D-BMAP Models For
Performance Evaluation
Of ATM Networks

John A. Silvester, Nelson L.S. Fonseca
and Stanley S. Wang

Department of Electrical Engineering - Systems
University of Southern California
Los Angeles, California 90089-2562
(213)740-4579

D-BMAP Models for Performance Evaluation of ATM Networks

John A. Silvester, Nelson L. S. Fonseca and Stanley S. Wang

Department of Electrical Engineering-Systems
University of Southern California
Los Angeles, CA 90089-2562

Abstract

The fixed cell size used in ATM networks naturally leads to the use of discrete-time models for performance evaluation. In this paper, we describe our work on estimating network performance based on *Discrete-time Batch Markovian Arrival Process* (D-BMAP) models. Many recent studies make use of multi-state Markov modulated arrival processes to model different types of multimedia traffic. In the continuous time domain, the performance of an ATM multiplexer loaded with this kind of traffic has been approximated by reducing the arrival processes to two-state models. We extended these results to the discrete-time domain with accurate results and fast computation. We are also interested in network-wide performance evaluation and propose models that make use of similar two-state Markov models to represent both external and internal traffic. For given input characteristics, we determine the parameters of the output process for a switch and also specify how to modify the parameters when splitting and joining of flows occurs. This allows us to model the network-wide performance (delay and loss). Comparison with simulation shows good agreement, which suggests that these fast models have potential application to real-time traffic management as well as network design procedures.

1 Introduction

Multi-state Markov modulated arrival processes have been widely used in the literature to model different types of traffic such as the *on-off process* used in [5] for voice sources and the one-dimensional and two-dimensional *Markov chains* used in [12] and [17] respectively for video sources. In [19], we proposed a two-state *Discrete-time Batch Markovian Arrival Process* (D-BMAP) as an approximation for the aggregation of different types of traffic which are modeled by various multi-state Markov modulated arrival processes. In this paper, we review the single node approximation techniques and then go on to study networks of queues by making use of a uniform representation of traffic flows based on D-BMAP.

First, we review the definition of D-BMAP. The *Batch Markovian Arrival Process* (BMAP) [11], which is equivalent to *Neuts' versatile Markov process* [13]. It is a very rich class of point processes that contains many well known processes such as *PHase-type processes* (PH) and *Markov Modulated Poisson Processes* (MMPP). D-BMAP, which was originally formulated in [3], is the discrete-time analog of BMAP. It can be defined by a two-dimensional Markov chain, $\{ (R_k, J_k), \quad k = 1, 2, 3, \dots \}$, where R_k is the total number of arrivals by the end of slot k and J_k is the phase of the D-BMAP in slot k. The transition probability matrix of the Markov chain for m phases is given by:

$$T = \begin{bmatrix} D_0 & D_1 & D_2 & D_3 & \dots \\ 0 & D_0 & D_1 & D_2 & \dots \\ 0 & 0 & D_0 & D_1 & \dots \\ \cdot & \cdot & \cdot & \cdot & \dots \\ \cdot & \cdot & \cdot & \cdot & \dots \end{bmatrix}$$

where $(D_k)_{i,j}, 0 \le i, j \le m, k = 0, 1, 2, \dots$ is the probability that there is a phase change from phase i to phase j accompanied by an arrival of size k. Note that we let the probability of bulk size being 0 be the probability of no arrival and normalize the bulk size probability mass function accordingly, i.e, the probability of a bulk arrival is 1 for every phase.

Consider a D-BMAP/D/1/K queueing system where we observe the system state at the end of each time slot just before the cell in the server (if there is one) leaves the system. Then, the steady state joint probability distribution of the queue length (including the server) and the phase of the D-BMAP at the observation time instants can be obtained by solving the invariant probability vector (denoted $L = \{L_0, L_1, \dots, L_K\}$, where the jth element of the vector L_i, $0 \le i \le m$, is the

steady state joint probability of phase j and a queue length of i) of the following matrix [3]:

$$
Q = \begin{bmatrix}
D_0 & D_1 & D_2 & \dots & D_{K-1} & \sum_{i=K}^{\infty} D_i \\
D_0 & D_1 & D_2 & \dots & D_{K-1} & \sum_{i=K}^{\infty} D_i \\
0 & D_0 & D_1 & \dots & D_{K-2} & \sum_{i=K-1}^{\infty} D_i \\
\cdot & \cdot & \cdot & \dots & \cdot & \cdot \\
0 & 0 & 0 & \dots & D_0 & \sum_{i=1}^{\infty} D_i
\end{bmatrix}
\tag{1}
$$

Once L is obtained the loss probability can be calculated as follows [19]:

$$
P_{loss} = \frac{1}{\rho} \sum_{i=0}^{K} \sum_{j=1}^{\sim} max\{j - (K-i) - 1 + \delta_i, 0\} L_i D_j e
\tag{2}
$$

where ρ is the *traffic intensity* (to be defined later); e is a column vector of all 1's; and

$$
\delta_i \equiv \begin{cases} 1, & \text{if } i = 0 \\ 0, & \text{otherwise} \end{cases}
$$

The above analytical model has been successfully applied in [4] and [19] to study the performance of an ATM multiplexer and forms the basis for the analysis of ATM networks discussed in this paper.

2 Single node models

In this section, we first introduce the approximation proposed in [19]. We then give examples to demonstrate the effectiveness of the approach.

2.1 Traffic models

We model a voice source by a discrete-time *ON-OFF* process in which the voice source alternates between geometrically distributed *ON* and *OFF* periods. Cells are generated with a constant arrival probability (instead of constant interarrival time) during the *ON* periods and no cells are generated during the *OFF* periods. In discrete time, this point process is called an *Interrupted Bernoulli Process* (IBP), which can be defined by three parameters: the transition

probabilities from ON to OFF (β) and from OFF to ON (α) and the constant arrival probability (ω). To model the superposition of N independent voice sources, we further assume that during any time slot only one voice source can change its state (either from ON to OFF or from OFF to ON). This is a reasonable assumption, since the values of α and β are typically in the order of 10^{-5} (or less). For example: a channel capacity of 44.736 Mbps (standard DS-3 data rate) would set α and β to 1.46×10^{-5} and 2.63×10^{-5} respectively (for an average ON duration of 360 msec and an average OFF period of 650 msec [5]) and the higher the channel capacity is the smaller the values of α and β are. The superposition of N such independent voice sources can be represented by an (N+1)-state D-BMAP where the states represent the number of voice sources in the ON state and the bulk size distribution is *binomial* with parameters k and ω, denoted $\mathcal{B}(k, \omega)$, for phase k, $0 \leq k \leq N$.

We adopt the discrete-time version of the model originally proposed by Maglaris *et al.* [12] to model video sources with uniform activity level. In this model, it is assumed that there are no sudden movements in the video scenes, e.g., a video scene from a videotelephone connection showing a person talking in front of the camera. This model can be viewed as the superposition of M mini-sources each of which is an IBP with a cell arrival probability of, say, η and transition probabilities of, say, a and b (see [12] and [19] for details). Thus, the expected number of cells arriving in a slot from all video sources can be fitted into $M + 1$ equal-distance discrete levels, 0, η, 2η,..., $M\eta$. As in our voice model, for level i, $0 \leq i \leq M$, the bulk size distribution is *binomial* with parameters i and η, $\mathcal{B}(i, \eta)$. Transitions are assumed to take place only to adjacent levels where the transition probabilities are obtained by matching the statistical characteristics of the process to that of the video sources.

Combining the models for voice and video sources results in a discrete-time *two-dimensional Markov chain*, where the states represent the level of expected number of arrivals for video sources and the number of active voice sources. Again, we have assumed that a state change for a voice source and a level change for a video source cannot occur in a single slot. (This is reasonable due to the fact that both α, β and a, b are very small.) The steady-state distribution of this two-dimensional Markov chain is given by:

$$\pi_{ij} = \binom{N}{j}\binom{M}{i}\left(\frac{a}{a+b}\right)^i \left(\frac{b}{a+b}\right)^{M-i}\left(\frac{\alpha}{\alpha+\beta}\right)^j \left(\frac{\beta}{\alpha+\beta}\right)^{N-j}, \quad 0 \leq i \leq M, 0 \leq j \leq N \quad (3)$$

It can be verified that, in state (i, j), the bulk size distribution is multinomial (more precisely, a trinomial distribution) with the following two parameter sets: i, η and j, ω.

We assume that packet arrivals from aggregated data sources form a *Bernoulli process* (single packet) with a probability of arrival μ, i.e., the packet interarrival time is geometrically distributed with the same parameter. We assume that the size of the data packets has a general probability mass function $s_k = Pr\{$data packet is comprised of k cells$\}$, $k = 0, 1, 2, \ldots$ and a mean of $\bar{s}$, and that it is non-trivial, i.e., $s_k > 0$ for some $k > 0$. Note that, in order to facilitate the integration of different types of traffic, we have normalized the packet size distribution in a way that the event of no arrival is presented in the form of an arrival with packet size 0, i.e., $s_0 = 1 - \mu$.

2.2 The approximation technique

In the combined model for video and voice traffic integration, we say a state (i, j) is overloaded if $(\eta i + \omega j) > 1 - \mu\bar{s}$; otherwise, we say that it is underloaded. Let S_u, S_o be the sets and θ_u, θ_o be the numbers of the underload states and overload states; σ_i, $0 \le i \le M$ be the number of voice sources which the system can support given that the video sources are currently in level i; and $F = max_i\{\sigma_i + i\}$ be the largest possible number of active mini-sources among underload states. Note that $\sigma_i = \lfloor (1 - \mu\bar{s} - \eta i)/\omega \rfloor$ and is constrained to $0 \le \sigma_i \le N$.

We then propose (presented in [19]) the following procedure for matching the four parameters (the transition probabilities, p_1 and p_2, and the expected numbers of arrivals, υ_1 and υ_2, for phase 1 and 2 respectively) of a two-state D-BMAP as an approximation of the video and voice traffic integration:

i. $\quad p_2 = \left(\sum_{i=0}^{M} (\sigma_i + 1) / \sum_{i=0}^{M} (N - \sigma_i) \right)(\beta + b)$

ii. $\quad q_1(k) \sim \mathcal{B}(F, \frac{\upsilon_1}{F})$, where $\upsilon_1 = \sum_{(i,j) \in S_u} (\eta i + \omega j) \left(\frac{\pi_{ij}}{\Pi_u} \right)$ with $\Pi_u = \sum_{(x,y) \in S_u} \pi_{xy}$

iii. $\quad q_2(k) \sim \mathcal{B}(N + M, \frac{\upsilon_2}{N + M})$, where $\upsilon_2 = \sum_{(i,j) \in S_o} (\eta i + \omega j) \left(\frac{\pi_{ij}}{\Pi_o} \right)$ with $\Pi_o = \sum_{(x,y) \in S_o} \pi_{xy}$

iv. $\quad p_1 = p_2 \dfrac{(\omega\phi_1 + \eta\phi_2) - \upsilon_1}{\upsilon_2 - (\omega\phi_1 + \eta\phi_2)}$, where $\phi_1 = \dfrac{N\alpha}{\alpha + \beta}$ and $\phi_2 = \dfrac{Ma}{a + b}$

It can be readily seen that the integration of the data sources (modeled by a *Bernoulli process* with bulk arrivals) and video and voice sources (approximated by the above two-state D-BMAP) is just another two-state D-BMAP with:

$$p'_i = p_i, i = 1, 2 \tag{4}$$

$$q'_i(k) = (s \otimes q_i)(k), \ i = 1, 2, \ k = 0, 1, 2, \ldots \tag{5}$$

where $\otimes$ denotes *convolution*.

Using (4) and (5) and the definition of D-BMAP, we can easily get the following for the two-state D-BMAP representing the integration of video, voice and data sources:

$$D_k = \begin{bmatrix} q'_1(k)(1-p_1) & q'_1(k)p_1 \\ q'_2(k)p_2 & q'_2(k)(1-p_2) \end{bmatrix}, \quad k = 0, 1, 2, \ldots \tag{6}$$

Also, the transition probability matrix for the phase process is given by:

$$D \equiv \sum_{k=0}^{\infty} D_k = \begin{bmatrix} 1-p_1 & p_1 \\ p_2 & 1-p_2 \end{bmatrix} \tag{7}$$

2.3 Examples

In the following numerical examples, we characterize a voice source by a cell arrival rate of 1/6 cells/msec (assume 64 Kbps PCM coding with speech activity detector and a standard 48-octet payload size per cell) and average *ON* and *OFF* durations of 360 msec and 650 msec correspondingly (as concluded by [5]). We use the same set of parameters for the video sources as used by [4], [12] and [19], i.e., a video source is characterized by: an average bit rate of 3.9 Mbps, a peak bit rate of 10.58 Mbps, a standard deviation of the bit rate of 1.73 Mbps and a parameter for the autocorrelation function of 3.9. The total number of levels for video sources is assumed to be 10 times the number of video sources (as suggested by [12]). Aggregated data traffic is assumed to have a packet arrival rate of $\mu = 20N_d/3$ packets per second, where N_d is the number of data sources. We assume that the packet size is geometrically distributed with an average of 5 cells/packet (note that this has to be adjusted for the probability of bulk size being 0 before it can be applied to our model).

We use Fig. 1, one of the numerical examples presented in [19], to demonstrate the accuracy of the above approximation technique in modeling systems with integrated voice and data traffic. In this example, there is no video source and we fix the number of voice and data sources to 20 and 30 respectively. The average number of cells in the system under a specific *traffic intensity* (defined to be $\rho = \mu\bar{s} + \omega\phi_1$) is obtained by solving the invariant probability vector of matrix Q

defined in (1) for various of buffer sizes. Note that, since we observe the system at the end of each slot, the average delay is the same as the average number of cells in the system. In Fig. 2, we plot the loss probability of a system with 1 video, 10 voice, and 50 data sources for various traffic intensities (defined to be $\rho = \mu \bar{s} + \omega \phi_1 + \eta \phi_2$).

In these examples, the exact model can be solved for only a limited buffer size. This is due to the large state space involved in the exact model. For example, it can be solved roughly up to a buffer size of only 100 for the first example; while a buffer size of only 23 (an equivalent state space of 2873×2873) can be solved for the second example (programmed in *MATLAB*™ and running on a *SPARC*™-*600* workstation). On the other hand, the approximate model, the state space of which is independent of the number of sources in the system, can be solved for a much larger buffer size (roughly up to a buffer size of 1000 running on the same workstation). In order to verify the accuracy of the approximation with an interesting buffer size, we have also included some simulation results in the second example. In Fig. 2, the solid line (representing the analytical result from the exact model) stops at buffer size of 23 due to the reason stated above. Results for larger buffer sizes are presented using simulation results (dotted line).
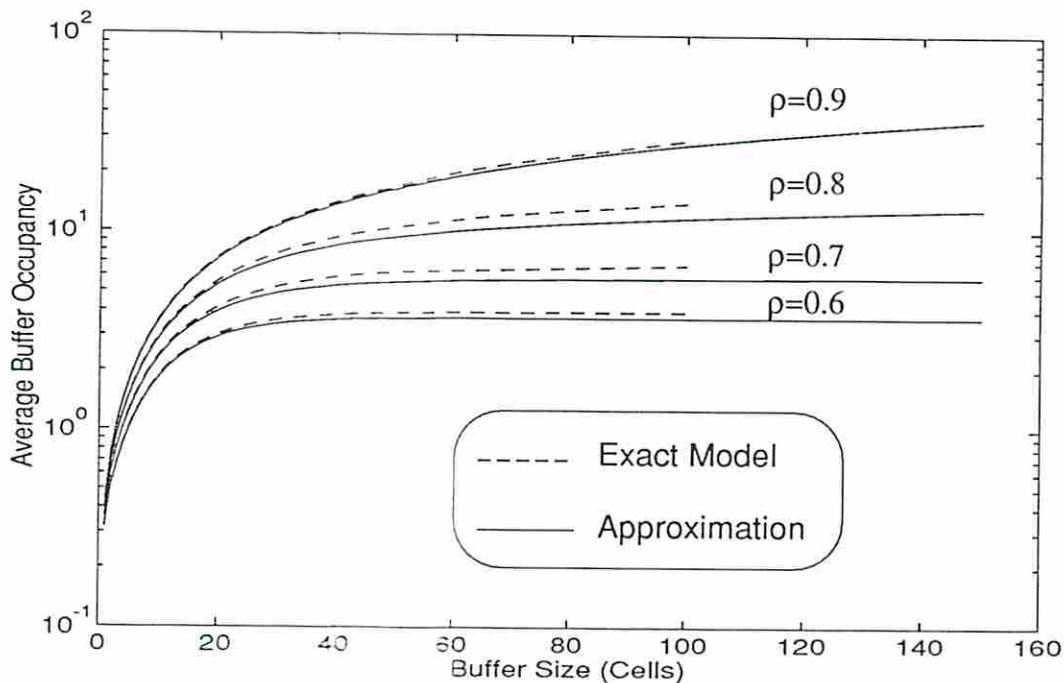

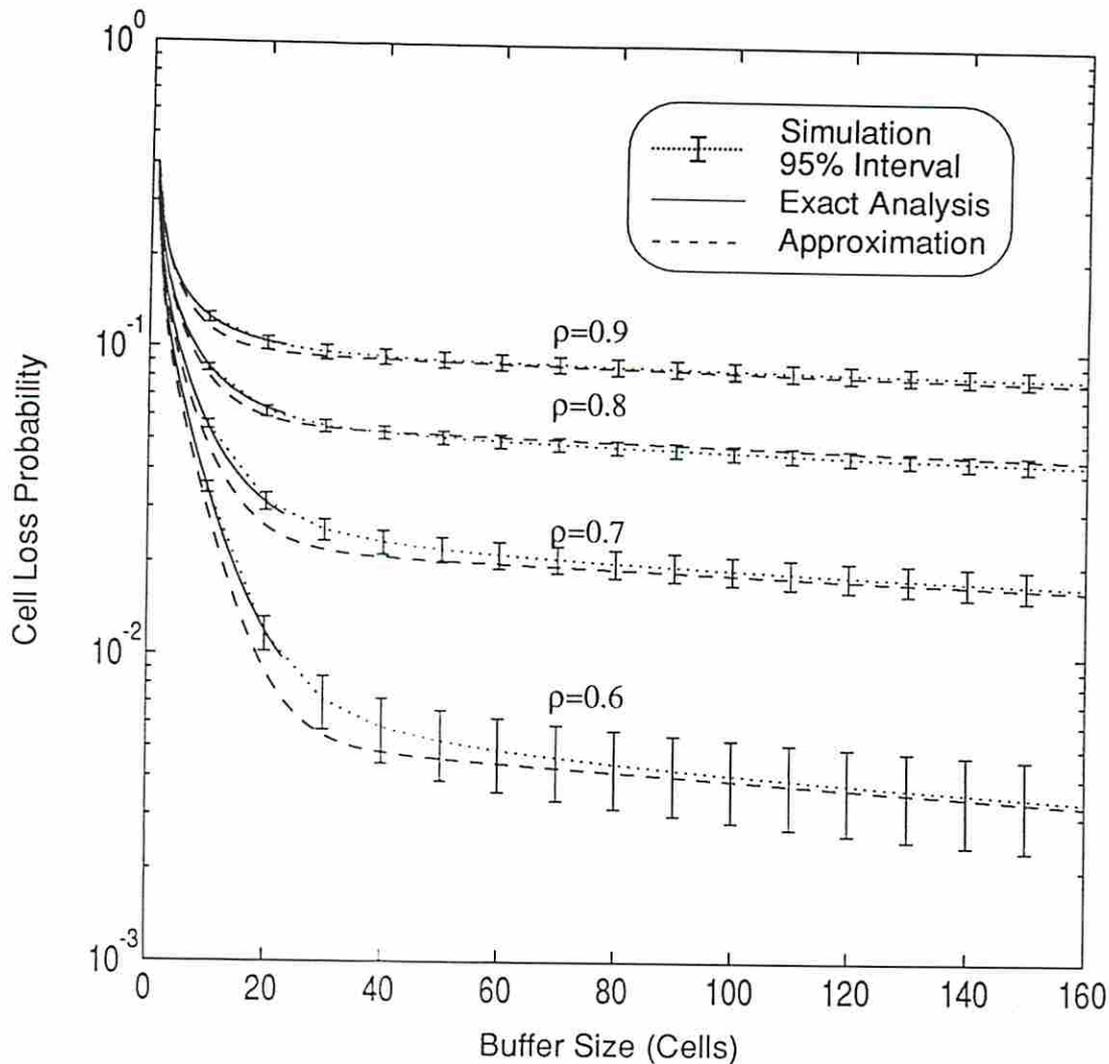Fig. 1. Average number of cells in the system for various of buffer sizes.

Fig. 2. Cell loss probability for different traffic intensity with 1 video, 10 voice and 50 data calls.

Because of the fast computation of the proposed approximation, one of its potential applications is real-time traffic management such as admission control. In the next example, we assume a fixed buffer size of 200 and a fixed channel capacity of 44.736 Mbps, i.e., standard DS-3 rate, and show (in Fig. 3) the trade-off between the number of voice sources and video sources that the system can support given a fixed number of data sources in the background. Similarly, Fig. 4 shows the trade-off between the number of data sources and video sources that the system can support given a certain number of voice sources in the background. Using these figures, we can easily find, given a specific loss probability requirement, the maximum number of voice sources, for example, that the system can support with a certain number of video sources.
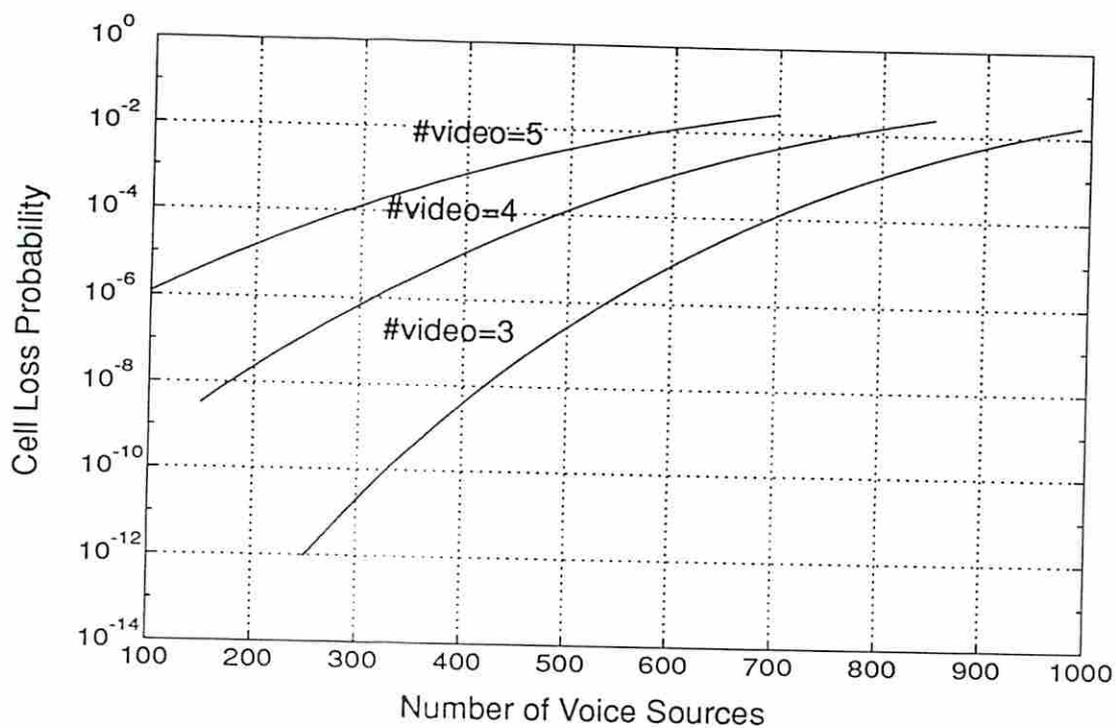
Fig. 3. Cell loss probability for different number of video calls with 500 data calls as a function of the number of voice sources.
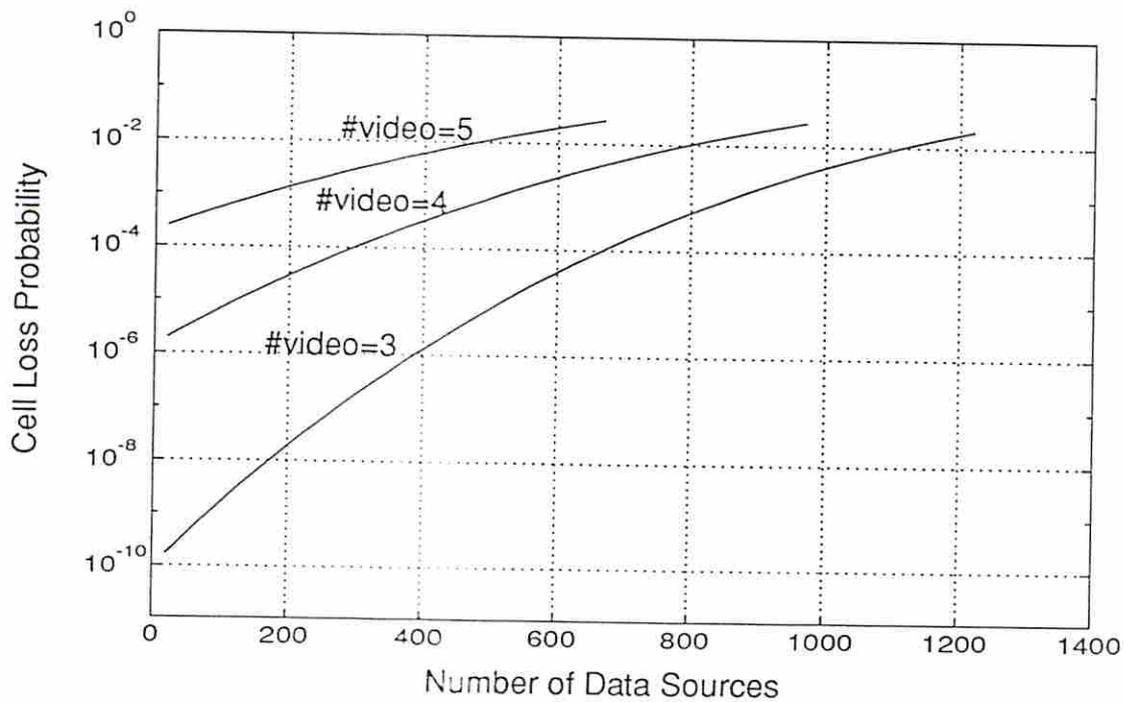

Fig. 4. Cell loss probability for different number of video calls with 600 voice calls as a function of the number of data sources.

3 Network models

In this section, we now consider a network of ATM switches and the corresponding queueing network. In order to solve this queueing network with non-renewal flows, we employ the parametric decomposition approximation [10]. Parametric decomposition evaluates the queues in the network as if they were stochastically independent. The queues are analyzed in isolation after the input flow parameters are approximated [10], [20]. The elementary network operations including *i*) output process, *ii*) joining, and *iii*) splitting as well as a computational procedure for estimating the end-to-end performance will be discussed in this section.

3.1 The output process of a D-BMAP/D/1/K queue

Previous studies have shown that the output process of a D-BMAP/D/1/K queue is also correlated, and neglecting its correlations leads to inaccurate results [16], [18]. At each time slot, at most one cell may depart from a queue, so the *Markov Modulated Bernoulli Process* (MMBP), a correlated process with single arrivals, is a good candidate for modelling the output process of a D-BMAP/D/1/K queue. We, therefore, developed a procedure for matching the statistics of the output process with the statistics of a two-state MMBP [6]. Moreover, by modelling the output process as a two-state MMBP, we are able to represent all the flows in the network as D-BMAP process.

Before showing how to match the statistics of the output process with the statistics of a two-state MMBP, we need to characterize the output process itself. Having exactly one departure at each time instant of a busy period suggests that we can represent the output process as a MMBP in which the matrices D'_1 and D'_0 correspond respectively to busy and idle periods. In order to capture the behavior of busy/idle periods, we need to associate each state of the D-MAP with the phase of the arrival process and with the number of enqueued cells at the end of each time slot [3]. If we have a gated server (i.e., if a cell finds the server empty at its arrival slot, it can only be transmitted at the next slot) then, the output process is given by [3]:

$$D'_0 = \begin{bmatrix} D_0 & D_1 & D_2 & \dots & D_{K-1} & \sum_{i=K}^{\infty} D_i \\ 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix}$$

$$D'_1 = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ D_0 & D_1 & D_2 & \dots D_{K-1} & \displaystyle\sum_{i=K}^{\infty} D_i \\ 0 & D_0 & D_1 & \dots D_{K-2} & \displaystyle\sum_{i=K-1}^{\infty} D_i \\ . & . & . & \dots & . & . \\ 0 & 0 & 0 & \dots & D_0 & \displaystyle\sum_{i=1}^{\infty} D_i \end{bmatrix}$$

The index of dispersion time curve completely defines the correlation structure of a counting process. Consequently, to accurately approximate the output process, it is important to provide a good match with the index of dispersion time curve. In our procedure, we chose to match the long-term index of dispersion for counts and the covariance of the number of arrivals at lag 1 and lag 2 (as below):

$$\text{output}_{\text{mean}} = \text{2-MMBP}_{\text{mean}}$$

$$\text{output}_{\text{variance}} = \text{2-MMBP}_{\text{variance}}$$

$$\text{output}_{\text{covariance lag=1}} = \text{2-MMBP}_{\text{covariance lag=1}}$$

$$\text{output}_{\text{covariance lag=2}} = \text{2-MMBP}_{\text{covariance lag = 2}}$$

In [3], it was demonstrated that for D-BMAP the mean arrival rate, the variance of number of arrivals and the covariance at lag k are given by:

$$\lambda = \pi \left(\sum_{n=1}^{\infty} n D_n \right) \grave{e}$$

$$var = \pi \left(\sum_{n=1}^{\infty} n^2 D_n \right) \grave{e} - \lambda^2$$

$$cov(X_1, X_i) = \pi \left(\sum_{n=1}^{\infty} n D_n \right) D^{k-2} \left(\sum_{n=1}^{\infty} n D_n \right) \grave{e} - \lambda^2$$

where $\vec{e}$ is the unit column vector and π is the steady state probability of the underlying Markov chain, i.e, $\pi D = \pi$ (where D is defined in (7)) and $\pi\vec{e} = 1$.

The matching procedure has been validated in [6]. Errors below 7% and 10% were found respectively in the estimation of the delay and loss rate at the second queue in a two-queue tandem network when the output process of the first queue is replaced by a two-state MMBP. The matching procedure is reasonably accurate and our results are consonant with a similar study done by Park and Perros [14].

3.2 Joining

The superposition of two D-BMAP processes with m_1, m_2 states and n_1, n_2 maximum batch sizes is also a D-BMAP with $m_1 \times m_2$ states and $n_1 + n_2$ maximum batch size. Assume that the two D-BMAP's are defined by D_k^1, $0 \leq k \leq n_1$, and D_k^2, $0 \leq k \leq n_2$. Then, the (i, j) th element of matrix D_k, $0 \leq k \leq n_1 + n_2$, the probability that the superposed process makes a transition from state i to state j accompanied by an arrival of size k, can be computed by:

$$D_k = \sum_{j=0}^{min(k, n_1)} D_j^1 \otimes D_{k-j}^2 \tag{8}$$

3.3 Splitting

We assume that routing is state independent which means that the probability of a cell departing from one node and going to another node is fixed. When characterizing the flow between two nodes, we represent the output process of the first queue as a two-state MMBP process, and then model the flow that goes to the second queue as a two-state MMBP with parameters $(p_{ij} \times p_1, p_{ij} \times p_2, \alpha_1, \alpha_2)$ where p_{ij} is the probability that a cell leaves node i and goes to node j and p_n and α_n, $1 \leq n \leq 2$, are respectively the arrival probabilities and the transition probabilities in state n.

3.4 The Computational Procedure

To compute the end-to-end performance of an ATM virtual path, we make use of the parametric decomposition approximation, i.e., each queue in the queueing network is analyzed in isolation after its input process is fully characterized. In this approach, the dependencies among the queues are approximated by the flow parameters. We concentrate on ATM networks whose topology can be described as an acyclic directed graph. We assume that there are two distinct sets

of nodes: sets E and I. The elements of set E receive only input (external) traffic to the network, i.e., elements of set E are the entry points of the network. The elements of set I are nodes whose input is composed of the output process of other nodes and possibly input traffic to the network, i.e., nodes belonging to set I are network internal nodes which can also receive external traffic. We define S_k as the set of nodes whose input traffic can be determined only at iteration k of the computational procedure. In other words, nodes belonging to S_k have at least one input link whose flow parameters can only be computed at step k-1. We compute the occupancy distribution of all nodes of S_k at step k; and we denote a link whose traffic parameters have been determined as a *marked link*. The computational procedure can be summarized as follows (by assuming a feed-forward topology, we guarantee that the procedure terminates):

1. $k = 1$ & $S_1 = E$, the input process (for each session) to a node in S_1 is determined from approximation given in section 2 based on the given traffic load.

2. While $S_k \neq \varnothing$ do:

   2.1. $\forall$ nodes $i \in S_k$ do:

      2.1.1. Characterize the input process for node i by performing a joining operation of all inputs (from links and external sessions).

      2.1.2. Compute the steady state queue length distribution for node i.

      2.1.3. Compute the mean delay and loss probability at node i.

      2.1.4. Characterize the output process for node i by matching the statistics of the output process to a corresponding two-state MMBP.

      2.1.5. Characterize the process on each outgoing link from node i by performing a spliting operation.

      2.1.6. Mark node i.

   2.2. $k = k + 1$

3.5 Numerical Results

We have validated our queueing network framework for various scenarios including tandem and feed-forward networks [7], [8]. Percentage errors of the delay estimation and of the loss rate

computation were below 10% and 13% for networks with 20 nodes in tandem. In Table 1 and Table 2, we show examples of the delay and loss rate respectively for a five-node tandem network with a buffer size of 100 in each node. The input process and the interfering process at each node are both two-state D-BMAP's with the same transition probability in each state and with Poisson distributed batch size with means $(1 + c)\rho$ and $(1 - c)\rho$ where ρ is the average cell arrival rate (which is equal to the traffic intensity) and c is a parameter (as in [18]).

ρ	(c, α)	analytical	simulation	error (%)
0.775	0.1, 0.9	34	31 ($\pm$ 3)	9.1
0.8	0.1, 0.9	80	73 ($\pm$ 6)	8.8
0.76	1.5, 0.9	391	371 ($\pm$ 17)	5.4
0.8	1.5, 0.9	433	415 ($\pm$ 17)	4.3
0.775	1.75, 0.9	467	448 ($\pm$ 20)	4.2

Table 1: Average delay for tandem network (interfering process $(\rho, c, \alpha) = (0.05, 0.1, 0.95)$).

ρ	(c, α)	analytical	simulation	error (%)
0.7	1.25, 0.9	2.7×10^{-1}	$2.5 (0.3) \times 10^{-1}$	7.1
0.75	0.9, 0.9	2.9×10^{-2}	$2.7 (0.2) \times 10^{-2}$	7.7
0.675	1.125, 0.9	4.8×10^{-3}	$4.4 (0.8) \times 10^{-3}$	8.6
0.6125	0.9, 0.9	4.2×10^{-6}	$3.7 (0.7) \times 10^{-6}$	11.7
0.6	0.9, 0.9	2.8×10^{-7}	$2.5 (0.5) \times 10^{-7}$	12.6

Table 2: Loss probability for tandem network (interfering process $(\rho, c, \alpha) = (0.05, 0.1, 0.95)$).

4 Numerical examples for an ATM network

We consider 3 experiments:

4.1 Experiment 1

To demonstrate the proposed framework for a more general network, we study the ATM network shown in Fig. 5 with the traffic sessions specified in Table 3 (where R is the channel data

rate). In this example, all links are assumed to have a data rate of $R = 22.5$ Mbps. and all traffic sessions are assumed to support a similar traffic mix of (roughly) 40%, 40% and 20% from video, voice and data sources respectively. The corresponding (open) queueing network for this example is given by Fig. 6. We assume a buffer size of 100 for each queue in Fig. 6 and model this ATM network using the framework introduced in the previous sections. More specifically, for analytical results, we approximate each traffic session as a two-state D-BMAP using the model described in section 2. Output processes for all intermediate queues, traffic splitting and traffic joining are modeled using the technique introduced in section 3. To evaluate the accuracy of the approximation, we first compare the analytical results with that of a simulation where the traffic sources are exactly represented and routing of cells is done according to the session (whereas the analytical model uses random traffic splitting). Note that at node $\overline{AB}$ in Fig. 6, we assume that cells from different sessions are served alternately. The average delay and loss probability for each node are presented in Table 4 and Table 5 respectively. In this simulation, we also measure average delay and loss probability on a per session basis which is also shown in the tables. We see that the delay results from our model are in very good agreement with the simulation. The loss figures also show reasonable agreement but not quite as good as the delay results. We find that the analytical model underestimates delay and loss especially at links $\overline{BC}$ and $\overline{BE}$. It is precisely for these links that the effect of the random splitting of the output traffic from node B would be expected to manifest itself. The significance of the correlation in the output streams was noted in [16] and [18]. The results are somewhat preliminary in that we do not have very tight confidence intervals.

End-to-end performance results on a per session basis are given in Table 6 and Table 7. In the analytical results, the average delay of a session is calculated as the sum of the average delay of all queues on the path of the session and the loss probability is calculated as:

$$1 - \prod_{k \in \zeta} (1 - P_k) \tag{9}$$

where ζ is the set of queues alone the path of the session and P_k is the loss probability (over all sessions for the analytical results) at queue k. We find surprisingly good agreement except in the loss probabilities of sessions γ_4 and γ_5. We believe this to be due to the fact that, as seen in the simulation, sessions γ_1 and γ_2 traffic suffer no loss (since it is given priority over new traffic) here but the analytical model lumps the loss of sessions γ_4 with γ_1 at link $\overline{BC}$ and γ_5 with γ_2 at link $\overline{BE}$. We expect better results with a more complex topology and traffic pattern (so that more mixing occurs).
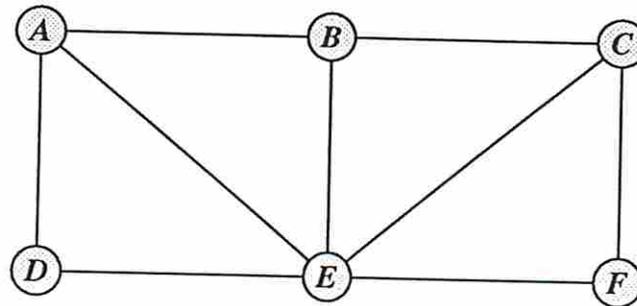
Fig. 5 An ATM network with 6 nodes connected by OC-3 rate connections.

traffic session	load	source	destination	routing
γ_1	0.4335R	A	C	via B
γ_2	0.4335R	A	E	via B
γ_3	0.8671R	D	F	via E
γ_4	0.4335R	B	C	- - -
γ_5	0.4335R	B	E	- - -

Table 3: Traffic sessions considered in the experiment 1 ($R = 22.5$ Mbps).
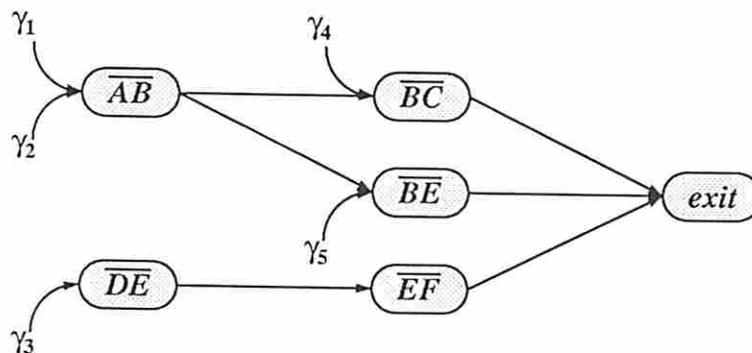
Fig. 6 Corresponding queueing network for the example.

	analytical	simulation	error(%)
Link $\overline{AB}$ Session γ_1 Session γ_2	24.0	25.3 ($\pm$ 0.7) 25.3 ($\pm$ 0.6) 25.4 ($\pm$ 0.7)	5.1
Link $\overline{DE}$ Session γ_3	24.0	25.2 ($\pm$ 0.6) 25.2 ($\pm$ 0.6)	4.5
Link $\overline{BC}$ Session γ_1 Session γ_4	20.0	21.3 ($\pm$ 0.8) 21.0 ($\pm$ 0.8) 21.6 ($\pm$ 0.8)	6.3
Link $\overline{BE}$ Session γ_2 Session γ_5	20.0	21.4 ($\pm$ 0.5) 21.2 ($\pm$ 0.5) 21.9 ($\pm$ 0.5)	6.7
Link $\overline{EF}$ Session γ_3	1.0	1.0 1.0	0.0

Table 4: Delay: approximate model and simulation of exact traffic model ($R = 22.5$ Mbps).

	analytical	simulation	error(%)
Link $\overline{AB}$ Session γ_1 Session γ_2	7.1×10^{-3}	$8.1 (\pm 0.7) \times 10^{-3}$ $8.0 (\pm 0.6) \times 10^{-3}$ $8.3 (\pm 0.6) \times 10^{-3}$	13.0
Link $\overline{DE}$ Session γ_3	7.1×10^{-3}	$8.2 (\pm 0.5) \times 10^{-3}$ $8.2 (\pm 0.5) \times 10^{-3}$	14.4
Link $\overline{BC}$ Session γ_1 Session γ_4	1.5×10^{-3}	$1.3 (\pm 0.8) \times 10^{-3}$ 0.0 $2.5 (\pm 0.2) \times 10^{-3}$	19.6
Link $\overline{BE}$ Session γ_2 Session γ_5	1.5×10^{-3}	$1.3 (\pm 0.5) \times 10^{-3}$ 0.0 $2.2 (\pm 0.6) \times 10^{-3}$	17.8
Link $\overline{EF}$ Session γ_3	0.0	0.0 0.0	0.0

Table 5: Loss: approximate model and simulation of exact traffic model ($R = 22.5$ Mbps).

session	analytical	simulation	error(%)
γ_1	44.0	46.3	5.2
γ_2	44.0	47.3	6.8
γ_3	25.0	26.1	4.2
γ_4	20.0	21.6	7.5
γ_5	20.0	21.2	5.5

Table 6:   End-to-end delay ($R = 22.5$ Mbps).

session	analytical	simulation	error(%)
γ_1	8.5×10^{-3}	8.0×10^{-3}	7.3
γ_2	8.5×10^{-3}	8.3×10^{-3}	2.7
γ_3	7.1×10^{-3}	8.2×10^{-3}	14.3
γ_4	1.5×10^{-3}	2.5×10^{-3}	39.2
γ_5	1.5×10^{-3}	2.2×10^{-3}	31.4

Table 7:   End-to-end loss probability ($R = 22.5$ Mbps).

4.2 Experiment 2

In order to determine whether a detailed representation of the traffic is necessary, we simulate a network with the external traffic generated by a two-state D-BMAP for each session (with parameters corresponding to those used in the approximate analytical model). These results can be interpreted in two ways: i) if we compare the D-BMAP simulation against the approximate analytical model, we can investigate the inaccuracies introduced by the simplification of the queueing network model; b) if we compare the D-BMAP simulation against the accurate traffic model we can investigate whether the reduced traffic model is reasonable as a simplification. The main advantage of the reduced model being a significant reduction in the run times of the simulation. We show these results in Table 8 and Table 9. We see that the approximate analytical model overestimates delay and loss compared to the two-state source simulation. This is probably due to tandem queue effects in this simple network. As might be expected, the two-state traffic model underestimates delay and loss compared to the exact traffic simulation (due to the reduced

burstiness). The results are preliminary in that the confidence intervals are quite large.

link	analytical	two-state simulation	detailed simulation
$\overline{AB}$	24.0	23.2 ($\pm$ 0.6)	25.3 ($\pm$ 0.7)
$\overline{DE}$	24.0	23.0 ($\pm$ 0.7)	25.2 ($\pm$ 0.6)
$\overline{BC}$	20.0	18.6 ($\pm$ 0.9)	21.3 ($\pm$ 0.9)
$\overline{BE}$	20.0	18.6 ($\pm$ 0.6)	21.4 ($\pm$ 0.5)
$\overline{EF}$	1.0	1.0	1.0

Table 8: Delay: two-state source models. ($R = 22.5$ Mbps).

Link	analytical	two-state simulation	detailed simulation
$\overline{AB}$	7.1×10^{-3}	$6.6 (\pm 0.8) \times 10^{-3}$	$8.1 (\pm 0.7) \times 10^{-3}$
$\overline{DE}$	7.1×10^{-3}	$6.3 (\pm 0.6) \times 10^{-3}$	$8.2 (\pm 0.5) \times 10^{-3}$
$\overline{BC}$	1.5×10^{-3}	$1.4 (\pm 0.4) \times 10^{-3}$	$1.3 (\pm 0.9) \times 10^{-3}$
$\overline{BE}$	1.5×10^{-3}	$1.2 (\pm 0.7) \times 10^{-3}$	$1.3 (\pm 0.5) \times 10^{-3}$
$\overline{EF}$	0.0	0.0	0.0

Table 9: Loss: two-state source models. ($R = 22.5$ Mbps).

4.3 Experiment 3

In this experiment, we study a higher capacity network with link rates of $R = 155.52$ Mbps (standard OC-3 data rate). The traffic mix is shown in Table 10. We show results using the analytical model and the two-state traffic model (Table 11 and Table 12). We see better agreement between the analytical model and the simulation from which we can conclude that the errors introduced by the simplification necessary to solve the queueing network tend to reduce as the network is scaled up.

traffic session	load	source	destination	routing
γ_1	0.5R	A	C	via B
γ_2	0.3R	A	E	via B
γ_3	0.6R	D	F	via E
γ_4	0.35R	B	C	- - -
γ_5	0.55R	B	E	- - -

Table 10: Traffic sessions considered in the second example (R = 155.52 Mbps, OC-3 rate).

link	analytical	simulation	error(%)
$\overline{AB}$	7.13	6.72 ($\pm$ 0.001)	5.3
$\overline{DE}$	4.84	4.51 ($\pm$ 0.003)	6.1
$\overline{BC}$	6.18	5.77 ($\pm$ 0.016)	6.1
$\overline{BE}$	7.57	7.12 ($\pm$ 0.003)	5.3
$\overline{EF}$	1.00	1.0	0.0

Table 11: Delay: larger network example (R = 155.52 Mbps, OC-3 rate).

link	analytical	simulation	error(%)
$\overline{AB}$	4.8×10^{-5}	$4.4\ (\pm 0.2) \times 10^{-5}$	7.8
$\overline{DE}$	2.0×10^{-5}	$1.9\ (\pm 0.04) \times 10^{-5}$	8.8
$\overline{BC}$	1.9×10^{-5}	$1.6\ (\pm 0.3) \times 10^{-5}$	8.6
$\overline{BE}$	4.3×10^{-5}	$4.0\ (\pm 0.3) \times 10^{-5}$	8.1
$\overline{EF}$	0.0	0.0	0.0

Table 12: Loss: larger network example (R = 155.52 Mbps, OC-3 rate).

5 Conclusions

In this paper we describe our research investigating the use of discrete time reduced state Markov modulated processes to model the performance of ATM networks. We first describe the approximation technique that we use to represent a complex combination of different sources (voice, video, and data) by a 2-state *discrete-time batch Markovian arrival process*, and demonstrate the effectiveness of this approach in comparison to simulation. We then extend the results by approximating the output process of a D-BMAP/D/1 queue in a similar way. By defining splitting and joining processes we develop a network of queues model based on this uniform representation of traffic flows. We present examples that compare the analytical results with simulation at different levels of detail showing reasonable agreement. We are currently exploring a wider range of examples. The results are encouraging and the speed of the computations indicate that this approach should provide useful in network management. We also have results for priority queues in isolation and are working on extensions of the network model to incorporate priority.

References

[1]   A. Baiocchi *et al.*, "Loss Performance Analysis of an ATM Multiplexer Loaded with High-Speed On-Off Processes," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 388-393, Apr. 1991.

[2]   C. Blondia, "The N/G/1 Finite Capacity Queue," *Stochastic Models*, vol. 5, no. 2, pp. 273-294, 1989.

[3]   C. Blondia, "A Discrete-Time Batch Markovian Arrival Process as B-ISDN Traffic Model", *Belgian J. of Oper Res., Stat. and Comp. Science*, vol. 32 (3), pp. 3-23, 1992.

[4]   C. Blondia and O. Casales, "Statistical Multiplexing of VBR Sources: A Matrix-Analytic Approach," *Performance Evaluation*, vol. 16, pp. 5-20, 1992.

[5]   P. T. Brady, "A Statistical Analysis of On-Off pattern in 16 Conversations," Bell Systems Technical Journal, pp. 73-91, Jan. 1968.

[6]   N. L. S. Fonseca and J. A. Silvester, "Modelling the Output Process of an ATM Multiplexer with Markov Modulated Arrivals", in *Proc. IEEE ICC'94*, pp. 721-725.

[7]   N. L. S. Fonseca and J. A. Silvester, "Estimating End-to-End Delay in ATM Virtual Paths", *submitted for publication*.

[8]   N. L. S. Fonseca and J. A. Silvester, "On the Computation of End-to-End Delay in Feed-Forward ATM Networks, to appear on *Proc. IEEE ITS '94*, Aug. 1994.

[9]   H. Heffes and D. M. Lucantoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance," *IEEE J. Select. Areas Commun.*, vol. 4, pp. 856-868, Sep. 1986.

[10]   P. J. Kuehn, "Approximative Analysis of General Queueing Networks by Decomposition", *IEEE Trans. Commun.*, vol COM-27, 1, pp. 113-126, 1979.

[11]   D. Lucantoni, "New Results on the Single Server Queue with a Batch Markov Arrival Process," *Stochastic Models*, vol. 7, no. 1, pp. 1-46, 1991.

[12]   B. Maglaris *et al.*, "Performance Models of Statistical Multiplexing in Packet Video Communications," *IEEE Trans. Commun.*, vol. 36, pp. 834-844, Jul. 1988.

[13]   M. F. Neuts, "A Versatile Markovian Point Process," *J. Applied Prob.*, vol. 16, pp.764-779, Dec. 1979.

[14]   D. Park and H. G. Perros, "Approximative Analysis of Discrete-Time Tandem Queueing Networks with Bursty and Correlated Input Traffic and Customers Loss", *Technical Report*, Department of Computer Science, NCSU, 1992.

[15]   V. Ramaswami, "The N/G/1 Queue and Its Detailed Analysis," *Adv. Appl. Prob.*, vol. 12, pp. 222-261, Mar. 1980.

[16]   H. Saito, "The Departure Process of an N/G/1 Queue", *Performance Evaluation* vol. 11, pp. 241-251, 1990.

[17]   P. Sen *at al.*, "Models for Packet Switching of Variable-Bit-Rate Video Sources," IEEE J. Select. Areas Commun., vol. 7, pp. 865-869, Jun. 1989.

[18]   T. Takine, T. Suda and T. Hasegawa, "Cell loss and output process analyses of finite buffer discrete time queueing system with correlated arrivals", in *Proc. IEEE INFOCOM '93*, pp 1259-1268, 1993.

[19]   S. S. Wang and J. A. Silvester, "A Discrete-Time Performance Model for Integrated Service ATM Multiplexers," in *Proc. IEEE Globecom '93*, pp. 757-761 (a detail version is available as *Technical Report*, CEng 93-05, EE-Systems, USC).

[20]   W. Whitt, "The Queueing Network Analyzer", *Bell Systems Technical Journal*, vol. 62, pp. 2779-2815, Nov. 1983.