A numerically efficient method for the MAP/D/1/K queue via rational approximations

Nail Akar[1]

Computer Science Telecommunications, University of Missouri – Kansas City, 5100 Rockhill Road, Kansas City, MO 64110, USA

Erdal Arıkan

Electrical and Electronics Eng. Dept., Bilkent University, 06533 Ankara, Turkey

Received 21 April 1994; revised 20 July 1995

The Markovian Arrival Process (MAP), which contains the Markov Modulated Poisson Process (MMPP) and the Phase-Type (PH) renewal processes as special cases, is a convenient traffic model for use in the performance analysis of Asynchronous Transfer Mode (ATM) networks. In ATM networks, packets are of fixed length and the buffering memory in switching nodes is limited to a finite number K of cells. These motivate us to study the MAP/D/1/K queue. We present an algorithm to compute the stationary virtual waiting time distribution for the MAP/D/1/K queue via rational approximations for the deterministic service time distribution in transform domain. These approximations include the well-known Erlang distributions and the Padé approximations that we propose. Using these approximations, the solution for the queueing system is shown to reduce to the solution of a linear differential equation with suitable boundary conditions. The proposed algorithm has a computational complexity independent of the queue storage capacity K. We show through numerical examples that, the idea of using Padé approximations for the MAP/D/1/K queue can yield very high accuracy with tractable computational load even in the case of large queue capacities.

Keywords: Performance analysis of ATM networks, Markovian arrival process, finite buffer queues, loss probability, state-space representations, Padé approximations.

1. Introduction

In an ATM network, all information such as voice, data, and video is segmented into fixed-size packets, called *cells*. The share of common network resources among individual connections is made on a statistical multiplexing basis. The performance analysis of a statistical multiplexer whose input consists of a superposition of several packetized sources is in general difficult. This difficulty is

[1] This work was done when the author was with the Bilkent University, Ankara, Turkey and the research was supported by TÜBİTAK under Grant No. EEEAG-93.

mostly due to the number of arrivals in adjacent time intervals possessing a positive correlation. A common approach is to approximate this complex nonrenewal input process by an analytically tractable one.

Neuts [29] introduced a versatile Markovian point process, called *N-process*, which is analytically tractable and which is convenient for approximation of these complicated nonrenewal processes. This class of processes includes the MMPP, PH-renewal processes and a wide range of other processes as special cases, e.g., see Heffes and Lucantoni [19], Kuczura [23], Lucantoni et al. [27]. Lucantoni [25] introduced the Batch Markovian Arrival Process (BMAP), which is equivalent to the N-process but which has a simpler unifying notation. For a BMAP, arrivals are allowed to occur in batches where different types of arrivals may have different batch size distributions. If batch arrivals are not allowed, BMAP reduces to the Markovian Arrival Process (MAP) which is still a rich class of processes that contains MMPP and PH-renewal processes as subcases. Furthermore, stationary MAP's have the significant property that they are dense in the set of all stationary point processes, see Asmussen and Koole [9].

A detailed study of the N/G/1 queue is made by Ramaswami [31] in the context of M/G/1 type Markov chains where the queueing problem is shown to reduce to finding the minimal nonnegative solution for a certain nonlinear matrix equation. Variants of the algorithm in [31] for computing the minimal matrix solution have been proposed in Ramaswami [32], Gün [18], Lucantoni [25] and Lucantoni et al. [26] which require less computational effort. QBD (Quasi-Birth-and-Death) chains are special cases of M/G/1 type Markov chains and include the MAP/PH/1 queue as a subcase. Latouche and Ramaswami [24] have presented a logarithmic reduction algorithm for finding the matrix-geometric rate matrix for QBD chains with a quadratic convergence rate. The extension of the N/G/1 queueing model to the case of limited buffering memory is studied by Blondia [12] for which the computational load strictly depends on the queue capacity and the method is therefore computationally intractable especially for large buffer sizes. For the subcase of finite QBD chains, we note the techniques proposed by Ye and Li [41, 42] in order to analyze multi-media traffic queues by which significant reductions in computational load and space requirements have been achieved.

Many forms of data, voice, and image based communications in ATM networks are expected to have an on-off type behavior. On-off sources generate traffic during activity periods alternating with silence periods during which there is no traffic generation. The cell arrival process from an individual on-off source may be highly complicated (e.g., packetized voice) and exact analysis of systems offered with a superposition of such sources is generally difficult. One basic approach is to approximate the superposition by fitting certain parameters of the original process to those of a 2-state MMPP, a subcase of the MAP, proposed by Heffes and Lucantoni [19]. The MMPP/G/1 queueing model is shown in [19] to approximate the first two moments of delays as well as the tail probabilities with high accuracy. In Ide [20], the individual on-off source is characterized by an

Interrupted Poisson Process (IPP) which is indeed a special case of the MMPP. The MMPP is also used to model packetized video traffic by Saito et al. [33] and Skelly, Schwartz and Dixit [34]. Other special cases of the MAP/G/1 queue have appeared in the telecommunications literature in the context of PH/G/1 queues. A general treatment of which, with its special cases, can be found in Neuts [30]. For recent work on applications of the MAP in traffic modeling and control in ATM networks, we refer the reader to works by Choudry et al. [14] and Whitt [38].

In this paper, we examine a queueing system for which the incoming arrival process is modeled by a MAP which is simple but general enough to cover many teletraffic models used for ATM source characterization. We assume that the service times are deterministic due to cell-based transport in ATM networks. Since buffering memory in switching nodes is limited, the loss probability as well as the waiting times turns out to be an important performance measure of the system especially for real-time services. These motivate us to study the MAP/D/1/K queue whereas particular emphasis is given to the computation of the cell loss rate.

We propose an approximate method to compute the important performance measures of the system rather than an exact solution. The proposed exact solution algorithms by Blondia [12] and Lucantoni [25] either suffer from low convergence rates or they become computationally intractable especially when the number of phases of the MAP or the buffer sizes are large. Our solution technique consists of two main stages. At the first stage, we present Padé approximations for the deterministic service time distribution in transform domain. Although these approximations are not necessarily associated with probability distribution functions (pdf), they are shown to be more effective in capturing the queue dynamics compared with Erlang distributions. The second stage consists of solving a linear differential equation with suitable boundary conditions. The computational effort reduces to efficiently computing a matrix exponential of size md, where m is the number of phases of the MAP and d is a parameter based on whichever approximation for the service time is employed. We show through numerical examples that a Padé approximation with parameter $d = 3$ suffices for most of the applications.

From the mathematical formulation and computational complexity point of view, we believe that our work is closest to the techniques proposed by Baiocchi [10] and Baiocchi and Blefari-Melazzi [11] except that they are based on root finding algorithms whereas in our case, the solution is given in terms of a matrix exponential form. Besides the simplicity of our algorithm and the resulting form of the expression for the virtual waiting time distributions we have obtained, there is more flexibility in ways of evaluating matrix exponentials (see Moler and Van Loan [28]) which include simple rational approximations at the expense of some loss of accuracy (see Golub and Van Loan [17, pp. 555–560]). Furthermore, we make use of Padé approximations for the particular but important subcase of deterministic service times, making it numerically tractable to solve for the MAP/D/1/K queues even when the MAP consists of a large number of phases.

The remainder of the paper is organized as follows. In section 2, we define the MAP and present the virtual waiting time expression in a MAP/G/1 queue. We also present a novel exact solution methodology for the MAP/PH/1 queue that is extendable to the MAP/D/1/K system. Section 3 concentrates on rational approximations for the deterministic service time distribution which consist of the classical Erlang distributions and the Padé approximations. The problem formulation and an approximate solution for the MAP/D/1/K queue is presented in section 4. The final section includes numerical examples to demonstrate the performance of the proposed algorithm mainly in terms of the cell loss rate.

2. The MAP/G/1 queue

The Markovian Arrival Process (MAP) is introduced by Lucantoni et al. [27] in which the reader can find a detailed description of the concept of MAP and related issues. This section is devoted to a brief discussion of the Markovian arrival process and virtual waiting time expression in a MAP/G/1 queue.

The Markovian arrival process generalizes the Poisson process by allowing interarrival times which are not exponential but still maintaining its Markovian structure. In the case of a Poisson process with rate λ, the counting process $\{N(t)\}$, (number of arrivals in $(0, t]$), is a Markov process on the state-space $\{i: \ i \in \mathcal{Z}\}$ ($\mathcal{Z}$ denotes the set of nonnegative integers). The infinitesimal generator matrix of this process, Q, has the form

$$Q = \begin{bmatrix} d_0 & d_1 & & & \cdots \\ & d_0 & d_1 & & \cdots \\ & & d_0 & d_1 & \cdots \\ & & & & \ddots \end{bmatrix}, \tag{1}$$

where $d_0 = -\lambda$, $d_1 = \lambda$. In the case of a MAP, there is the additional phase process $\{J(t)\}$ assuming values in $\{1, 2, \ldots, m\}$. The two-dimensional Markov process $\{N(t), J(t)\}$ is then modeled as a Markov process on the state-space $\{(i, j): i \in \mathcal{Z}, \ 1 \leq j \leq m\}$ whose infinitesimal generator matrix Q can be represented in block form as

$$Q = \begin{bmatrix} D_0 & D_1 & & & \cdots \\ & D_0 & D_1 & & \cdots \\ & & D_0 & D_1 & \cdots \\ & & & & \ddots \end{bmatrix}. \tag{2}$$

Here, D_0, D_1 are $m \times m$ matrices, D_0 has negative diagonal elements and non-negative off-diagonal elements, D_1 is non-negative, and $D \triangleq D_0 + D_1$ is an irreducible infinitesimal generator. The matrix D_0 is stable implying D_0 to be nonsingular and the sojourn time in each state (i, j) to be finite with probability 1. The evolution of the process is as follows. Assume that the Markov process (phase process) with generator D is in some state j, $1 \leq j \leq m$. After an exponentially distributed time interval with parameter $-(D_0)_{jj}$, there occurs either a transition to another state $k \neq j$ without an arrival with probability $\frac{(D_0)_{jk}}{-(D_0)_{jj}}$ or to a state l (possibly the same state) with an arrival with probability $\frac{(D_1)_{jl}}{-(D_0)_{jj}}$. Let π be the stationary probability vector of the phase process with generator D so that π satisfies

$$\pi D = 0, \ \pi e = 1, \tag{3}$$

where e is a column vector of ones. The mean arrival rate denoted by $\bar{\lambda}$ is given by

$$\bar{\lambda} = \pi D_1 e. \tag{4}$$

The MAP includes MMPP, PH-renewal processes and superpositions of these processes as special cases. The MMPP (see Heffes and Lucantoni [19]) with an infinitesimal generator R and rate matrix $\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ is a MAP with $D_0 = R - \Lambda$ and $D_1 = \Lambda$. The PH-renewal process (see Lucantoni et al. [27]) with representation (α, T), is a MAP with $D_0 = T$ and $D_1 = -Te\alpha$. This rich class includes superpositions of the Erlangian (E_k) and Hyperexponential (H_k) distributions. We refer to [27] for a general treatment of the subcases of the MAP.

Let us now consider a single server queue offered with a MAP characterized by the matrices D_0 and D_1. For the time being, let the service time have an arbitrary distribution function B, with Laplace–Stieltjes Transform (LST), $\hat{B}$. Hereafter, we assume that the parameters of the incoming MAP are normalized so that the mean service time is unity. We also assume a stable queue, i.e., $\bar{\lambda} < 1$.

We now restate the results for the virtual waiting time distribution in the MAP/G/1 queue given by Ramaswami [31] and Lucantoni [25]. An alternative proof for the same expression is developed by Akar and Arıkan [5] for the subcase of an MMPP/G/1 queue. For this purpose, we first define

$$W(x) = [\, W_1(x) \quad W_2(x) \quad \cdots \quad W_m(x) \,],$$

where $W_j(x)$ is the stationary probability that at an arbitrary time the arrival process is in phase j and the unfinished work at that time is at most x. The virtual waiting time cumulative distribution function (cdf) is denoted by $w(x) = W(x)e$. We define $\hat{W}(s)$ and $\hat{w}(s)$ to be the Laplace Transforms (LT) of $W(x)$ and $w(x)$, respectively, where in our LT definition, the lower limit of integration is 0^- allowing impulsive functions located at the origin. Ramaswami [31] has shown that

$$\hat{W}(s) = y_0[sI + D_0 + D_1\hat{B}(s)]^{-1}, \tag{5}$$

from which

$$\hat{w}(s) = \hat{W}(s)e,$$

where I is the identity matrix of size m and $y_0 = [y_{01} \quad y_{02} \quad \dots \quad y_{0m}]$ is such that y_{0j} is the stationary probability that at an arbitrary time the arrival process is in phase j and the queue is empty. The vector

$$g = \frac{1}{1-\lambda} y_0$$

is shown by Lucantoni [25] to be the stationary probability vector of G described implicitly via

$$G = \int_0^\infty e^{[(D_0+D_1G)x]}\, dB(x), \tag{6}$$

that is, given G,

$$gG = g, \quad ge = 1. \tag{7}$$

An iterative algorithm has been proposed by Lucantoni [25] for computing the matrix G in the BMAP framework which allows batch arrivals and includes MAP as a special case. This algorithm starts with $G_0 = 0$ and G can be computed by successively iterating in the following recursion:

$$H_{n+1,k} = [I + \theta^{-1}(D_0 + D_1G_k)]H_{n,k},$$

$$G_{k+1} = \sum_{n=0}^{\infty} \gamma_n H_{n,k},$$

where $H_{0,k} = I$, $\theta = \max_i\{(-D_0)_{ii}\}$, and

$$\gamma_n = \int_0^\infty e^{-\theta x} \frac{(\theta x)^n}{n!} dB(x), \quad \text{for } n \geq 0.$$

Whichever performance measure of the queueing system one is interested in finding, computing G is the essential part of the overall algorithm. Once the matrix G is determined, one can compute g (or, equivalently y_0) in (7) and then calculate the associated moments of the waiting time distribution which are explicitly given by Lucantoni [25]. If distributions are sought, inversion of the transform expression in (5) is required for which easily implementable and computationally efficient numerical algorithms are available in the literature, e.g., see Abate and Whitt [2, 3]. We note that a significant portion of the procedure outlined above is devoted to the computation of the matrix G.

Below, we give an alternative exact solution method for the unfinished work distribution in a MAP/PH/1 queue that has the following features:

(i) We compute the unknown boundary probability vector y_0 without the need for calculating the matrix G and write the virtual waiting time distribution in terms of a simple matrix exponential form.

(ii) Via simple extensions based on rational Padé approximations, this methodology can be used to obtain accurate approximations for the solution of the MAP/D/1/K queue with a computational complexity independent of the queue storage capacity, K.

Since the service time distribution is now assumed to be of phase type, the LST of the service time distribution, $\hat{B}(s)$, is a rational function of the indeterminate s. In other words,

$$\hat{B}(s) = \frac{\hat{R}(s)}{\hat{Q}(s)},$$

where the polynomials $\hat{R}(s)$ and $\hat{Q}(s)$ are assumed to have degrees n and l, respectively, and $n \leq l$, see Neuts [30]. We assume that the highest degree coefficient of $\hat{Q}$ is unity without any loss of generality. Then the LST of the unfinished work distribution in the MAP/PH/1 queue in (5) can be rewritten as

$$\begin{aligned} \hat{W}(s) &= y_0 \left[sI + D_0 + D_1 \frac{\hat{R}(s)}{\hat{Q}(s)} \right]^{-1} \\ &= y_0 \hat{Q}(s)[(sI + D_0)\hat{Q}(s) + D_1 \hat{R}(s)]^{-1} \\ &= y_0 \hat{Q}(s) \hat{H}(s)^{-1} \end{aligned} \tag{8}$$

In the above expression, the polynomial matrix

$$\hat{H}(s) = (sI + D_0)\hat{Q}(s) + D_1 \hat{R}(s)$$

has degree

$$d = l + 1, \tag{9}$$

that is, $\hat{H}$ can be written as

$$\hat{H}(s) = H_d s^d + H_{d-1} s^{d-1} + \cdots + H_1 s + H_0, \tag{10}$$

for some constant matrices H_i, $i = 0, 1, \ldots, d$. Similarly, the polynomial $\hat{Q}(s)$ is of the form

$$\hat{Q}(s) = q_{d-1}s^{d-1} + q_{d-2}s^{d-2} + \cdots + q_1 s + q_0, \tag{11}$$

since $deg(\hat{Q}(s)) = l = d - 1$. We note that q_{d-1} is the highest degree coefficient of $\hat{Q}$ and is equal to $q_{d-1} = 1$ which then yields $H_d = I$.

One can view the polynomial matrix fractional description given in (8) as the expression for the output of an m-output, linear, finite-dimensional, continuous-time system excited by its initial condition y_0 (see Chen [13]). In regard of this, the input-output relationship (8) can equivalently be represented by a vector-differential equation of size $deg(det(\hat{H}(s))) = md$ and of the form

$$\begin{aligned} &\frac{d}{dx}u(x) = u(x)A, \quad u(0) = y_0 B, \\ &W(x) = u(x)C, \end{aligned} \tag{12}$$

via an md-dimensional state vector $u(\cdot)$ and A, B, C being constant matrices of size $md \times md$, $m \times md$, and $md \times m$, respectively, e.g., see Chen [13] and Kailath [21]. The choice of the suitable matrices that yield

$$B(sI - A)^{-1}C = \hat{Q}(s)\hat{H}(s)^{-1}$$

is called a state-space realization of (8) [13]. Now, we will obtain a natural state-space realization of (8) through the following mathematical formulation. For this purpose, we define

$$\begin{aligned} &\hat{W}^1(s) = \hat{W}(s), \\ &\hat{W}^i(s) = s\hat{W}^{i-1}(s) - W^{i-1}(0), \quad i = 2, 3, \ldots, d, \end{aligned}$$

where $W^i(x)$ and $\hat{W}^i(s)$ form a LT-pair. Actually,

$$W^i(x) = \frac{d^{i-1}}{dx^{i-1}}W(x), \quad i = 1, 2, \ldots, d, \quad x \geq 0.$$

We note by the initial value theorem on Laplace transforms that

$$\lim_{s \to \infty} s\hat{W}^i(s) = W^i(0), \tag{13}$$

must be a bounded vector.

It is now easy to see that

$$s\hat{W}^i(s) = \hat{W}^{i+1}(s) + W^i(0), \quad i = 1, 2, \ldots, d - 1. \tag{14}$$

One can also show by using (8) and by algebraic manipulations the following expression for $s\hat{W}^d(s)$:

$$\begin{aligned} s\hat{W}^d(s) &= -\sum_{i=0}^{d-1} \hat{W}^{i+1}(s)H_i \\ &\quad + y_0 q_0 - \sum_{j=1}^{d-1} W^j(0)H_j \\ &\quad + \sum_{i=1}^{d-1} s^{d-i}\left(-W^i(0) + y_0 q_{d-i} - \sum_{j=1}^{i-1} W^j(0)H_{d-i+j}\right). \end{aligned} \tag{15}$$

By (13), the last term on the RHS of (15) should vanish, that is, to make $\lim_{s\to\infty} s\hat{W}^d(s)$ bounded, $W^i(0)$, $i = 1, 2, \ldots, d-1$ should satisfy

$$W^i(0) = y_0 q_{d-i} - \sum_{j=1}^{i-1} W^j(0)H_{d-i+j}, \quad i = 1, 2, \ldots, d-1. \tag{16}$$

Furthermore, since $\lim_{s\to\infty} \hat{W}^i(s) = 0$ for each i, $W^d(0)$ satisfies

$$W^d(0) = y_0 q_0 - \sum_{j=1}^{d-1} W^j(0)H_j. \tag{17}$$

We now iteratively define

$$B_1 = I,$$

and for $i = 2, 3, \ldots, d$

$$B_i = q_{d-i}I - \sum_{j=1}^{i-1} B_j H_{d-i+j},$$

so that one can now write

$$W^i(0) = y_0 B_i, \quad i = 1, 2, \ldots, d. \tag{18}$$

Let us define the concatenated vectors

$$\hat{W}_c(s) = \begin{bmatrix} \hat{W}^1(s) & \hat{W}^2(s) & \cdots & \hat{W}^d(s) \end{bmatrix},$$

and

$$B = [B_1 \quad B_2 \quad \cdots \quad B_d],$$

One can then make use of (14), (15), and (18) to obtain

$$\hat{W}_c(s)(sI - A) = y_0 B,$$
$$\hat{W}(s) = \hat{W}_c(s)C,$$

where

$$A = \begin{bmatrix} 0 & 0 & \cdots & 0 & -H_0 \\ I & 0 & \cdots & 0 & -H_1 \\ 0 & I & \cdots & 0 & -H_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & I & -H_{d-1} \end{bmatrix}, \quad C = \begin{bmatrix} I \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Or, equivalently,

$$\frac{d}{dx} W_c(x) = W_c(x)A, \quad x \geq 0 \tag{19}$$

and

$$W_c(0) = y_0 B, \quad W(x) = W_c(x)C.$$

Having obtained the state space realization of the queueing system in (19), the solution to the linear differential equation takes the matrix exponential form

$$W(x) = y_0 B e^{Ax} C. \tag{20}$$

Recall that the asymptotic behavior of $W(x)$ given by the matrix analytic form above is governed by the largest negative real eigenvalue of A, which we denote by σ,

$$1 - w(x) = 1 - W(x)e = ke^{\sigma x} + o(e^{\sigma x}), \quad \text{as } x \to \infty,$$

where k is a positive constant. It is actually straightforward to show that

$$\sigma = -pf(D_0 + D_1 \hat{B}(\sigma));$$

where $pf(\cdot)$ refers to the Perron-Frobenius eigenvalue and $-\sigma$ is called the *asymptotic decay rate* by Abate et al. [1]. Actually, the eigenvalues of the matrix A can be shown to coincide with the singularities of the matrix polynomial $\hat{H}(s)$.

Given the matrix exponential form of the virtual waiting time distribution

(20), what remains is to determine the boundary probability vector y_0. As described before, the unknown vector y_0 in the expression (20) can be determined by solving the stationary probability vector of the matrix G, the unique minimal nonnegative solution of the equation (6). Another alternative we will outline below is to use the spectral decomposition techniques of Akar [4] and Akar and Arıkan [5] which are based on determining y_0 by imposing that no unstable mode of the dynamical system (19) be excited together with the constraint $W_c(0) = y_0 B$.

It can be shown that the matrix A has $m - 1$ eigenvalues in the open right half plane, one at the origin and the remaining $m(d-1)$ eigenvalues in the open left half plane when $\bar{\lambda} < 1$. We define the $m(d-1) \times md$ matrix S_A whose rows are composed of the left eigenvectors of the matrix A associated with its $m(d-1)$ eigenvalues lying in the open left half plane. In other words, let u_i, $1 \leq i \leq m(d-1)$ be such that

$$u_i \sigma_i = u_i A, \quad \text{Re}\ \sigma_i < 0.$$

The matrix S_A is then defined as

$$S_A = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{m(d-1)} \end{bmatrix}.$$

The rows of S_A form a basis for the stable subspace of A (see Wonham [40]) where the term "stable" is inherited from the stability of differential systems.

For the infinite buffer case, the initial condition y_0 should be chosen so that no unstable mode of the matrix A should be excited. Otherwise, the solution for $W(x)$ in (20) blows up as $x \to \infty$. In mathematical terms, this is equivalent to saying $W_c(0) - W_c(\infty)$ should lie in the row space of S_A, i.e.,

$$W_c(0) - W_c(\infty) = x S_A,$$

for some $1 \times m(d-1)$ vector x, or equivalently,

$$y_0 B - x S_A = W_c(\infty).$$

Then one can solve the linear square system below:

$$[y_0 \quad x] \begin{bmatrix} B \\ -S_A \end{bmatrix} = W_c(\infty) = [\pi \quad 0 \quad \cdots \quad 0],$$

for y_0 and x. The last equality comes from the fact that $W(\infty) = \pi$ and the higher order derivatives of $W(x)$ should vanish as $x \to \infty$.

We note that, in the above formulation one can replace S_A by any matrix $\bar{S}_A$ whose row space is equal to the former. The two companion papers by Akar and Sohraby [6, 7] include fast and numerically reliable algorithms to compute a basis for the row space of S_A without the need for solving the eigenvalues and eigenvectors of A in the more general framework of M/G/1 and G/M/1 type Markov chains.

The emphasis here is introducing a new mathematical framework for the MAP/PH/1 queue that is extendable to the MAP/D/1/K queue through rational approximations rather than making a comparison of the existing algorithms to compute the boundary probability vector y_0 for the infinite buffer case (see Akar and Sohraby [7]), which is outside the scope of this paper. The next sections address to how this extension is made possible.

3. Rational approximations for the deterministic service time

In this section, we consider rational approximations for the deterministic service time distribution to allow computational analysis. The deterministic service time being unity, we have

$$\hat{B}(s) = \int_0^\infty e^{-sx} dB(x) = e^{-s}. \tag{21}$$

In the case of MAP/D/1 queue, the expression for the stationary unfinished work distribution turns out to be

$$\hat{W}(s) = y_0[sI + D_0 + D_1 e^{-s}]^{-1}, \tag{22}$$

which is an irrational transform. The irrational term (i.e., e^{-s} in the denominator matrix in (22)) may in general be difficult to handle if the vector of empty queue probabilities y_0 is sought. Therefore, we seek appropriate rational approximations for the irrational transform e^{-s} so as to compute the unfinished work distribution. A rational approximant of e^{-s} is denoted by $\hat{B}_a(s)$.

One alternative is to use phase-type distributions to approximate the distribution B. Indeed, a general distribution G of a nonnegative random variable can be approximated arbitrarily closely by phase-type distributions (see Wolff [39]). Consequently, if G has finite rth moment $(1 \leq r \leq \infty)$, one can find a phase-type distribution H for which the first r moments are arbitrarily close to those of G. The k-stage Erlang distribution is a special case of the phase-type distribution that is commonly used in the ATM literature in references by Saito et al. [33], Choudry et al. [14] and Skelly et al. [34] to approximate the deterministic service-time distribution. In the case of a k-stage Erlang distribution approximation, the rational approximant,

$\hat{B}_a(s)$, becomes

$$\hat{B}_a(s) = \left(\frac{k}{s+k}\right)^k. \tag{23}$$

There are two main disadvantages of this kind of approximations: first, there is the need for a large number of stages in the Erlang distribution to adequately match the original distribution (see Kleinrock [22]). Second, no matter how large a k we choose, we cannot capture the rth moment ($r \geq 2$), b_r, of the original distribution exactly. Actually,

$$b_r = \frac{(k+r-1)!}{(k-1)!\, k^r},$$

for a k-stage Erlang distribution and converges to unity as $k \to \infty$ with a linear convergence rate. We note that this low convergence rate may be intolerable for particular applications.

Alternatively, we propose here to use rational Padé approximations of the term e^{-s}. A Padé approximation with parameters n and l is a rational function

$$\hat{P}_{n,l}(s) = \frac{\hat{R}_n(s)}{\hat{Q}_l(s)},$$

where $\hat{R}_n(s)$ and $\hat{Q}_l(s)$ are polynomials of order n and l, respectively, and the first $(n+l+1)$ terms of its Taylor series expansion equal to those of the Taylor expansion of $\exp(-s)$, or equivalently the first $(n+l)$ moments of the original service time distribution match with those of the Padé approximation. A closed form expression for $\hat{P}_{n,l}$ exists and is given by Vlach and Singhal [37]

$$\hat{P}_{n,l}(s) = \frac{\sum\limits_{i=0}^{n} (l+n-i)!C(n,i)(-1)^i s^i}{\sum\limits_{i=0}^{l} (l+n-i)!C(l,i)s^i}, \tag{24}$$

where

$$C(n,i) = \frac{n!}{i!(n-i)!}.$$

Note that the inverse Laplace transform of $\hat{P}_{n,l}(s)$, say $P_{n,l}(x)$, is not necessarily a pdf. Removal of the restriction of approximating a distribution by another distribution brings one more degree of freedom in that the first r moments are exactly matched with a convenient choice of a Padé approximation. Although the use of Padé approximants is not restricted to the deterministic service time and may be used for general service time distributions, the focus of this paper is on approximants of the form given in (24). The main disadvantage in using Padé approximants lies under the fact that we might no longer be in the framework of probability distribution functions that causes a lack of physical interpretation of the underlying process. However, as far as accurate computational analysis of queueing systems is concerned, we believe that such approximations will serve an important role. As a final note on this issue, consider an MMPP/D/1 queue with the MMPP having the infinitesimal generator matrix R and the rate matrix Λ. In case $\hat{P}_{1,0}(s) = 1 - s$ is employed as a rational approximation for e^{-s}, the transform of the unfinished work distribution turns out to be

$$\begin{aligned}\hat{W}(s) &= y_0[sI + R - \Lambda + \Lambda\hat{P}_{1,0}(s)]^{-1}\\ &= y_0[sI + R - \Lambda + \Lambda(1 - s)]^{-1}\\ &= y_0[s(I - \Lambda) + R]^{-1},\end{aligned}$$

which is in fact equivalent to the expression suggested for the unfinished work distribution for the well-known Markov modulated fluid sources by Anick et al. [8]. Although $P_{1,0}(x)$ does not correspond to a probability distribution function, there is a wide-spread use of stochastic fluid flow models for the performance analysis of statistical multiplexers in the ATM context (e.g., see Anick et al. [8], Stern and Elwalid [35] and Elwalid and Mitra [16]).

In the next section, a mathematical framework is presented to solve the MAP/D/1/K system in case an arbitrary rational approximation $\hat{B}_a(s)$ is imposed. Then performance assessment of Erlang and Padé approximations in the analysis of the MAP/D/1/K queue is demonstrated via the use of numerical examples.

4. Analysis of the MAP/D/1/K queue

Let an arbitrary rational transform

$$\hat{B}_a(s) = \frac{\hat{R}_a(s)}{\hat{Q}_a(s)}$$

be imposed as an approximation of e^{-s}. The polynomials $\hat{R}_a$ and $\hat{Q}_a$ are assumed to have degrees n and l, respectively, where $n \leq l$. The case of $n > l$ is omitted since in this case it is no longer possible to interpret the vector y_0 as the equilibrium probability vector associated with empty queue lengths. We also assume the highest degree coefficient of $\hat{Q}_a(s)$ is unity without loss of generality. Defining $d = l + 1$, let us write

$$\hat{Q}_a(s) = s^l + q_{a,l-1}s^{l-1} + \cdots + q_{a,1}s + q_{a,0},$$
$$\hat{H}_a(s) = (sI + D_0)\hat{Q}_a(s) + D_1\hat{R}_a(s) = s^dI + H_{a,d-1}s^{d-1} + \cdots H_{a,1}s + H_{a,0}.$$

Also let the queue storage capacity be denoted by K. When a new arrival finds fewer than K cells in the queue waiting to be served, it is admitted to the system. Following the schemes of Tucker [36] and Elwalid and Mitra [15] used for Markov modulated fluid sources and noting that in the MAP/PH/1 analysis we only made use of the fact that the LT of the service time distribution is a rational function, one can show that the following differential equation is valid in the interval $0 \leq x \leq K$:

$$\begin{aligned} \frac{d}{dx}W_c(x) &= W_c(x)A_1, \quad W_c(0) = y_{0,K}B_1, \\ W(x) &= W_c(x)C_1, \quad 0 \leq x \leq K, \end{aligned} \tag{25}$$

where $y_{0,K}(j)$ refers to the stationary probability of the incoming MAP being in phase j and the unfinished work being zero for the case the buffer size equals K. In the above differential equation,

$$A_1 = \begin{bmatrix} 0 & 0 & \cdots & 0 & -H_{a,0} \\ I & 0 & \cdots & 0 & -H_{a,1} \\ 0 & I & \cdots & 0 & -H_{a,2} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & I & -H_{a,d-1} \end{bmatrix}, \quad C_1 = \begin{bmatrix} I \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{26}$$

and

$$B_1 = \begin{bmatrix} B_{1,1} & B_{1,2} & \cdots & B_{1,d} \end{bmatrix} \tag{27}$$

is such that

$$B_{1,1} = I,$$

and for $i = 2, 3, \ldots, d$

$$B_{1,i} = q_{a,d-i}I - \sum_{j=1}^{i-1} B_{1,j}H_{a,d-i+j}. \tag{28}$$

On the other hand, if an arrival occurs at time t and the instantaneous queue length at that time is above K, the packet associated with that arrival is dropped.

From the queue length point of view, it is convenient to visualize the incoming MAP characterized by the matrix pair (D_0, D_1) to change to another MAP described by the matrix pair $(D, 0)$ whenever the number of packets in the queue is K. This is equivalent to assuming that no arrivals will occur and the MAP will be constituted of only its phase process. Also note that the queue length cannot exceed $K + 1$ since there is one deterministic server. Then one can obtain as in (19) the following differential equation in the interval $K \leq x < K + 1$:

$$\frac{d}{dx} W_c(x) = W_c(x) A_2, \ K \leq x < K + 1. \tag{29}$$

In this equation, the matrix A_2 is of the form

$$A_2 = \begin{bmatrix} 0 & 0 & \cdots & 0 & -G_{a,0} \\ I & 0 & \cdots & 0 & -G_{a,1} \\ 0 & I & \cdots & 0 & -G_{a,2} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & I & -G_{a,d-1} \end{bmatrix}, \tag{30}$$

where

$$\hat{G}_a(s) = (sI + D)\hat{Q}_a(s) = s^d I + G_{a,d-1} s^{d-1} + \cdots + G_{a,1} s + G_{a,0}.$$

We are now ready to compute the virtual waiting time distribution in a MAP/D/1/K system except for the boundary conditions. The boundary condition at $x = K + 1$ is easy to write since (i) queue length cannot exceed $K + 1$, (ii) stationary probability of the queue length being $K + 1$ is zero, i.e., there may not be a jump in the unfinished work cdf vector at $x = K + 1$. Based on these two observations, one can write

$$W^1(K + 1) = W(K + 1) = \pi. \tag{31}$$

Making use of the continuity of the solution of the two differential equations (25) and (29) at $x = K$, one can rewrite (31) as

$$y_{0,K} B_1 e^{A_1 K} e^{A_2} C_1 = \pi. \tag{32}$$

The unknown vector $y_{0,K}$ can be solved through the linear matrix equation (32) of size m. At this stage, any algorithm for computing matrix exponentials given by Moler and Van Loan [28] can be used to compute the left hand side of (32). In particular, to evaluate the matrix exponential in our numerical experimentation, we compute the eigenvalues and the eigenvectors of the matrices through converting them to Hessenberg form using orthogonal similarity transformations and then using the QR method. The details of the procedure above can be found in the paper by Golub and Van Loan [17]. Actually, in case the incoming MAP is made up of a superposition of many independent MAP's, these eigenvalue-eigenvector pairs can be computed via simpler partial problems via the extension of the techniques used by Stern and Elwalid [35] in the context of stochastic fluid models. We also note that the extension of the proposed technique in [35] to the particular case of a superposition of 2-state MMPP's is examined by Akar [4].

Once the boundary vector $y_{0,K}$ is computed, the solution to the differential equations for W_c is easy to write:

$$\begin{aligned} W_c(x) &= y_{0,K}B_1 e^{A_1 x}, \quad 0 \le x \le K, \\ &= W_c(K)e^{A_2(x-K)}, \quad K \le x \le K+1. \end{aligned}$$

The stationary unfinished work cdf vector $W(x)$ is then expressed as

$$W(x) = W^1(x) = W_c(x)C_1. \tag{33}$$

Cell losses occur when arrivals find K cells waiting in the buffer. The cell loss rate, p_{loss}, is therefore described by the following expression:

$$p_{loss} = \frac{(\pi - W(K))D_1 e}{\bar{\lambda}}. \tag{34}$$

We now give the simple-to-implement step-by-step procedure of the overall algorithm for the MAP/D/1/K queue for convenience of implementation. The time unit is the deterministic service time and the MAP characterized by the two matrices D_0 and D_1 is assumed to be normalized with respect to this time unit.

PROCEDURE

1. Choose the Padé approximation $\hat{B}_a(s) = \hat{R}_a(s)/\hat{Q}_a(s)$ based on (24) with numerator and denominator degrees being n and l, respectively, and highest degree coefficient of the denominator being unity.
2. Write $d = l + 1$ and

$$\hat{Q}_a(s) = s^l + q_{a,l-1}s^{l-1} + \cdots + q_{a,1}(s) + q_{a,0},$$

$$\hat{H}_a(s) = (sI + D_0)\hat{Q}_a(s) + D_1\hat{R}_a(s) = s^d I + H_{a,d-1}s^{d-1} + \cdots H_{a,1}s + H_{a,0},$$

$$\hat{G}_a(s) = (sI + D)\hat{Q}_a(s) = s^d I + G_{a,d-1}s^{d-1} + \cdots + G_{a,1}s + G_{a,0}.$$

3. Define the matrices A_1, B_1, and C_1 as in (26)–(28).
4. Define the matrix A_2 as in (30).

5. Find the stationary probability vector π of the generator $D = D_0 + D_1$.
6. Solve for $y_{0,K}$ out of the following linear equation of size m:

$$y_{0,K} B_1 e^{A_1 K} e^{A_2} C_1 = \pi.$$

7. Write the stationary virtual waiting time cdf vector as

$$\begin{aligned} W(x) &= y_{0,K} B_1 e^{A_1 x} C_1, \ 0 \le x \le K \\ &= y_{0,K} B_1 e^{A_1 K} e^{A_2 (x-K)} C_1, \ K \le x \le K+1 \end{aligned}$$

and the cell loss probability p_{loss} as in (34).

5. Numerical examples

In this section, we present some numerical examples to demonstrate the performance assessment of the proposed algorithm based on Padé approximations.

We first consider the M/D/1 infinite capacity queue so as to clarify the concept of rational approximations for the deterministic service time. In this case,

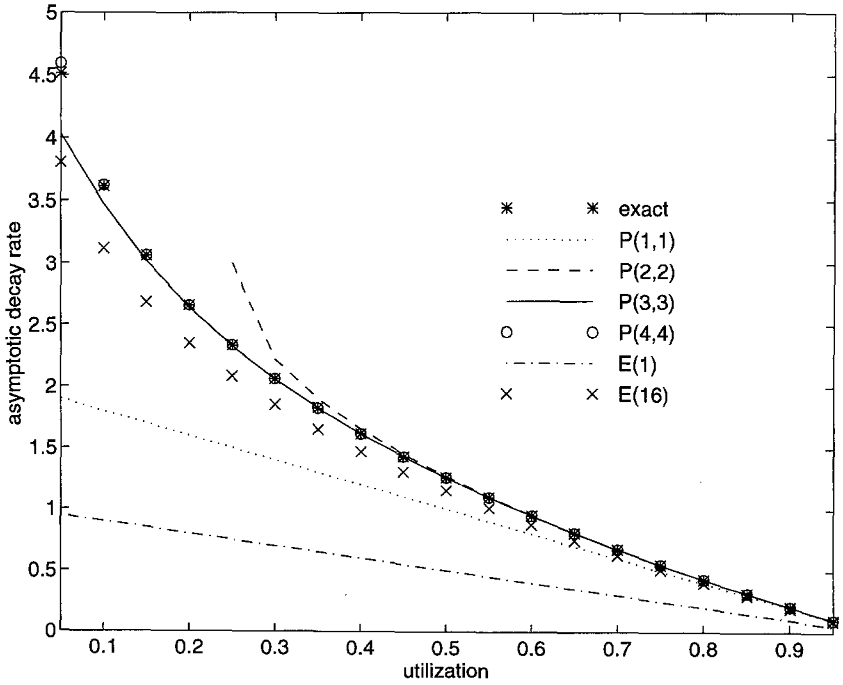

Fig. 1. Asymptotic decay rate for the M/D/1 queue.

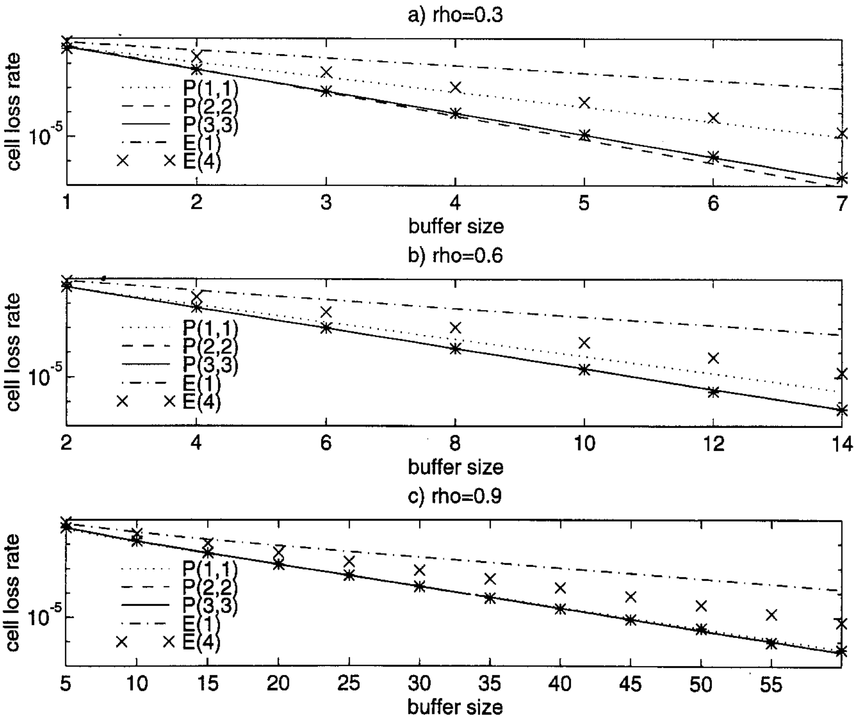

Fig. 2. Cell loss rate approximations for the M/D/1/K queue with (a) $\rho = 0.3$, (b) $\rho = 0.6$ and (c) $\rho = 0.9$ ($*$ denotes the simulation results).

the transform of the unfinished work cdf reduces to

$$\hat{w}(s) = \frac{(1-\rho)}{s - \rho + \rho e^{-s}}, \tag{35}$$

where ρ is the utilization of the system. Note that

$$w(x) = 1 - ke^{\sigma x} + o(e^{\sigma x}), \quad x \to \infty,$$

where $\sigma < 0$ is the largest negative real root of the denominator of (35) and plays a key role in the performance of the queueing system. We now compare the asymptotic decay rates obtained via the Padé approximations and the Erlang distributions with the numerical values of σ we have obtained through root finding algorithms. Rather than presenting the approach as an approximation for determining the asymptotic decay rate which can easily be computed using standard numerical techniques, our aim is to show the performances of various related approximations in terms of one important parameter of the queueing system. The k-stage Erlang distributions and Padé approximations are used to approximate the asymptotic decay

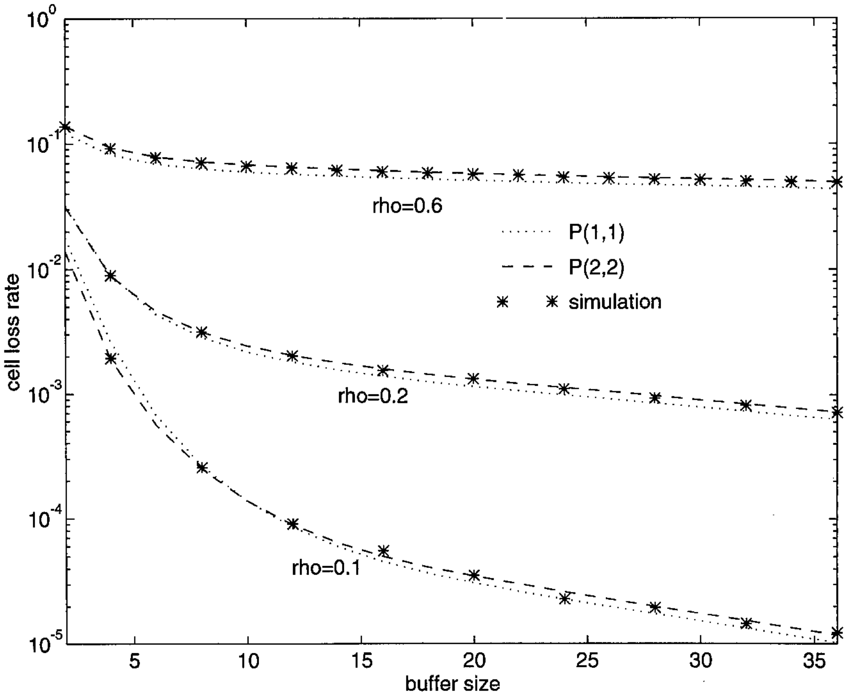

Fig. 3. Cell loss rate approximations for the MMPP/D/1/K queue with individual source parameters $\alpha^{-1} = 4363.63$, $\beta^{-1} = 436.36$, and $P = 0.275$ and with three different utilizations.

rate in an M/D/1 queue and the performance results of these approximations are presented in fig. 1 with respect to the utilization in the system.

The notation $E(k)$ is used to denote when the k-stage Erlang distribution is imposed. Similarly, we use the notation $P(n, l)$ to denote the case of a Padé approximation $\hat{P}_{n,l}(s)$. Throughout the examples we only focus on the Padé approximations of the type $P(l, l)$ since it is clear that $P(l - l_0, l)$ $(l_0 > 0)$ and $P(l, l)$ yield the same computational load whereas the former can match fewer moments than the latter and is not considered here. As far as the results in fig. 1 are concerned, there is a key observation, the rate of convergence (as $l \rightarrow \infty$) of the Padé approximations to the exact asymptotic decay rate is fast whereas this convergence rate in the case of k-stage Erlang approximations is rather slow. Besides, for heavy loads $(\rho > 0.5)$, the simple $P(2, 2)$ works as well as higher order Padé approximations which makes it well-suited for use in the ATM environment due to its simplicity. However, there is the drawback of using approximations which are not themselves probability distributions which is demonstrated by the break in the $P(2, 2)$ curve at $\rho = 0.25$ which indicates there is no largest negative real root below that utilization. This is problematic (e.g., negative probabilities may result) but we recommend the use of higher order Padé approximations in the light load case for which we have not

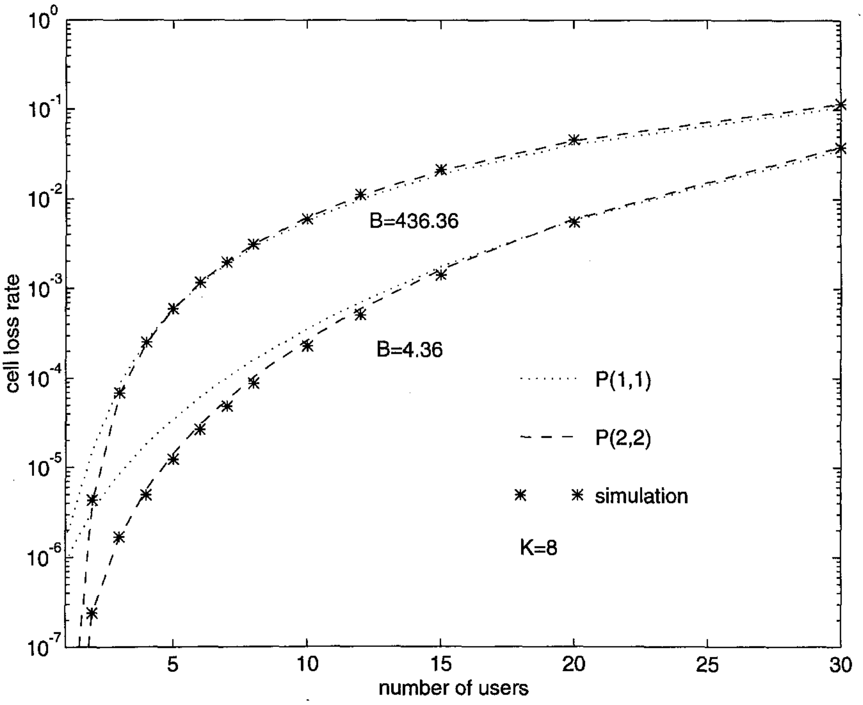

Fig. 4. Cell loss rate approximations for the MMPP/D/1/K queue for a small buffer size, $K = 8$, with two different burst lengths of the individual source.

observed any such brake for the range $\rho > 0.05$. In terms of the asymptotic decay rate of the unfinished work distribution, the Padé approximations work better than the Erlang approximations in the sense that to get the same degree of accuracy a significantly higher degree Erlang distribution is needed. This statement is also true for the general MAP/D/1/K queue as will be demonstrated by the following examples.

Figure 2 is devoted to the cell loss rate approximations with respect to the buffer size (in cells) in an M/D/1/K queue. Three different cases are examined with ρ being 0.3, 0.6, and 0.9, respectively. $P(3,3)$ captures the simulation curve for all the cases whereas $P(2,2)$ though being indistinguishable from $P(3,3)$ in the latter two cases, exhibits a slight deviation in the $\rho = 0.3$ case. Even the simplest $P(1,1)$ works better than the $E(4)$ for all the cases whereas its performance for the heavy load case (e.g., $\rho = 0.9$) is quite satisfactory. Here, we recall that the key parameter that determines the computational load is the denominator degree d defined in (9) which is $k + 1$ for a k-stage Erlang distribution approximation and $l + 1$ for the Padé approximation $\hat{P}_{l,l}(s)$.

We present our results for the MMPP/D/1/K queue in fig. 3. The input arrival process is assumed to be a superposition of N identical and independent 2-state IPP sources (users). In the silence state, the source generates no traffic whereas in the

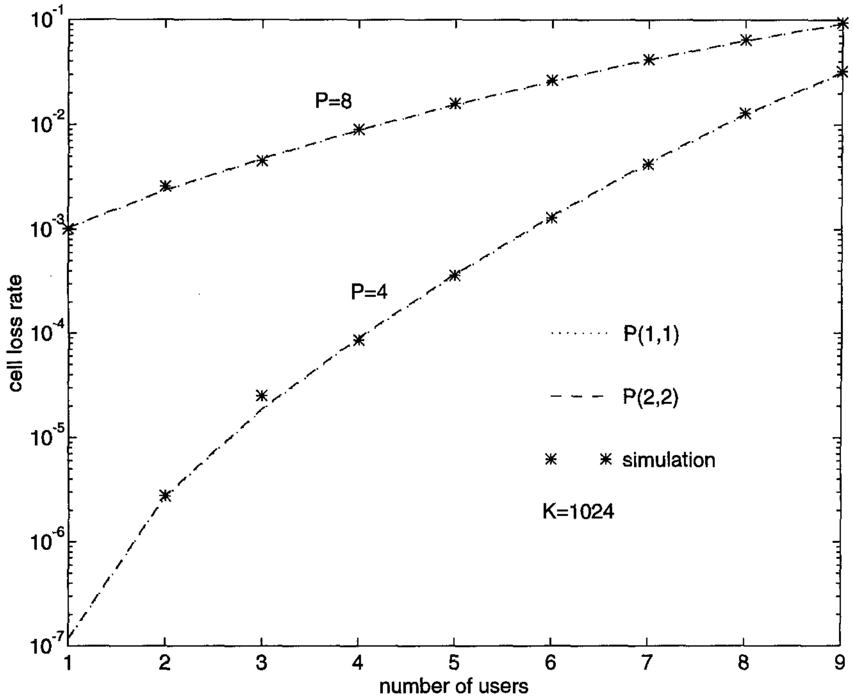

Fig. 5. Cell loss rate approximations for the MMPP/D/1/K queue for a large buffer size, $K = 1024$, with two different user peak rates.

activity state arrivals occur according to a Poisson process with rate P. The silence and the activity times are exponentially distributed with parameters α and β, respectively. The cell loss rates based on two Padé approximants ($P(1,1)$ and $P(2,2)$) for the case of

$$\alpha^{-1} = 4363.63, \quad \beta^{-1} = 436.36, \quad P = 0.275$$

are plotted in fig. 3 with respect to the buffer size for three different utilizations. Note that with the parameters above $\rho = 0.025N$ where N is the number of sources. We observe that $P(2,2)$ captures the simulation curve for the three traffic regimes accurately irrespective of the buffer size. $P(1,1)$ overestimates the cell loss rate for small buffer sizes and underestimates that for large buffer sizes but it is still convenient for use for applications that can tolerate a small amount of error in accuracy with the advantage of requiring less computation. We have omitted the $P(l,l)$ approximations with $l > 2$ in the figure due to the fact that they are almost identical to the $P(2,2)$ curve.

We then fix the buffer size to $K = 8$ in the next example and assess the performance of the approximations with respect to the number of users in fig. 4. Two sources are treated, the source of the previous example with the burst length $B = 436.36$ and this source with the mean activity and silence times changed to 1/100 of the previous source ($B = 4.3636$ in this case). While $P(2,2)$ gives accurate results as in the previous examples irrespective of the burst length and the load, we observe an overestimation of $P(1,1)$ for small loads and a slight underestimation for moderate to heavy traffic.

The final example attempts to demonstrate the performance of the approximations for large buffer sizes. We fix $K = 1024$ and we let each individual source (modeled as an IPP) to have the following parameters:

$$\alpha^{-1} = 780, \quad \beta^{-1} = 20, \quad P = 4.$$

Note that each source introduces a 10% load. The results are given in fig. 5 where the cell loss rate is plotted with respect to the number of users. We also change the peak rate of the individual user P to 8 as well as change the mean silence time of the individual user to 1580 and present the associated results. The observation is that for the case of large buffers, both two approximations $P(1,1)$ and $P(2,2)$ capture the simulation curve accurately regardless of the load.

References

[1] J. Abate, G.L. Choudhury and W. Whitt, Asymptotics for steady-state tail probabilities in structured Markov queueing models, Stoch. Models 10(1) (1994).

[2] J. Abate and W. Whitt, The Fourier-series method of inverting transforms of probability distributions, Queueing Systems 10 (1992) 5–88.

[3] J. Abate and W. Whitt, Numerical inversion of Laplace transforms of probability distributions, ORSA J. Comp. 7 (1995) 36–43.

[4] N. Akar, Performance analysis of an asynchronous transfer mode multiplexer with Markov modulated inputs, Ph.D. Thesis, Bilkent University, Ankara, Turkey (1994).

[5] N. Akar and E. Arıkan, Padė approximations in the analysis of the MMPP/D/1 system, *Proc. ITC Spon. Sem.*, Bangalore (1993) pp. 137–143.

[6] N. Akar and K. Sohraby, An invariant subspace approach in M/G/1 and G/M/1 type Markov chains, submitted to Commun. Stat. Stoch. Models.

[7] N. Akar and K. Sohraby, On computational aspects of the invariant subspace approach to teletraffic problems and comparisons, submitted to Commun. Stat. Stoch. Models.

[8] D. Anick, D. Mitra and M.M. Sondhi, Stochastic theory of a data handling system with multiple sources, Bell Syst. Tech. J. 61 (1982) 1871–1894.

[9] S. Asmussen and G. Koole, Marked point processes as limits of Markovian arrival streams, J. Appl. Prob. 30 (1993) 365–372.

[10] A. Baiocchi, Analysis of the loss probability of MAP/G/1/K queue, Part I: Asymptotic theory, Commun. Stat. Stoch. Models 10(4) (1994) 867–8793.

[11] A. Baiocchi and N. Blefari-Melazzi, Analysis of the loss probability of the MAP/G/1/K queue, Part II: Approximations and numerical results, Commun. Stat. Stoch. Models 10(4) (1994) 895–925.

[12] C. Blondia, The N/G/1 finite capacity queue, Commun. Stat. Stoch. Models 5(2) (1989) 273–294.

[13] C.T. Chen, *Linear System Theory and Design* (Holt, Rinehart and Winston, New York, 1984).

[14] G.L. Choudhury, D.M. Lucantoni and W. Whitt, Squeezing the most out of ATM, to appear in IEEE Trans. Commun.

[15] A.I. Elwalid and D. Mitra, Fluid models for the analysis and design of statistical multiplexing with loss priorities on multiple classes of bursty traffic, *INFOCOM'92* (1992) pp. 415–425.

[16] A.I. Elwalid and D. Mitra, Effective bandwidth of general Markovian traffic sources and admission control of high speed networks, IEEE/ACM Trans. Networking 1(3) (1993) 329–343.

[17] G.H. Golub and C.F. Van Loan, *Matrix Computations* (The Johns Hopkins University Press, Baltimore, 1989).

[18] L. Gün, Experimental techniques on matrix-analytical solution techniques – extensions and comparisons, Commun. Stat. Stoch. Models 5(4) (1989) 669–682.

[19] H. Heffes and D.M. Lucantoni, A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, IEEE JSAC 4(6) (1986) 856–868.

[20] I. Ide, Superposition of interrupted Poisson processes and its application to packetized voice multiplexers, *Proc. ITC-12* (1988).

[21] T. Kailath, *Linear Systems* (Prentice-Hall, Englewood Cliffs, NJ, 1980).

[22] L. Kleinrock, *Queueing Systems, Vol. 1: Theory* (Wiley-Interscience, 1975).

[23] A. Kuczura, The interrupted Poisson process as an overflow process, Bell Syst. Tech. J. 52 (1973) 437–448.

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

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

[26] D.M. Lucantoni, G.L. Choudhury and W. Whitt, The transient BMAP/G/1 queue, Commun. Stat. Stoch. Models 10(1) (1994) 145–182.

[27] D.M. Lucantoni, K.S. Meier-Hellstern and M.F. Neuts, A single-server queue with server vacations and a class of non-renewal arrival processes, Adv. Appl. Prob. 22 (1990) 676–705.

[28] C.B. Moler and C. Van Loan, Nineteen dubious ways to compute the matrix exponential, SIAM Rev. 20 (1978) 801–836.

[29] M.F. Neuts, A versatile Markovian point process, J. Appl. Prob. 16 (1979) 764–779.

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

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

[32] V. Ramaswami, Nonlinear matrix equations in applied probability – solution techniques and open problems, SIAM Rev. 30 (1988) 256–263.

[33] H. Saito, M. Kawarasaki and H. Yamada, An analysis of statistical multiplexing in an ATM transport network, IEEE JSAC 9(3) (1991) 359–367.

[34] P. Skelly, M. Schwartz and S. Dixit, A histogram-based model for video traffic behavior in an ATM multiplexer, IEEE/ACM Trans. Networking 1(4) (1993) 446–459.

[35] T.E. Stern and A.I. Elwalid, Analysis of separable Markov-modulated rate models for information-handling systems, Adv. Appl. Probl. 23 (1991) 105–139.

[36] R.C.F. Tucker, Accurate method for analysis of a packet-speech multiplexer with limited delay, IEEE Trans. Commun. 36(4) (1988) 479–483.

[37] J. Vlach and K. Singhal, *Computer Methods for Circuit Analysis and Design* (Van Nostrand Reinhold, New York, 1983).

[38] W. Whitt, Tail probabilities with statistical multiplexing and effective bandwidths for multi-class queues, Telecom. Syst. 2 (1993) 71–107.

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

[40] W.M. Wonham, *Linear Multivariable Control: A Geometric Approach* (Springer, New York, 1974).

[41] J. Ye and S.Q. Li, Analysis of multi-media traffic queues with finite buffer and overload control, Part I: Algorithm, *Proc. IEEE INFOCOM* (1991) pp. 1464–1474.

[42] J. Ye and S.Q. Li, Analysis of multi-media traffic queues with finite buffer and overload control, Part II: Applications, *Proc. IEEE INFOCOM* (1992) pp. 848–859.