11

Buffer requirements in ATM-related queueing models with bursty traffic : an alternative approach.

B. Steyaert
SMACS Research Group, Lab. for Commun. Engineering, University of Ghent
Sint–Pietersnieuwstraat 41, B–9000 Gent, Belgium, bart.steyaert@lci.rug.ac.be

Y. Xiong
INRS – Télécommunications
16 Place du Commerce, Verdun Quebec, Canada H3E1H6, xiong@inrs–telecom.uquebec.ca

Abstract

During the past couple of years, a lot of effort has been put into solving all kinds of Markov Modulated discrete–time queueing models, which occur, almost in a natural way, in the performance analysis of slotted systems, such as ATM multiplexers and switching elements. However, in most cases, the practical application of such solutions is limited, due to the large state space that is usually involved. In this paper, we try to set a first step towards obtaining approximate solutions for a discrete–time multiserver queueing model with a general heterogeneous Markov Modulated cell arrival process, that allow accurate predictions concerning the behavior of the buffer occupancy in such a model, and still remains tractable, both from an analytical and a computational point–of–view. We first introduce a solution technique which leads to a closed–form expression for the joint probability generating function of the buffer occupancy and the state of the arrival process, from which an expression for V(z), the probability generating function of the buffer occupancy is easily derived. Based on this result, for the single–server case, we propose an approximation for the boundary probabilities, that reduces all calculations to an absolute minimum. In addition, we show how accurate data for the distribution of the buffer occupancy can be obtained, by using multiple poles of V(z) in the geometric–tail approximation of the distribution.

Keywords

discrete–time queueing, Markov–modulated models, generating functions, tail distribution, buffer dimensioning.

1 INTRODUCTION

As the basic information units to be transferred in ATM (asynchronous transfer mode) based communication networks are fixed–length cells (De Prycker (1991)), buffers in multiplexers and switches can in general be modeled as a discrete–time queueing system where new cells are generated by a superposition of individual traffic sources. The service time of a cell equals its transmission time, which is one slot. Analyzing such a queueing system is essential in the design and evaluation of ATM networks. However, this can be a difficult task, due to the fact that the traffic sources to be connected to the same buffer may have different traffic characteristics (like voice, data and video), and the time–correlated behavior that each of the individual sources might exhibit.

To facilitate the queueing analysis, a traffic source with variable bit rate (VBR) in ATM is usually modeled as a Markov–modulated arrival process (e.g., Markov–modulated Bernoulli process, Markov–modulated Poisson process, etc...). The problem is then reduced to analyzing a queueing system with heterogeneous Markov–modulated arrival streams generating subsequent cell arrivals. Even so, the related queueing analysis is still complicated and requires solving a multi–dimensional Markov chain. In this heterogeneous traffic environment, the matrix–geometric solution technique

S. Fdida et al. (eds.), *Data Communications and their Performance*

(Neuts (1989)), which has been widely used in the performance analysis of various types of related problems, is only suitable for analyzing small systems because of the large state space (Blondia (1992)).

A general solution technique, called the matrix spectral decomposition method, was developed in Li (1991b) (and extended in Zhang (1991)) to analyze the above queueing system. This solution technique is based on a generating–functions approach and uses the properties of Kronecker products to decompose the problem of solving a global system with multiple traffic types into the problem of solving subsystems, each of which consisting of one single traffic type. Thus, calculating the poles of the probability generating function of the buffer occupancy depends only on the traffic source parameters and is independent of the system size and the number of traffic types. The main computational limitation in this method is the memory size required to solve the set of linear equations for the boundary probabilities (Li (1991b)). Furthermore, in this general solution technique, the superposed arrival processes is expressed in a Kronecker products form. The whole derivation is quite complicated and the final results are not easy to apply. Another main solution technique is the fluid–flow approximation (Anick (1982), Stern (1991)), in which a traffic source is described by a Markov–modulated continuous flow process. It equally uses properties of Kronecker products and sums in the decomposition of the overall problem into smaller 'sub– problems'. Consequently, similar comments as above also hold for the fluid–flow method.

The purpose of this paper is two fold : first to present an alternative solution technique, based on a generating–functions approach, for discrete–time queueing analysis in ATM, and secondly to give a good approximation for the tail distribution of the buffer occupancy, an important performance measure that allows an accurate estimate of the required buffer space, crucial for dimensioning purposes in practical engineering. Compared to the matrix spectral decomposition method, the solution technique to be presented below is relatively simple and has following properties: (1) it uses straightforward analysis, again based on a generating–functions approach, instead of Kronecker products, to represent the superposition of arrival processes; (2) no sophisticated computational matrix manipulations are required and the whole derivation is easy to follow; (3) the final results are relatively easy to use. Furthermore, we found, via comparison of a large number of numerical examples, that the tail distribution of the buffer occupancy can be well approximated when only considering a few poles (i.e., the ones with the smallest modulus) of the probability generating function of the buffer occupancy. This paper is an extension of the work presented in Steyaert (1992).

2 TRAFFIC SOURCE DESCRIPTION

Consider a multiplexer model fed by several independent traffic sources, which, according to their traffic characteristics, are grouped together into K distinct classes or types, each class having N_k, $1 \leq k \leq K$, identical and independent sources. A source belonging to class k is modeled as an L_k–state Markov Modulated arrival process, where the states will be labeled by $S_{i,k}$, $1 \leq i \leq L_k$, and where the $L_k x L_k$ probability generating matrix (as in Sohraby (1992))

$$\mathbf{Q}_k(z) = \begin{bmatrix} q_{11,k}(z) & q_{12,k}(z) & \cdots & q_{1L_k,k}(z) \\ q_{21,k}(z) & q_{22,k}(z) & \cdots & q_{2L_k,k}(z) \\ \vdots & & \ddots & \vdots \\ q_{L_k1,k}(z) & q_{L_k2,k}(z) & \cdots & q_{L_kL_k,k}(z) \end{bmatrix} , \tag{1}$$

characterizes the cell arrival process. It is assumed that transitions between states occur at slot boundaries, and let us denote by $p_{ij,k}$, $1 \leq i,j \leq L_k$, the one–step transition probability that a source from the k–th traffic class transits from state $S_{i,k}$ to state $S_{j,k}$ at the end of a slot during which it was in state $S_{i,k}$. Then, the elements $q_{ij,k}(z)$ in the above matrix are given by

$$q_{ij,k}(z) \triangleq G_{ij,k}(z)\, p_{ij,k} \quad , \tag{2}$$

where $G_{ij,k}(z)$, $1 \leq i,j \leq L_k$, is the probability generating function describing the number of cells generated during a slot by a source from class k, given that the source is in state $S_{j,k}$ during the tagged slot and was in state $S_{i,k}$ during the preceding slot. For the Markov Modulated Bernoulli Process (MMBP), the number of cell arrivals generated by a source during any slot is either zero or equal to one, which is reflected by the property that each of the probability generating functions $G_{ij,k}(z)$ is a linear function of z, meaning that they can be written as

$$G_{ij,k}(z) = 1 - g_{ij,k} + z g_{ij,k} \quad , \tag{3}$$

for some parameters $g_{ij,k}$ satisfying $0 \leq g_{ij,k} \leq 1$. Although attention is focused on this specific arrival model, the theory developed here is far more general, and can also be applied when the $G_{ij,k}(z)$'s have a more complex form than given by (3).

The aggregate cell arrival process is fully determined, once the probability generating matrices $\mathbf{Q}_k(z)$, $1 \leq k \leq K$, have been specified for each individual traffic class. Let us define $e_k(n)$ as the total number of cell arrivals generated by the N_k sources of class k during slot n, and $a_{i,k}(n)$, $1 \leq i \leq L_k$, as the total number of sources of class k that are in state $S_{i,k}$ during slot n. Note that the latter random variables satisfy

$$\sum_{i=1}^{L_k} a_{i,k}(n) = N_k \quad , \tag{4}$$

for any value of n. We will denote by $\boldsymbol{x}_k$ the L_kx1 vector with elements $x_{i,k}$, $1 \leq i \leq L_k$. Let us also define the L_kx1 vector $\boldsymbol{B}_k(\boldsymbol{x}_k,z)$ with elements $B_{i,k}(\boldsymbol{x}_k,z)$, as the matrix product $\mathbf{Q}_k(z)\boldsymbol{x}_k$. Then, with the previous definitions, it is not difficult to show that the joint generating function of the random variables $e_k(n+1)$ and $a_{i,k}(n+1)$, $1 \leq i \leq L_k$, can be written in terms of the joint generating function of the random variables $a_{i,k}(n)$:

$$\mathbf{E}\left[z^{e_k(n+1)} \prod_{i=1}^{L_k} x_{i,k}^{\,a_{i,k}(n+1)} \right] = \mathbf{E}\left[\prod_{i=1}^{L_k} B_{i,k}(\boldsymbol{x}_k,z)^{a_{i,k}(n)} \right] \quad , \tag{5}$$

(where E[.] denotes the expected value of the argument) an important relation that des– cribes the number of cells generated during consecutive slots by the N_k sources of class k.

We define $A_k(\boldsymbol{x}_k)$ as the joint probability generating function of the number of sources of class k in state $S_{i,k}$, $1 \leq i \leq L_k$, during an arbitrary slot in the steady state, i.e.,

$$A_k(\boldsymbol{x}_k) \triangleq \lim_{n\to\infty} \mathbf{E}\left[\prod_{i=1}^{L_k} x_{i,k}^{\,a_{i,k}(n)} \right] = A_k(\mathbf{Q}_k(1)\boldsymbol{x}_k) \quad , \tag{6}$$

assuming that the cell arrival process indeed reaches a stochastic equilibrium, and the latter limit exists. It readily follows from (4,5) that $A_k(\boldsymbol{x}_k)$ indeed should satisfy the above property. Furthermore, if we define $\sigma_{i,k}$ as the steady–state probability that a source of class k is in state $S_{i,k}$ during a slot, and $\boldsymbol{\sigma}_k$ as the L_kx1 column vector with $\sigma_{i,k}$ as its i–th element, which is the solution of the matrix equations

$$\boldsymbol{\sigma}_k^T = \boldsymbol{\sigma}_k^T \mathbf{Q}_k(1) \quad , \quad \boldsymbol{\sigma}_k^T \boldsymbol{1}_k = 1 \quad , \tag{7}$$

($\boldsymbol{1}_k$ is the L_kx1 column vector with all elements equal to 1, and $(.)^T$ represents the matrix transposition operation), then $A_k(\boldsymbol{x}_k)$ equals

$$A_k(\boldsymbol{x}_k) = (\boldsymbol{\sigma}_k^T \boldsymbol{x}_k)^{N_k} \quad , \tag{8}$$

and with (7), it is easily verified that this generating function indeed satisfies (6).

3 THE BUFFER OCCUPANCY : A FUNCTIONAL EQUATION

Due to the extremely low cell–loss ratios that will occur in B–ISDN communication networks, the multiplexer buffer could be considered having infinite storage–capacity, meaning that all arriving cells are accepted and temporarily stored to await their transmission. The multiplexer has c output lines via which cells are transmitted, thus allowing up to a maximum of c cells to leave the multiplexer buffer during each slot. Let us observe the system at the end of a slot (i.e., just after new arrivals, but before departures, if any), say slot n, and denote by the random variable v_n the buffer contents at that time instant; this is the number of cells in the multiplexer buffer, not including the cells that have been transmitted during slot n. From the previous, it is then clear that this quantity evolves according to the system equation

$$v_{n+1} = (v_n - c)^+ + \sum_{k=1}^{K} e_k(n+1) \quad , \tag{9}$$

where $(.)^+ \triangleq \max\{.,0\}$. Since we consider an infinite storage–capacity buffer, the system reaches a stochastic equilibrium only if the equilibrium condition, requiring that the mean number of cell arrivals per slot must be less than c, is satisfied. If we denote by p the mean number of cells carried by each output link per time slot, then, in view of the cell arrival model described in the previous section, it follows that this quantity equals

$$p = \frac{1}{c} \sum_{k=1}^{K} N_k \sum_{i=1}^{L_k} \sigma_{i,k} \sum_{j=1}^{L_k} G'_{ij,k}(1) p_{ij,k} \quad , \tag{10}$$

(where primes denote derivatives with respect to the argument) and $p<1$ is the necessary requirement for reaching a steady state.

The evolution of the (L+1)–th dimensional Markov chain $\{\boldsymbol{a}_k(n) \mid 1 \leq k \leq K\} \cup \{v_n\}$ (where $\boldsymbol{a}_k(n)$, $1 \leq k \leq K$, represents the set of random variables $\{a_{i,k}(n) \mid 1 \leq i \leq L_k\}$, and L is the sum of all L_k's) throughout consecutive slots completely determines the buffer behavior of the discrete–time queueing system previously described. Let us therefore define their joint generating function as

$$P_n(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K, z) \triangleq E\left[z^{v_n} \prod_{k=1}^{K} \prod_{i=1}^{L_k} x_{i,k}^{a_{i,k}(n)} \right] \quad .$$

Combining this definition with system equation (9), together with (5), it follows that

$$P_{n+1}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K, z) = E\left[z^{(v_n - c)^+} \prod_{k=1}^{K} \left[\prod_{i=1}^{L_k} B_{i,k}(\boldsymbol{x}_k, z)^{a_{i,k}(n)} \right] \right] \quad .$$

Again, we assume that the system reaches a steady–state after a sufficiently long period of time (implying that the equilibrium condition $p<1$ must be satisfied), and that $P_n(.)$ has a steady–state limit, which will be denoted by P(.). Then, with the definition of $\boldsymbol{B}_k(\boldsymbol{x}_k, z)$, and using some standard probabilistic techniques, we find that this generating function must satisfy

$$P(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K, z) = z^{-c}\{P(\mathbf{Q}_1(z)\boldsymbol{x}_1, \mathbf{Q}_2(z)\boldsymbol{x}_2, \ldots, \mathbf{Q}_K(z)\boldsymbol{x}_K, z) + R(\mathbf{Q}_1(z)\boldsymbol{x}_1, \mathbf{Q}_2(z)\boldsymbol{x}_2, \ldots, \mathbf{Q}_K(z)\boldsymbol{x}_K, z)\} \quad , \tag{11.a}$$

where $R(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K, z)$ is given by

$$R(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K, z) \triangleq \sum_{j=0}^{c-1} (z^c - z^j) E\left[\prod_{k=1}^{K} \left[\prod_{i=1}^{L_k} x_{i,k}^{a_{i,k}} \right] \middle| v=j \right] \mathrm{Prob}[v=j] \quad . \tag{11.b}$$

In the right–hand side of this equation, the random variable v denotes the buffer contents at the end of an arbitrary slot, while $a_{i,k}$, $1 \leq k \leq K$, $1 \leq i \leq L_k$, is the number of sources of traffic class k that are in state $S_{i,k}$ during this slot. Equations (11.a,b) define a functional equation for P(.), (the joint probability generating function of these random variables), which contains all information concerning the buffer behavior of the queueing model with heterogeneous traffic under study. In the next section, we will describe a technique for solving this functional equation. Also note that the function R(.) contains a number of unknown probabilities, still to be determined. Throughout the following sections, it will also become clear how these unknowns can be computed.

4 SOLVING THE FUNCTIONAL EQUATION

4.1 Homogeneous Traffic

Let us first focus attention on the case where the cell arrival process is homogeneous, i.e., the N traffic sources generating cell arrivals all have the same traffic characteristics, and can be modeled as an L–state MMBP. In the following, the subscript k, $1 \leq k \leq K$, that reflected the traffic class type in the previous sections, will be omitted. The functional equation (11.a,b) then becomes

$$P(\boldsymbol{x},z) = z^{-c}\{P(\mathbf{Q}(z)\boldsymbol{x},z) + R(\mathbf{Q}(z)\boldsymbol{x},z)\} \quad , \tag{12.a}$$

where, $\mathbf{Q}(z)$ is the LxL probability generating matrix describing the arrival process per traffic source, given by (1), and $R(\boldsymbol{x},z)$ now becomes

$$R(\boldsymbol{x},z) \triangleq \sum_{j=0}^{c-1} (z^c - z^j) E\left[\prod_{i=1}^{L} x_i^{a_i} \,\middle|\, v=j\right] \text{Prob}[v=j] = \sum_{j=0}^{c-1} (z^c - z^j) \sum_{\boldsymbol{\ell}} \left[\prod_{i=1}^{L} x_i^{\ell_i}\right] p(\boldsymbol{\ell},j) \quad , \tag{12.b}$$

where $\boldsymbol{a} \triangleq \{a_i \mid 1 \leq i \leq L\}$, is the set of ramdom variables representing the number of input sources in state S_i during an arbitrary slot, and where

$$p(\boldsymbol{\ell},j) \triangleq \text{Prob}[\boldsymbol{a} = \boldsymbol{\ell}, v=j] = \text{Prob}[a_1=\ell_1, \dots ,a_L=\ell_L,v=j] \quad . \tag{12.c}$$

In (12.b,c), $\boldsymbol{\ell}$ is the set of positive integers $\{\ell_i \mid 1 \leq i \leq L \text{ and } \ell_i \geq 0\}$ that satisfies

$$\sum_{i=1}^{L} \ell_i = N \quad , \tag{12.d}$$

and the sum for $\boldsymbol{\ell}$ includes all possible sets $\{\ell_i \mid 1 \leq i \leq L \text{ and } \ell_i \geq 0\}$.

The matrix $\mathbf{Q}(z)$ will be diagonizable under quite general circumstances, and let $\lambda_i(z)$ be the i–th eigenvalue of $\mathbf{Q}(z)$, and $\boldsymbol{w}_i(z)$ ($\boldsymbol{u}_i(z)$) the left row (right column) eigenvector of $\mathbf{Q}(z)$ with respect to $\lambda_i(z)$. Define the diagonal eigenvalue matrix

$$\Lambda(z) \triangleq \text{diag}[\lambda_1(z), \lambda_2(z), \dots ,\lambda_L(z)]$$

and eigenvector matrices

$$\mathbf{W}(z) \triangleq (\boldsymbol{w}_1(z)\ \boldsymbol{w}_2(z)\ \dots\ \boldsymbol{w}_L(z))^T \quad , \quad \mathbf{U}(z) \triangleq (\boldsymbol{u}_1(z)\ \boldsymbol{u}_2(z)\ \dots\ \boldsymbol{u}_L(z)) \quad . \tag{13.a}$$

From this definition, we have

$$\Lambda(z)\,\mathbf{W}(z) = \mathbf{W}(z)\,\mathbf{Q}(z) \quad , \quad \mathbf{U}(z)\,\Lambda(z) = \mathbf{Q}(z)\,\mathbf{U}(z) \quad . \tag{13.b}$$

For each value of i, $1 \leq i \leq L$, equation (13.b) determine the left row eigenvector and the right column eigenvector of $\mathbf{Q}(z)$ corresponding to $\lambda_i(z)$ upon some constant factor, which is uniquely determined when requiring that

$$U(z)\,\mathbf{1} = \mathbf{1} \quad \text{and} \quad \mathbf{W}(z)\,U(z) = \mathbf{I} \Rightarrow \mathbf{W}(z)\,\mathbf{1} = \mathbf{1} \quad , \tag{13.c}$$

where **I** is the LxL identity matrix, and as defined before, $\mathbf{1}$ is the Lx1 column vector with all elements equal to 1. Equation (13.c) implies that $\mathbf{Q}(z)$ can be written as

$$\mathbf{Q}(z) = U(z)\,\Lambda(z)\,W(z) = \sum_{i=1}^{L} \lambda_i(z)\,\boldsymbol{u}_i(z)\,\boldsymbol{w}_i(z) \quad . \tag{13.d}$$

Let us now go back to equation (12.a), from which we can derive that

$$P(\boldsymbol{x},z) = z^{-Hc}P(\mathbf{Q}(z)^H\boldsymbol{x},z) + \sum_{h=1}^{H} z^{-hc}R(\mathbf{Q}(z)^h\boldsymbol{x},z) \quad , \tag{14}$$

From (13.c,d), it is clear that $\mathbf{Q}(z)^h\boldsymbol{x} = U(z)\Lambda(z)^hW(z)\boldsymbol{x}$. Then, letting H approach infinity, in a similar way as was explained in Steyaert (1992), the right hand side of the above equation can be further worked out, leading to an expression for $P(\boldsymbol{x},z)$, the joint generating function of the buffer occupancy at the end of an arbitrary slot, and $\boldsymbol{a}$, the set of random variables describing the L–state MMBP arrival process

$$P(\boldsymbol{x},z) = \sum_{\boldsymbol{m}} \frac{\prod_{i=1}^{L}\left\{\lambda_i(z)\,\boldsymbol{w}_i(z)\,\boldsymbol{x}\right\}^{m_i}}{z^c - \prod_{i=1}^{L}\lambda_i(z)^{m_i}} \sum_{\boldsymbol{\ell}} F_{\boldsymbol{\ell m}}(z) \sum_{j=0}^{c-1}(z^c - z^j)\,p(\boldsymbol{\ell},j) \quad . \tag{15}$$

The functions $F_{\boldsymbol{\ell m}}(z)$ are defined by the relation

$$\prod_{i=1}^{L}\left[\sum_{j=1}^{L} u_{ij}(z)\,x_j\right]^{\ell_i} \triangleq \sum_{\boldsymbol{m}} F_{\boldsymbol{\ell m}}(z)\left[\prod_{i=1}^{L} x_i^{m_i}\right] \quad , \tag{16}$$

where, similarly to $\boldsymbol{\ell}$, $\boldsymbol{m}$ represents a set of positive integers $\{m_i \mid 1 \le i \le L \text{ and } m_i \ge 0\}$ that satisfy (12.d)), and where $u_{ij}(z)$ is the i–th element of the column vector $\boldsymbol{u}_j(z)$. These functions can be calculated in terms of the $u_{ij}(z)$'s by identifying the appropriate coefficients in both hand sides of the above equation. It is worth noting that, in Section 5.2, we propose an approximation for the boundary probabilities, which in the mean time avoids the calculation of these functions.As in most applications, we are mainly interested in the distribution of the buffer occupancy, or, equivalently, the probability generating function V(z) of the buffer occupancy. Since V(z) equals $P(\mathbf{1},z)$, we find

$$V(z) = \sum_{\boldsymbol{m}} \frac{\prod_{i=1}^{L}\lambda_i(z)^{m_i}}{z^c - \prod_{i=1}^{L}\lambda_i(z)^{m_i}} \sum_{\boldsymbol{\ell}} F_{\boldsymbol{\ell m}}(z) \sum_{j=0}^{c-1}(z^c - z^j)\,p(\boldsymbol{\ell},j) \quad , \tag{17}$$

where we have used the property $\boldsymbol{w}_i(z)\mathbf{1} = 1$, which follows from (13.c). This expression for the probability generating function of the buffer occupancy at the end of an arbitrary slot still contains the unknown probabilities $p(\boldsymbol{\ell},j)$. These can be calculated by exploiting the property that V(z) is analytic inside the complex unit disk, which implies that the zeros inside the unit disk of the denominators in the right–hand side of (17) must also be zeros of the numerators. It can be shown that each denominator in (17) has c zeros inside the unit disk, and we thus find a total of $J = c.(N+L-1)i!/(N!(L-1)!)$ zeros inside the unit disk (including z=1, which leads to no additional equation for the unknowns).

Together with the normalization condition V(1)=1, in general, this is the number of linear equations we obtain for the same number of unknown probabilities, and this set of linear equations has a unique solution.

Once these unknowns have been calculated, all major characteristics concerning the buffer occupancy, such as mean value, variance, and tail distribution, can be calculated from (17). In this paper, we concentrate our efforts mainly on the tail distribution, which plays an important role in buffer dimensioning. However, as one observes from the value of J, the number of unknown probabilities can become quite large, as N and L increase, thus requiring solving a large set of linear equations. In order to avoid this, in Section 5, we discuss some techniques for approximating these unknown probabilities, that lead to accurate estimates of the tail probabilities, as will be shown by various numerical examples.

4.2 Heterogeneous Traffic

The derivation of the probability generating function in the case of heterogeneous traffic evolves along similar lines as in the case of homogeneous traffic, and adds no particularly new insights to the analysis. The final result for V(z), the probability generating function of the buffer occupancy at the end of an arbitrary slot, can be written as

$$V(z)=\sum_{\boldsymbol{m}_1}..\sum_{\boldsymbol{m}_K}\left[\frac{\prod_{k=1}^{K}\prod_{i=1}^{L_k}\lambda_{i,k}(z)^{m_{i,k}}}{z^c-\prod_{k=1}^{K}\prod_{i=1}^{L_k}\lambda_{i,k}(z)^{m_{i,k}}}\right]\sum_{\boldsymbol{\ell}_1}..\sum_{\boldsymbol{\ell}_K}\left\{\prod_{k=1}^{K}F_{\boldsymbol{\ell}_k\boldsymbol{m}_k}(z)\right\}\sum_{j=0}^{c-1}(z^c-z^j)p(\boldsymbol{\ell}_1..\boldsymbol{\ell}_K,j)\ . \tag{18}$$

As before, $\boldsymbol{\ell}_k$ ($\boldsymbol{m}_k$), $1 \le k \le K$, represents the set of positive integers $\{\ell_{i,k} | 1 \le i \le L_k$ and $\ell_{i,k} \ge 0\}$ ($\{m_{i,k}| 1 \le i \le L_k$ and $m_{i,k} \ge 0\}$) and the sums in (18) for $\boldsymbol{\ell}_k$, ($\boldsymbol{m}_k$) include all such sets that, as a consequence of (4), satisfy

$$\sum_{i=1}^{L_k}\ell_{i,k} = N_k\ ,\quad \sum_{i=1}^{L_k}m_{i,k} = N_k\ . \tag{19.a}$$

Similarly to the homogeneous cell–arrivals case, $\{\lambda_{i,k}(z) \mid 1 \le i \le L_k\}$ is the set of eigenvalues of $\mathbf{Q}_k(z)$ (defined in (1)), and $\boldsymbol{\Lambda}_k(z)$ is the $L_k \times L_k$ diagonal matrix with $\lambda_{i,k}(z)$ on the intersection of the i–th row and the i–th column. In addition, $\boldsymbol{u}_{i,k}(z)$, $1 \le i \le L_k$, represent the right column eigenvectors corresponding to $\lambda_{i,k}(z)$, which is the i–th column of $\mathbf{U}_k(z)$, the $L_k \times L_k$ matrix that can be calculated from

$$\mathbf{U}_k(z)\,\boldsymbol{\Lambda}_k(z) = \mathbf{Q}_k(z)\,\mathbf{U}_k(z) \quad \text{and} \quad \mathbf{U}_k(z)\boldsymbol{1}_k = \boldsymbol{1}_k\ . \tag{19.b}$$

Finally, extending (16), the $F_{\boldsymbol{\ell}_k\boldsymbol{m}_k}(z)$'s that occur in (18) are implicitly defined by

$$\prod_{i=1}^{L_k}\left[\prod_{j=1}^{L_k}u_{ij,k}\,x_j\right]^{\ell_{i,k}} \triangleq \sum_{\boldsymbol{m}_k}F_{\boldsymbol{\ell}_k\boldsymbol{m}_k}(z)\left[\prod_{i=1}^{L_k}x_i^{m_{i,k}}\right]\ , \tag{19.c}$$

(with $u_{ij,k}(z)$ the j–th element of $\boldsymbol{u}_{ij,k}(z)$) and can be obtained in terms of the $u_{ij,k}(z)$'s by identifying the appropriate coefficients in both hand sides of this expression. The unknown probabilities

$$p(\boldsymbol{\ell}_1..\boldsymbol{\ell}_K,j) \triangleq \mathrm{Prob}[\boldsymbol{a}_1=\boldsymbol{\ell}_1,..,\boldsymbol{a}_K=\boldsymbol{\ell}_K,v=j]\ , \tag{19.d}$$

(where $\boldsymbol{a}_k$ represents the set of random variables $\{a_{i,k} \mid 1 \le i \le L_k\}$, $a_{i,k}$ being the number of traffic sources of class k in state $S_{i,k}$ during an arbitrary slot) that occur in the right–hand side of (18), can be calculated by expressing that the zeros inside the unit disk of the denominators must also be zeros of the numerators. In general, this will involve solving a set of

$$J \triangleq c \prod_{k=1}^{K} \begin{bmatrix} N_k+L_k-1 \\ L_k-1 \end{bmatrix} , \tag{20}$$

linear equations for the same number of unknowns.

5. TAIL DISTRIBUTION OF THE BUFFER OCCUPANCY

In this section, we consider the tail distribution of the buffer occupancy, a performance measure of considerable interest for dimensioning purposes. We try to establish an approximation for the tail distribution of the buffer occupancy, that is both accurate, and easy to calculate, from a computational point–of–view.

5.1 The Multiple Poles Approximation

It has been observed in many cases that approximating the tail distribution of the buffer contents by a geometric form is quite accurate, if the poles of V(z) have a different modulus and multiplicity equal to one, which is, in general, the case. As in Steyaert (1992), we improve this kind of approach by considering a mixture of geometric terms in the approximation for the tail distribution of the buffer contents. In particular, in order to obtain accurate results, we claim that, in a first approximation, it is sufficient to merely consider multiple real and positive poles of V(z) in the series expansion of this function. Approximating the distribution of the buffer contents by a mixture of geometric terms (say M) corresponds to approximating V(z) by

$$V(z) \cong \sum_{m=1}^{M} \frac{\theta_m}{z - z_{o,m}} = - \sum_{m=1}^{M} z_{o,m}^{-1} \theta_m \sum_{s=0}^{\infty} \left[\frac{z}{z_{o,m}}\right]^s , \tag{21.a}$$

where we are particularly interested in sufficiently large values of s. In all cases considered further on, $z_{o,m}$, $1 \leq m \leq M$, are the M real and positive poles of V(z) with smallest modulus (which, of course, lay outside the unit disk).

The poles of V(z) correspond to the zeros outside the unit disk of the denominators in the right–hand side of expression (18) for V(z). Depending on the arrival model, the exact number of zeros outside the unit disk of each of the denominators varies, and calculating all the zeros can become a complicated numerical task. Nevertheless, for a wide variety of arrival models (among which those considered in Section 5.3), in all cases it has been observed that, the denominators which occur in the right–hand side of (18), in general, have a real and positive zero outside the unit disk, which, of course is a pole of V(z). The above expression for V(z) leads to the following approximation for the tail distribution of the system contents :

$$\mathrm{Prob}[v>s] \cong - \sum_{m=1}^{M} \frac{\theta_m z_{o,m}^{-s-1}}{z_{o,m} - 1} , \quad s \geq 0 , \tag{21.b}$$

and this approximation improves for increasing values of s and M. Furthermore, using the residue theorem, the quantity $\theta_{\{m\}}$ in (21.a,b) can be shown to be equal to

$$\theta_m = \lim_{z \to z_{o,m}} (z-z_{o,m}) V(z) . \tag{21.c}$$

From expression (18) for V(z) and using de l'Hôpitals rule, these quantities can be easily calculated. The accuracy of the approximation for the buffer contents distribution presented here will be confirmed in Section 5.3 through comparison with the exact distribution.

5.2 Boundary Probabilities Approximation

As became clear in Section 4, a drawback of the technique presented here is the potentially huge number of boundary probabilities that must be calculated. Therefore, it is essential to find good approximations for these quantities. One possible approach for this problem in the single–server case is presented in this section. Denote by e the random variable describing the number of cell arrivals during a slot, whereas v, as before, indicates the buffer contents at the end of this slot. Using similar notations as in (19.d), let us also define the joint probability

$$q(\ell_1 \cdot\cdot \ell_K, 0) \triangleq \text{Prob}[\boldsymbol{a}_1=\ell_1, \cdot\cdot, \boldsymbol{a}_K=\ell_K, e=0] \quad .$$

Obviously, v=0 implies that there have been no cell arrivals during the tagged slot, i.e., $v=0 \Rightarrow e=0$. Consequently, it is clear that the following inequality between the latter quantities and the unknown probabilities $p(\ell,0)$ holds :

$$q(\ell_1 \cdot\cdot \ell_K, 0) > p(\ell_1 \cdot\cdot \ell_K, 0) \quad .$$

In the next section, we will show through some numerical examples that approximating the conditional unknown probabilities by

$$p(\ell_1 \cdot\cdot \ell_K, 0)/(1-p) \cong q(\ell_1 \cdot\cdot \ell_K, 0)/\text{Prob}[e=0] \quad , \tag{22}$$

when calculating the tail probabilities of the buffer contents, yields an excellent upper bound for the latter quantities.

The values of the q(.,0)'s could be calculated from the traffic parameters. Indeed, combining the steady–state limit of (5) with (8), and using the statistical independence of different sources, we obtain that

$$\sum_{\ell_1} \cdot\cdot \sum_{\ell_K} \left\{ \prod_{k=1}^{K} \prod_{i=1}^{L_k} x_{i,k}^{\ell_{i,k}} \right\} q(\ell_1 \cdot\cdot \ell_K, 0) = \prod_{k=1}^{K} (\sigma_k^T \, Q_k(0) \, \boldsymbol{x}_k)^{N_k} \quad , \tag{23.a}$$

and where Prob[e=0] which occurs in (22) obviously satisfies

$$\text{Prob}[e=0] = \prod_{k=1}^{K} (\sigma_k^T \, Q_k(0) \, \boldsymbol{1}_k)^{N_k} \quad . \tag{23.b}$$

An additional advantage of the approximation proposed in this section, is that it avoids the explicit calculation of the $q(.\ell)$'s, as well as the calculation of the functions $F_{\ell_k \boldsymbol{m}_k}(z)$ from (16) that occur in expression (18) for V(z) (which will be reflected in the calculation of the constants θ_m in (21.c)). From definition (19.c), it is not difficult to show, with the q(.,0)'s satisfying (23.a), that

$$\sum_{\ell_1} \cdot\cdot \sum_{\ell_K} \left\{ \prod_{k=1}^{K} F_{\ell_k \boldsymbol{m}_k}(z) \right\} q(\ell_1 \cdot\cdot \ell_K, 0) = \prod_{k=1}^{K} C_{\boldsymbol{m}_k}^{N_k} \prod_{i=1}^{L_k} (\sigma_k^T Q_k(0) \boldsymbol{u}_{i,k}(z))^{m_{i,k}} \quad , \tag{24.a}$$

where $\boldsymbol{u}_{i,k}(z)$, as already mentioned, is the right column eigenvector with respect to $\lambda_{i,k}(z)$, which is obtained from solving (19.b), and where

$$C_{\boldsymbol{m}_k}^{N_k} \triangleq \frac{N_k}{m_{1,k}! \, \cdots \, m_{L,k}!} \quad . \tag{24.b}$$

It is clear that (24.a) considerably reduces the numerical calculations, when using approximation (22) for the boundary probabilities in the evaluation of (21.c).

5.3 Numerical Examples

In this subsection, we compare the tail approximations derived above with the exact buffer contents distribution. The numerical examples to be shown below are based on the MMBP traffic model, especially the well–studied 2–state MMBP. The exact buffer contents distribution is obtained by just using the simple repeated substitution algorithm.

The homogeneous MMBP traffic model with L states and one–step transition probabilities p_{ij}, $1 \leq i,j \leq L$, has been described in Section 2. When a source is in state S_j, it will generate either one or no cell, with probabilities g_{ij} and $1-g_{ij}$ respectively. It is thus clear that the sojourn time (in slots) of a source in state S_{ij} is geometrically distributed, with mean value $T_i = 1/(1-p_i)$, $1 \leq i \leq L$.

For simplicity, we consider the following special case in our numerical examples : (1) upon leaving state S_i, the source will transit to the other states with equal probability, i.e., $p_{ij} = (1-p_{ii})/(L-1)$, if $j \neq i$; (2) the number of cells sent by the source during a slot only depends on the source state in this slot and is independent of the source state in the previous slot, i.e., $g_{ij} \equiv g_j$. It is further assumed without loss of generality that $g_j \geq g_i$ if $j \geq i$, for all $1 \leq i,j \leq L$.

So the MMBP traffic model we are going to use can be completely described by the mean sojourn time T_i and the average cell arrival rate g_i in state S_i ($1 \leq i \leq L$). In this case the steady–state probability of a traffic source being in state S_i is equal to

$$\sigma_i = T_i / \left[\sum_{i=1}^{L} T_i \right] .$$

From (10), the average traffic load on each outgoing link can be written as

$$p = \frac{N}{c} \left[\sum_{i=1}^{L} \sigma_i g_i \right] .$$

Note that in the following examples, we always take $g_1=0$, which implies that no cells are sent during the state 1 period. For the 2–state MMBP with $g_1=0$, g_2 is usually called the "mean peak rate" and T_2 is the "average burst length".

Now consider a queueing system fed by N identical MMBP traffic sources as described above. It is clear that when the number of states of each source L=2, the queueing performance is determined by the parameter set (N, p, c, g_2, T_2), as $g_1=0$. Let us first concentrate on the single–server case (c=1) and look at the impact of different parameters on the tail approximations. Fig. 1 compares the exact buffer contents distribution with its tail approximations for the traffic load p=0.4 and 0.8. The tail approximations are calculated using (21.b,c), where the boundary probabilities can be derived by solving the set of linear equations, obtained when expressing that the zeros of the denominators in the right hand side of (17) inside the unit disk are also zeros of the numerators. For this arrival model, we found that there are in total N–1 positive poles of V(z), the probability generating function of the buffer contents. One can observe from Fig. 1 that for high traffic load, the tail distribution can be well approximated by the geometric term of the smallest pole of V(z) (i.e., M=1), which will be referred to as the asymptotic queueing behavior. However, for low traffic loads, it is necessary to add more geometric terms corresponding to larger poles of V(z) in order to approximate the tail distribution more accurately. Fig. 1 illustrates that M=5 geometric terms are sufficient in approximating the tail distribution in the region of low probabilities (e.g., <10E–6) of interest. Of course, increasing the number of geometric terms M will eventually lead to better approximation in the high probability region (see M=10). In general, we found that the whole distribution of the buffer contents, except for small buffer contents (e.g., < 10 cells), can be accurately approximated by taking into account all the positive poles of V(z) (in this case M=19).

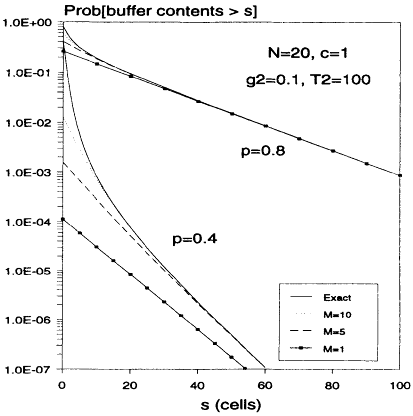

Fig. 1 : Buffer–contents distr. and approx.; traffic load p=0.4 and 0.8.

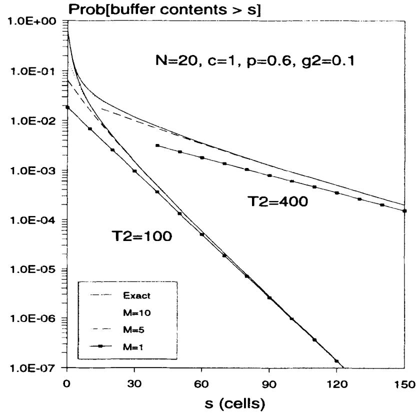

Fig. 2 : Buffer–contents distr. and approx.; mean burst length T_2=100 and 400.

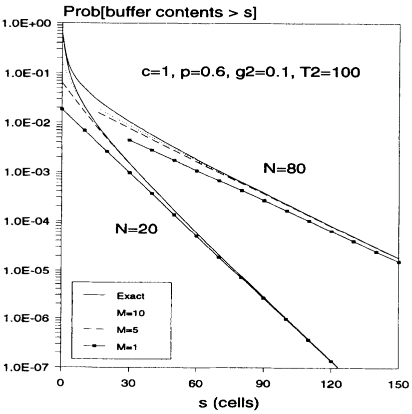

Fig. 3 : Impact of the number of traffic sources on the tail approximations.

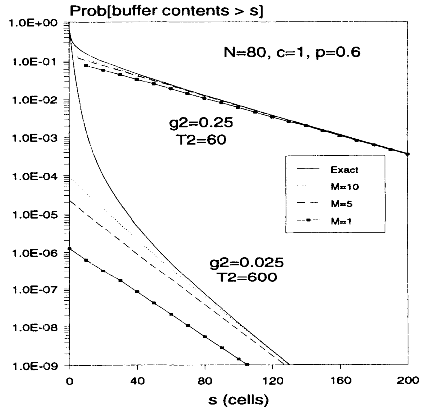

Fig. 4 : Impact of the mean peak rate on the tail approximations.

Similar results can also be observed when changing the values of the parameters. Fig. 2 shows an example where the average burst length T_2=100 and 400 slots, respectively. Fig. 3 gives another example in which the number of traffic sources N=20 and 80. It is interesting to see from the latter figure that although the total number of positive poles of V(z) (= N–1) increases linearly with N, the number of sources, the tail distribution of the buffer contents seems be dominated by a few geometric terms related to the smallest positive poles of V(z). The impact of the mean peak rates g_2 on the tail approximations is illustrated in Fig. 4. This figure reveals that for rather low source peak rate, more geometric terms might be required to get accurate approximations for the tail distribution. From Figs. 1–4, we also see that the tail distribution of the buffer contents cannot always be well approximated by only taking into account its asymptotic behavior.

The above results based on the 2–state MMBP's and single server case (c=1) also hold for an L–state MMBP's (L>2) and the multiple servers case (c>1). Fig. 5 shows an example when multiplexing of N=10 identical 3–state MMBP traffic sources. The buffer–contents distribution of a multiserver (c=4) queueing system with 2–state MMBP

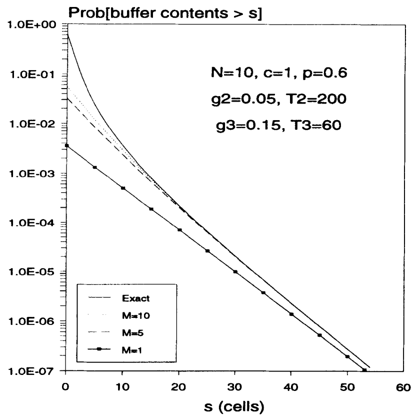

Fig. 5 : Tail approx. for a multiplex of 3–state MMBP traffic sources.

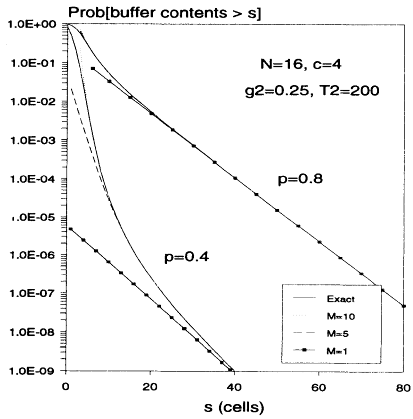

Fig. 6 : Tail approximations in the multiserver case.

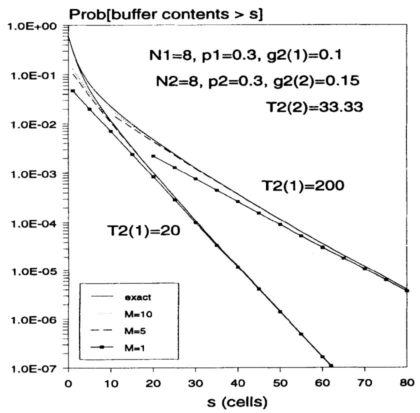

Fig. 7 : Heterogeneous traffic : Buffer–contents distribution and its approximations for $T_2(1) = 20$ and 200.

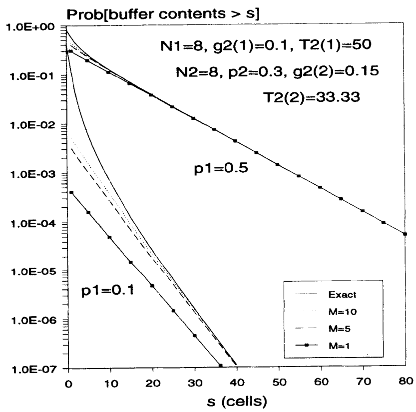

Fig. 8 : Heterogeneous traffic : Buffer–contents distribution and its approximations for $p_1 = 0.1$ and 0.5.

traffic sources is compared with its tail approximations in Fig. 6. Note that this multiserver queueing system can be used to model an output port of an ATM switching element, which has been used as a building block to construct a large ATM switching network (Henrion (1990, 1993)).

The above traffic descriptors for the homogeneous traffic case, are also well suited for describing heterogeneous traffic. Focusing attention on the case of a single–server queue fed by 2–state MMBP heterogeneous arrivals, the parameter set (N_k, p_k, $g_2(k)$, $T_2(k)$), $1 \leq k \leq K$, can then be used for characterizing each of the K traffic classes. Setting K=2, we have plotted some results in Figs. 7–8 for changing values of the average burst length (Fig. 7) and the offered load (Fig. 8) of the first traffic class, while keeping the traffic parameters of the other class constant. The conclusions that can be drawn here are basically identical to the case of homogeneous traffic : (1) considering even a relatively small number of terms (for instance M=5 or 10) in the geometric–tail approximation already leads to very accurate results as far as the tail behavior of the buffer contents

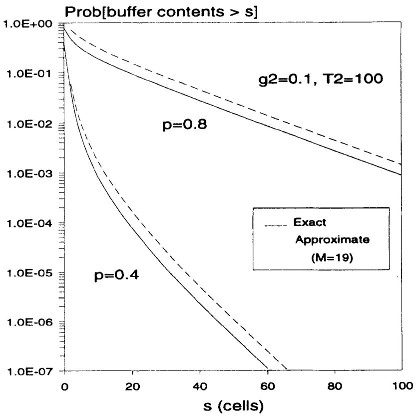

Fig. 9 : Upper–bound tail approximations for different traffic loads.

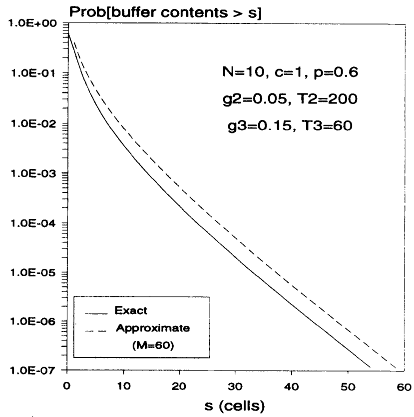

Fig. 10 : Upper–bound tail approximation for 3–state MMBP traffic sources.

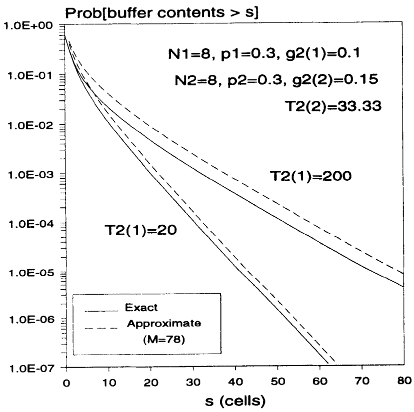

Fig. 11 : Heterogeneous traffic : upper–bound tail approximation.

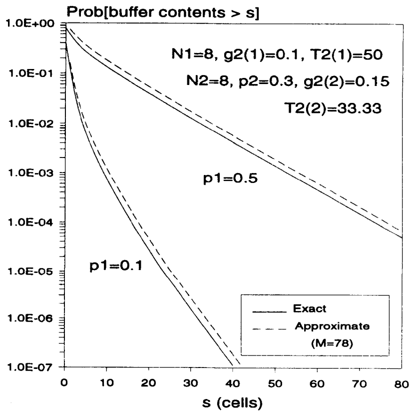

Fig. 12 : Heterogeneous traffic : upper–bound tail approximation.

distribution is concerned; (2) it is advisable to consider multiple terms in the above mentioned tail approximation, especially when the offered load is low.

As we discussed before, for large real systems with heterogeneous traffic, it is infeasible to obtain the unknown boundary probabilities due to the huge memory space requirement. Finding good approximations for the boundary probabilities is thus very important in ATM queueing analysis. In Section 5.2, we proposed a simple approximation for the boundary probabilities, from which all the geometric terms can be easily calculated. We found via numerous numerical results that this approximation leads to a good upper bound for the tail distribution of the buffer contents. Examples for homogeneous 2–state and 3–state MMBP's are shown in Figs 9 and 10, which use the same parameters as in Figs. 1 and 5 respectively; the heterogeneous arrivals case is illustrated in Figs. 11 and 12, with parameter sets that are identical as in Figs 7 and 8. In these figures, all the geometric terms related to the positive poles of V(z) are taken into account. An important property one observes from these curves is that the slopes of the tail distribution and its upper–bound approximation are identical. This is because they both contain the same positive poles of V(z), and merely differ in the value of the θ_m's (see (21.a–c)). Furthermore, the observed differences between exact and approximate results are small, thus leading to the conclusion that the approximation method proposed in Section 5.2 yields sufficiently accurate results.

6 CONCLUDING REMARKS

In this paper, we have presented an alternative solution technique for analyzing discrete–time queueing systems with general heterogeneous Markov–modulated arrival processes, which is relatively simple and easy to use compared to the matrix spectral decomposition method. We found via numerous numerical results that the tail distribution of the buffer contents can be well approximated by using only a few geometric terms related to the smallest positive poles of V(z), the probability generating function of the buffer contents. Moreover, an approximation for the boundary probabilities is given in the single server case, from which a good upper bound for the tail distribution is obtained, which is one of the main contributions of the paper. This upper bound is certainly quite useful in practical engineering (e.g., buffer dimensioning), and the calculation of this result is not limited by the system size and/or the number of traffic types.

Regarding the tail approximations, the main difficulty in this solution technique (as well as in the other methods) is the calculation of the poles of V(z) when the number of states L of each multiplexed source gets large (e.g., L>3). Finding an efficient way to calculate the poles of V(z) is one of the issues currently under study. As an initial result, we found a simple algorithm to calculate the smallest pole of V(z) for large value of L (Xiong (1994)). This smallest pole determines the asymptotic behavior of the tail distribution. Another issue that needs further investigation is finding efficient approximations for the boundary probabilities, in particular in the multiple servers case.

REFERENCES.

Anick, D.; Mitra, D. and Sondhi, M.M (1982) "Stochastic theory of a data– handling system with multiple sources", *Bell Syst. Tech. J.*, vol. 61, no. 8, pp. 1871–1894.

Blondia, C. and Casals, O., (1992) "Statistical multiplexing of VBR sources : A matrix–analytic approach", *Performance Evaluation*, vol. 16, pp. 5–20.

De Prycker, M. (1991) *Asynchronous Transfer Mode – Solution for Broadband ISDN*, Ellis Horwood Ltd. (England).

Henrion, M.; Schrodi, K.; Boettle, D.; De Somer, M. and Dieudonné, M. (1990) "Switching network architecture for ATM based broadband communications", Proc. of *ISS'90* (Stockholm), vol. V, pp. 1–8.

Henrion, M.; Eilenberger, G.; Petit, G. and Parmentier, P. (1993) "A multipath self–routing switch", *IEEE Communications Magazine*, pp. 46–52.

Hirano, M. and Watanabe, N. (1990) "Traffic characterization and a congestion control scheme for an ATM network", *Int. J. Digit. and Analog Commun. Syst.*, vol. 3, pp. 211–217.

Le Boudec, J.–Y. (1991) "An efficient solution method for Markov models of ATM links with loss priorities", *IEEE J. Select. Areas Commun.*, vol. 9, no. 3, pp. 408–417.

Li, S. and Sheng, H. (1991a) "Discrete queueing analysis of multi–media traffic with diversity of correlation and burstiness properties", Proc. of *IEEE INFOCOM'91*, pp. 368–381.

Li, S. (1991b) "A general solution technique for discrete queueing analysis of multimedia traffic on ATM", *IEEE Trans. Commun.*, vol. 39, no. 7, July 1991, pp. 1115– 1132.

Neuts, M.F. (1989) *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker Inc. (New York).

Sohraby, K. (1992) "Heavy traffic multiplexing behavior of highly–bursty heterogeneous sources and their admission control in high–speed networks", Proc. of *IEEE GLOBECOM'92* (Orlando), pp. 1518–1523.
Stavrakakis, I., (1991) "Efficient modeling of merging and splitting processes in large networking structures", *IEEE J. Select. Areas Commun.*, vol. 9, no. 8, pp. 1336–1347.
Stern, and Elwalid, A.I. (1991) "Analysis of separable Markov–modulated rate models for information–handling systems", *Adv. Appl. Prob.*, vol. 23, pp. 105–139.
Steyaert, B. and Bruneel, H. (1992) "Analysis of the buffer behavior of an ATM switch with cell arrivals generated by a general correlated process : a novel approach", Proc. of *Digital Communication Networks Management Seminar* (St. Petersburg), pp. 199–217.
Xiong, Y. and Bruneel, H. (1993) "A tight upper bound for the tail distribution of the buffer contents in statistical multiplexers with heterogeneous MMBP traffic sources", Proc. of *IEEE GLOBECOM'93* (Houston), pp. 767–771.
Xiong, Y. and Bruneel, H. (1994) "On the asymptotic behavior of discrete–time single–server queueing systems with general Markov–modulated arrival processes", Proc. of *ITC'14* (Antibes Juan–les–Pins), pp. 179–189.
Zhang, Z. (1991) "Analysis of a discrete–time queue with integrated bursty inputs in ATM networks", *Int. J. Digit. and Analog Commun. Syst.*, vol. 4, pp. 191–203.

BIOGRAPHY

Bart Steyaert was born in Roeselare, Belgium, in 1964. He received the degrees of Licentiate in Physics and Licentiate in Computer Science in 1987 and 1989, respectively, all from the University of Ghent, Gent, Belgium. Since 1990, he has been working as a PhD student at the Laboratory for Communications Engineering as a member of the SMACS Research Group. His main research interests include performance evaluation of discrete–time queueing models, applicable in B–ISDN networks.

Yijun Xiong received the B.S. and M.S. degrees from Shangai Jiao–Tong University, China, in 1984 and 1987 respectively, and the Ph.D. degree from the University of Ghent, Belgium, in 1994, all in electrical engineering.

He is now a postdoctoral fellow at INRS–Telecommunications, Montreal, Canada. Before this, he worked at the Laboratory for Communications Engineering, University of Ghent, from Sept. 1992 to June 1994. From Aug 1989 to Aug 1992, he was with the Research Center of Alcatel Bell Telephone, Antwerp, Belgium. His current research interests include multicast ATM switches, traffic control, and recourse management of high speed networks.