)\\
&=\;(1-\rho)\;\boldsymbol{\pi}\frac{d}{ds}\;\frac{s}{\omega_0(s)}\Big|_{s=0} \;-\; N\overline{m}(\lambda_0-\lambda_1)\frac{q_0-q_1}{(q_0+q_1)^2} \\
&\quad+\;(-\nu)^{-N}\mathbf{g}\left(\boldsymbol{\psi}_0'(0)\boldsymbol{\phi}_0(0)+\boldsymbol{\psi}_0(0)\boldsymbol{\phi}_0'(0)\right)\;\mathbf{1}\;. \qquad (38)
\end{aligned}$$

Indeed, the term (37) vanishes. To see this, use formula (17) for $\mathbf{A}(0)$ coupled with the fact that $\mathbf{A}(0) = \mathbf{Q}$ and $\omega_0(0) = 0$. One may write:

$$\begin{aligned}
\boldsymbol{\Psi}(0)\begin{pmatrix}1 & & & \\ & \omega_1(0) & & \\ & & \ddots & \\ & & & \omega_N(0)\end{pmatrix}\boldsymbol{\Phi}(0) \;&=\; (-\nu)^N\mathbf{A}(0)+(1-\omega_0(0))\boldsymbol{\psi}_0(0)\boldsymbol{\phi}_0(0) \\
&=\; (-\nu)^N\;\;(\mathbf{Q}+\mathbf{1}\boldsymbol{\pi})\,.
\end{aligned}$$

Consequently,

$$\sum_{k=1}^{N}\boldsymbol{\psi}_k(0)\boldsymbol{\phi}_k(0)\frac{1}{\omega_k(0)} \;=\; (-\nu)^N\left((\mathbf{Q}+\mathbf{1}\boldsymbol{\pi})^{-1}-\mathbf{1}\boldsymbol{\pi}\right),$$

which is zero when post-multiplied by $\mathbf{1}$.

In order to evaluate (35), it is necessary to perform a Taylor expansion of ω_0. After straightforward calculations, one obtains

$$\omega_0(s) \;=\; s(1-\rho)\;+\;s^2\left(\rho\frac{\overline{m}^{(2)}}{2\overline{m}}\;+\;N\overline{m}^2\frac{(\lambda_0-\lambda_1)^2}{(q_0+q_1)^3}q_0q_1\right)\;+\;o(s^2)\;.$$

In order to evaluate (36), we differentiate the product $\boldsymbol{\psi}_0(s)\boldsymbol{\phi}_0(s)\mathbf{1}$, using the fact that $\boldsymbol{\phi}_0(s)\mathbf{1} = (1-\sigma_2(s))^N$ (Remark 2.2). The result is:

$$(\boldsymbol{\psi}_0\boldsymbol{\phi}_0\mathbf{1})'(0) \;=\; (-\nu)^N \left(\overline{m}\frac{\lambda_0-\lambda_1}{q_0+q_1}\,\mathbf{j} \;-\; N\overline{m}q_1\frac{\lambda_0-\lambda_1}{(q_0+q_1)^2}\mathbf{1} \right) \;.$$

Replacing all of these values in (38) leads to (34). ■

4.3.2 Superpositions of heterogeneous binary sources

We assume now that the source is the superposition of homogeneous binary sources, and we use the notation of Section 3.2. The calculation of Section 4.3.1 now yields the following result:

Proposition 4.2 *When the source is the superposition of heterogeneous binary sources, the expected virtual waiting time in the MMPP/GI/1 is:*

$$\begin{aligned} w \;=\; & \frac{1}{1-\rho}\left(\rho\frac{\overline{m}^{(2)}}{2\overline{m}} + \overline{m}^2 \sum_{k=1}^{K} N_k \frac{(\lambda_0^{(k)}-\lambda_1^{(k)})^2}{(q_0^{(k)}+q_1^{(k)})^3} q_0^{(k)} q_1^{(k)} \right) \\ & + \overline{m} \sum_{k=1}^{K} N_k q_0^{(k)} \frac{\lambda_0^{(k)}-\lambda_1^{(k)}}{(q_0^{(k)}+q_1^{(k)})^2} \;-\; \overline{m} \sum_{i_1,\ldots,i_N} g_{i_1,\ldots,i_N} \sum_{k=1}^{K} i_k \frac{\lambda_0^{(k)}-\lambda_1^{(k)}}{q_0^{(k)}+q_1^{(k)}} \;. \end{aligned} \tag{39}$$

4.4 The Case of Exponential Service Times

In this paragraph, we consider the MMPP/M/1 queue with a source which is the superposition of binary sources. We show that the Laplace transform $\mathbf{W}^*(s)$ can be formally inverted, resulting in formulas (42) and (44) which are readily computed from known functions and quantities. In this section, we shall only discuss the scalar distribution W.

We first study the case of a homogeneous superposition, then generalize to the case of a heterogeneous superposition.

4.4.1 Superposition of homogeneous binary sources

We now have $1 - H^*(s) = s/(s+\mu)$. Let us call $\delta(s)$ the determinant of the matrix

$$\tilde{\mathbf{A}}(s) \;=\; (s+\mu)\;\mathbf{A}(s) \;=\; s(s+\mu)\mathbf{I} \;+\; (s+\mu)\;\mathbf{Q} \;-\; s\boldsymbol{\Lambda} \;.$$

The degree of $\delta(s)$ is exactly $2(N+1)$, because all elements of $\tilde{\mathbf{A}}(s)$ are polynomials of s with degree 2 on the diagonal, and at most 1 elsewhere. This conclusion can also be reached using (18).

It is plain from (4) that $W^*(s)$ is a rational function, which can be written as

$$\begin{aligned} W^*(s) \;&=\; (1-\rho)\frac{s}{\delta(s)}\;(s+\mu)\;\mathbf{g}\;[\tilde{\mathbf{A}}(s)]^a\;\mathbf{1} \\ &=\; (1-\rho)\;s\;(s+\mu)\;\frac{\nu(s)}{\delta(s)}\;, \end{aligned}$$

where $[\tilde{\mathbf{A}}(s)]^a$ is the adjoint matrix of $\tilde{\mathbf{A}}(s)$, and $\nu(s)$ a polynomial. It is already known from the above analysis that s and the $s - s_k$, $1 \leq k \leq N$, are factors of $\delta(s)$. The latter ones are also factors of $\nu(s)$ and cancel out, according to (24).

The asymptotic analysis of the functions $\omega_k(s)$ in the vicinity of $s = -\mu$ reveals that

$$\omega_k(s) \;=\; ((N-k)\;\max\{\lambda_0,\lambda_1\} \;+\; k\;\min\{\lambda_0,\lambda_1\})\;\frac{1}{s+\mu} \;+\; C_k \;+\; o(1)\;, \tag{40}$$

with $C_k \neq 0$. Therefore, when $s \to -\mu^+$, $\omega_k(s) \to +\infty$, *except* when $N = k$ and $\min\{\lambda_0, \lambda_1\}$. It was shown in Section 4.1 that $\omega_k(0) < 0$, $k \neq 0$ and that $\omega_0(0) = 0$ but $\omega_k'(0) > 0$ under the stability condition. Then, $\omega_k(s)$ has a root $t_k \in (-\mu, 0)$ for $0 \leq k < N$. If $\lambda_0, \lambda_1 > 0$, then $\omega_N(s)$ has a root t_N in the same interval. Necessarily, the $(s - t_k)$ are factors of $\delta(s)$.

We have determined the $2N + 2$ factors of $\delta(s)$ for the case $\lambda_0, \lambda_1 > 0$. If $\lambda_0 = 0$ or $\lambda_1 = 0$, the missing factor is $(s + \mu)$. Indeed, it is easily seen that this term is a factor of each element in the first row of the matrix $\widetilde{\mathbf{A}}(s)$. We set $t_N = -\infty$ in this case.

At this stage, we have proven that the denominator of $W^*(s)$ is the product of the terms $(s - t_k)$. Therefore,

$$W(s) \;=\; 1 - \rho \;+\; \sum_{k=0}^{N} \frac{a_k}{s - t_k}\;,$$

where a_k is the residue of $-t_k/(s - t_k)$ at $s = t_k$. Using now (32), it is readily seen that:

$$a_k \;=\; (1-\rho)(\sigma_2(t_k) - \sigma_1(t_k))^{-N}\;\frac{1}{\omega_k'(t_k)}\;\mathbf{g}\;\boldsymbol{\psi}_k(t_k)\;\boldsymbol{\phi}_k(t_k)\;\mathbf{1}\;. \tag{41}$$

Once t_k has been determined, all quantities in this formula are known or computable,

We have thus proven the following result.

Proposition 4.3 *In the MMPP/M/1 queue, the distribution of the workload is given by*

$$\mathbb{P}(W \leq x) \;=\; 1 \;-\; \sum_{k=0}^{N} a_k\; e^{x t_k}, \quad x > 0, \tag{42}$$

where the t_k, $k \leq 0 \leq N$ are the negative roots of the equations $s + \omega_k(s) = 0$ (with the convention $t_N = -\infty$ if $\lambda_0 = 0$ or $\lambda_1 = 0$), and a_k is given in (41).

4.4.2 Superposition of heterogeneous binary sources

We assume here that the source is the superposition of homogeneous binary sources, and we use the notation of Section 3.2. The reasoning of Section 4.4.1 applies with the obvious modifications.

In the case where $\lambda_0^{(k)}, \lambda_1^{(k)} > 0$ for *some* k, then every function $\omega(i_1, \ldots, i_N)$ has a root $t_{i_1,\ldots,i_N}$ in the interval $(-\mu, 0)$. In the case that there is a $\lambda_i^{(k)} = 0$ for all k we adapt the convention, $t_{N_1,\ldots,N_K} = -\infty$.

The residue $a_{i_1,\ldots,i_N}$ takes the form

$$a_{i_1,\ldots,i_N} = (1-\rho)\prod_{k=1}^{N}(\sigma_2^{(k)}(t_{i_1,\ldots,i_N}) - \sigma_1^{(k)}(t_{i_1,\ldots,i_N}))^{-N_k} \frac{1}{\omega'_{i_1,\ldots,i_N}(t_{i_1,\ldots,i_N})}\, \mathbf{g}\, \boldsymbol{\psi}_k(t_{i_1,\ldots,i_N})\, \boldsymbol{\phi}_k(t_{i_1,\ldots,i_N})\, \mathbf{1}\ . \tag{43}$$

The value of the stationary workload distribution is then given by

$$\mathbb{P}(W \le x) = 1 - \sum_{i_1,\ldots,i_N} a_{i_1,\ldots,i_N}\, e^{x\ t_{i_1,\ldots,i_N}}, \quad x > 0. \tag{44}$$

5 Bounds on the Workload Distribution for MMPP/GI/1 Queue

Consider an MMPP/GI/1 queue as described in Section 2 with generator $\mathbf{Q}$ and rate matrix $\mathbf{\Lambda}$. Recall that $H(x)$ and $H^*(s)$ are the probability distribution and the Laplace transform of the service times, respectively, that $\overline{m}$ denotes the mean service time and that $\mathbf{p}$ is the invariant measure of the Markov chain embedded at arrivals epochs (cf. Section 2).

In the following $pf(\mathbf{A})$ will denote the Perron-Frobenius eigenvalue of a matrix $\mathbf{A}$ [11].

Recall the definition of the matrix $\mathbf{F}^*(s)$ in (3). For $s \in \mathcal{D} := \{s : H^*(-s)\ \mathbf{F}^*_{i,j}(s) < \infty\}$, let $\mathbf{z}(s)$ be the left-eigenvector of the matrix $H^*(-s)\,\mathbf{F}^*(s)$ associated with the eigenvalue $pf(H^*(-s)\,\mathbf{F}^*(s))$. We assume that $\mathbf{z}(s)$ is normalized so that its components sum up to 1.

Let W be the stationary workload in this MMPP/GI/1 queue.

The following result is shown in [13]:

Proposition 5.1 *Assume that the set $\mathcal{D}$ is open and the stability condition $\rho < 1$ holds. Then, there exist constants B and C such that*

$$B\, e^{-s^* x} \le \mathbb{P}(W > x) \le C\, e^{-s^* x}, \quad \forall x \ge 0 \tag{45}$$

where s^* is the unique solution in $(0,\infty) \cap \mathcal{D}$ of the equation

$$H^*(-s)\, pf(\mathbf{F}^*(s)) = 1. \tag{46}$$

The constants B and C are given by $B = \inf_{x\ge 0,\, 0\le i\le N}\ g_i(x)$ and $C = \sup_{x\ge 0,\, 0\le i\le N}\ g_i(x)$ where

$$g_i(x) = \frac{\mathbf{p}\left(\int_x^\infty dH(u)(\mathbf{I} - \exp(\mathbf{Q}-\mathbf{\Lambda})(u-x))(\mathbf{\Lambda}-\mathbf{Q})^{-1}\mathbf{\Lambda}\right)\mathbf{e}_i}{\mathbf{z}(s^*)\left(\int_x^\infty e^{s^*(u-x)}\int_u^\infty dH(y)(\mathbf{I} - \exp((\mathbf{Q}-\mathbf{\Lambda})(y-u)))\,du\mathbf{\Lambda}\right)\mathbf{e}_i} \tag{47}$$

where $\mathbf{e}_i$ is the vector whose components are 0 except the i-th one which is equal to 1.

The assumption that the set $\mathcal{D}$ is open is satisfied (in particular) for all service times with phase-type distribution [13].

The aim of this section is to propose an efficient algorithm for computing the bounds in (45) in the case that the arrival process is the superposition of K independent MMPP's $(\mathbf{Q}^{(k)}, \mathbf{\Lambda}^{(k)})$, $k = 1, 2, \ldots, K$. Under this assumption it is well known that the resulting input process is again an MMPP given by [9]

$$(\mathbf{Q}, \mathbf{\Lambda}) = \left(\oplus_{k=1}^{K} \mathbf{Q}^{(k)}, \oplus_{k=1}^{K} \mathbf{\Lambda}^{(k)} \right). \tag{48}$$

In order to compute the bounds in (45) we first need to evaluate three quantities: the optimal decay rate s^*, the eigenvector $\mathbf{z}(s^*)$ and the invariant vector $\mathbf{p}$.

We start with some preliminary remarks that connect the present analysis with that in the previous sections.

Define $\rho_1(s) = \log\, pf(\mathbf{F}^*(-s))$ and $\rho_2(s) := pf((e^s - 1)\ \mathbf{\Lambda} + \mathbf{Q})$. It is known that [19, Section 4 and Proposition 14]

$$\rho_2(s) = -\rho_1^{-1}(-s). \tag{49}$$

Hence,

$$\begin{aligned} pf((H^*(-s^*) - 1)\ \mathbf{\Lambda} + \mathbf{Q}) &= \rho_2(\log(H^*(-s^*))) \\ &= \rho_2(-\rho_1(-s^*)) \qquad (50) \\ &= s^* \qquad (51) \end{aligned}$$

where (50) and (51) follow from (46) and (49), respectively.

This result, in conjunction with (48) and the identity $pf(\mathbf{A}_1 \oplus \mathbf{A}_2) = pf(\mathbf{A}_1) + pf(\mathbf{A}_2)$ [10], yields

$$s^* = \sum_{k=1}^{K} pf\left((H^*(-s^*) - 1)\mathbf{\Lambda}^{(k)} + \mathbf{Q}^{(k)} \right). \tag{52}$$

We now focus on the computation of the eigenvector $\mathbf{z}(s^*)$. Let $\boldsymbol{\phi}$ be the left eigenvector of the matrix $(H^*(-s^*) - 1)\ \mathbf{\Lambda} + \mathbf{Q}$ associated with its Perron-Frobenius eigenvalue. We deduce from (51) that

$$\begin{aligned} &\boldsymbol{\phi}\ ((H^*(-s^*) - 1)\ \mathbf{\Lambda} + \mathbf{Q}) \ =\ \rho_2(\log(H^*(-s^*)))\boldsymbol{\phi} \ =\ s^*\ \boldsymbol{\phi} \\ &\iff \boldsymbol{\phi}\ \mathbf{\Lambda}\ H^*(-s^*) \ =\ \boldsymbol{\phi}\ (s^*\mathbf{I} + \mathbf{\Lambda} - \mathbf{Q}) \\ &\implies \boldsymbol{\phi}\ \mathbf{\Lambda}\ H^*(-s^*)\ \mathbf{F}^*(s^*) \ =\ \boldsymbol{\phi}\ \mathbf{\Lambda}. \end{aligned}$$

From the above and the uniqueness of the normalized eigenvector $\mathbf{z}(s^*)$ we deduce that

$$\mathbf{z}(s^*) = \boldsymbol{\phi}\ \mathbf{\Lambda}/|\boldsymbol{\phi}\ \mathbf{\Lambda}| \tag{53}$$

where $|\mathbf{v}|$ denotes the sum of the composants of any vector $\mathbf{v}$.

By using now the property that $\mathbf{v}_1 \otimes \mathbf{v}_2$ is an eigenvector of $\mathbf{A}_1 \oplus \mathbf{A}_2$ associated with $pf(\mathbf{A}_1 \oplus \mathbf{A}_2)$ if $\mathbf{v}_i$ is an eigenvector of $\mathbf{A}_i$ associated with $pf(\mathbf{A}_i)$, $i = 1, 2$ (see [10]) we get from (53), (52), and (48)

$$\mathbf{z}(s^*) = c_0 \left(\otimes_{k=1}^K \boldsymbol{\phi}^{(k)}\right) \left(\otimes_{k=1}^K \boldsymbol{\Lambda}^{(k)}\right) \tag{54}$$

with $c_0^{-1} := |(\otimes_{k=1}^K (\boldsymbol{\phi}^{(k)}) \ (\otimes_{k=1}^K \boldsymbol{\Lambda}^{(k)})|$, where $\boldsymbol{\phi}^{(k)}$ is the eigenvector of $(H^*(-s^*) - 1) \ \boldsymbol{\Lambda}^{(k)} + \mathbf{Q}^{(k)}$ associated with the Perron-Frobenius eigenvalue of this matrix.

A similar analysis yields (set $s^* = 0$ in (54))

$$\mathbf{p} = c_1 \left(\otimes_{k=1}^K \boldsymbol{\pi}^{(k)}\right) \left(\oplus_{k=1}^K \boldsymbol{\Lambda}^{(k)}\right) \tag{55}$$

where $\boldsymbol{\pi}^{(k)}$ is the invariant measure associated with the generator $\mathbf{Q}^{(k)}$ and c_1 is a normalization constant.

We now specialize formulas (52)-(55) to the cases when the input process is the superposition of independent *homogeneous* (resp. *heterogeneous*) binary sources.

Superposition of homogeneous binary sources We assume that the input process of the MMPP/GI/1 queue is the superposition of N independent, homogeneous binary sources (see definitions and notation in Section 4.1.1).

As already discussed in Section 4.1.1, the MMPP resulting from the superposition of N independent homogeneous binary sources may be seen as a MMPP with $N + 1$ states with generator $\mathbf{Q} = \mathbf{M}(N; q_0, q_1)$ and rate matrix $\boldsymbol{\Lambda} = N\lambda_0 \ \mathbf{I}(N) + (\lambda_1 - \lambda_0) \ \mathbf{J}(N)$, respectively. Therefore,

$$\mathbf{Q} - h\,\boldsymbol{\Lambda} = \mathbf{M}(N; q_0, q_1) - hN\lambda_0\,\mathbf{I}(N) + (\lambda_0 - \lambda_1)h\,\mathbf{J}(N) \tag{56}$$

with $h := 1 - H^*(-s^*)$.

The Perron-Frobenius eigenvalue of $\mathbf{Q} - h\boldsymbol{\Lambda}$ is obtained by letting $k = 0$ in (9). Together with (51) this implies that s^* is the unique solution in $(0, \infty) \cap \mathcal{D}$ of the equation

$$s^* = \frac{N}{2}\left(\sqrt{(h(\lambda_0 - \lambda_1) + q_0 - q_1)^2 + 4q_0q_1} - h(\lambda_0 + \lambda_1) - q_0 - q_1\right). \tag{57}$$

On the other hand, the left eigenvector $\boldsymbol{\phi} = (\phi(0), \phi(1), \ldots, \phi(N))$ of $\mathbf{Q} - h\,\boldsymbol{\Lambda}$ associated with the Perron-Frobenius eigenvalue of this matrix is also obtained by letting $k = 0$ in (12), so that

$$\begin{aligned} \phi(i) &= [x^i]\,(x - \sigma_2)^N \\ &= \binom{N}{i}(-\sigma_2)^{N-i}, \quad i = 0, 1, \ldots, N \end{aligned} \tag{58}$$

with $\sigma_2 = \left((\lambda_0 - \lambda_1)h + q_0 - q_1 - \sqrt{((\lambda_0 - \lambda_1)h + q_0 - q_1)^2 + 4q_0q_1}\right)/2q_0$. We may then conclude from (53), (58) and the definition of $\boldsymbol{\Lambda}$ that

$$\begin{aligned} \mathbf{z}(s^*) &= c_0 \ \boldsymbol{\phi}\,\boldsymbol{\Lambda} \\ &= c_0 \ \text{diag} \ \left(\binom{N}{i}(-\sigma_2)^{N-i}(N\lambda_0 + (\lambda_1 - \lambda_0)i),\ i = 0, 1, \ldots, N\right) \end{aligned} \tag{59}$$

with $c_0^{-1} = \sum_{i=0}^{N} \binom{N}{i} (-\sigma_2)^{N-i} (N\lambda_0 + (\lambda_1 - \lambda_0)i)$.

Finally, the invariant vector $\mathbf{p}$ is obtained by letting $h = 0$ in the derivation of $\mathbf{z}(s^*)$. This gives

$$\mathbf{p} = c_1 \text{ diag } \left(\binom{N}{i} (q_1/q_0)^{N-i} (N\lambda_0 + (\lambda_1 - \lambda_0)i),\ i = 0, 1, \ldots, N \right) \tag{60}$$

with $c_1^{-1} = \sum_{i=0}^{N} \binom{N}{i} (q_1/q_0)^{N-i} (N\lambda_0 + (\lambda_1 - \lambda_0)i)$.

Superposition of heterogeneous binary sources Assume now that the MMPP source is the superposition of K types of independent binary sources and let N_k be the number of sources of type $k = 1, 2, \ldots, K$. The shorthand MMPP$^{N_1,\ldots,N_K}$ will be used to denote this particular MMPP.

For a source of type $k = 1, 2, \ldots, K$, let $q_0^{(k)}$ (resp. $q_1^{(k)}$) be the rate out of state 0 (resp. state 1) and let $\lambda_0^{(k)}$ (resp. $\lambda_1^{(k)}$) be the generation rate in state 0 (resp. state 1).

From (52) and (57) we find that s^* is the unique solution in $(0, \infty) \cap \mathcal{D}$ of the equation

$$s^* = \sum_{k=1}^{K} \frac{N_k}{2} \left(\sqrt{(h(\lambda_0^{(k)} - \lambda_1^{(k)}) + q_0^{(k)} - q_1^{(k)})^2 + 4q_0^{(k)} q_1^{(k)}} - h(\lambda_0^{(k)} + \lambda_1^{(k)}) - q_0^{(k)} - q_1^{(k)} \right). \tag{61}$$

The eigenvector $\mathbf{z}(s^*)$ and the invariant vector $\mathbf{p}$ are computed from (54), (58) and from (55), (58) (with $h = 0$), respectively. We find

$$\mathbf{z}(s^*) = c_0 \left(\otimes_{k=1}^{K} \boldsymbol{\phi}^{(k)} \right) \left(\oplus_{k=1}^{K} \boldsymbol{\Lambda}^{(k)} \right) \tag{62}$$

$$\mathbf{p} = c_1 \left(\otimes_{k=1}^{K} \boldsymbol{\pi}^{(k)} \right) \left(\oplus_{k=1}^{K} \boldsymbol{\Lambda}^{(k)} \right) \tag{63}$$

with

$$\boldsymbol{\Lambda}^{(k)} = \text{diag} \left(N_k \lambda_0^{(k)} + (\lambda_1^{(k)} - \lambda_0^{(k)})\, i,\ i = 0, 1, \ldots, N_k \right) \tag{64}$$

$$\boldsymbol{\phi}^{(k)} = \left(\binom{N_k}{i} (-\sigma_2^{(k)})^{N_k - i},\ i = 0, 1, \ldots, N_k \right) \tag{65}$$

$$\boldsymbol{\pi}^{(k)} = \left(\binom{N_k}{i} (q_1^{(k)}/q_0^{(k)})^{N_k - i},\ i = 0, 1, \ldots, N_k \right) \tag{66}$$

and

$$\sigma_2^{(k)} = \frac{(\lambda_0^{(k)} - \lambda_1^{(k)})h + q_0^{(k)} - q_1^{(k)} - \sqrt{((\lambda_0^{(k)} - \lambda_1^{(k)})h + q_0^{(k)} - q_1^{(k)})^2 + 4q_0^{(k)} q_1^{(k)}}}{2q_0^{(k)}} \tag{67}$$

for $k = 1, 2, \ldots, K$. The constants c_0 and c_1 in (62)-(63) must be chosen so that $|\mathbf{z}(s^*)| = |\mathbf{p}| = 1$.

To conclude this section, we briefly discuss the computation of the constants B and C in (45) when W is the stationary workload of an MMPP$^{N_1,\ldots,N_K}$/Erlang(S)/1 queue. It is shown in [13]

that (with $\mu := 1/\overline{m}$ and $(0,\infty) \cap \mathcal{D} = (0,\mu)$)

$$g_i(x) = \left(\frac{\mu - s}{\mu}\right)^{S+1} \frac{\mathbf{p}\ \mathbf{\Gamma}(x,\mu,S)\ \mathbf{e}_i}{\mathbf{z}(s^*)\ \mathbf{\Gamma}(x,\mu - s^*,S)\ \mathbf{e}_i}, \quad i = 0,1,\ldots, \prod_{k=1}^{K}(N_k+1) - 1 := N \tag{68}$$

where s^*, $\mathbf{z}(s^*)$ and $\mathbf{p}$ are given in (61), (62) and (63), respectively, and where $\mathbf{\Gamma}(x,\mu,S) = \sum_{r=0}^{S-1}\sum_{j=0}^{r}((x\mu)^j/j!)(\mu\mathbf{\Delta})^{S-r}$ with

$$\mathbf{\Delta} = -\left(\oplus_{k=1}^{K}\mathbf{A}^{(k)}\right)^{-1} \tag{69}$$

where

$$\mathbf{A}^{(k)} := \mathbf{M}\left(N_k, q_0^{(k)}, q_1^{(k)}\right) - \left(\mu + N_k\lambda_0^{(k)}\right)\mathbf{I}(N_k) + \left(\lambda_0^{(k)} - \lambda_1^{(k)}\right)\mathbf{J}(N_k). \tag{70}$$

In direct analogy with the derivation of (28) we obtain from Lemma 2.1 and Lemma 2.3 that

$$\mathbf{\Delta} \;=\; -\prod_{k=1}^{K}(\sigma_2^{(k)} - \sigma_1^{(k)})^{-N_k} \sum_{(j_1,\ldots,j_K)} \frac{\left(\otimes_{k=1}^{K}\boldsymbol{\psi}_{j_k}^{(k)}\right)\cdot\left(\otimes_{k=1}^{K}\boldsymbol{\phi}_{j_k}^{(k)}\right)}{\sum_{k=1}^{K}\omega_{j_k}^{(k)}} \tag{71}$$

where $\omega_j^{(k)}$ $(0 \le j \le N_k)$ and $\sigma_r^{(k)}$ $(r = 1,2)$ are given in (9) and in (10), respectively, after setting $k = j$, $N = N_k$, $\lambda = q_0^{(k)}$, $\mu = q_1^{(k)}$, $a = -(\mu + N_k\lambda_0^{(k)})$ and $b = \lambda_0^{(k)} - \lambda_1^{(k)}$ for $k = 1,2,\ldots,K$. In (71), the vector $\boldsymbol{\phi}_j^{(k)} = (\phi_j^{(k)}(i),\ 0 \le i \le N_k)$ is given by $\phi_j^{(k)}(i) = [x^i](x - \sigma_1^{(k)})^j\,(x - \sigma_2^{(k)})^{(N_k - j)}$ and the vector $\boldsymbol{\psi}_j^{(k)} = (\psi_j^{(k)}(i),\ 0 \le i \le N_k)$ is given by $\psi_j^{(k)}(i) = \phi_i^{(k)}(j)\,(\sigma_2^{(k)})^{i+j+N_k}$ for $0 \le j \le N_k$, $1 \le k \le K$.

The complexity of computing B (resp. C) is dominated by the search for the value of x that yields the infimum (resp. supremum) in the expression for B (resp. C) in Proposition 5.1. It is easily shown that no more that $S-1$ values of x (possibly including $x = 0$ and $x = \infty$) need to be checked and that, except for $x = 0$ and $x = \infty$, they are the positive real roots of the polynomial

$$\begin{aligned} p_i(x) \;:=\; \mathbf{p}\Bigg(&\frac{d\mathbf{\Psi}(x,\mu,S)}{dx}\ \mathbf{\Delta}\ \mathbf{e}_i\ \mathbf{z}(s^*)\ \mathbf{\Psi}(x,\mu - s^*,S) \\ &-\ \mathbf{\Psi}(x,\mu,S)\ \mathbf{\Delta}\ \mathbf{e}_i\ \mathbf{z}(s^*)\ \frac{d\mathbf{\Psi}(x,\mu - s^*,S)}{dx}\Bigg)\mathbf{\Delta}\ \mathbf{e}_i \end{aligned} \tag{72}$$

which can be shown to be a degree $2\,(S-2)$.

For the case $S = 1$ (MMPP$^{N_1,\ldots,N_K}$/M/1 queue), $g_i(x)$ does not depend on x and we have

$$C = \left(\frac{\mu - s^*}{\mu}\right)\max_{i=0,1,\ldots,N}\frac{\mathbf{p}\,\mathbf{\Delta}\,\mathbf{e}_i}{\mathbf{z}(s^*)\,\mathbf{\Delta}\,\mathbf{e}_i}, \qquad B = \left(\frac{\mu - s^*}{\mu}\right)\min_{i=0,1,\ldots,N}\frac{\mathbf{p}\,\mathbf{\Delta}\,\mathbf{e}_i}{\mathbf{z}(s^*)\,\mathbf{\Delta}\,\mathbf{e}_i}. \tag{73}$$

For the case $S = 2$, $p_i(x)$ is a constant and we have been able to establish that the infimum (resp. supremum) of $g_i(x)$ over x in $[0, \infty)$ is always achieved at $x = 0$ (resp. $x = \infty$).

For the case $S > 2$ we conjecture that $p_i(x)$ has no positive real roots. We further conjecture that the infimum (resp. supremum) of $g_i(x)$ over x in $[0, \infty)$ is always reached at $x = 0$ (resp. $x = \infty$).

An explicit expression for $g_i(x)$ is also given in [13] for the MMPP$^{N_1,\ldots,N_K}$/D/1 queue.

Let us now briefly summarize the algorithm for computing the bounds for MMPP$^{N_1,\ldots,N_K}$/GI/1 queues. The following steps must be followed:

1. Evaluate s^* as the unique solution of (61) in $(0, \infty) \cap \mathcal{D}$.
2. Compute $\mathbf{z}(s^*)$ by using (62).
3. Compute $\mathbf{p}$ by using (63).
4. Compute B and C via (47). More specifically, if the service time distribution is an Erlang distribution then use (68); if the services are constant, then use (3.14) in [13].

6 Numerical Results

In this section, we present some preliminary results using the previously described algorithms. The results were obtained with software implementing the algorithms of Section 4.2. This software is written in C and uses the Meschach library for numerical linear algebra [18]. It will be made publicly available.

In these experiments, the source consists of a superposition of two types of binary sources having the following characteristics

$$\begin{array}{llllllll} q_0^{(1)} = 1.5384, & q_1^{(1)} = 2.8409, & \lambda_0^{(1)} = 0, & \lambda_1^{(1)} = 0.064, \\ q_0^{(2)} = 1.25, & q_1^{(2)} = 5.0, & \lambda_0^{(2)} = 0, & \lambda_1^{(2)} = 0.32. \end{array}$$

These numbers are taken from [3], and are derived from voice traffic data. The source is the superposition of twelve type 1 and six type 2 sources ($N_1 = 12$, $N_2 = 6$). The service time distribution is two phase Erlang with an average adjusted to achieve different loads. Figures 1, 2 and 3 compare the exact value of $\mathbb{P}(W > x)$ with the bounds in (45). The Figures differ according to the average service time which is taken as 0.6, 1.0 and 1.35, resulting in load factors of 0.392, 0.654 and 0.883, respectively.

The main conclusions of the numerical experiments are:

- The algorithms and their implementation turn out to be quite efficient, in the sense that problems of reasonable size can be solved within a few minutes. For instance, when building the figures, ten data points for $\mathbb{P}(W > x)$ were typically obtained in less than one minute CPU time on a SUN Ultra 170 workstation. The construction of the $\mathbf{g}$ vector is immediate, and the principal part of the computation time is spent on inversing the Laplace transform $\mathbf{W}^*(s)$.

  The sources used in the reported experiment are of moderate size ($7 \times 13 = 91$ states). We have performed experiments with superpositions of two or three groups of binary sources, leading

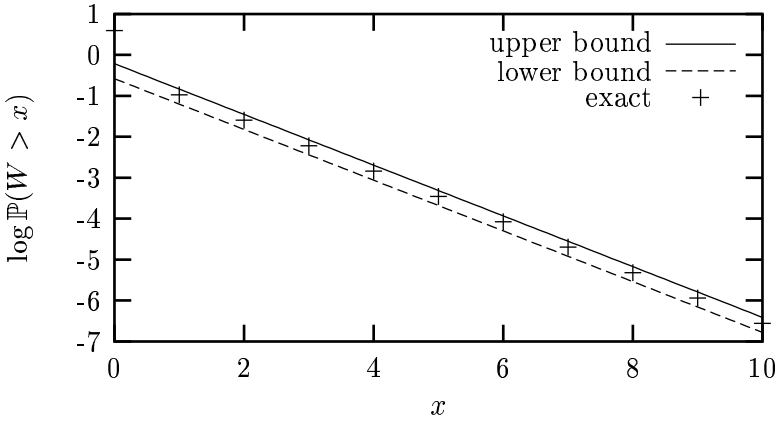

Figure 1: Bounds and exact values for $\rho \simeq 0.392$

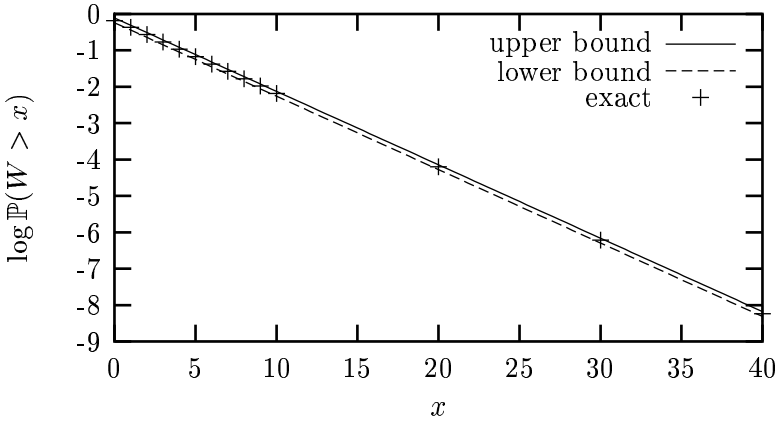

Figure 2: Bounds and exact values for $\rho \simeq 0.654$

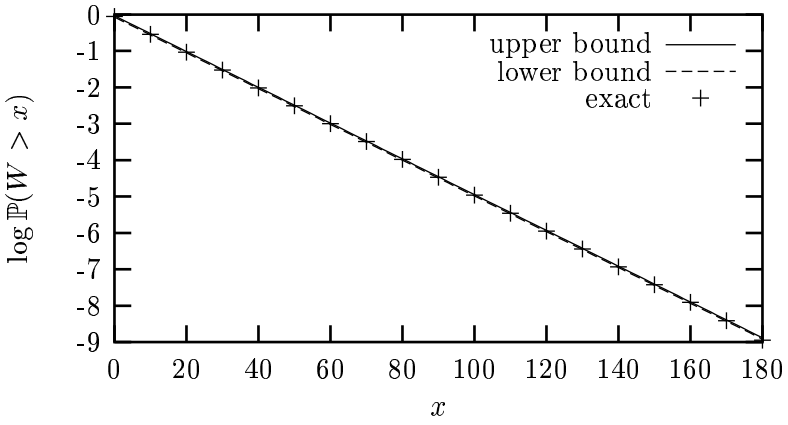

Figure 3: Bounds and exact values for $\rho \simeq 0.883$

to up to 1690 states (a superposition of three groups of 4, 12 and 25 binary sources). For 400 states, computing ten points of $\mathbb{P}(W > x)$ takes about 15 minutes in the same conditions. For 1000 states, computing one value of $\mathbb{P}(W > x)$ takes about 10 minutes. For 1600 states, computing the vector $\mathbf{g}$ alone takes about 2,5 minutes.

- In the case of exponential service times (still with binary sources), where the inversion of the Laplace transform is not necessary, the computation times reduce to a few seconds. This is also the case when computing expected waiting times with (39).

- Numerical precision problems prevent the exact computations of probabilities less than $10^{-7}/10^{-8}$ in the case of low traffic. This problem can be addressed using an increased precision, as for instance with the MAPLE[1]software, at the cost of a much increased running time.

- The potential drawback of using the linear system (24) instead of the $\mathbf{G}$ matrix is that the $\mathbf{\Xi}$ matrix is not stochastic. It may contain numbers of arbitrary sign and magnitude, and the solution of (24) is expected to yield numerical instability, especially at high loads. This problem has indeed been observed, though, quite surprisingly, not before the size of the problem reaches several hundred states. For instance, computations involving sources with 300 states and with a load equal to 0.995 have been found to be stable.

7 Concluding Remarks

We have presented a new computation scheme for the analysis of the stationary workload in the MMPP/GI/1 queue when the MMPP is reversible. The technique parallels that of Stern and Elwalid [17] for the analysis of MMRP processes. The basic ideas are to diagonalize the matrix $\mathbf{A}(s)$ (cf. (21)) and to develop its spectral expansion which yields a new formula (23) for the Laplace transform

[1]MAPLE is a trademark of Waterloo Maple Inc.

of the workload $\mathbf{W}^*(s)$. The singularity analysis of this formula in turn allows us to establish a linear system whose solution gives the probability vector $\mathbf{g}$ and $\mathbf{W}^*(s)$.

The diagonalization of the matrix $\mathbf{A}(s)$ also simplifies the computation when the input is the superposition of independent sources. Indeed, owing to the Kronecker algebra, $\mathbf{A}(s)$, its inverse and its exponential are easily obtained by Kronecker sums and Kronecker products of those of composing elements, see Lemma 2.1 and Corollary 2.2.

These properties have allowed us to devise a new computation algorithm which is particularly suitable for superposition of sources. It is even more interesting when superpositions of binary sources are to be analyzed. In this case, closed-form expressions have been obtained for the diagonalization of $\mathbf{A}(s)$. When the service times are exponentially distributed, the workload distribution has a closed-form expression as well. We have applied this method to the computation of both exact results and bounds for the workload distribution.

We are aware that the applicability of any algorithm based on the exact solution of Markov Modulated models is limited, and that very large problems will still stay out of reach. However, to the best of our knowledge, results concerning sources with a thousand states have not been reported in the literature. Limitations originate from algorithmic complexity and numerical inaccuracies. The first problem is not specific to our approach. The second problem arises typically when the system is highly loaded. Interestingly enough, this situation corresponds to the case when the bounding approach is the most efficient (see Figure 3 and [13]).

It will be interesting to further investigate the symbolic inversion of $\mathbf{W}^*(s)$ when the service time has a rational Laplace transform. Other future research directions include the extension of the results to the case of non-reversible Markov chains, and the investigation of appropriate data structures for Kronecker algebra to obtain additional computation time savings.

A Proof of and Remarks on Lemma 2.3

Proof (of Lemma 2.3.)

The fact that $\mathbf{\Omega\Phi} = \mathbf{\Phi A}$ is a direct application of the analysis of [2] and [15]. Indeed, a pair $(\omega, \boldsymbol{\phi})$ is a solution of

$$\omega\boldsymbol{\phi} \;=\; \boldsymbol{\phi}\mathbf{A} \tag{74}$$

if and only if:

$$\boldsymbol{\phi}\left(-b\mathbf{J}(N) + (\omega - a)\mathbf{I}(N)\right) \;=\; \boldsymbol{\phi}\mathbf{M}(N;\lambda,\mu)\ .$$

In [15, eq. (4.3i) *et seq.*], Mitra solves the problem of finding the numbers z and the vectors $\boldsymbol{\phi}$ solution of:

$$z\boldsymbol{\phi}\ \left(c\mathbf{J}(N) - \nu\mathbf{I}(N)\right) \;=\; \boldsymbol{\phi}\mathbf{M}(N;\lambda,\mu)\ . \tag{75}$$

He shows that the values z for which a solution exists are such that:

$$\nu z - \frac{N}{2}(cz + \lambda + \mu) + (\frac{N}{2} - k)\sqrt{(cz - \lambda + \mu)^2 + 4\lambda\mu} \;=\; 0, \tag{76}$$

for some $k \in \{0, \ldots, N\}$. The corresponding eigenvectors ϕ are given by:

$$\phi(i) \;=\; [x^i](x - \tau_1(z_k))^k (x - \tau_2(z_k))^{N-k},$$

where $\tau_{1,2}(z)$ are the roots of the polynomial:

$$\lambda X^2 + (cz - \lambda + \mu)X - \mu \;=\; 0 \;,$$

Here, we have a problem of type (75) with:

$$c \;=\; -b \;, \quad \nu \;=\; -(\omega - a) \;, \quad \text{and} \quad z = 1 \;. \tag{77}$$

Consequently, the pair (ω, ϕ) is a solution of (74) if and only if the pair $(1, \phi)$ is solution of (75). Using (76) and (77), we find that ω must satisfy:

$$0 \;=\; -(\omega - a) - \frac{N}{2}(-b + \lambda + \mu) + (\frac{N}{2} - k)\sqrt{(b + \lambda - \mu)^2 + 4\lambda\mu} \;,$$

This gives the values (9). All these values are distinct provided that (16) holds. The form of the eigenvectors $\boldsymbol{\phi}_k$ follows from (12).

In order to establish (17), we prove that:

$$\boldsymbol{\Phi}\,\boldsymbol{\Psi} \;=\; (\sigma_2 - \sigma_1)^N\;\mathbf{I} \;. \tag{78}$$

By definition, we have:

$$\begin{aligned}
\sum_{k=0}^{N} \phi_{ik}\psi_{kj} \;&=\; \sum_{k=0}^{N} [x^k](x - \sigma_1)^i (x - \sigma_2)^{N-i} \phi_{kj} \sigma_2^{k+j-N} \\
&=\; \sigma_2^{j-N} \sum_{k=0}^{N} [x^k](x - \sigma_1)^i (x - \sigma_2)^{N-i} [y^j](y - \sigma_1)^k (y - \sigma_2)^{N-k} \sigma_2^k \\
&=\; \sigma_2^{j-N} [y^j](y - \sigma_2)^N \;(x - \sigma_1)^i (x - \sigma_2)\big|_{x = \sigma_2(y-\sigma_1)/(y-\sigma_2)} \\
&=\; \sigma_2^{j-N} [y^j](y - \sigma_2)^N \frac{(\sigma_2(y - \sigma_1) - \sigma_1(y - \sigma_2))^i \sigma_2^{N-i} (y - \sigma_1 - y + \sigma_2)^{N-i}}{(y - \sigma_2)^N} \\
&=\; \sigma_2^{j-i} (\sigma_2 - \sigma_1)^N [y^j] y^i \\
&=\; (\sigma_2 - \sigma_1)^N \;\delta_{i=j} \;.
\end{aligned}$$

From (78), we can conclude (17) provided that $\sigma_1 - \sigma_2 \neq 0$. According to (10) or (11), this is true as long as

$$(-b + \lambda + \mu)^2 + 4\lambda\mu \neq 0 \;,$$

which is equivalent to (16). ■

Remark A.1 It is easily seen from (9) that the eigenvalues ω_k are placed in a line of the complex plane, and regularly spread out. The spectrum is symmetrical with respect to the point $a + N(b - \lambda - \mu)$. Due to the symmetry between k and $N - k$, one may choose any determination for the square root in (9). This symmetry is also the reason why the characteristic polynomial of $\mathbf{A}$ is actually a polynomial of all its variables, despite the square roots of appearing in (18).

Remark A.2 (Degenerate cases) Lemma 2.3 holds true in case $\lambda = 0$, $b \neq \mu$. In that case, however, the definitions (10), (12) and (14) "degenerate" into: $\sigma_1 = -\mu/(b - \mu)$ and:

$$\phi_k(i) \;=\; [x^i](x - \sigma_1)^k \;, \qquad \psi_k(i) \;=\; [x^i](x + \sigma_1)^k \;.$$

Remark A.3 (Analyticity) In applications, the matrix $\mathbf{A}$ is often considered as a function of its (complex) parameters. It is therefore useful to state on the analyticity of the decomposition (17) with respect to a or b.

The functions $\sigma_{1,2}$ defined in (10) are analytic in each of its parameters λ, μ, a, b as long as condition (16) holds. Obviously, the parameter a does not play a role in the analyticity.

For λ and μ fixed, $\sigma_{1,2}$ are analytic in the variable b in domains slit along lines going from the points $b_{1,2} = -(\sqrt{\lambda} \pm i\sqrt{\mu})^2$ to infinity. See Figure 4. Observe that these points lie in the right-hand half plane if $\lambda < \mu$, in the left-hand half plane if $\lambda > \mu$, and on the imaginary axis if $\lambda = \mu$.

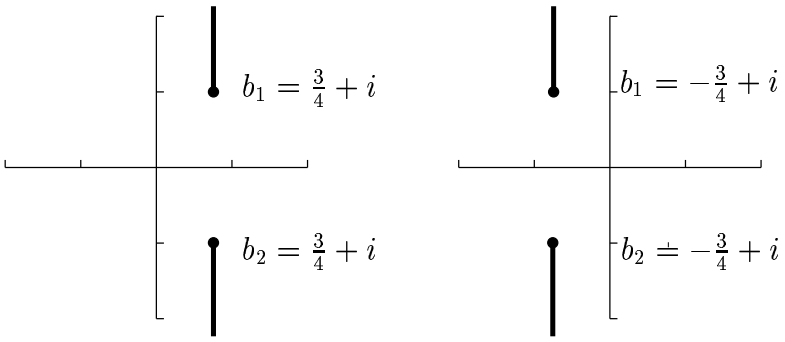

Figure 4: Domain of analyticity for b

The point $\lambda = 0$ is singular for the decomposition (17), although $\mathbf{A}$ is normally diagonalizable there (see Remark A.2). The matrix $\mathbf{\Psi}$ defined in (14) may fail to be analytic when σ_2 vanishes. According to (11), this can happen only when $\mu = 0$.

Finally, note that although each of the elements $\sigma_1, \sigma_2, \mathbf{\Phi}$ and $\mathbf{\Psi}$ may have a restricted domain of analyticity, their combinations may be regular on a larger domain. For instance, the product in (17) is $\mathbf{A}$, which is an entire function of all parameters. Also, $\mathbf{A}^{-1}$ is given by (19), but is also a rational function of its parameters. Therefore, its domain of analyticity is the entire complex plane, minus the zeroes of $\det(A)$. An instance of this phenomenon appears in Section 4.

References

[1] J. Abate and W. Whitt. The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems*, 10:5–88, 1992.

[2] D. Anick, D. Mitra, and M.M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell Sys. Tech. J.*, 61:1871–1894, October 1982.

[3] D. Artiges and P. Nain. Upper and lower bounds for the multiplexing of multiclass Markovian on/off sources. *Performance Evaluation*, **27&28**, pp. 673–698, 1996.

[4] J. W. Brewer, Kronecker products and matrix calculus in system theory. *IEEE Trans. on Circuit and Syst.*, **25**, 9, pp. 772-781, 1978.

[5] A.I. Elwalid and D. Mitra. Markovian arrival and service communication systems: Spectral expansions, separability and Kronecker-product forms. In W.J. Stewart, editor, *Computations in the Markov Chains*, pages 507–546. Kluwer, 1995.

[6] A.I. Elwalid, D. Mitra, and T.E. Stern. Statistical multiplexing of markov modulated sources:theory and computational algorihms. In A. Jensen and V.B. Iversen, editors, *Proc. 13th International Teletraffic Congress*, pages 495–500, Copenhagen, 1991. Elsevier Science.

[7] A.I. Elwalid, D. Mitra, and T.E. Stern. A theory of statistical multiplexing of markov modulated sources: Spectral expansions and algorihms. In W.J. Stewart, editor, *Numerical solution of Markov Chains*, 1991.

[8] A.I. Elwalid and D. Mitra. Statistical multiplexing with loss priorities in rate-based congestion control of high speed networks. *IEEE Trans. Comm.*, 42(11):2989–3002, November 1994.

[9] W. Fischer and K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18:149–171, 1992.

[10] A. Graham, *Kronecker Products and Matrix Calculus with Applications*. Ellis Horwood, Chichester, 1981.

[11] R.A. Horn and C.R Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[12] J. Keilson. *Markov Chain Models – Rarity and Exponentiality*. Springer Verlag, New York, 1979.

[13] Z. Liu, P. Nain, and D. Towsley. Exponential bounds with applications to call admission. *JACM*, 44 (2):366–394, 1997.

[14] D.M. Lucantoni, G.L. Choudhury, and W. Whitt. The transient $BMAP/G/1$ queue. *Commun. Statist.-Stochastic Models*, 10(1):145–182, 1994.

[15] D. Mitra. Stochastic theory of a fluid models of producers and consumers coupled by a buffer. *Adv. Appl. Prob.*, 20:646–676, 1988.

[16] M.F. Neuts. The fundamental period of a queue with Markov-modulated arrivals. In *Probability, Statistics and Mathematics: papers in honour of Samuel Karlin*. Academic Press, New York, 1989.

[17] T.E. Stern and A.I. Elwalid. Analysis of separable Markov-modulated rate models for information-handling systems. *Adv. Appl. Prob.*, 23:105–139, 1991.

[18] D. E. Stewart. Meschach: Matrix Computations in C. Centre for Mathematics and its Applications, School of Mathematical Sciences, Australian National University, Canberra, 1994. CMA Proceedings #32.

[19] W. Whitt. Tail probabilities with statistical multiplexing and effective bandwidth. *Telecommun. Syst.*, 3:71–107.