Analysis of Separable Markov-Modulated Rate Models for Information-Handling Systems

Thomas E. Stern; Anwar I. Elwalid

Advances in Applied Probability, Vol. 23, No. 1. (Mar., 1991), pp. 105-139.

Stable URL:
http://links.jstor.org/sici?sici=0001-8678%28199103%2923%3A1%3C105%3AAOSMRM%3E2.0.CO%3B2-3

Advances in Applied Probability is currently published by Applied Probability Trust.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/apt.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Fri Oct 13 15:45:05 2006

Adv. Appl. Prob. **23**, 105–139 (1991)
Printed in N. Ireland
© *Applied Probability Trust* 1991

ANALYSIS OF SEPARABLE MARKOV-MODULATED RATE MODELS FOR INFORMATION-HANDLING SYSTEMS

THOMAS E. STERN* AND
ANWAR I. ELWALID, **Columbia University*

Abstract

In many communication and computer systems, information arrives to a multiplexer, switch or information processor at a rate which fluctuates randomly, often with a high degree of correlation in time. The information is buffered for service (the server typically being a communication channel or processing unit) and the service rate may also vary randomly. Accurate capture of the statistical properties of these fluctuations is facilitated by modeling the arrival and service rates as superpositions of a number of independent finite state reversible Markov processes. We call such models separable Markov-modulated rate processes (MMRP).

In this work a general mathematical model for separable MMRPs is presented, focusing on Markov-modulated continuous flow models. An efficient procedure for analyzing their performance is derived. It is shown that the 'state explosion' problem typical of systems composed of a large number of subsystems, can be circumvented because of the separability property, which permits a decomposition of the equations for the equilibrium probabilities of these systems. The decomposition technique (generalizing a method proposed by Kosten) leads to a solution of the equilibrium equations expressed as a sum of terms in Kronecker product form. A key consequence of decomposition is that the computational complexity of the problem is vastly reduced for large systems. Examples are presented to illustrate the power of the solution technique.

KRONECKER PRODUCT FORM; CONTINUOUS FLOW MODEL

1. Introduction

In many communication and computer systems, information arrives to a multiplexer, switch or information processor at a rate which fluctuates randomly, often with a high degree of correlation in time. The information is buffered for service (the server typically being a communication channel or processing unit) and the service rate may also vary randomly. Accurate capture of the statistical properties of these fluctuations is facilitated by modeling the arrival and service rates as Markov processes. We call such models Markov-modulated rate processes (MMRP). In

Received 7 June 1989; revision received 27 February 1990

* Postal address for both authors: Department of Electrical Engineering and Center for Telecommunications Research, Columbia University, New York, NY 10027, USA.

Part of this work was performed when the first author was on sabbatical leave at INRIA Centre de Sophia Antipolis 06565 Valbonne, France. This work was partially supported by the National Science Foundation under grant CDR-84-21402, and the office of Naval Research under grant N00014-85-k0371.

some applications (typically, realtime communications such as voice or video) the information being processed may be realistically modeled as a Markov-modulated continuous flow process, while in others, (e.g. computer data) a point process model (e.g. Markov-modulated Poisson) may be more appropriate.

In practical design and performance evaluation applications involving MMRPs, it is often necessary to obtain detailed information regarding buffer occupancy distributions. It is frequently the tails of these distributions which contain critical design and performance information. In these cases mean values are useless. The analytical difficulties are compounded by the fact that in realistic models of 'moderate size' the modulating process will often have thousands of states. This 'state explosion' problem poses various computational difficulties: enormous memory requirements, very long computation time, and complete breakdown of numerical algorithms.

The objective of this work is to present an analysis of these systems in as general a setting as possible, using a decomposition technique which yields insight into system behavior, gives complete information on buffer occupancy distributions, and which is computationally feasible for very large systems. We focus on the continuous flow model, which has been termed a fluid flow model by Anick et al. [1] or a process with chain-dependent growth rate by Keilson and Rao [9], [10]. However, the general approach can be extended to point processes as well. The literature abounds with examples of Markov-modulated rate processes. Existing analytical techniques for such systems generally fall into three categories: (1) Exact analysis of systems of relatively small size, Fischer [6] or with a special structure [22], [21], [18], [1], [15], [13], [14], [16], (2) analysis using simplifying approximations (typically approximating the modulating process as uncorrelated), and (3) asymptotic approaches, e.g. [19], [4], [24]. Of the work just cited, [1] and [13] are of special interest here. Anick et al. [1], showed that when the MMRP consists of the superposition of a finite number of independent identical indistinguishable continuous flow sources each modulated by a two-state Markov process, and the service capacity is constant, it is possible to derive closed-form solutions for the buffer-occupancy distributions, as well as very simple asymptotic approximations. They exhibited numerical results for systems with up to 767 states. Kosten [13] showed how [1] could be extended to systems composed of a superposition of several independent subsystems, each one of which is of the type considered in [1]. In more recent work [15], carried out independently and simultaneously with ours, Mitra extended [1] to systems with variable service capacity, and obtained certain results on the eigenvalues of reversible systems which are similar to some of ours. Our work was motivated by the idea that the basic approach of [1], [13] should be generalizable to much larger classes of systems. We show herein how the main features of both [1] and [13] carry over to MMRPs in which the underlying Markov process for arrivals and/or service are superpositions of any number of independent non-identical finite state reversible

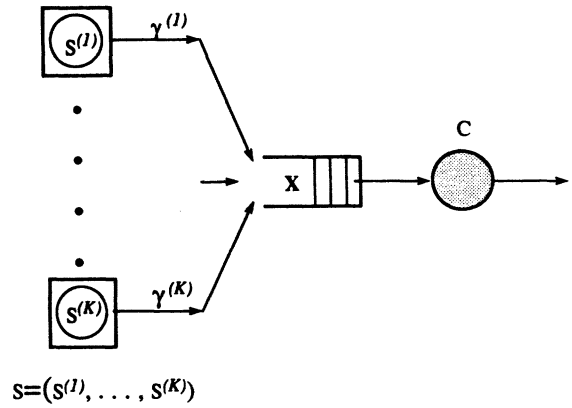

Figure 1. Markov-modulated sources model

processes of arbitrary structure. We call these *separable* MMRPs. To exhibit the relation between this work and its antecedents, we keep our notation consistent with that of [1], [13] wherever possible.

To fix ideas consider two typical problems.

1.1. *Problem* 1. Suppose that K independent information sources are served by a common communication channel of capacity C data units/s, where the data units (du) might be packets, bytes or bits (see Figure 1). The kth source generates information at a rate $\gamma^{(k)}(s^{(k)})$ data units/s, where $s^{(k)}$ is the state of a Markov process which 'modulates' the source arrival process. We assume that the data units arrive as a continuous flow while the source remains in a given state. (In another version of the problem the source arrival rate could be a Markov-modulated Poisson process, with γ being the state-dependent arrival rate.) Since the combined arrival rate may sometimes exceed C, temporarily overloading the channel, a buffer is provided, whose content X is a continuous random variable.

The equation for the evolution of the buffer occupancy X is

$$\frac{dX}{dt} = \gamma(\boldsymbol{S}(t)) - C, \qquad X > 0 \tag{1}$$

where γ is the combined arrival rate from all sources, and $\boldsymbol{S}(t)$ is the combined state $(s^{(1)}, s^{(2)}, \cdots, s^{(K)})$, of the K sources at time t. The buffer introduces a random waiting time X/C, which is a critical performance parameter in real time applications. Another important parameter is probability of buffer overflow. Thus, the analysis problem here consists of determining the complete probability distribution of the random variable X, from which these and other performance parameters of interest can be directly determined.

At this point we have not specified the functional form of the arrival rates $\gamma^{(k)}(s^{(k)})$ nor the structure of the underlying Markov process. In a typical

application, all sources might be identical, with each source modulated by an independent N-state process, representing the statistical features of a data terminal [1], or an encoded voice or video source. If $N=2$ we have the case treated in [1]. Models with $N>2$ are useful for modeling voice sources whose periods of silence have complex statistics, (for example, short silences between words/syllables and long silences while listening to the other speaker in a dialog). They are also useful as models of encoded video sources which produce information at several different rates, depending on visual activity (movement).

1.2. *Problem* 2. A communication channel of capacity C serves two classes of traffic: x and y (see Figure 2). Class x is managed by delay, that is, a buffer is provided to store all class x information that cannot be immediately served. Class y is managed by loss: any information that cannot be immediately served is discarded. This type of arrangement might be suitable, for example, in a situation where class x is computer data, for which loss cannot be tolerated, but some delay is tolerable, while class y is voice or video traffic, which must be transmitted with minimum delay, but for which some small probability of loss is acceptable. The service strategy is one of partial sharing of the communication channel, with a fraction, $(1-\alpha)$, reserved for class x and α for class y. If one class does not completely occupy its reserved portion of the channel, the residual capacity is available for the other class. Each traffic class is assumed to arrive as a continuous flow generated by a Markov-modulated source of the type described in Problem 1, with arrival rates $\gamma^{(x)}(s^{(x)})$ and $\gamma^{(y)}(s^{(y)})$ for the x and y sources respectively. The combined system state is $S=(s^{(x)}, s^{(y)})$.

Let X be the class x buffer content and L the accumulated loss of class y information. Then, according to the partial sharing rules indicated above, the equations of evolution of X and L are

$$\frac{dX}{dt}=\gamma^{(x)}(s^{(x)})+\min\left(\alpha C,\ \gamma^{(y)}(s^{(y)})\right)-C \qquad X>0, \tag{2}$$

$$\frac{dL}{dt}=\max\left(\gamma^{(y)}(s^{(y)})-\alpha C,\ 0\right) \qquad X>0, \tag{3}$$

$$=\max\left(\gamma^{(y)}(s^{(y)})+\gamma^{(x)}(s^{(x)})-C,\ 0\right) \qquad X=0. \tag{4}$$

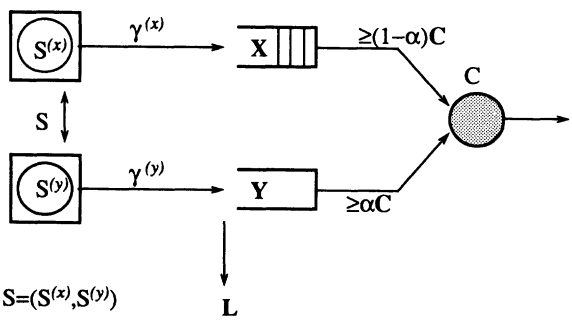

Figure 2. Example of a shared channel

Note from Equation (2) that the channel sharing strategy provides a minimum service rate $(1-\alpha)C$ to class x traffic at all times, and this rate may increase to the total channel capacity C if the class y arrival rate drops to zero. Whenever the X buffer is non-empty, class y traffic experiences a loss rate equal to the difference between its arrival rate and its reserved service rate, αC. If $X=0$, (which is only possible if the right-hand side of (2) is $\leqq 0$), the class y traffic has more capacity available to it, and the loss rate diminishes to zero as the total arrival rate decreases below C. The case $\alpha=1$ gives complete priority to class y traffic.

The arrival statistics of the two traffic classes in this problem can be quite general. Each of the classes might be modeled as a Markov-modulated source, with $N^{(x)}$ and $N^{(y)}$ states in the underlying processes for sources x and y respectively. If the system is large, it is likely that each source would be composed of the superposition of many, possibly identical smaller sources, each with a relatively small number of states.

As will be shown below, the models in the above problems are both examples of separable MMRPs. In each, the state explosion problem is evident. In Problem 1, if there are K independent sources, each with N states, the number of states for the complete system is N^K. A simplification results if all sources are identical and indistinguishable. In this case the number of states can be reduced to the number of ways of partitioning K things into N groups, which is $(K+N-1)!/K!\,(N-1)!$. Even in this simpler case, we have 5151 states for $K=100$, $N=3$. In Problem 2, each of the two traffic sources might have of the order of 100 states, yielding of the order of 10000 states for the complete system. The large numbers are of course the consequence of the fact that the number of states for a system composed of many independent subsystems is (except for special cases) equal to the product of the numbers of states in each subsystem. The complexity of the computation in these systems grows as the *cube* of the number of states. Thus, these numbers are clearly catastrophic.

In the exposition that follows, we propose a general procedure for the analysis of separable Markov-modulated rate processes. The approach is valid irrespective of the specific structure of the subsystems. Two restrictions are crucial to our objective:

- Independence of the underlying Markov processes.
- Separability of the arrival/departure rates.

An additional reversibility assumption will also be made, which is not crucial but considerably simplifies the computational problems. It is shown that, subject to these assumptions, the core of the analysis problem can be reduced to manageable size by decomposition. In order to reduce the computational problems remaining after decomposition, some simple bounds and approximations are obtained. Using this approach, the complexity of the main computation is of the order of $\sum (N^{(k)})^3$ rather than $(\prod N^{(k)})^3$, where $N^{(k)}$ is the number of states in the kth subsystem. Not only does this vastly reduce computational requirements but it also generally results in increased accuracy at critical points in the analysis. In addition and most importantly, the formalism yields considerable insight into the behavior of these

systems. In particular, it reveals the importance of the concept of reversibility in this application, and it explains the considerable power of the methods proposed in [1], [13].

In Section 2 we present the general mathematical model, derive the equations for the buffer occupancy distributions and give the form of their solutions. The general structure of the model is similar to those used in [10] and [15]. The equilibrium probabilities are expressed in terms of the solutions of a related differential equation. We demonstrate certain fundamental properties of the eigenvalues and eigenfunctions of this equation, which are the basis for the decomposition procedures which follow. Section 3 presents bounds and approximations for the equilibrium probabilities. Section 4 deals with separability of the equations of buffer occupancy distribution, yielding eigenvector solutions in the form of Kronecker (tensor) products, and the analytical procedure is applied to Problem 2 in Section 5. Section 6 presents conclusions and points out some extensions of this work which are currently underway.

2. The general model

Consider the system of Figure 3, where $\boldsymbol{S}$ is the state of a finite irreducible Markov process, C is the maximum service rate, henceforth called the 'capacity', $\gamma(\boldsymbol{S})$ is an arrival rate modulated by $\boldsymbol{S}$, $(C - \nu(\boldsymbol{S}))$ is a service rate modulated by $\boldsymbol{S}$, and X (a non-negative continuous random variable) is the buffer content. (Without loss of generality, C, γ and ν are assumed to be non-negative.) The behavior of $X(t)$ in the infinite buffer case is defined by

$$\frac{dX}{dt} = r(\boldsymbol{S}) - C, \qquad X > 0 \tag{5}$$

where

$$r(\boldsymbol{S}) = \gamma(\boldsymbol{S}) + \nu(\boldsymbol{S}).$$

Note that as far as the buffer content is concerned, it makes no difference whether the function $\nu(\boldsymbol{S})$ is considered to be a portion of capacity removed from the maximum capacity C, or arrival rate added to the rate γ. However, it is important to observe that the waiting time is different in the two cases. (If $\nu = 0$, the waiting time for arrivals to a buffer of length x is x/C. However, if $\nu \neq 0$ this waiting time is a random variable.) In what follows, we refer to $r(\boldsymbol{S})$ as the 'net arrival rate'.

Equation (5), together with the properties of the modulating process, defines the evolution in time of the state $(\boldsymbol{S}, X)$ of a bivariate Markov process, where $\boldsymbol{S}$ takes on

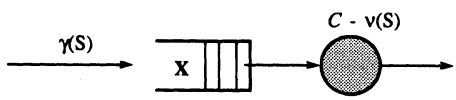

Figure 3. Model of a general Markov-modulated source/server

values $\boldsymbol{s}$ in the set

$$\mathcal{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_N\}$$

and X takes on values x in the set $[0, \infty)$ for the infinite buffer case, or $[0, X_{\max}]$ for the finite buffer case. In the latter case, (5) is valid for $0 < X < X_{\max}$.

Let

$$P(t, \boldsymbol{s}, x) = \Pr\{\boldsymbol{S}(t) = \boldsymbol{s}, X(t) \leqq x\}.$$

Because the modulating process is finite and irreducible its equilibrium probabilities

$$\pi(\boldsymbol{s}) = \lim_{t \to \infty} \Pr\{\boldsymbol{S}(t) = \boldsymbol{s}\}$$

exist, are all positive, and are independent of the initial state distribution. They are found from the properties of the modulating process alone.

An important system parameter derived from $\pi(\boldsymbol{s})$ is the mean value of the net arrival rate

$$\bar{r} = \sum_{\boldsymbol{s} \in \mathcal{S}} \pi(\boldsymbol{s}) r(\boldsymbol{s})$$

from which we obtain the system traffic intensity

$$\rho = \bar{r}/C.$$

A necessary and sufficient condition for the existence of equilibrium probabilities $F(\boldsymbol{s}, x)$ for the joint process $(\boldsymbol{S}, X)$ in the infinite buffer case is $\rho < 1$. We henceforth assume that this condition is fulfilled, in which case

$$F(\boldsymbol{s}, x) = \lim_{t \to \infty} P(t, \boldsymbol{s}, x).$$

For a system with infinite buffer

$$\pi(\boldsymbol{s}) = \lim_{x \to \infty} F(\boldsymbol{s}, x). \tag{6}$$

For a finite buffer system, there will generally be a discontinuity in the function $F(\boldsymbol{s}, x)$ at $X_{\max}$, so that

$$\pi(\boldsymbol{s}) = P_{\mathrm{f}}(\boldsymbol{s}) + F(\boldsymbol{s}, X_{\max^-}) \tag{7}$$

where

$$F(\boldsymbol{s}, X_{\max^-}) = \lim_{x \to X_{\max}} F(\boldsymbol{s}, x)$$

and the full buffer equilibrium probabilities $P_{\mathrm{f}}(\boldsymbol{s})$ are defined as

$$P_{\mathrm{f}}(\boldsymbol{s}) = \lim_{t \to \infty} \Pr\{\boldsymbol{S}(t) = \boldsymbol{s}, X(t) = X_{\max}\}.$$

As indicated previously, the modulating process is a finite state Markov process. Let $M(\boldsymbol{s}, \boldsymbol{u})$ be the transition rate from state $\boldsymbol{u}$ to state $\boldsymbol{s}$ for this process, $\boldsymbol{s} \neq \boldsymbol{u}$, and

define

$$M(s, s) = -\sum_{u \neq s} M(u, s).$$

We partition the states into an 'underload' set, $\mathcal{S}_u$, and an 'overload' set, $\mathcal{S}_o$, where

$$\mathcal{S}_u = \{s_j : r(s_j) < C\}, \qquad \mathcal{S}_o = \{s_j : r(s_j) > C\}.$$

(For reasons stated below, we assume throughout, without loss of generality, that $r(s_j) \neq C$ for all s_j.) In this section we assume further that the states are labelled so that

$$0 \leqq r(s_1) \leqq r(s_2) \leqq \cdots \leqq r(s_\omega) < C < r(s_{\omega+1}) \leqq \cdots r(s_N) \tag{8}$$

where ω and $(N - \omega)$ are respectively the cardinality of the underload and overload sets. This gives

$$\mathcal{S}_u = \{s_1, \cdots s_\omega\}, \qquad \mathcal{S}_o = \{s_{\omega+1}, \cdots s_N\}.$$

As can be seen from (1), the buffer occupancy decreases during underload and increases during overload. The Kolmogorov differential equation defining the function $P(t, s, x)$ for this system is

$$\frac{\partial P}{\partial t} + d(s)\frac{\partial P}{\partial x} = \sum_u M(s, u)P(t, u, x) \tag{9}$$

where

$$d(s) = r(s) - C. \tag{10}$$

To find the equilibrium probabilities we set $\partial P/\partial t = 0$ in (9) to obtain

$$d(s)F'(s, x) = \sum_u M(s, u)F(u, x) \tag{11}$$

where $(\cdot)'$ denotes differentiation with respect to x. Equation (11) represents a set of N linear ordinary differential equations which, with suitable boundary conditions, can be solved uniquely for F. Note that if $r(s_j) = C$ for some s_j, the set (11) is singular. In that case, the corresponding equation and the state s_j can be eliminated yielding a lower order system. We avoid this case by imposing the two strict inequalities stated in (8). Note also that if F' is set equal to zero in (11) we obtain N homogeneous linear equations for the equilibrium probabilities $\pi(s)$ of the modulating process.

Denoting the states as integers:

$$s_j = j, \qquad j = 1, 2, \cdots, N$$

and letting

$$\boldsymbol{F}(x) = [F(1, x), F(2, x), \cdots, F(N, x)]^t$$
$$\boldsymbol{R} = \text{diag}\,\{r(j)\}$$
$$\boldsymbol{D} = \text{diag}\,\{d(j)\}$$
$$\boldsymbol{M} = [M(i, j)]$$

where $(\cdot)^t$ denotes transpose, (11) becomes

$$\boldsymbol{D}\boldsymbol{F}' = \boldsymbol{M}\boldsymbol{F} \tag{12}$$

where $\boldsymbol{M}$, an $N \times N$ matrix, is the infinitesimal generator matrix for the underlying Markov process. The solution of (12) takes the form

$$\boldsymbol{F}(x) = \sum_n a_n \boldsymbol{\varphi}_n \exp(z_n x) \tag{13}$$

where

$$\begin{aligned}\boldsymbol{\varphi}_n &= (\varphi_n(s_1), \cdots, \varphi_n(s_N))^t \\ &= (\varphi_n(1), \cdots, \varphi_n(N))^t\end{aligned}$$

and each pair $\{z_n, \boldsymbol{\varphi}_n\}$ now satisfies the generalized eigenvalue problem

$$\boldsymbol{D}z\boldsymbol{\varphi} = \boldsymbol{M}\boldsymbol{\varphi}. \tag{14}$$

(While (14) could be replaced by a standard eigenvalue problem for the matrix $\boldsymbol{D}^{-1}\boldsymbol{M}$, the development that follows is based on the special properties of $\boldsymbol{D}$ and $\boldsymbol{M}$, which would be obscured in the standard eigenvalue formulation.)

Because $\boldsymbol{M}$ is the generator matrix of an irreducible Markov process, it has one and only one zero eigenvalue, (also a solution of (14)), which will be denoted z_1. Its eigenvector $\boldsymbol{\varphi}_1$, normalized so that its elements are positive and sum to unity, is

$$\boldsymbol{\varphi}_1 = \boldsymbol{\pi} = (\pi(s_1), \pi(s_2), \cdots, \pi(s_N))^t$$

the vector of equilibrium probabilities.

Up to this point, we are dealing with a standard problem and solution of (11) subject to given boundary conditions is in theory straightforward. However, it becomes intractable in most cases of practical interest simply because of its size. As mentioned earlier, the number of equations can easily range from hundreds to tens of thousands. In such cases, brute force techniques are generally useless. Computing time, memory requirements numerical accuracy, and even the specification of the equations themselves can all be serious problems. While large computers may make speed and memory requirements minor considerations, inherent problems associated with ill-conditioning and numerical instability will generally cause conventional numerical approaches to break down. The techniques which follow are designed to mitigate problems associated with large N, by exploiting the special structure of the system equations.

2.1. *Reversible processes: properties of the eigenvalues.* Henceforth we make the additional assumption that the underlying Markov process obeys the Kolmogorov condition

$$M(s, u)\pi(u) = M(u, s)\pi(s). \tag{15}$$

Equation (15) is a necessary and sufficient condition for reversibility [11]. Although this assumption is not required for separability, it greatly simplifies the subsequent development. A consequence of reversibility [11], which follows directly from condition (15), is that the infinitesimal generator matrix $\boldsymbol{M}$ in (12) is symmetrized by

the similarity transformation,

$$\tilde{\boldsymbol{M}} = \boldsymbol{E}^{-1}\boldsymbol{M}\boldsymbol{E} \tag{16}$$

where

$$\boldsymbol{E} = \text{diag}\,\{e(s_j)\}: \qquad j = 1, 2, \cdots, N$$

and

$$e(s_j) = \sqrt{\pi(s_j)}.$$

The matrix $-\tilde{\boldsymbol{M}}$ is symmetric positive semi-definite. From (16) we see that the change of variable

$$\boldsymbol{F} = \boldsymbol{E}\tilde{\boldsymbol{F}}$$

reduces (12) to the symmetric form

$$\boldsymbol{D}\tilde{\boldsymbol{F}}' = \tilde{\boldsymbol{M}}\tilde{\boldsymbol{F}}$$

and (14) to the symmetric form

$$\boldsymbol{D}z\tilde{\boldsymbol{\varphi}} = \tilde{\boldsymbol{M}}\tilde{\boldsymbol{\varphi}} \tag{17}$$

where

$$\boldsymbol{\varphi} = \boldsymbol{E}\tilde{\boldsymbol{\varphi}}.$$

We refer to $\tilde{\boldsymbol{\varphi}}$ as a symmetrized eigenvector.

If $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ are real N-vectors we denote their inner product as

$$(\boldsymbol{\varphi}, \boldsymbol{\psi}) = (\boldsymbol{\psi}, \boldsymbol{\varphi}) = \boldsymbol{\varphi}^{\mathrm{t}}\boldsymbol{\psi}.$$

One of the useful consequences of symmetry in (17) is that the eigenvalues z_i are real and a set of N real linearly independent symmetrized eigenvectors $\tilde{\boldsymbol{\varphi}}_n$ can be found which obey the following generalized orthogonality relations:

$$(\tilde{\boldsymbol{\varphi}}_i, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_j) = 0, \qquad i \neq j \tag{18}$$

$$(\tilde{\boldsymbol{\varphi}}_1, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_1) = (\bar{r} - C)(\tilde{\boldsymbol{\varphi}}_1, \tilde{\boldsymbol{\varphi}}_1) \tag{19}$$

$$(\tilde{\boldsymbol{\varphi}}_i, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_i) = (1/z_i)(\tilde{\boldsymbol{\varphi}}_i, \tilde{\boldsymbol{M}}\tilde{\boldsymbol{\varphi}}_i), \qquad i = 2, 3, \cdots, N \tag{20}$$

(see Section A.1 for proof).

Note that $(\tilde{\boldsymbol{\varphi}}_i, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_i)$ is negative for $z_i > 0$ and positive for $z_i < 0$. Henceforth we assume that the $\{\tilde{\boldsymbol{\varphi}}\}$ are normalized to unity, i.e.

$$(\tilde{\boldsymbol{\varphi}}_i, \tilde{\boldsymbol{\varphi}}_i) = 1, \qquad i = 1, 2, \ldots, N.$$

For $\tilde{\boldsymbol{\varphi}}_1$, this normalization gives

$$(\tilde{\boldsymbol{\varphi}}_1, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_1) = \bar{r} - C = C(\rho - 1)$$

and, assuming the appropriate sign for the normalization constant

$$\boldsymbol{\varphi}_1 = \boldsymbol{E}\tilde{\boldsymbol{\varphi}}_1 = \boldsymbol{\pi}.$$

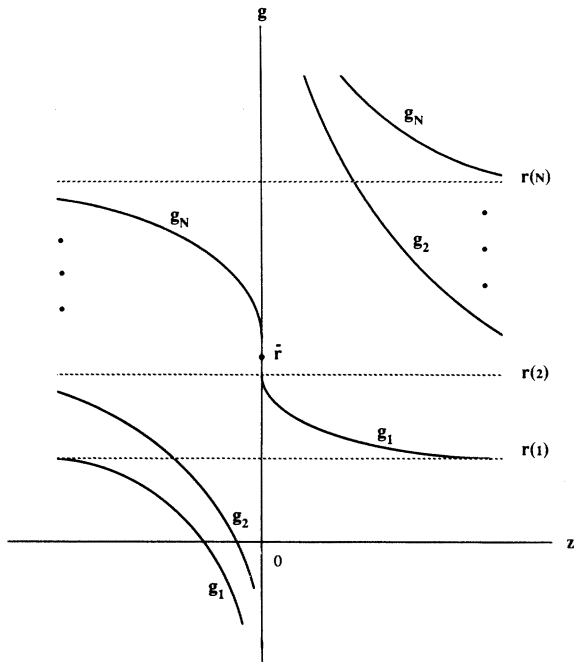

Figure 4. The functions $g_n(z)$

The decomposition techniques described in the subsequent sections are based on certain key relations among the $\{z_n\}$, the $\{\boldsymbol{\varphi}_n\}$ and the system parameters, particularly the capacity C. These relations, proved in Section A.2, are as follows.

For any $C > \bar{r}$ there exists a set of N real solutions, $\{z_n, \boldsymbol{\varphi}_n\}$, of (14), where $z_1 = 0$, z_n is the unique solution of the equation

$$C = g_n(z), \qquad n = 2, 3, \cdots, N \tag{21}$$

and the set $\{\boldsymbol{\varphi}_n\}$ is linearly independent. The function $g_n(z)$ is the nth eigenvalue of the matrix

$$\boldsymbol{A}(z) = \boldsymbol{R} - \boldsymbol{M}/z, \qquad z \neq 0. \tag{22}$$

Letting $\boldsymbol{\varphi}(n; z)$ denote an eigenvector corresponding to $g_n(z)$, the vector $\boldsymbol{\varphi}(n; z_n)$ satisfies (14), so that $\boldsymbol{\varphi}_n = \boldsymbol{\varphi}(n; z_n)$. (The form of the family $\{g_n\}$ is illustrated in Figure 4). Each function $g_n(z)$ is continuous and is continuously differentiable. At each z at which $\boldsymbol{A}(z)$ has distinct eigenvalues its derivative is

$$g_n'(z) = (\tilde{\boldsymbol{\varphi}}(n; z)\tilde{\boldsymbol{M}}\tilde{\boldsymbol{\varphi}}(n; z))/z^2 < 0, \qquad n = 1, 2, \cdots, N$$

where

$$\tilde{\boldsymbol{\varphi}}(n; z) = \boldsymbol{E}^{-1}\boldsymbol{\varphi}(n; z)$$

and

$$(\tilde{\boldsymbol{\varphi}}(n; z), \tilde{\boldsymbol{\varphi}}(n; z)) = 1.$$

The functions g_n are indexed so that their asymptotic behavior is as follows:

$$\lim_{z\to 0+} g_n(z) = +\infty, \qquad n = 2, 3, \cdots, N$$

$$\lim_{z\to 0-} g_n(z) = -\infty, \qquad n = 1, 2, \cdots, N-1$$

$$\lim_{z\to \pm\infty} g_n(z) = r(n), \qquad n = 1, 2, \cdots, N$$

$$\lim_{z\to 0+} g_1(z) = \lim_{z\to 0-} g_N(z) = \bar{r}.$$

Note from (8) that the functions $g_n(z)$ are indexed to be increasing functions of n for sufficiently large $|z|$. For future reference this will be termed 'asymptotically ascending order'. (If $\boldsymbol{A}(z)$ has distinct eigenvalues for all $z \neq 0$ then the ascending order is maintained for all z, as shown in Figure 4.)

The above development is simply an alternative way of determining solutions of (14). Thus, the generalized eigenvalue problem (14) has been replaced by two other problems: a standard eigenvalue problem involving $\boldsymbol{A}(z)$ followed by functional inversion. While this approach may appear convoluted, its usefulness will be apparent in Section 4.

It follows from the above that the $\{z_n\}$ satisfy

$$\begin{aligned} z_n &> 0, \qquad n = 2, 3, \cdots, \omega \\ z_n &< 0, \qquad n = \omega+1, \omega+2, \cdots, N. \end{aligned} \tag{23}$$

If $\boldsymbol{A}(z)$ has distinct eigenvalues for all z then the z_n are ordered as follows:

$$z_{\omega+1} \leqq \cdots \leqq z_N < 0 = z_1 < z_2 \leqq \cdots \leqq z_\omega \tag{24}$$

with ω defined in (8). Hence, for each value of C satisfying (8), the general solution of (12) given in (13), is composed of a set of $(N-\omega)$ 'stable' modes (those terms in (13) with negative z_n), $(\omega-1)$ 'unstable' modes (those with positive z_n) and the 'equilibrium' mode, $\boldsymbol{\varphi}_1$, independent of x.

2.2. *Boundary conditions.* By using known properties of the function $\boldsymbol{F}(x)$ at the boundaries $x = 0$ and $X_{\max}$ (or $+\infty$), it is possible to determine the coefficients a_n in (13). To this end it is convenient to write (13) in the more compact form

$$\boldsymbol{F}(x) = \boldsymbol{\Phi} e^{\mathbf{Z}x} \boldsymbol{a} \tag{25}$$

where

$$\boldsymbol{\Phi} = (\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \cdots, \boldsymbol{\varphi}_N) = (\boldsymbol{\Phi}_u, \boldsymbol{\Phi}_s)$$

$$\boldsymbol{\Phi}_u = (\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \cdots, \boldsymbol{\varphi}_\omega)$$

$$\boldsymbol{\Phi}_s = (\boldsymbol{\varphi}_{\omega+1}, \boldsymbol{\varphi}_{\omega+2}, \cdots, \boldsymbol{\varphi}_N)$$

$$e^{\mathbf{Z}x} = \operatorname{diag}\{\exp(z_n x) : n = 1, 2, \cdots, N\}$$

$$\boldsymbol{a} = (a_1, a_2, \cdots, a_N)^t = (a_u^t, a_s^t)^t.$$

The partitions Φ_u and Φ_s represent the eigenvectors associated with unstable and stable modes respectively, where the equilibrium mode $\varphi_1 = \pi$ is included with the unstable modes. The vector a is partitioned similarly. Throughout this section the subscripts u (unstable) and s (stable) will be used to denote partitions of a matrix consisting of the first ω columns and last $(N - \omega)$ columns respectively. Similarly, the subscripts u (underload) and o (overload) will be used to denote partitions of a matrix consisting of the first ω rows and last $(N - \omega)$ rows respectively.

Let

$$f = F(0) = \Phi a \tag{26}$$

be the equilibrium empty buffer probability vector, where

$$f = (f(s_1), f(s_2), \cdots, f(s_N))^t$$

and

$$f(s) = F(s, 0).$$

We partition f as follows:

$$f = (f_u^t, f_o^t)^t.$$

The partitions f_u and f_o correspond to the underload and overload states respectively. We note that the buffer is never empty while in overload so that

$$f_o = 0. \tag{27}$$

In the infinite buffer case, because $F(x)$ is bounded, the coefficients of the unstable modes must vanish (with the exception of $a_1 = 1$), i.e.,

$$a_u = (1, 0, 0, \cdots, 0)^t. \tag{28}$$

Substituting this condition into (26) yields

$$f = \pi + \Phi_s a_s.$$

Letting

$$\pi = (\pi_u^t, \pi_o^t)^t$$

$$\Phi_s = (\Phi_{us}^t, \Phi_{os}^t)^t$$

we have

$$\Phi_{os} a_s = -\pi_o \tag{29}$$

which is a set of $(N - \omega)$ linear equations determining the unknown coefficients of the stable modes, a_s. This approach to finding the coefficients is effective when the number of overload states is relatively small. In the opposite case, where ω is relatively small, the following alternative approach is more suitable. We write (26) in the form

$$a = \Phi^{-1} f \tag{30}$$

where the inverse of $\boldsymbol{\Phi}$ can be expressed explicitly using the orthogonality relations (20). From (26) we have

$$a_i = (\tilde{\boldsymbol{\varphi}}_i, \boldsymbol{D}\boldsymbol{E}^{-1}\boldsymbol{f})/(\tilde{\boldsymbol{\varphi}}_i, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_i) \tag{31}$$

or

$$(\Phi^{-1})_{ij} = \tilde{\varphi}_j(\boldsymbol{s}_i)d(\boldsymbol{s}_j)/e(\boldsymbol{s}_j)(\tilde{\boldsymbol{\varphi}}_i, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_i). \tag{32}$$

The last $(N-\omega)$ rows of (30) give the stable mode coefficients, $\boldsymbol{a}_s$, in terms of the empty buffer probabilities $\boldsymbol{f}$, of which those for the underload states, $\boldsymbol{f}_u$, remain to be determined. These are found using the first ω rows of (30):

$$(\boldsymbol{\Phi}^{-1})_{uu}\boldsymbol{f}_u = \boldsymbol{a}_u \tag{33}$$

where $\boldsymbol{a}_u$ is given in (28). Equation (33) represents a set of ω linear equations for the empty buffer probabilities, $\boldsymbol{f}_u$, the solution of which must be substituted into (30) to find $\boldsymbol{a}_s$.

Since solution of L linear equations requires of the order of L^3 operations, the computational complexity of the above procedures is dominated by the solution of (29) in the former case, or (33) in the latter case.

In the finite buffer case, all coefficients $\boldsymbol{a}_i$ are generally present in the solution (13). Furthermore, the coefficient of the constant term in (13) is no longer unity. Instead, we have (7) relating $\pi(\boldsymbol{s})$ to the full buffer probability $P_f(\boldsymbol{s})$.

Let

$$\boldsymbol{f}_{max} = (F(\boldsymbol{s}_1, X_{max}-), F(\boldsymbol{s}_2, X_{max}-), \cdots, F(\boldsymbol{s}_N, X_{max}-))^t$$

$$\boldsymbol{P}_f = (P_f(\boldsymbol{s}_1), P_f(\boldsymbol{s}_2), \cdots, P_f(\boldsymbol{s}_N))^t.$$

Then in addition to the boundary condition (26), which holds for all systems, we have from (26) and (7)

$$\boldsymbol{f}_{max} = \boldsymbol{\Phi} \exp(\boldsymbol{Z}X_{max})\boldsymbol{a} = \boldsymbol{\pi} - \boldsymbol{P}_f. \tag{34}$$

Noting that the buffer can never be full while in an underload state, we have

$$P_f(\boldsymbol{s}) = 0, \qquad \boldsymbol{s} \in \mathcal{S}_u.$$

These additional boundary conditions lead to $2N$ linear equations to be solved for $\boldsymbol{a}$.

In Section 3 we derive some approximations based on these relations, which greatly reduce the computational burden.

3. Coefficient bounds, approximations

In most problems involving large systems a determination of the exact probability distribution $\boldsymbol{F}(x)$ is unnecessary and/or computationally infeasible. In such cases it is useful to obtain simple approximations using only a few terms in the summation (13). We consider this problem now. Our results are illustrated with a numerical

example in Section 5. The performance parameters of interest in systems of the type described in Section 1 are typically scalar functions of $F(\boldsymbol{s}, x)$. Two functions which are particularly important are

$$G(x) = \Pr\{X > x\} = (\mathbf{1}, \boldsymbol{\pi} - \boldsymbol{F}) \tag{35}$$

where

$$\mathbf{1} = (1, 1, \cdots, 1)^t$$

and

$$H(x) = (\mathbf{1}, \boldsymbol{R}(\boldsymbol{\pi} - \boldsymbol{F}))/\bar{r}. \tag{36}$$

In most well-designed systems the size $X_{\max}$ of the (finite) buffer is chosen so that $(\mathbf{1}, \boldsymbol{P}_f)$, the probability of a full buffer, is small. In such cases $G(X_{\max}) \approx (\mathbf{1}, \boldsymbol{P}_f)$, and $G(X_{\max})$ overbounds $(\mathbf{1}, \boldsymbol{P}_f)$, so that the infinite buffer model can be used to determine the probability of buffer overflow in these cases. The function $H(x)$ defined in (36) is the fraction of flow that arrives to a buffer whose content exceeds x, or equivalently, the probability that an arriving data unit sees a buffer of content exceeding x. Thus, $H(X_{\max})$ approximates the fraction of arriving flow that is 'blocked' by a finite buffer. For a system with constant service rate ($v = 0$), $H(CT)$ represents the fraction of flow that is delayed a time exceeding T, an important performance parameter. We seek approximations to G and H which can be obtained with a minimal amount of computation.

Consider first $G(x)$. In view of the above remarks we limit the discussion to the infinite buffer case and assume that the $\{z_n\}$ are ordered as in (24) with repeated eigenvalues permitted. Equation (13) can then be rewritten in the form

$$\boldsymbol{F}(x) = \sum_{n=\omega+1}^{N} a_n \boldsymbol{\varphi}_n \exp(z_n x) + \boldsymbol{\pi}. \tag{37}$$

Substituting (35) into (37) we have

$$G(x) = \sum_{n=\omega+1}^{N} b_n \exp(z_n x) \tag{38}$$

where

$$b_n = -a_n(\mathbf{1}, \boldsymbol{\varphi}_n). \tag{39}$$

If the number of overload states is large, the computation of (38) may be both onerous and unnecessary. Instead, a good approximation of $G(x)$ may usually be obtained by truncating the series to a small number of terms. In the context of well-designed systems, mentioned above, it is the larger values of x that are of interest. For these values the dominating terms are those for large n, since the modes with lower indices generally decay rapidly with x. The truncated series retaining m terms of highest index will be denoted

$$G_m(x) = \sum_{n=N-m+1}^{N} b_n \exp(z_n x), \qquad 1 \leqq m \leqq N - \omega. \tag{40}$$

The exponents z_n in (40) are solutions of (14), which are readily computed for separable systems using the decomposition methods of Section 4. In some cases they can be expressed in closed form. The exponent z_N, in the single exponential approximation ($m = 1$), is the most important parameter in characterizing system performance, since it indicates the asymptotic behavior of all performance measures of interest for large x. (The function relating z_N to ρ in matrix-geometric systems, termed the 'caudal characteristic', is studied by Neuts [19].) Of course, if behavior for small values of x is of interest then the rapidly decaying modes are important [24], [17]. The coefficients b_n in (40) are found by substituting the expression for a_n in (31) into (39) to give

$$\begin{aligned} b_n &= (\mathbf{1}, \boldsymbol{\varphi}_n)(\tilde{\boldsymbol{\varphi}}_n, -\boldsymbol{D}\tilde{\boldsymbol{f}})/(\tilde{\boldsymbol{\varphi}}_n, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_n) \\ &= (\mathbf{1}, \boldsymbol{\varphi}_n)(\tilde{\boldsymbol{\varphi}}_n, -\boldsymbol{D}\tilde{\boldsymbol{f}})z_n/(\tilde{\boldsymbol{\varphi}}_n, \tilde{\boldsymbol{M}}\tilde{\boldsymbol{\varphi}}_n) \end{aligned} \tag{41}$$

where

$$\tilde{\boldsymbol{f}} = \boldsymbol{E}^{-1}\boldsymbol{f}.$$

In what follows we assume that the sign of the normalization constant of $\tilde{\boldsymbol{\varphi}}_n$ is chosen so that $(\mathbf{1}, \boldsymbol{\varphi}_n) \geqq 0$, which implies that $(\mathbf{1}, \boldsymbol{\varphi}_n)/(\tilde{\boldsymbol{\varphi}}_n, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_n) \geqq 0$. (Our previous assumption that $\tilde{\boldsymbol{\varphi}}_n$ is normalized to unity still leaves the sign undetermined.) Because of the dependence of b_n on the empty buffer probability vector $\boldsymbol{f}$, which is unknown, the coefficients b_n cannot be computed on a term by term basis as is the case for the exponents z_n. If it happens that the methods of Section 2.2.1 for obtaining a_n (either directly, or indirectly via $\boldsymbol{f}$) are unsuitable because of the size of the system, it is possible to obtain some good upper and lower bounds on b_n by using known constraints on $\boldsymbol{f}$. Recall from Section 2 that

$$0 \leqq f(\boldsymbol{s}) \leqq \pi(\boldsymbol{s}), \qquad \boldsymbol{s} \in \mathcal{S}_u, \tag{42}$$

$$f(\boldsymbol{s}) = 0, \qquad \boldsymbol{s} \in \mathcal{S}_o. \tag{43}$$

Also, it is easily shown (see Section A.3) that

$$\sum_{\boldsymbol{s}} -d(\boldsymbol{s})f(\boldsymbol{s}) = C(1-\rho). \tag{44}$$

Now let

$$\begin{aligned} \bar{b}_n &= \frac{(\mathbf{1}, \boldsymbol{\varphi}_n)}{(\tilde{\boldsymbol{\varphi}}_n, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_n)} \max_{\tilde{\boldsymbol{f}}} (\tilde{\boldsymbol{\varphi}}_n, -\boldsymbol{D}\tilde{\boldsymbol{f}}) \\ &= \frac{(\mathbf{1}, \boldsymbol{\varphi}_n)}{(\tilde{\boldsymbol{\varphi}}_n, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_n)} \max_{f} \sum_{\boldsymbol{s} \in \mathcal{S}_u} \frac{-\varphi_n(\boldsymbol{s})d(\boldsymbol{s})f(\boldsymbol{s})}{\pi(\boldsymbol{s})} \end{aligned} \tag{45}$$

$$\begin{aligned} \boldsymbol{b}_n &= \frac{(\mathbf{1}, \boldsymbol{\varphi}_n)}{(\tilde{\boldsymbol{\varphi}}_n, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_n)} \min_{\tilde{\boldsymbol{f}}} (\tilde{\boldsymbol{\varphi}}_n, -\boldsymbol{D}\tilde{\boldsymbol{f}}) \\ &= (\mathbf{1}, \boldsymbol{\varphi}_n) \min_{f} \sum_{\boldsymbol{s} \in \mathcal{S}_u} \frac{-\boldsymbol{\varphi}_n(\boldsymbol{s})d(\boldsymbol{s})f(\boldsymbol{s})}{\pi(\boldsymbol{s})} \end{aligned} \tag{46}$$

where the maximization (minimization) is effected subject to the constraints (42), (43), (44). The coefficients $\bar{b}_n$ and $\underline{b}_n$ are respectively upper and lower bounds on b_n. To evaluate $\bar{b}_n$, the terms in the summation in (45) are arranged so that the quantities $\varphi_n(\boldsymbol{s})/\pi(\boldsymbol{s})$ appear in descending order, and $f(\boldsymbol{s})$ is assigned its maximum possible value (which is $\leqq\pi(\boldsymbol{s})$) in each term until the condition (44) is satisfied. Similarly, the minimum $\underline{b}_n$, is attained when the maximum possible value of $f(\boldsymbol{s})$ is assigned to each term in (46) in *ascending* order of the quantities $\varphi_n(\boldsymbol{s})/\pi(\boldsymbol{s})$ until condition (44) is satisfied. A slightly looser upper bound $\hat{b}_n$, on b_n can be found more simply without reordering the summation:

$$\hat{b}_n = \frac{(\mathbf{1}, \boldsymbol{\varphi}_n)}{(\tilde{\boldsymbol{\varphi}}_n, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_n)} \sum_{\boldsymbol{s}\in\mathcal{S}_u} -\max[0, \varphi_n(\boldsymbol{s})]d(\boldsymbol{s}). \tag{47}$$

In (47) the upper bound is found by assigning $f(\boldsymbol{s})$ its maximum value in each term while ignoring the equality constraint (44). Good approximations to $G(x)$, which will overbound and underbound it for sufficiently large x, are found by replacing the coefficients b_n in (38) by $\bar{b}_n$ or $\underline{b}_n$ respectively. Thus, define

$$\bar{G}_m(x) = \sum_{n=N-m+1}^{N} \bar{b}_n \exp(z_n x), \tag{48}$$

$$\underline{G}_m(x) = \sum_{n=N-m+1}^{N} \underline{b}_n \exp(z_n x). \tag{49}$$

The advantage of using this technique is that the computation can proceed one term at a time, until a satisfactory approximation is obtained. It also explicitly exhibits the exponents, which dominate in determining system performance.

Approximations for $H(x)$ are obtained in a manner exactly paralleling those for $G(x)$ by replacing each occurrence of the quantity $(\mathbf{1}, \boldsymbol{\varphi}_n)$ in the above development by $(\mathbf{1}, \boldsymbol{R}\boldsymbol{\varphi}_n)/\bar{r}$. Note that this requires replacing the condition, $(\mathbf{1}, \boldsymbol{\varphi}_n) \geqq 0$, by $(\mathbf{1}, \boldsymbol{R}\boldsymbol{\varphi}_n) \geqq 0$, which may require reversing the signs of some of the normalization constants.

If $G(x)$ and $H(x)$ are to be evaluated at given set of values of x, the above approximation can be refined [23].

4. Separability and decomposition

As mentioned earlier, a direct application of the formalism presented in Section 2 leads to formidable computational problems when the system size is very large. However, large systems are very often composed of combinations of independent and 'separable' smaller systems. We now show that in this case the bulk of the computation can be performed at the level of each subsystem. This can produce orders of magnitude reduction in the computational burden and increased accuracy in the numerical algorithms.

Let the system (14) be composed of K independent subsystems, where $s^{(k)}$ denotes the state of the kth subsystem, and the K-tuple

$$\boldsymbol{s} = (s^{(1)}, s^{(2)}, \cdots, s^{(K)})$$

denotes the combined or 'global' system state. The kth subsystem ($k = 1, \cdots, K$) is assumed to be a $N^{(k)}$-state, irreducible, reversible Markov process whose states will be denoted by the integers

$$s^{(k)} \in \{1, 2, \cdots, N^{(k)}\} = \mathcal{S}^{(k)}$$

where

$$N = \prod_k N^{(k)}$$

so that each state $\boldsymbol{s}_j$ is an integer K-tuple. The indices will be ordered lexicographically, i.e.

$$\begin{aligned} \boldsymbol{s}_1 &= (1, 1, \cdots, 1) \\ \boldsymbol{s}_2 &= (1, 1, \cdots, 2) \\ &\vdots \\ \boldsymbol{s}_N &= (N^{(1)}, N^{(2)}, \cdots, N^{(K)}). \end{aligned}$$

Let $\boldsymbol{M}^{(k)}$ denote the matrix of transition rates for the kth subsystem. The buffer content X is now the accumulation of arrivals from the K subsystems, and the net arrival rate $r(\boldsymbol{s})$ in (1) is composed of the superposition of net arrival rates $r^{(k)}$ for each subsystem:

$$r(\boldsymbol{s}) = \sum_k r^{(k)}(s^{(k)}) \tag{50}$$

where

$$r^{(k)}(s^{(k)}) = \gamma^{(k)}(s^{(k)}) + \nu^{(k)}(s^{(k)}) \tag{51}$$

and we define a matrix of net arrival rates for each subsystem as

$$\boldsymbol{R}^{(k)} = \text{diag}\,\{r^{(k)}(1), r^{(k)}(2), \cdots, r^{(k)}(N^{(k)})\}. \tag{52}$$

It is assumed that the states of each subsystem are labelled in such a way that

$$r^{(k)}(1) \leqq r^{(k)}(2) \leqq \cdots \leqq r^{(k)}(N^{(k)}), \qquad k = 1, 2, \cdots, K.$$

The lexicographic ordering of the global states will not necessarily label them so that condition (8) is satisfied. Nevertheless, all of the development of the previous sections still holds, subject to a minor modification of notation in Section 2.2. There, in order to maintain the proper row partitioning of the various matrices, corresponding to underload and overload states, it may be necessary to relabel some of the global states.

Independence of the subsystems implies that the allowable state transitions correspond to changes in only one element of the state K-tuple at a time, with the state transition rates for the kth subsystem defined by the matrix $\boldsymbol{M}^{(k)}$. We will refer to condition (50) together with the independence assumption as 'separability conditions'. Subject to the above separability conditions it is easily verified that the matrices $\boldsymbol{D}$ and $\boldsymbol{M}$ in the global equilibrium equation (12) and the related eigenvalue equation (14) are given by

$$\boldsymbol{M} = \boldsymbol{M}^{(1)} \oplus \boldsymbol{M}^{(2)} \oplus \cdots \oplus \boldsymbol{M}^{(K)} \tag{53}$$

and

$$\boldsymbol{D} = \boldsymbol{R} - \boldsymbol{CI} \tag{54}$$

where

$$\boldsymbol{R} = \boldsymbol{R}^{(1)} \oplus \boldsymbol{R}^{(2)} \oplus \cdots \oplus \boldsymbol{R}^{(K)}. \tag{55}$$

In the above equations and those that follow $\oplus$ and $\otimes$ denote the Kronecker sum and product respectively (see Section A.4). Also, $\boldsymbol{I}$ denotes an identity matrix of appropriate size.

The Kronecker sum form of (53), and (55) suggests a decomposition method for obtaining solutions of (14) by working at the level of the subsystems. The method includes as special cases results appearing in [13] and [15].

Paralleling the development of Section 2, we assume that each subsystem is an irreducible reversible Markov process. For each k, the zero eigenvalue of $\boldsymbol{M}^{(k)}$, will be denoted z_1. Its eigenvector $\boldsymbol{\varphi}_1^{(k)}$, (assumed to be normalized so that its elements are positive and sum to unity), is

$$\boldsymbol{\pi}^{(k)} = (\pi^{(k)}(1), \pi^{(k)}(2), \cdots, \pi^{(k)}(N^{(k)}))'$$

the vector of equilibrium probabilities. The equilibrium vector $\boldsymbol{\pi}$ for the global process $\boldsymbol{M}$ is then given by

$$\boldsymbol{\pi} = \boldsymbol{\pi}^{(1)} \otimes \boldsymbol{\pi}^{(2)} \otimes \cdots \otimes \boldsymbol{\pi}^{(K)}.$$

In terms of these equilibrium probabilities, we define the mean net arrival rate for the kth subsystem as

$$\bar{r}^{(k)} = \sum_j \pi^{(k)}(j) r^{(k)}(j),$$

so that for the global system

$$\bar{r} = \sum_k \bar{r}^{(k)}.$$

Letting

$$e^{(k)}(j) = \sqrt{\pi^{(k)}(j)}, \qquad j = 1, 2, \cdots, N^{(k)}$$

$$\boldsymbol{E}^{(k)} = \text{diag}\,\{e^{(k)}(1),, \cdots, e^{(k)}(N^{(k)})\}$$

we may symmetrize the matrix $\boldsymbol{M}^{(k)}$ by the transformation

$$\tilde{\boldsymbol{M}}^{(k)} = (\boldsymbol{E}^{(k)})^{-1}\boldsymbol{M}^{(k)}\boldsymbol{E}^{(k)}. \tag{56}$$

We are now prepared to state the relations among the $\{z_n\}$, the $\{\boldsymbol{\varphi}_n\}$ and the system parameters, in terms of the characteristics of the subsystems. (See Section A.5 for proof.)

For any $C > \bar{r}$ there exists a set of N real solutions, $\{z_n, \boldsymbol{\varphi}_n\}$, of (14), where $z_1 = 0$, z_n is the unique solution of the equation

$$C = g_n(z), \qquad n = 2, 3, \cdots, N \tag{57}$$

and the set $\{\boldsymbol{\varphi}_n\}$ is linearly independent. The functions $g_n(z)$ are defined as

$$g_n(z) = \sum_{k=1}^{K} g^{(k)}(i_n^{(k)}; z), \qquad n = 1, 2, \cdots, N \tag{58}$$

where $g^{(k)}(i; z)$ is the ith eigenvalue of the matrix,

$$\boldsymbol{A}^{(k)}(z) = \boldsymbol{R}^{(k)} - \boldsymbol{M}^{(k)}/z, \qquad z \neq 0 \tag{59}$$

indexed in asymptotically ascending order, and $\boldsymbol{\varphi}^{(k)}(i; z)$ is a corresponding eigenvector. Letting $\boldsymbol{\varphi}_n^{(k)}$ denote $\boldsymbol{\varphi}^{(k)}(i_n^{(k)}; z_n)$, the global eigenvectors are expressed as

$$\boldsymbol{\varphi}_n = \boldsymbol{\varphi}_n^{(1)} \otimes \boldsymbol{\varphi}_n^{(2)} \otimes \cdots \otimes \boldsymbol{\varphi}_n^{(K)}. \tag{60}$$

The g_n are indexed in lexicographic order of the arguments $i_n^{(k)}$, i.e.

$$\begin{aligned}
g_1(z) &= g^{(1)}(1; z) + g^{(2)}(1; z) + \cdots + g^{(K)}(1; z) \\
g_2(z) &= g^{(1)}(1; z) + g^{(2)}(1; z) + \cdots + g^{(K)}(2; z) \\
&\vdots \\
g_N(z) &= g^{(1)}(N^{(1)}; z) + g^{(2)}(N^{(2)}; z) + \cdots + g^{(K)}(N^{(k)}; z).
\end{aligned}$$

Each function $g^{(k)}(i; z)$ is continuous and is continuously differentiable. At each z at which $\boldsymbol{A}^{(k)}(z)$ has distinct eigenvalues its derivative is

$$g^{(k)}(i; z)' = (\tilde{\boldsymbol{\varphi}}^{(k)}(i; z), \tilde{\boldsymbol{M}}^{(k)}\tilde{\boldsymbol{\varphi}}^{(k)}(i; z))/z^2, \qquad i = 1, 2, \cdots, N^{(k)} \tag{61}$$

where

$$\tilde{\boldsymbol{\varphi}}^{(k)}(i; z) = (\boldsymbol{E}^{(k)})^{-1}\boldsymbol{\varphi}^{(k)}(i; z).$$

The asymptotic behavior of $g^{(k)}$ is as follows:

$$\lim_{z \to 0+} g^{(k)}(i; z) = +\infty, \qquad i = 2, 3, \cdots, N^{(k)}, \tag{62}$$

$$\lim_{z \to 0-} g^{(k)}(i; z) = -\infty, \qquad i = 1, 2, \cdots, N^{(k)} - 1, \tag{63}$$

$$\lim_{z \to \pm\infty} g^{(k)}(i; z) = r^{(k)}(i), \qquad i = 1, 2, \cdots, N^{(k)}, \tag{64}$$

$$\lim_{z \to 0+} g^{(k)}(1; z) = \lim_{z \to 0-} g^{(k)}(N^{(k)}; z) = \bar{r}^{(k)}. \tag{65}$$

In the above development, each choice of index, $n>1$, corresponds to the selection of a unique set of terms $g^{(k)}(i_n^{(k)}; z)$ on the right-hand side of (58), whose sum is the function $g_n(z)$, which must be inverted to yield z_n. With z_1 defined to be zero, the remaining $N-1$ combinations of terms are accounted for as n runs from 2 to N.

The sign of the $\{z_n\}$ can be predicted in a manner analogous to that used when the system is analyzed in its global form. Identifying the K-tuple of indices $i_n^{(k)}$ with the integer K-tuple representation of a global state $\boldsymbol{s}_n$:

$$\boldsymbol{s}_n = (i_n^{(1)}, i_n^{(2)}, \cdots, i_n^{(K)})$$

it follows from the asymptotic behavior of $g_n(z)$ that the non-zero $\{z_n\}$ satisfy

$$\begin{aligned} z_n &> 0, \qquad \boldsymbol{s}_n \in \mathcal{S}_u \\ z_n &< 0, \qquad \boldsymbol{s}_n \in \mathcal{S}_o. \end{aligned} \tag{66}$$

However, an ordering of the type of (24) cannot be predicted in terms of the properties of the subsystems when decomposition is used.

To make the relation between the global system and the subsystems more explicit, note that $g^{(k)}(i_n^{(k)}; z_n)$ is an eigenvalue and $\boldsymbol{\varphi}_n^{(k)}$ an associated eigenvector for $\boldsymbol{A}^{(k)}(z_n)$, so that

$$g^{(k)}(i_n^{(k)}; z_n)\boldsymbol{\varphi}_n^{(k)} = \boldsymbol{A}^{(k)}(z_n)\boldsymbol{\varphi}_n^{(k)}. \tag{67}$$

Using (59) to rearrange terms in (67) we have

$$\boldsymbol{D}_n^{(k)} z_n \boldsymbol{\varphi}_n^{(k)} = \boldsymbol{M}^{(k)} \boldsymbol{\varphi}_n^{(k)}, \qquad k = 1, 2, \cdots, K \tag{68}$$

where

$$\boldsymbol{D}_n^{(k)} = \boldsymbol{R}^{(k)} - g^{(k)}(i_n^{(k)}; z_n)\boldsymbol{I} \tag{69}$$

and

$$\boldsymbol{D} = \boldsymbol{D}_n^{(1)} \oplus \boldsymbol{D}_n^{(2)} \oplus \cdots \oplus \boldsymbol{D}_n^{(K)}, \qquad n = 2, 3, \cdots, N.$$

Note from (69) that $g^{(k)}(i_n^{(k)}; z_n)$ plays the role of a (fictitious) channel capacity in the kth subsystem.

It can be seen from (68) that $\{z_n, \boldsymbol{\varphi}_n^{(k)}\}$ satisfies a generalized eigenvalue problem analogous to (14) at the level of the kth subsystem, so that $\boldsymbol{\varphi}_n^{(k)}$ is in fact a 'subsystem eigenvector'. The value z_n, however, is a *global* eigenvalue, common to all subsystems as well as the global system. Certain relations between global and subsystem inner products are useful for reducing the computational burden. Assuming that the symmetrized subsystem eigenvectors,

$$\tilde{\boldsymbol{\varphi}}_n^{(k)} = (\boldsymbol{E}^{(k)})^{-1} \boldsymbol{\varphi}_n^{(k)}$$

are normalized to unity:

$$(\tilde{\boldsymbol{\varphi}}_i^{(k)}, \tilde{\boldsymbol{\varphi}}_i^{(k)}) = 1, \qquad i = 1, 2, \cdots, N^{(k)}$$

they obey the following relations:

$$\boldsymbol{\varphi}_1^{(k)} = \boldsymbol{E}^{(k)}\tilde{\boldsymbol{\varphi}}_1^{(k)} = \boldsymbol{\pi}^{(k)} \tag{70}$$

$$(\tilde{\boldsymbol{\varphi}}_i^{(k)}, \boldsymbol{D}_i^{(k)}\boldsymbol{\varphi}_i^{(k)}) = z_i^{-1}(\tilde{\boldsymbol{\varphi}}_i^{(k)}, \tilde{\boldsymbol{M}}^{(k)}\tilde{\boldsymbol{\varphi}}_i^{(k)}), \qquad i = 2, 3, \cdots, N^{(k)}. \tag{71}$$

Furthermore, it follows from the properties of Kronecker sums and products (Section A.4) that the global inner product [20] can be decomposed into a sum of subsystem inner products:

$$\begin{aligned}(\tilde{\boldsymbol{\varphi}}_i, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_i) &= \sum_k (\tilde{\boldsymbol{\varphi}}_i^{(k)}, \boldsymbol{D}_i^{(k)}\tilde{\boldsymbol{\varphi}}_i^{(k)}) \\ &= \sum_k (\tilde{\boldsymbol{\varphi}}_i^{(k)}, \boldsymbol{R}_i^{(k)}\tilde{\boldsymbol{\varphi}}_i^{(k)}) - C, \qquad i = 2, 3, \cdots, N.\end{aligned} \tag{72}$$

To clarify the decomposition procedure and to compare its complexity to the 'brute force' technique of analyzing the global system directly, let us examine the computational steps in more detail. We note first that a solution of (57) for z cannot generally be found in closed form. Thus some sort of approximation procedure is required. The fact that the function g_n is well-behaved, with its derivative directly available, suggests that a second-order technique such as Newton's method is appropriate. Letting $z_n(i)$ be the ith iterate for the solution z_n of (57), Newton's method is defined as

$$z_n(i+1) = z_n(i) - \frac{g_n(z_n(i)) - C}{g_n'(z_n(i))}. \tag{73}$$

Each term $g^{(k)}(i_n^{(k)}; z_n(i))$ in $g_n(z_n(i))$ is an eigenvalue of the matrix $\boldsymbol{A}^{(k)}(z_n(i))$ of the appropriate order, as specified above. The derivative g_n' is evaluated as a sum of terms $g^{(k)}(i_n^{(k)}; z_n(i))'$, where each term is defined in (61). Evaluation of the right-hand side of (73) for a given index n and iterate $z_n(i)$ thus requires the solution of K (small) eigenvalue problems, one for each subsystem. A single pair, $\{g^{(k)}, \tilde{\boldsymbol{\varphi}}^{(k)}\}$, is selected from each subsystem to evaluate g_n and g_n'. At convergence, $z_n(i) \approx z_n$ and $\tilde{\boldsymbol{\varphi}}^{(k)}(i_n^{(k)}; z_n(i)) \approx \tilde{\boldsymbol{\varphi}}_n^{(k)}$, which are the solutions of (68). The eigenvector $\boldsymbol{\varphi}_n$ for the global system is found as the Kronecker product of the subsystem eigenvectors $\boldsymbol{\varphi}_n^{(k)}$ as prescribed in (60). To avoid the singularity of g_n at the origin, the initial iterate must be chosen of the correct sign, predicted by (66). Newton's method exhibits quadratic convergence, so that very few iterations are required given a reasonable initial trial. (In some typical numerical experiments 1–4 iterations were required for convergence.)

To compare computational complexity we consider the infinite buffer case only, and recall that $N - \omega$ modes must be calculated for an exact solution. However, as pointed out in Section 3, it is normally sufficient to compute only a few terms in the summation for $\boldsymbol{F}(x)$. In this case a fair comparison of the brute force method versus decomposition can be based on the comparative complexity of computing a single mode in (13). (This excludes from consideration the coefficients a_n, which require

very little computation if they are handled using the bounds of Section 3.) The complexity of the eigenvalue problem for the kth subsystem is of the order of $(N^{(k)})^3$, so that complexity of computation for one mode using decomposition is of the order of $\sum (N^{(k)})^3$. This is to be compared to a complexity of the order of $\prod (N^{(k)})^3$ by brute force. As an example, taking $K = 3$ and $N^{(k)} = 10$ for each subsystem, the comparison is 3×10^3 versus 10^9. Decomposition therefore results in dramatic savings, especially when N is large. It should be noted, however, that the brute force computation gives *all* modes with a complexity of $\prod (N^{(k)})^3$, while decomposition requires of the order of $N \sum (N^{(k)})^3$ operations to obtain all N modes. Thus, if all modes are to be computed the comparison is 3×10^6 versus 10^9, still very favorable to decomposition.

5. An application

The analytical and computational procedures discussed in Sections 2, 3 and 4 are best illustrated through an example. In this section we apply the decomposition technique to the model of Problem 2 (Section 1.2) and obtain various performance curves. However, our main objective here is to demonstrate the performance of the *algorithms* rather than the performance of the system. In computing the performance curves we compare exact methods to approximation and bounding techniques, and thereby obtain information on the tightness of the bounds.

The buffer dynamics for the model of Problem 2 are defined in (2). To proceed with the analysis it is necessary to specify the statistical characteristics of the traffic sources. Recall that there are two traffic classes: class x, which is buffered, and class y, which is blocked. It will be assumed that each class is made up of a collection of independent sources which alternate randomly between active and inactive periods, where the activity process for each source is generated by a two-state Markov process.

Let

N_x = the number of class x sources
N_y = the number of class y sources
i_x = the number of active class x sources
i_y = the number of active class y sources
γ_x = the traffic generated by an active source in class x (data units/s)
γ_y = the traffic generated by an active source in class y
λ_x = the transition rate from the inactive to the active state for a source in class x (s^{-1})
λ_y = the transition rate from the inactive to the active state for a source in class y
μ_x = the transition rate from the active to the inactive state for a source in class x
μ_y = the transition rate from the active to the inactive state for a source in class y

This system can be separated into two independent subsystems where, using the

terminology of Section 4, the system state is defined as

$$s = (s^{(1)}, s^{(2)}).$$

The subsystem states are

$$s^{(1)} = i_x + 1, \qquad s^{(2)} = i_y + 1$$

with

$$N^{(1)} = N_x + 1, \qquad N^{(2)} = N_y + 1, \qquad N = (N_x + 1)(N_y + 1).$$

Equation (2) then translates to the form of (5), where

$$r(s) = r^{(1)}(s^{(1)}) + r^{(2)}(s^{(2)})$$

and

$$r^{(1)}(s^{(1)}) = \gamma^{(1)}(s^{(1)}) = (s^{(1)} - 1)\gamma_x, \qquad r^{(2)}(s^{(2)}) = \min[\alpha C, (s^{(2)} - 1)\gamma_y].$$

Equations (3) and (4) become

$$\frac{dL}{dt} = \max[(s^{(2)} - 1)\gamma_y - \alpha C, 0], \qquad X > 0$$

$$= \max[(s^{(2)} - 1)\gamma_y + (s^{(1)} - 1)\gamma_x - C, 0], \qquad X = 0.$$

Letting

$$\lambda^{(1)} = \lambda_x, \qquad \lambda^{(2)} = \lambda_y, \qquad \mu^{(1)} = \mu_x, \qquad \mu^{(2)} = \mu_y,$$

the non-zero elements of the transition rate matrices for the two modulating processes (which are tridiagonal) are

$$M^{(k)}(i, i+1) = \mu^{(k)} i$$

$$M^{(k)}(i+1, i) = \lambda^{(k)}(N^{(k)} - i), \qquad k = 1, 2, \quad \text{and} \quad i = 1, 2, \cdots, N^{(k)} - 1$$

$$M^{(k)}(i, i) = \lambda^{(k)}(i - N^{(k)}) + (1 - i)\mu, \qquad k = 1, 2, \quad \text{and} \quad i = 1, 2, \cdots, N^{(k)}.$$

It is worth noting that the first subsystem (but not the second) is itself separable. It is, in fact, a system of the form analyzed in [1], whose eigenvalues and eigenvectors can be expressed in closed form.

The above parameters completely define the model in a form suitable for application of the decomposition methods of Section 4. A program was written in FORTRAN to execute the various steps in the algorithm for computing the equilibrium probabilities and evaluating system performance. Quantities of interest in this application are the functions G and H (defined in Section 3). Recall that the function $G(x)$ represents the probability that the class x buffer occupancy exceeds a value x and $H(x)$ represents the fraction of class x traffic that arrives to a buffer whose occupancy exceeds x. Also of interest is the fractional loss f_y for the traffic in class y. This quantity is defined as

$$f_y = \frac{E\{dL/dt\}}{N_y \lambda_y \gamma_y / (\lambda_y + \mu_y)}$$

where $E\{.\}$ denotes expectation. The parameter α partitions the channel, reserving

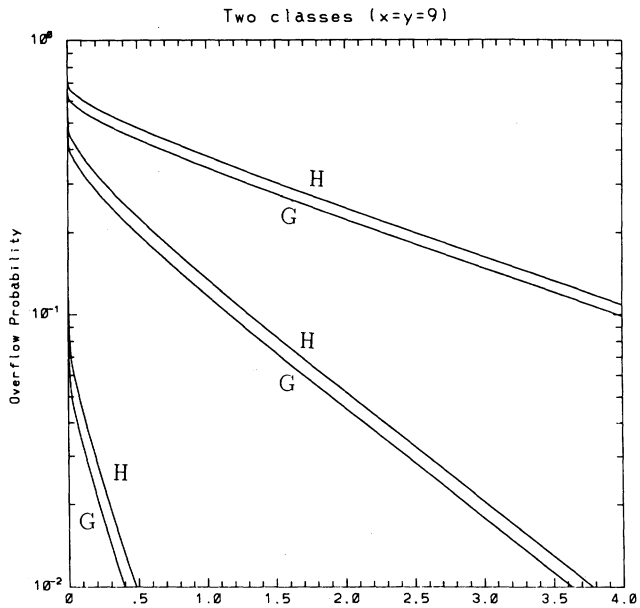

Figure 5. Probabilities of overflow G and H vs. buffer size for $N_x = N_y = 9$. For $\alpha = 1$, 0·5, and 0·3, the traffic intensity ρ is set to 0·91, 0·87, and 0·74, respectively

a fraction $(1-\alpha)$ of the capacity for class x and the rest for class y. Thus it is of interest to explore the tradeoff between delay for class x and blocking for class y as a function of α. This is done through a family of curves in Figures 5 and 6, with system parameters

$$C = 9{\cdot}9$$
$$N^{(1)} = N^{(2)} = 10$$
$$\lambda^{(1)} = \lambda^{(2)} = \mu^{(1)} = \mu^{(2)} = \gamma^{(1)} = \gamma^{(2)} = 1.$$

Note that as α increases from 0 (complete priority to class x) to 1 (complete priority to class y), the traffic intensity in the class x buffer increases, thereby increasing the buffer content (Figure 5), but reducing the fraction of blocked class y traffic (Figure 6). Because the higher arrival rates are positively correlated with higher buffer occupancies, the values of H in Figure 5 are consistently larger than G. The quantities $H(CT)$ and $H((1-\alpha)CT)$ give lower and upper bounds respectively on the probability that buffering delay exceeds T. However, exact calculation of buffering delay requires an additional (and difficult) analytical step, since the capacity available to the class x traffic is a random variable.

The bulk of the computation in this problem is the determination of the equilibrium probabilities. Table 1 shows the essential quantities involved in this computation for the case $\alpha = 0{\cdot}3$ and $C = 9{\cdot}9$ with the remaining system parameters as stated above. The (class x) traffic intensity is $\rho = 0{\cdot}744$, and the fractional loss for

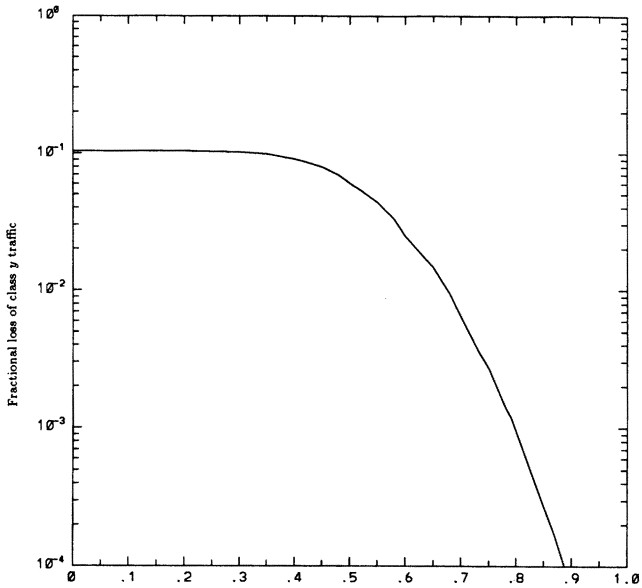

Figure 6. Fractional loss of class y traffic vs α for $N_x = N_y = 9$

TABLE 1

n	z_n	$\underline{b}_n$	b_n	$\bar{b}_n$	$\hat{b}_n$
100	−0·3167E + 01	0·2990E − 01	0·3125E − 01	0·3273E − 01	0·3273E − 01
99	−0·4776E + 01	0·6575E − 04	0·1545E − 03	0·2547E − 03	0·3981E − 03
98	−0·6338E + 01	0·4122E − 05	0·3696E − 04	0·5291E − 04	0·8742E − 04
92	−0·7801E + 01	0·2128E − 01	0·2294E − 01	0·2440E − 01	0·2440E − 01
97	−0·7893E + 01	0·1574E − 05	0·1463E − 04	0·2145E − 04	0·3191E − 04
96	−0·9436E + 01	0·5655E − 06	0·7285E − 05	0·1073E − 04	0·1564E − 04
91	−0·1046E + 02	0·1190E − 03	0·3324E − 03	0·6176E − 03	0·8334E − 03
95	−0·1095E + 02	0·1210E − 06	0·3995E − 05	0·6267E − 05	0·8163E − 05
94	−0·1240E + 02	−0·5361E − 07	0·1938E − 05	0·3090E − 05	0·4307E − 05
90	−0·1328E + 02	−0·1011E − 04	0·1066E − 03	0·1766E − 03	0·2745E − 03
89	−0·1616E + 02	−0·6972E − 05	0·4864E − 04	0·8426E − 04	0·1194E − 03
93	−0·1624E + 02	0·1866E − 05	0·3162E − 04	0·3907E − 04	0·4303E − 04
88	−0·1906E + 02	−0·4133E − 05	0·2482E − 04	0·4321E − 04	0·6674E − 04
87	−0·2189E + 02	−0·3189E − 05	0·1234E − 04	0·2630E − 04	0·3752E − 04
86	−0·2456E + 02	−0·2589E − 05	0·4521E − 05	0·1069E − 04	0·1884E − 04
83	−0·1310E + 03	0·3903E − 01	0·4529E − 01	0·4969E − 01	0·4969E − 01
82	−0·1713E + 03	0·1592E − 03	0·1176E − 02	0·2834E − 02	0·3233E − 02
85	−0·1783E + 03	0·3669E − 05	0·5295E − 03	0·6583E − 03	0·6583E − 03
84	−0·1783E + 03	−0·1613E − 07	0·1336E − 04	0·1831E − 04	0·1831E − 04
81	−0·2134E + 03	−0·3332E − 03	0·4104E − 03	0·9463E − 03	0·1334E − 02
80	−0·2564E + 03	−0·2121E − 03	0·1986E − 03	0·5018E − 03	0·7040E − 03
79	−0·2994E + 03	−0·1180E − 03	0·1051E − 03	0·2651E − 03	0·4512E − 03
78	−0·3414E + 03	−0·8606E − 04	0·5279E − 04	0·2056E − 03	0·3998E − 03
77	−0·3765E + 03	−0·4833E − 04	0·1871E − 04	0·8964E − 04	0·1225E − 03

class y traffic is $f_y = 0{\cdot}102$. In this case there are a total of 100 states of which 24 are overload states, $(\omega = 76)$, giving the same number of stable modes. Their eigenvalues are shown in the second column of Table 1, using the numbering convention of Section 2. (To exhibit the dominant eigenvalues, the entries are listed in decreasing order of z_n. Note that this does not correspond to decreasing order of n.) The remaining columns compare the exact values of the coefficients b_n in the expression for $G(x)$ (Equation (38)) with the bounds $\underline{b}_n$, $\bar{b}_n$, $\hat{b}_n$ defined in (46), (45), and (47) respectively.

Note that there are only three coefficients, b_{83}, b_{92}, b_{100}, of significant magnitude, and the bounds on each are fairly tight. While the bounds on the smaller coefficients are poor in a relative sense, they are still quite good in an absolute sense. One can use the coefficient bounds $\underline{b}_n$ and $\bar{b}_n$ of Table 1 to obtain (approximate) bounds on $G(x)$. Figure 7 shows comparisons of $\bar{G}_{10}(x)$ and $\underline{G}_{10}(x)$ as defined in (48) and (49) respectively, with the exact value $G(x)$ for the system of Figure 5. Note that these expressions bound G_{10} in (40) using upper and lower bounds on the coefficients of each exponential. Thus Figure 7 illustrates the combined effects of using a subset of the modes and of bounding their coefficients. The expressions do not require the calculation of any system modes beyond those contained in the summation. Furthermore, they do not require solution of any linear equations for the coefficients.

To illustrate the computational techniques in the setting of a larger system we treat the case $N^x = 29$, $N^y = 29$ (900 states) in Figure 8. This corresponds to a tripling of

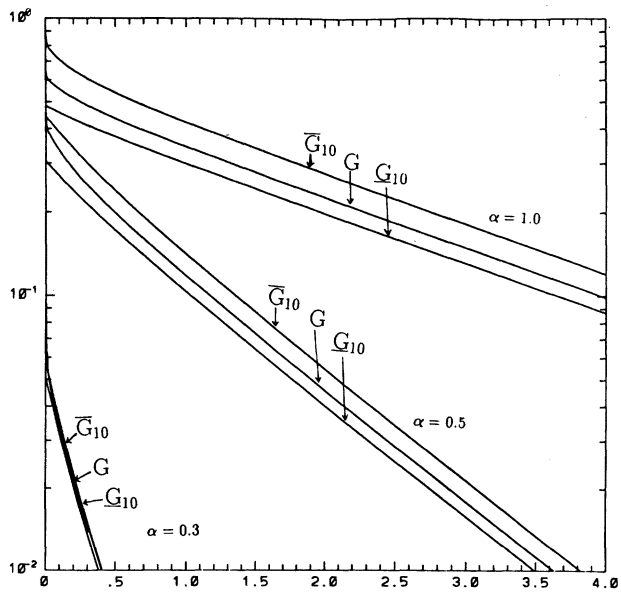

Figure 7. Probability of overflow G and its approximations $\bar{G}_{10}$ and $\underline{G}_{10}$ vs. buffer size for $N_x = N_y = 9$. For $\alpha = 1$, 0·5, and 0·3, the traffic intensity ρ is set to 0·91, 0·87, and 0·74, respectively

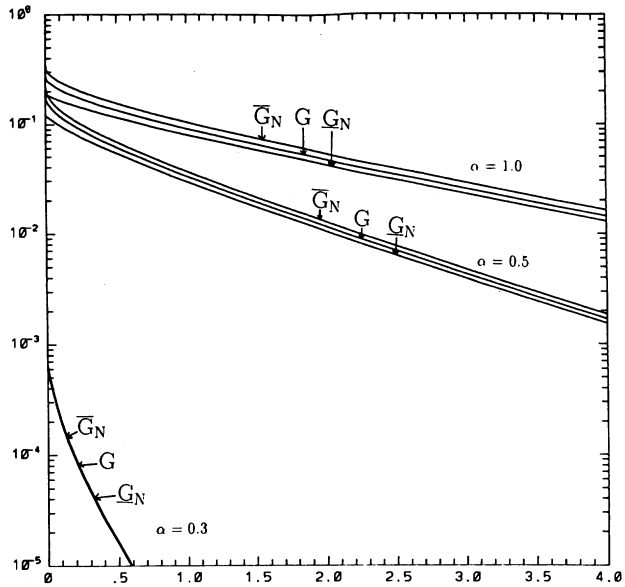

Figure 8. Probability of overflow G and its approximations $\bar{G}_N$ and $\underline{G}_N$ vs buffer size for $N_x = N_y = 29$. For $\alpha = 1$, 0·5, and 0·3, the traffic intensity ρ is set to 0·91, 0·87, and 0·74, respectively

the size of the original system, with an increase of a factor of 9 in the number of states. The capacity C is adjusted to maintain approximately the same traffic intensities as in the smaller system of Figure 7. The curves show $\bar{G}_N(x)$ and $\underline{G}_N(x)$ as defined in (48) and (49) for comparison with the exact value of $G(x)$. Note that this corresponds to using all of the modes with exact coefficients replaced by upper and lower bounds respectively. The results give very tight bounds on the exact solution, especially in the case $\alpha = 0{\cdot}3$. It can be seen that the buffer occupancy for this system decreases markedly as system size increases while keeping the traffic intensity fixed. One would expect this as a consequence of the laws of large numbers.

We observe from these examples that the algorithm is numerically stable, performs efficiently (computation time was relatively short), and the coefficient bounds are well suited to engineering approximations.

6. Conclusions

This work has examined a broad class of buffered information-handling systems, modeled as continuous flow processes modulated by an underlying reversible Markov process. We showed that if the modulating process is separable, considerable reduction in computational complexity is possible using a decomposition method. This avoids problems of 'state explosion' and numerical inaccuracy in very large systems. Further reductions in computational complexity were shown to be possible using simple bounds on coefficients in the expression for the buffer

occupancy distribution. The behavior of the various computational algorithms was illustrated in an example wherein the performance of a multiplexer involving channel sharing was computed. While the largest system illustrated have had 900 states, our computational experiments give us confidence that much larger systems can be handled.

We have been extending this work in a number of directions. As mentioned in Section 1, when the system represented by (5) is composed of the superposition of a large number of identical subsystems the total number of states can be significantly reduced by aggregating the states of the global system. Work is in progress on adapting the results presented here to this special case. Work is also partly completed on carrying over the essential features of these results to the case of point process models: i.e., Markov modulated Poisson processes [5]. In this case, the models are similar to those treated in [18], with the additional features of reversibility and separability. An objective of this effort is to show how certain computational problems associated with matrix-geometric methods can be avoided using decomposition.

Appendix

A.1. Proof of Equations (18)–(20). Because the matrices $\boldsymbol{D}$ and $\tilde{\boldsymbol{M}}$ are both Hermitian, it can be shown that (17) has real eigenvalues and a set of N (real) linearly independent eigenvectors. The proof is based on simultaneous diagonalization of two quadratic forms. (See, for example, [7], p. 106 or the appendix of [15].)

To prove (18) consider two pairs $\{z_j, \tilde{\boldsymbol{\varphi}}_i\}$ and $\{z_j, \tilde{\boldsymbol{\varphi}}_j\}$ satisfying (17). Taking inner products on each side of (17) we have

$$z_j(\tilde{\boldsymbol{\varphi}}_i, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_j) = (\tilde{\boldsymbol{\varphi}}_i, \tilde{\boldsymbol{M}}\tilde{\boldsymbol{\varphi}}_j) = (\tilde{\boldsymbol{\varphi}}_j, \tilde{\boldsymbol{M}}\tilde{\boldsymbol{\varphi}}_i) = z_i(\tilde{\boldsymbol{\varphi}}_j, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_i) = z_i(\tilde{\boldsymbol{\varphi}}_i, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_j)$$

or

$$(z_i - z_j)(\tilde{\boldsymbol{\varphi}}_i, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_j) = 0,$$

which implies (18) when the two eigenvalues are distinct. In the case of a set of identical eigenvalues, any basis for the subspace spanned by their eigenvectors can be orthogonalized with respect to the matrix $\boldsymbol{D}$ using the Gram–Schmidt orthogonalization procedure.

To prove (19) we expand the left side to

$$(\tilde{\boldsymbol{\varphi}}_1, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_1) = (\tilde{\boldsymbol{\varphi}}_1, \boldsymbol{R}\tilde{\boldsymbol{\varphi}}_1) - C(\tilde{\boldsymbol{\varphi}}_1, \tilde{\boldsymbol{\varphi}}_1).$$

From the definition of $\boldsymbol{E}$ and the fact that $\boldsymbol{\varphi}_1 \sim \boldsymbol{\pi}$ we have

$$(\tilde{\boldsymbol{\varphi}}_1, \boldsymbol{R}\tilde{\boldsymbol{\varphi}}_1) = \bar{r}(\tilde{\boldsymbol{\varphi}}_1, \tilde{\boldsymbol{\varphi}}_1),$$

from which (19) follows directly.

Finally, to prove (20) we take inner products on each side of (17) using the same eigenvectors, to give

$$z_i(\tilde{\boldsymbol{\varphi}}_i, \boldsymbol{D}\tilde{\boldsymbol{\varphi}}_i) = (\tilde{\boldsymbol{\varphi}}_i, \tilde{\boldsymbol{M}}\tilde{\boldsymbol{\varphi}}_i)$$

from which (20) follows.

A.2. Properties of the eigenvalues. To prove the properties of the eigenvalues stated in Section 2.1, we need to relate the eigenvalues and eigenvectors of (14) to those of the matrix $\boldsymbol{A}(z)$ defined in (22). We first prove that for $z \neq 0$, $\det[\boldsymbol{D}z - \boldsymbol{M}] = 0$ if and only if $\det[C\boldsymbol{I} - \boldsymbol{A}(z)] = 0$. The equation satisfied by the eigenvalues $\{z_n\}$ can be written

$$\det[\boldsymbol{D}z - \boldsymbol{M}] = \det[\boldsymbol{R}z - Cz\boldsymbol{I} - \boldsymbol{M}] = (-z)^N \det[C\boldsymbol{I} - \boldsymbol{A}(z)] = 0.$$

Thus, there must be exactly $(N-1)$ non-zero values of z for which C is an eigenvalue of $\boldsymbol{A}(z)$, corresponding to the $(N-1)$ non-zero eigenvalues of (14).

Now, to exhibit the correspondence between eigenvectors, let $g_n(z)$ denote the nth eigenvalue of $\boldsymbol{A}(z)$, with eigenvector $\boldsymbol{\varphi}(n; z)$; i.e.

$$\boldsymbol{A}(z)\boldsymbol{\varphi}(n; z) = g_n(z)\boldsymbol{\varphi}(n; z) \qquad n = 1, 2, \cdots, N, \quad z \neq 0, \tag{74}$$

which can be rearranged to

$$[\boldsymbol{R} - g_n(z)\boldsymbol{I}]z\boldsymbol{\varphi}(n; z) = \boldsymbol{M}\boldsymbol{\varphi}(n; z). \tag{75}$$

Suppose there exists a non-zero z_n such that

$$C = g_n(z_n). \tag{76}$$

Then (75) becomes

$$\boldsymbol{D}z_n\boldsymbol{\varphi}(n; z_n) = \boldsymbol{M}\boldsymbol{\varphi}(n; z_n)$$

implying that z_n is a solution of (14) with eigenvector

$$\boldsymbol{\varphi}_n = \boldsymbol{\varphi}(n; z_n).$$

We now show that the functions $g_n(z)$ can be defined so that for $n = 2, \cdots, N$, (76) has a unique solution z_n for any $C > \bar{r}$ and C not equal to any of the $r(n)$. The following result will be used. (See [20], p. 44.)

Let $\boldsymbol{H}(z)$ be an $N \times N$ Hermitian matrix, continuously differentiable with respect to the parameter z, and let its derivative $\boldsymbol{H}'(z)$ be Hermitian. Then there exist real functions $\mu(n; z)$, $n = 1, 2, \cdots, N$, continuously differentiable in z such that

$$\boldsymbol{H}\boldsymbol{\xi}(n; z) = \mu(n; z)\boldsymbol{\xi}(n; z)$$

where $\{\boldsymbol{\xi}(n; z)\}$ is a properly chosen orthonormal system of vector functions. At every point at which $\boldsymbol{H}(z)$ has distinct eigenvalues, their derivatives are given by

$$\mu'(n; z) = (\boldsymbol{\xi}(n; z), \boldsymbol{H}'(z)\boldsymbol{\xi}(n; z)). \tag{77}$$

In the case at hand, we let

$$\boldsymbol{H}(z) = z\tilde{\boldsymbol{A}}(z) = z\boldsymbol{E}^{-1}\boldsymbol{A}\boldsymbol{E} = \boldsymbol{R}z - \tilde{\boldsymbol{M}}$$

which is Hermitian, so that $\mu(n; z)$ is an eigenvalue of $z\tilde{\boldsymbol{A}}(z)$ for $n = 1, 2, \cdots, N$, $\{\boldsymbol{\xi}(n; z)\}$ is an orthonormal set of eigenvectors of $z\tilde{\boldsymbol{A}}(z)$ as well as $\tilde{\boldsymbol{A}}(z)$ and $\mu'(n; z) = (\boldsymbol{\xi}(n; z), \boldsymbol{R}\boldsymbol{\xi}(n; z))$.

To define an ordering of the $\{\mu(n; z)\}$, let

$$0 = \lambda_1 < \lambda_2 \leqq \cdots \leqq \lambda_N$$

be the eigenvalues of $-\boldsymbol{M}$, and let $\mu(n; 0) = \lambda_n$. First, consider the case where $z\tilde{\boldsymbol{A}}(z)$ has distinct eigenvalues for all z. Then by continuity, the ascending order of the $\{\mu(n; z)\}$ must be maintained for all z. Now let

$$g_n(z) = \begin{cases} \mu(n; z)/z, & z > 0, \\ \mu(N - n + 1; z)/z, & z < 0. \end{cases} \tag{78}$$

For each $z \neq 0$, $g_n(z)$ is the nth eigenvalue of $\tilde{\boldsymbol{A}}(z)$ (in ascending order), with eigenvector

$$\tilde{\boldsymbol{\varphi}}(n; z) = \boldsymbol{E}^{-1}\boldsymbol{\varphi}(n; z) = \begin{cases} \boldsymbol{\xi}(n; z), & z > 0, \\ \boldsymbol{\xi}(N - n + 1; z), & z < 0. \end{cases} \tag{79}$$

It follows that each $g_n(z)$ is continuously differentiable for all $z \neq 0$ with

$$g_n'(z) = (\tilde{\boldsymbol{\varphi}}(n; z), \tilde{\boldsymbol{M}}\tilde{\boldsymbol{\varphi}}(n; z))/z^2 \leqq 0. \tag{80}$$

Since $\tilde{\boldsymbol{M}}$ is negative semi-definite with null-space spanned by $\tilde{\boldsymbol{\varphi}}_1$, $g_n'(z) = 0$ if and only if $\tilde{\boldsymbol{\varphi}}(n; z) = c\tilde{\boldsymbol{\varphi}}_1$ for some constant c. But this implies

$$\tilde{\boldsymbol{A}}\tilde{\boldsymbol{\varphi}}_1 = (\boldsymbol{R} - \tilde{\boldsymbol{M}}/z)\tilde{\boldsymbol{\varphi}}_1 = \boldsymbol{R}\tilde{\boldsymbol{\varphi}}_1 = g_n(z)\tilde{\boldsymbol{\varphi}}_1.$$

Since all elements of $\tilde{\boldsymbol{\varphi}}_1$ are non-zero, this relation can only hold if $\boldsymbol{R}$ is a scalar matrix, contradicting the inequalities in (8). Thus we conclude that the inequality in (80) is strict for all $z \neq 0$.

It follows directly from the limiting values of $\boldsymbol{A}(z)$ at 0 and ∞, and the continuity and ordering of the set $\{g_n(z)\}$ that

$$\lim_{z \to 0+} g_n(z) = +\infty, \qquad n = 2, 3, \cdots, N,$$

$$\lim_{z \to 0-} g_n(z) = -\infty, \qquad n = 1, 2, \cdots, N - 1,$$

$$\lim_{z \to \pm\infty} g_n(z) = r(n), \qquad n = 1, 2, \cdots, N.$$

To find the remaining limits we use l'Hospital's rule to give

$$\lim_{z \to 0+} g_1(z) = \lim_{z \to 0-} g_N(z) = \mu'(1; 0) = (\tilde{\boldsymbol{\varphi}}(1; 0), \boldsymbol{R}\tilde{\boldsymbol{\varphi}}(1; 0)) = (\tilde{\boldsymbol{\varphi}}_1, \boldsymbol{R}\tilde{\boldsymbol{\varphi}}_1) = \bar{r}.$$

The above development holds with minor variations for the case where $z\tilde{\boldsymbol{A}}(z)$ has repeated eigenvalues for some or all values of z. In this case, some of the curves $\mu(n; z)$ may intersect, or coincide over the entire range of z. If they intersect, the ascending order of the $\mu(n; z)$ may not be maintained for all z, which means that with the mapping from μ to g defined in (78), the limits at ∞ of $g_n(z)$ defined in (81) may occur in a different order; e.g.

$$\lim_{z\to+\infty} g_n(z) = r(i), \qquad i \neq n.$$

To maintain the correct order of the limiting values, the mapping in (78) must then be replaced by

$$g_n(z) = \begin{cases} \mu(i(n); z)/z, & z > 0, \\ \mu(j(n); z)/z, & z < 0 \end{cases} \tag{81}$$

where $i(n)$ and $j(n)$ are appropriate permutations of the integers, $1, 2, \cdots, N$. Continuous differentiability of the eigenvalues still holds in this case, but one must be more careful about defining the derivatives. (See [20].)

The properties of the family $\{g_n(z)\}$ derived above and illustrated in Figure 4 show that for $n = 2, 3, \cdots, N-1$, $g_n(z)$ is a one–one mapping of $(0, \infty)$ onto $(r(n), \infty)$ and $(-\infty, 0)$ onto $(-\infty, r(n))$, while $g_N(z)$ is a one–one mapping of $(0, \infty)$ onto $(r(N), \infty)$ and $(-\infty, 0)$ onto $(\bar{r}, r(N))$. We therefore conclude that the equation

$$C = g_n(z)$$

has a unique solution, z_n, for any $C > \bar{r}$ and not equal to any of the $r(n)$. Furthermore, with ω defined in (8), it is clear from the form of $g_n(z)$ that

$$\begin{aligned} z_n &> 0, \qquad 2 \leqq n \leqq \omega \\ &< 0, \qquad \omega < n \leqq N. \end{aligned}$$

A.3. Proof of Equation (44). Substituting (19) into (31) for the case $i = 1$, we have

$$1 = (\tilde{\boldsymbol{\varphi}}_1, \boldsymbol{D}\boldsymbol{E}^{-1}\boldsymbol{f})/(\bar{r} - C)$$

or

$$\sum_s \tilde{\varphi}_1(s)e^{-1}(s)d(s)f(s) = \sum_s -d(s)f(s) = C(1-\rho).$$

A.4. Kronecker product and sum: definitions and properties. The Kronecker product $\boldsymbol{A} \otimes \boldsymbol{B}$ of the matrix $\boldsymbol{A}$ of dimension $p \times q$ and the matrix $\boldsymbol{B}$ of dimension $m \times n$ is the matrix of dimension $pm \times qn$ obtained by replacing each element a_{ij} of the matrix $\boldsymbol{A}$ by the full matrix $a_{ij}\boldsymbol{B}$. (See for example, [2].)

The Kronecker sum of $\boldsymbol{A}$ $(n \times n)$ and $\boldsymbol{B}$ $(m \times m)$ denoted by $\boldsymbol{A} \oplus \boldsymbol{B}$ is defined as

$$\boldsymbol{A} \oplus \boldsymbol{B} = \boldsymbol{A} \otimes \boldsymbol{I}_m + \boldsymbol{I}_n \otimes \boldsymbol{B}$$

where I_m and I_n are the identity matrices of order m and n respectively. The operation $\otimes$ is associative but not commutative, and the same holds true for $\oplus$.

The following properties, which are proven in [3], [2], [8], are used in this paper. All matrices (vectors) are assumed to be of appropriate order.

1. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$
2. $(A \otimes B, C \otimes D) = (A, C)(B, D)$
3. Let $\lambda_1, \lambda_2, \cdots, \lambda_n$ be the eigenvalues of the matrix A with corresponding eigenvectors $\alpha_1, \alpha_2, \cdots, \alpha_n$ and let $\mu_1, \mu_2, \cdots, \mu_m$ be the eigenvalues of B with corresponding eigenvectors $\beta_1, \beta_2, \cdots, \beta_m$ then the eigenvalues of $A \oplus B$ are the nm sums $\lambda_i + \mu_j$ with corresponding eigenvectors $\alpha_i \otimes \beta_j$ $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m$.

A.5. Decomposition. In this section we prove relations (57)–(66). It suffices to show that the relations stated in (21) and (22) and proved in Section A.2 are expressible in the decomposed form of (58), (59), and (60). Recall from Section A.2 that for non-zero z_n, $\{z_n, \varphi_n\}$ is a solution of (14) if and only if it satisfies

$$C\varphi_n = A(z_n)\varphi_n. \tag{82}$$

Now the separability conditions imply that

$$A(z) = A^{(1)}(z) \oplus A^{(2)}(z) \oplus \cdots \oplus A^{(K)}(z). \tag{83}$$

Let $g(z)$ denote an eigenvalue of $A(z)$, with eigenvector $\varphi(z)$. Then from Property 3 in Section A.4 it follows that

$$g(z) = \sum_k g^{(k)}(i^{(k)}; z) \tag{84}$$

for some set of integers $i^{(1)}, i^{(2)}, \cdots, i^{(K)}$, where $g^{(k)}(i^{(k)}; z)$ represents the $i^{(k)}$th eigenvalue of $A^{(k)}(z)$ in asymptotically ascending order. Furthermore,

$$\varphi(z) = \varphi^{(1)}(i^{(1)}; z) \otimes \varphi^{(2)}(i^{(2)}; z) \otimes \cdots \otimes \varphi^{(K)}(i^{(K)}; z) \tag{85}$$

where

$$g^{(k)}(i^{(k)}; z)\varphi^{(k)}(i^{(k)}; z) = A^{(k)}(z)\varphi^{(k)}(i^{(k)}; z). \tag{86}$$

Conversely, for any set $\{i^{(k)}\}$ and any $z \neq 0$ a pair $\{g(z), \varphi(z)\}$ satisfying (84), (85), (86) must satisfy (82) with $C = g(z)$.

If the functions $g(z)$ obtained in this way are indexed lexicographically to define the set $\{g_n(z)\}_{n=2}^N$, then these functions are equivalent to those defined for the global system in (21). In particular, each function $g_n(z)$, $n = 2, 3, \cdots, N-1$ is a one–one mapping of $(0, \infty)$ onto $(r(s_n), \infty)$ and $(-\infty, 0)$ onto $(-\infty, r(s_n))$, while $g_N(z)$ is a one–one mapping of $(0, \infty)$ onto $(r(s_N), \infty)$ and $(-\infty, 0)$ onto $(\bar{r}, r(s_N))$. It therefore follows from Section A.2 that all non-zero eigenvalues of (14) are obtained as solutions of (57) with the corresponding eigenvectors expressed as the Kronecker products of (60). This proves (57)–(60). Properties (62)–(65) follow

directly from Section A.2, since each subsystem satisfies the same reversibility assumption as the global system. The identifications of the signs of the eigenvalues in (66) follows directly from the asymptotic behavior of $g_n(z)$, inherited from its component functions on the right-hand side of (58).

Acknowledgments

The authors acknowledge with thanks, helpful discussions with Christian Huitema and Philippe Nain of INRIA, and computational help from Clark Wang of Columbia. Thanks are also due to D. Mitra for making a pre-publication copy of [15] available.

References

[1] ANICK, D., MITRA, D. AND SONDHI, M. M. (1982) Stochastic theory of a data-handling system with multiple sources. *Bell System Tech. J.* **61,** 1871–1894.

[2] BELLMAN, R. (1960) *Introduction to Matrix Analysis.* McGraw-Hill, New York.

[3] BREWER, J. W. (1978) Kronecker products and matrix calculus in system theory. *IEEE Trans. Circuits Syst.* **25,** 772–781.

[4] BURMAN, D. Y. AND SMITH, D. R. (1986) An asymptotic analysis of a queueing system with markov-modulated arrivals. *Operat. Res.* **34,** 05-119.

[5] ELWALID, A. E., MITRA, D. AND STERN, T. E. (1990) A theory of statistical multiplexing of markovian sources: spectral expansions and algorithms. In *Numerical Solution of Markov Chains,* ed. W. J. Stewart, Marcel Dekker, New York.

[6] FISCHER, M. J. (1979) Data performance in a system where data packets are transmitted during voice silent periods—single channel case. *IEEE Trans. Comm.* **27,** 1371–1375.

[7] FRANKLIN, J. N. (1968) *Matrix Theory.* Prentice Hall, Englewood Cliffs, NJ.

[8] GRAHAM, A. (1981) *Kronecker Products and Matrix Calculus with Applications.* Ellis Horwood, Chichester.

[9] KEILSON, J. AND RAO, S. S. (1970) A process with chain-dependent growth rate. *J. Appl. Prob.* **7,** 699–711.

[10] KEILSON, J. AND RAO, S. S. (1971) A process with chain-dependent growth rate. part II: the ruin and ergodic problems. *Adv. Appl. Prob.* **3,** 315–338.

[11] KEILSON, J. (1979) *Markov Chain Models—Rarity and Exponentiality.* Springer-Verlag, New York.

[12] KELLY, F. P. (1979) *Reversibility and Stochastic Networks.* Wiley, Chichester.

[13] KOSTEN, L. (1984) Stochastic theory of data-handling systems with groups of multiple sources. In *Performance of Computer-Communication Systems,* ed. H. Ruding and W. Bux, Elsevier, Amsterdam, 321–331.

[14] KOSTEN, L. (1986) Liquid models for a type of information buffer problems. *Delft Progress Report* **11,** 71–86.

[15] MITRA, D. (1988) Stochastic theory of a fluid model of producers and consumers coupled by a buffer, *Adv. Appl. Prob.* **20,** 646–676.

[16] MITRANI, I. (1968) A many server queue with service interruptions. *Operat. Res.* **16,** 628–638.

[17] MORRISON, J. (1989) Asymptotic analysis of a data-handling system with many sources, *SIAM J. Appl. Math.* **49,** 617–637.

[18] NEUTS, M. (1981) *Matrix Geometric Solutions in Stochastic Systems.* Johns Hopkins University Press, Baltimore.

[19] NEUTS, M. (1986) The caudal characteristic curve of queues. *Adv. Appl. Prob.* **18,** 221–254.

[20] RELLICH, F. (1969) *Perturbation Theory of Eigenvalue Problems.* Gordon and Breach, New York.

[21] SCHWARTZ, M. (1987) *Telecommunication Networks, Protocols, Modeling and Analysis.* Addison-Wesley, Reading, Mass.

[22] STERN, T. E. (1983) A queueing analysis of packet voice. *Proc. GLOBECOM '83,* 71–76.

[23] STERN, T. E. AND ELWALID, A. I. (1989) Analysis of separable markov-modulated rate models for information-handling systems. CTR technical report No. 164–89-43, Columbia University.

[24] WEISS, A. (1986) A new technique for analyzing large traffic systems. *Adv. Appl. Prob.* **18,** 506–532.