ELSEVIER

Computer Networks and ISDN Systems 29 (1997) 293-313

COMPUTER NETWORKS and ISDN SYSTEMS

Multiplexing ATM traffic streams with time-scale-dependent arrival processes [1]

Randall Landry [a,2], Ioannis Stavrakakis [b,*]

[a] *Core Network Technology, DSP R & D Center, Texas Instruments Inc., Dallas, TX 75265, USA*

[b] *Department of Electrical and Computer Engineering, 409 Dana Research Building, Northeastern University Boston, MA 02115, USA*

Accepted 30 June 1996

Abstract

This paper presents a flexible traffic model which is capable of describing traffic dynamics on a variety of time-scales associated with broadband packet networks such as ATM. The model is applied to the study of Non-Deterministic Periodic (NDP) sources as well as Variable Bit Rate (VBR) video sources in which cells are delivered at a sub-frame level referred to as a *slice*. An efficient numerical technique is presented for the study of a finite-capacity multiplexer under an input process determined by the superposition of N of the proposed traffic models. Numerical results support the assertion that the time-scales at which variations in a cell arrival stream occur have a significant impact on multiplexer performance, thereby illustrating the relevance of the proposed modeling and analytical techniques.

Keywords: Time-scales; Traffic modeling; Queueing analysis; Periodic sources; ATM

1. Introduction

The desire to accommodate a diverse mixture of user traffic in a Broadband Integrated Services Digital Network (B-ISDN) has led to an intense concentration of research efforts on technologies which allow for traffic multiplexing, such as the Asynchronous Transfer Mode (ATM). Performance evaluation of statistical multiplexers under a variety of input processes and/or service policies (see [1-6] and the references therein) has been essential in determining the achievable effectiveness of statistical multiplexing.

The development of efficient congestion control and call admission procedures is an area in which multiplexing analysis and network traffic characterization continue to play an important role. Consequently, the formulation of accurate traffic models which lend themselves to analytical studies is a topic of ongoing research. In a high-speed integrated services networking environment, accurate traffic models should have the ability to model variations in the input cell arrival process on a variety of different time-scales. These time-scales may represent time-intervals (in network time slots) over which logical information blocks are delivered to the network.

* Corresponding author. Email: ioannis@cdsp.neu.edu.

[1] Research supported by the Advanced Research Project Agency under Grant F49620-93-1-0564 monitored by the Air Force Office of Scientific Research (AFOSR). I. Stavrakakis is on leave from the University of Vermont.

[2] E-mail: landry@hc.ti.com.

PII S0169-7552(96)00097-9

For instance, consider a Variable Bit Rate (VBR) video coder [7–9]. Bits generated over a frame form a logical block of information which could provide a basis for the description of the generated traffic as well as the desirable QoS. Notice that if the VBR video coder pre-buffers and uniformly distributes cells in each frame, then the cell activity over a frame can be described in terms of a fixed cell-arrival rate [9], and a relevant time-scale may be that of a video frame. If pre-buffering at the frame level is not feasible (or desirable), due to the induced frame delay, pre-buffering and uniform spreading over a shorter time-scale (referred to as a slice) may be necessary. In this case, the cell arrival rate over a slice will be constant and changes in the cell arrival rate process will occur within the frame at slice boundaries. Note that the absence of frame-level pre-buffering has been shown [10–13] to give rise to cell arrival autocorrelation functions which are *pseudo-periodic*, as opposed to the monotonic autocorrelation functions arising from frame-level pre-buffering schemes [7]. Packetized voice provides yet another example of a common traffic source time-scale when one considers the duration of time required to generate enough bits to form a cell. The resulting time-scale is then representative of the minimum cell inter-arrival time associated with the source.

When traffic sources such as those mentioned above are multiplexed onto a network link of capacity C, the time-scale T associated with each source can be seen as the one at which the network perceives events of interest (i.e. frame, slice, cell arrivals) to occur, and can be easily expressed in units of slots. For instance, when a voice source generating information at a peak rate of r bits/sec is multiplexed onto an ATM link with capacity C bits/sec, a relevant time-scale for the source is given by the ratio C/r.

In this paper, a stochastic traffic model is presented which is capable of capturing the cell arrival dynamics associated with a wide range of deterministic time-scales. The model is based on an underlying m-state periodic Markov chain in which state transitions occur every $T_s \geqslant 1$ time slots. While in a given state, cells are generated independently from slot to slot according to an arbitrary distribution. Two time-scales of length T_s and mT_s may be easily identified in this traffic model. A computationally tractable performance analysis of a finite-capacity multiplexer is developed under input traffic determined by the superposition of N such traffic sources. The technique involves the decomposition of a potentially large irreducible Markov chain into smaller irreducible Markov chains whose dimensionality is independent of m and T_s. Since the time-scales T_s and mT_s are expected to increase as the network speed increases, this decomposition is central to the tractability of the analysis approach in a high-speed networking environment. The model is applied to the study of Non-Deterministic Periodic (NDP) cell streams as well as to VBR video which is multiplexed at slice level.

In the following section, the proposed traffic model is described in detail in terms of a generalized periodic Markov chain and the finite capacity multiplexer is analyzed in discrete time. Subsequent sections will present the applications mentioned above, as well as a discussion of relevant past work.

2. Source characterization and queueing analysis

In the next subsection the proposed traffic model is described in terms of a generalized periodic Markovian source and the generated traffic is characterized. The resulting sub-period-level autocorrelation function for a source is derived in Section 2.2. In Section 2.3, the superposition of N such traffic sources is considered, and the aggregate traffic is described. Finally, a finite-capacity multiplexer fed by N generalized periodic Markovian sources is studied in the last subsection. The fundamental time unit is taken to be the transmission time of the fixed size cell, or the duration of a time slot on the output link as denoted by Δ (in seconds).

2.1. Description of the proposed traffic model

Let $\{m^k, T_s^k, f_j^k(\cdot) \mid 1 \leqslant j \leqslant m^k\}$ denote a set of parameters associated with the kth traffic source, $1 \leqslant k \leqslant N$, where N denotes the total number of traffic sources to be multiplexed and the parameters are defined as follows:

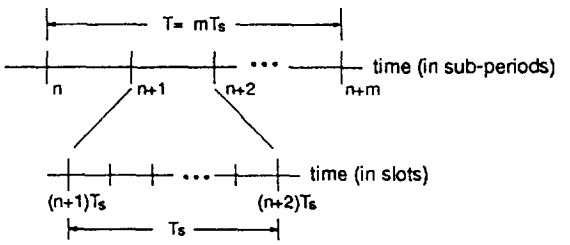

Fig. 1. The relevant time-scales associated with the GPM source.

m^k: The number of states in an underlying periodic Markov chain associated with the kth source. Let π_i^k denote the stationary probability that this periodic chain is in state i, and $p_{i,j}^{(n)}$ denote the n-step transition probability from state i to state j.

T_s^k: An integer constant (called sub-period) which represents the sojourn time (in slots) for each state in the periodic Markov chain. In other words, a transition occurs every T_s^k time slots (or $T_s^k\Delta$ seconds), and the kth source has a period of length $T^k = m^k T_s^k$ slots.

$f_j^k(\cdot)$: The probability mass function (PMF) for the number of cells generated in each slot by the kth source when that source is in state j, $1 \leqslant j \leqslant m^k$. λ_i^kj will denote the mean number of cell arrivals per slot for source k in state j, and N_j^k will denote the maximum number of such arrivals. N_j^k takes values from $\mathbb{Z}^0$, where $\mathbb{Z}^0$ denotes the set of non-negative integers.

A traffic source described in terms of $\{m^k, T_s^k, f_j^k(\cdot) | 1 \leqslant j \leqslant m^k\}$ will be referred to here as a Generalized Periodic Markov (GPM) source. It may easily be established that the mean arrival rate for source k is given by

$$\lambda^k = \frac{1}{m^k} \sum_{j=1}^{m^k} \lambda_j^k. \tag{1}$$

For the remainder of the paper, it will be assumed that all sources have the same sub-periods; hence, T_s^k will be denoted simply by T_s. In addition to the parameters defined above, let c^k (referred to as the initialization time) denote the time instant, defined at the beginning of a sub-period, at which the kth source was activated (from state 1), assuming that $1 \leqslant c^k \leqslant m^k$.

Let A_n^k denote the number of cells arrived over sub-period n, $n \geqslant 1$ (interval $[(n-1)T_s, nT_s]$); γ_n^k will denote the mean arrival rate over this interval. Arrivals are assumed to be declared at the end of the time slot over which they occur and the mean arrival rate over the nth sub-period is given by

$$\gamma_n^k \triangleq \mathrm{E}\{A_n^k\} = \lambda_j^k \quad \text{where } j = n \bmod (m^k + 1) - c^k + 1, \tag{2}$$

and n is assumed to be greater than or equal to c^k. Clearly, if $n < c^k$, then the source has not yet been activated, and $\gamma_n^k = 0$.

Note that the proposed source is capable of modeling the cell arrival dynamics of integrated services traffic sources at time-scales of one, T_s^k (sub-period) and T^k (period) slots (see Fig. 1). Since T_s can be greater than or equal to 1, the proposed traffic model is capable of capturing variations in the input process on many different time-scales. When $T_s = 1$, for instance, the proposed source model reduces to the periodic Markov chain model considered in [14], where the first and second moments of the queue length process are determined for an infinite-capacity queue in discrete time.

2.2. *The sub-period-level autocorrelation function*

In this section, the cell arrival autocorrelation function for a GPM source will be considered. In order to simplify notation, the source index k will be dropped for all quantities associated with the GPM model. The

normalized autocorrelation function $R(n)$, sometimes referred to as the correlation coefficient at lag n, where n is indexed at sub-period boundaries, is defined as

$$R(n) \triangleq \frac{\mathrm{E}\{A_{n'}A_{n'+n}\} - \mathrm{E}^2\{A_{n'}\}}{\mathrm{E}\{A_{n'}^2\} - \mathrm{E}^2\{A_{n'}\}}. \tag{3}$$

This function is derived in Appendix A for the case of Bernoulli arrival processes within each sub-period, and is given by the expression

$$R(n) = \frac{\frac{1}{m}\sum_{i=1}^{m}\lambda_i\lambda_j + \frac{1}{T_s}\left(\lambda - \frac{1}{m}\sum_{i=1}^{m}\lambda_i^2\right)\delta_n - \lambda^2}{\frac{1}{m}\sum_{i=1}^{m}\lambda_i^2 + \frac{1}{T_s}\left(\lambda - \frac{1}{m}\sum_{i=1}^{m}\lambda_i^2\right) - \lambda^2}, \tag{4}$$

where $j = (i + n - 1) \bmod(m) + 1$, and δ_n is the Kronecker delta.

The autocorrelation function for cells/sub-period will not be used to characterize the GPM sources until Section 3.2, where the arrival PMF $f_j(\cdot)$ for each sub-period is assumed to be Bernoulli. For arrival processes other than Bernoulli, the expression for ξ^2 given in (25) will be different and Eq. (4) will require slight modifications.

2.3. The superposition of N GPM traffic sources

Before considering a statistical multiplexer fed by a number of the proposed GPM sources, the aggregate input traffic, which is given by the superposition of the individual sources, will be characterized. A useful property of the GPM traffic model is that it is closed under superposition. That is, the superposition of N GPM sources, of equal sub-periods T_s, is also a GPM source with an underlying periodic Markov chain of some period M. Note that since the initialization time c^k is defined at sub-period boundaries, it is implicitly assumed that the sub-period boundaries are synchronized among all N sources. The sources are said to be completely synchronized when $c^1 = \cdots = c^N$.

It is easy to show that the period, M, of the GPM source associated with the aggregate source is given by

$$M = \mathrm{LCM}\{m^1, m^2, \ldots, m^N\}, \tag{5}$$

where LCM $\{\cdot\}$ denotes the least common multiple of its argument.

The PMF $\tilde{f}_j(\cdot)$ associated with state j of the aggregate GPM source, $1 \leqslant j \leqslant M$, is easily evaluated through the N-fold convolution

$$\tilde{f}_j = f_{\alpha^1}^1 * f_{\alpha^2}^2 * \cdots * f_{\alpha^N}^N. \tag{6}$$

The quantities α^k are functions of both j and the initialization constant c^k, and are given by

$$\alpha^k = \begin{cases} j + c^k - 1 & \text{if } 1 \leqslant j < c^k, \\ j - c^k + 1 & \text{if } c^k \leqslant j < c^k + m^k, \\ j - m^k - c^k + 1 & \text{if } c^k + m^k \leqslant j \leqslant M. \end{cases} \tag{7}$$

2.4. Multiplexer performance analysis

In this section a First-Come First-Served (FCFS) multiplexer fed by N GPM sources (Fig. 2) is analyzed in discrete time; its buffer capacity is assumed to be finite and is denoted by B.

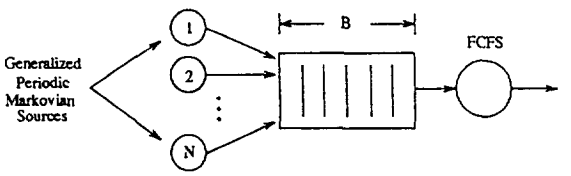

Fig. 2. The finite-capacity queueing system.

It is assumed that cell departures occur before new arrivals are counted at the end of each time slot. The system state is defined at sub-period boundaries in terms of the quantities (q_k, x_k), where q_k represents the queue length (multiplexer occupancy) at the end of sub-period k (time instant kT_s), and x_k denotes the state of the aggregate source at that time. It is easy to establish that the stochastic process $\{q_k, x_k\}_{k\geq 1}$ is a Markov chain with state space $\{(i, j) | 0 \leq i \leq B, 1 \leq j \leq M\}$, where i (j) denotes the current state of q_k (x_k). The probability that the Markov chain $\{q_k, x_k\}_{k\geq 1}$ moves from state (i, j) at time k (defined at sub-period boundaries) to state (i', j') at time $k+1$ will be denoted by $p(i, j, i', j')$. The corresponding transition matrix is sparse due to the simple periodic structure of the input process $\{x_k\}_{k\geq 1}$ and, in fact, a fraction of $(M-1)/M$ of the total number of elements are zero.

By mapping $p(i, j, i', j')$, for $0 \leq i, i' \leq B$ and $1 \leq j, j' \leq M$, into $p(l, l')$ through the transformation, $l = (j-1)(B+1)+i$ and $l' = (j'-1)(B+1)+i'$, the resulting transition matrix, $\mathbf{P}$, with elements $p(l, l')$, takes the canonical form of an irreducible periodic Markov chain of period M [15], given by

$$\mathbf{P} = \begin{bmatrix} 0 & \mathbf{P}_1 & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{P}_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{P}_{M-1} \\ \mathbf{P}_M & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{8}$$

The following proposition allows for the easy determination of the square matrices $\mathbf{P}_j$, $1 \leq j \leq M$, which are of order $(B+1)$.

Proposition 1. *Given some initial value for q_k at the beginning of the $(k+1)$st sub-period (system state is $(q_k = i, x_k = j)$), the finite queue will evolve exactly as an* Geo/D/1/B *queue, with i.i.d. batch arrivals given by $\tilde{f}_j$, over the $(k+1)$st sub-period (interval $[kT_s, kT_s + T_s]$).* □

In view of Proposition 1, it is easy to see that the matrices $\mathbf{P}_j$ are simply the T_s-step transition probability matrices for a finite queue, under the i.i.d. batch arrival process specified by $\tilde{f}_j(\cdot)$, where $1 \leq j \leq M$. The one-step transition probabilities $p_j(i, i')$, $0 \leq i, i' \leq B$, are given below.

For $0 \leq i \leq B$ and $\max(0, i-1) \leq i' \leq B-1$,

$$p_j(i, i') = \tilde{f}_j(i' - i + 1) + \tilde{f}_j(0) \cdot 1_{\{i+i'=0\}}, \tag{9}$$

where $1_{\{\cdot\}}$ denotes the indicator function.

For $0 \leq i \leq B$ and $i' = B$,

$$p_j(i, i') = \sum_{l=B-i+1}^{N_j} \tilde{f}_j(l), \tag{10}$$

where N_j denotes the maximum number of arrivals, as determined by (6), when the source is in state j. The elements of $\mathbf{P}_j$, denoted $p_j^{(T_s)}(i, i')$, can now be calculated by using the recursion

$$p_j^{(T_s)}(i, i') = \sum_l p_j^{(T_s-1)}(i, l) \cdot p_j(l, i'). \tag{11}$$

It is important to note that the dimensionality of the transition matrix $\mathbf{P}$, which is $M(B+1)\times M(B+1)$, prohibits the solution of systems with input processes having large periods, and consequently, large values of M. The following theorem, however, provides for the exact stationary solution of $\{q_k, x_k\}_{k\geqslant 1}$ by solving M Markov chains with transition matrices of order $(B+1)$ and thus, significantly reduces the numerical complexity involved in calculating the stationary probability vector of the transition matrix in (8). Its proof can be found in [16].

Theorem 2. *The finite irreducible periodic Markov chain, with transition matrix given by* (8), *has a unique stationary probability vector,* $\boldsymbol{\pi} = (\pi_{0,1} \cdots \pi_{B,M})$, *given by*

$$\boldsymbol{\pi} = \left(\frac{1}{M}\right)\begin{bmatrix} \boldsymbol{\pi}_1 \\ \boldsymbol{\pi}_2 \\ \vdots \\ \boldsymbol{\pi}_M \end{bmatrix}, \tag{12}$$

where $\boldsymbol{\pi}_i$ *is the unique stationary probability vector for* $\mathbf{A}_i = \mathbf{P}_i \cdots \mathbf{P}_M \cdot \mathbf{P}_1 \cdots \mathbf{P}_{i-1}$ $(i = 2, \ldots, M)$, *and* $\mathbf{A}_1 = \mathbf{P}_1 \cdots \mathbf{P}_M$. □

Upon solving for $\boldsymbol{\pi} = (\pi_{0,1} \cdots \pi_{B,M})$ using Theorem 2, the queue length probability distribution at sub-period boundaries is easily evaluated as

$$q(k) = \lim_{n\to\infty} \Pr\{q_n = k\} = \sum_{j=1}^{M} \pi_{k,j}, \quad k = 0, \ldots, B. \tag{13}$$

Let $\hat{\pi}_{k,j}$ denote the stationary probability that the system is in state (k, j) at an arbitrary time instant (slot boundary). These quantities are derived in Appendix B through the use of a renewal/regenerative theory-based approach. The queue length distribution $\hat{q}(k)$ at arbitrary time instants can now be computed as in (13). Note that when $T_s = 1$ (sub-period is equivalent to the slot), the queue length distribution at arbitrary times is simply $\hat{q}(k) = q(k)$ for $0 \leqslant k \leqslant B$.

The cell-loss probability L is also derived in Appendix B using the same approach followed for the derivation of $\hat{\pi}_{k,j}$, and is given by

$$\begin{aligned} L &\triangleq \Pr\{\text{arriving cell is dropped}\} \\ &= \frac{1}{\lambda T_s} \sum_{i'=0}^{B} \sum_{j'=1}^{M} \sum_{n=1}^{T_s} \sum_{i=I_1}^{B} \sum_{k=K_1}^{K_2} k p_{j'}^{(n-1)}(i', i)\, \tilde{f}_{j'}(B+1-i+k)\, \pi_{i',j'}, \end{aligned} \tag{14}$$

where $I_1 = \max(0, B+2-N)$, $K_1 = \max(0, i'-B-1)$, $K_2 = N-B-1+i'$ and λ denotes the total offered load, which is given by

$$\lambda = \sum_{k=1}^{N} \lambda^k. \tag{15}$$

3. Application 1: Non-Deterministic Periodic (NDP) sources

An important characteristic of high-speed packet networks such as ATM, which is not considered in the relevant performance studies of [1–6] is the potentially large disparity in the relative speeds of an output link and an input source. This disparity, or *speed-up factor* – commonly appearing at the network edges where a number of sources are being multiplexed onto an output link – results in successive cells from an active source

being spaced by the interval of time required to generate enough bits to form a cell. The speed-up factor $m^k = C/r^k$, where C is the link speed (in slots/sec) and r^k is the source speed is also relevant in inter-networking, where a number of slower packet networks of various speeds are connected to the B-ISDN through a common gateway. In addition, it has been shown in [17] that, in a networking environment where input and output lines are of the same speed, source periodicities like those mentioned above give rise to cell arrival streams within the network which also exhibit significant periodicities. This observation has been used in [18] to justify the adoption of periodic input cell arrival streams for the simulation study of an ATM switch.

One way to study the speed-up factor analytically is by considering the multiplexing of deterministic periodic sources [19,20], which implies that all of the input sources are Constant Bit Rate (CBR). A more realistic traffic model for high-speed packet networks such as ATM would be a Non-Deterministic Periodic (NDP) model in which cells are generated every m^k slots according to a stochastic process, thereby allowing for a potential statistical multiplexing gain.

A relevant question that arises when considering the multiplexing of NDP sources is whether or not the aggregation of such sources can be approximated by any of the commonly used traffic models. For instance, it is widely accepted that the superposition of a large number of uncorrelated cell arrival streams can be approximated by a Poisson process, which allows for a simple analytical solution to many queueing problems. However, this limiting assumption may not be valid when the cell arrival processes have a periodic structure. In fact, it will be shown that for the superposition of NDP sources, the Poisson assumption can be either pessimistic or optimistic, depending upon the operating conditions.

A correlated traffic model in which cells are generated at a minimum spacing of s^k slots is considered in [21] and [22], where the cell arrival process is governed by an underlying 2-state Markov chain. The speed-up factor s^k, which is equal to the unit step transition time of the Markov chain for source k, is shown to have a significant impact on multiplexer performance. Specifically, as s^k increases, the inter-cell correlations present at the source become less dominant and the multiplexer's performance improves accordingly. These observations suggest that, for sufficiently large values of s^k, it may be possible to drop the correlated source model in favor of a simpler, uncorrelated model without significantly impacting on the model's accuracy.

3.1 GPM traffic model for NDP sources

The proposed GPM source can be easily adopted to model such an uncorrelated source with speed-up factor m^k by setting T_s equal to 1 slot; the kth source is then described by the parameters $\{m^k, 1, f_j^k(\cdot)\}$. Note that this is equivalent to setting $T^k = m^k$. Source k is assumed to deliver cells every m^k time slots according to a Bernoulli PMF with rate λ^k. This is captured through the model by setting, for $j = c^k$,

$$f_j^k(i) = \begin{cases} \lambda^k & \text{for } i = 1, \\ 1 - \lambda^k & \text{for } i = 0, \end{cases} \tag{16}$$

and for $j \neq c^k$,

$$f_j^k(i) = \begin{cases} 0 & \text{for } i = 1, \\ 1 & \text{for } i = 0. \end{cases} \tag{17}$$

Note that the analysis presented in Section 2 provides results for a given vector of source start times, while the analyses of [19] and [20] assume random starting times among the sources. For the deterministic (periodic) source models assumed in [19] and [20], random starting times represent the only stochastic element of the resulting queueing system, without which the analysis would be trivial. A more meaningful treatment of the problem, which is adopted here, may be one in which specific – perhaps worst-case – starting times can be assumed. A stochastic cell generation process can then be considered at the deterministic time-scales m^k, which is more representative of realistic traffic generators such as packetized voice and network uplinks with speed mismatches.

3.2. Numerical results for NDP input processes

Although the analysis described in Section 2.4 allows for the superposition of input sources having different periods, the presentation and discussion of the numerical results will be facilitated by considering sources with identical periods denoted by m. Each source is also assumed to deliver cells according to a Bernoulli PMF with rate α, which results in a mean rate per source of α/m; the total offered load is then given by $\lambda = N\alpha/m$.

Fig. 3 illustrates the effect that the source period has on cell-loss probabilities as a function of the buffer capacity B when $N = 40$ sources are multiplexed at an offered load of $\rho = 0.9$ and the sources are assumed to be synchronized ($c^1 = \cdots = c^{40}$). Clearly, the source period m has a significant impact on queue performance when the total offered load is held constant. Note, however, that this impact is not necessarily negative as has been suggested in the past [18]. In fact, for "acceptable" loss probabilities on the order of 10^{-6} and smaller, the impact that increasing values of m have on the queue performance is positive. The curves for $m = 10$, 20, and 30 are also compared with the corresponding M/D/1/B approximation, and the results are quite enlightening. Fig. 3 reveals that the Poisson assumption can either overestimate or underestimate loss probabilities depending upon buffer capacity and the source period m, which implies that the solution to the M/D/1/B queue should not be used for bounding purposes when considering such periodic input streams. For a loss probability of 10^{-9}, the M/D/1/B approximation is extremely pessimistic, requiring a buffer capacity of $B = 90$, compared to a required buffer capacity of $B = 50$ when $m = 30$. In fact, the negative slope of the M/D/1/B curve is always smaller than the cases for $m > 1$, which implies that, beyond some value of B (which varies with m), the loss probabilities for the limiting Poisson traffic assumption upper bounds the actual loss probabilities for the superposition of non-deterministic periodic sources.

Fig. 4 demonstrates the relationship between cell-loss probabilities and offered load for a case in which the Poisson assumption is optimistic ($B = 30$) as well as a case in which the assumption is overly pessimistic ($B = 50$). The $N = 40$ NDP sources are assumed to have speed-up factors of $m = 30$ associated with them. When $B = 30$, the superposition of NDP input sources results in loss probabilities larger than 10^{-5} for the range of loads between 0.7 and 0.99. The corresponding M/D/1/30 queue, however, is overly optimistic and results in loss probabilities smaller than 10^{-9} for loads as high as 0.72. On the other hand, when $B = 50$, the M/D/1/50 approximation overestimates the loss probabilities for all loads. It is clear from Fig. 4 that the Poisson assumption performs poorly for low to moderate loads, and even for offered loads of 0.99 there is a substantial difference between the M/D/1/B queue performance and the corresponding case of NDP input sources.

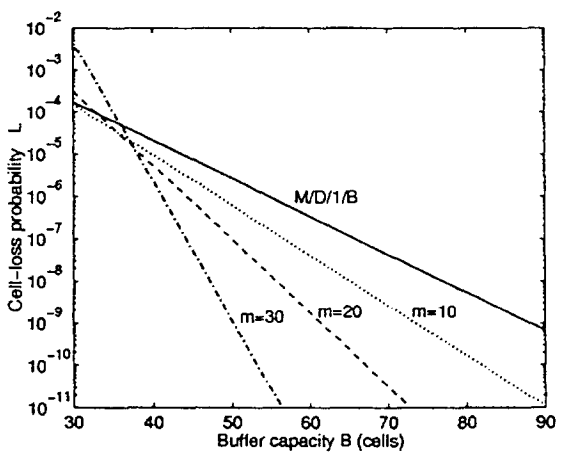

Fig. 3. Loss probability versus buffer capacity B for various m. The number of sources are $N = 40$, and the total load $\lambda = 0.9$.

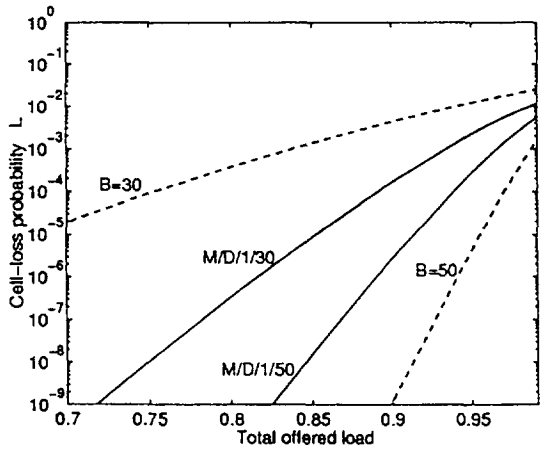

Fig. 4. Loss probability versus offered load for various B. The number of sources are $N = 40$, and speedup factors are set to $m = 30$.

Note that for the results presented in Figs. 3 and 4, it is assumed that the NDP sources are synchronized ($c^1 = \cdots = c^N$), which represents a worst-case scenario in terms of queue performance. Although it is difficult to define a meaningful distribution of initialization times c^k ($1 \leqslant k \leqslant M$) for the NDP sources, it is reasonable to expect that uniformly distributing these times will improve the queueing performance. This is illustrated in Fig. 5, where loss probabilities are plotted versus the number of multiplexed sources with $m = 20$ and $\rho = 0.9$ for synchronized and unsynchronized (initialization times uniformly distributed) sources. The uppermost curves represent results for a buffer capacity of $B = 20$, and the lower curves for $B = 50$. Similarly, the upper star on the right hand vertical axis denotes the M/D/1/20 result while the lower star denotes the M/D/1/50 result. For both cases ($B = 20$ and $B = 50$), the queue performance, in terms of cell-loss probabilities, approaches a limit for large values of N. However, even when the NDP sources are unsynchronized, the limiting results are significantly different than the M/D/1/B results.

The results presented thus far indicate that all of the parameters (B, m, N and λ) affect significantly the queue performance, and consequently, the validity of the M/D/1/B approximation. For each set of numerical results there seem to be two distinct regions of operation; one in which the source periodicities have a negative

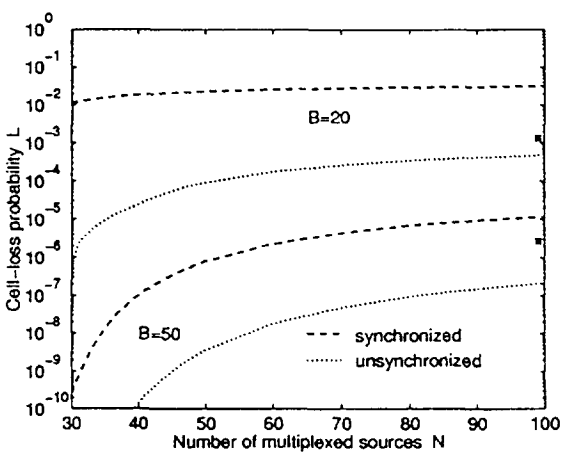

Fig. 5. Loss probability versus N for various B. The total offered load is $\lambda = 0.9$, and speedup factors are set to $m = 20$.

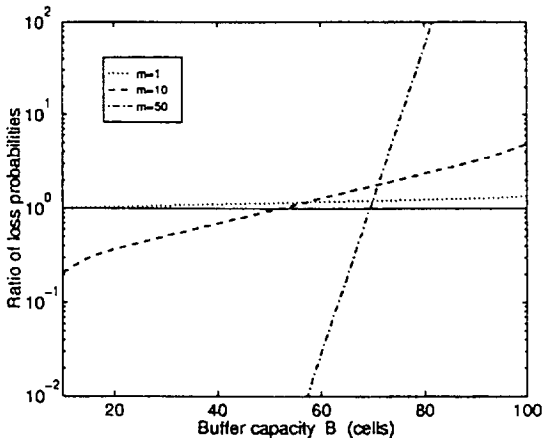

Fig. 6. L_P / L_{NDP} versus B for different values of m. The total offered load is $\lambda = 0.9$, and $N = 75$ sources are multiplexed.

influence on the queueing behavior, relative to a Poisson input source, and one in which the periodicities have a favorable impact. These operating regions are examined more closely in Fig. 6, where the ratio of the loss probability for the Poisson traffic source (L_P) to the loss probability for the NDP sources (L_{NDP}) is plotted versus buffer capacity B. When the ratio L_P/L_{NDP} is less than (greater than) 1, the M/D/1/B approximation underestimates (overestimates) the actual loss probability. In Fig. 6, $N = 75$ non-deterministic periodic sources are multiplexed at a total offered load of 0.9 for source periods of 1, 10 and 50. When $m = 1$, the input to the multiplexer consists of $N = 75$ Bernoulli sources, and the ratio L_P/L_{NDP} is close to 1 for all values of B. This result supports the assertion that the superposition of a large number of Bernoulli traffic streams can be approximated by a Poisson process (asymptotic behavior of the Binomial distribution), regardless of the buffer capacity. However, as the source periods increase, the two regions become more distinct. When $m = 10$ and $B = 10$, L_{NDP} is 5 times greater than L_P for the M/D/1/10 queue, and when $B = 100$, L_{NDP} is 5 times smaller than L_P. For this case, the M/D/1/B results are overly pessimistic for $B \geqslant 50$. When $m = 50$, the two regions are even more well-defined, and the Poisson assumption does not perform well in either region. It is interesting to note that, as was the case in Fig. 3, the region of operating conditions which produces loss probabilities considered acceptable in ATM (less than 10^{-6}) is the one in which the M/D/1/B queue provides overly pessimistic results ($L_P > L_{NDP}$).

4. Application 2: Multiplexing VBR video at slice-level

While the modeling of VBR video sources has recently received significant attention, there presently exists no widely accepted model which lends itself to mathematical analysis. Most existing models, like those in [7] and [8], assume that video is pre-buffered at the frame level. The purpose of pre-buffering is simply to provide

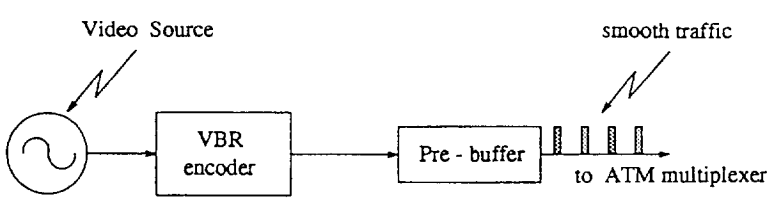

Fig. 7. Pre-buffering a VBR video source.

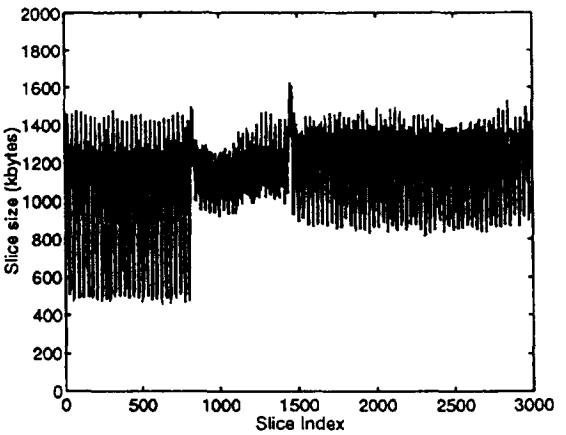

Fig. 8. Slice-level sample of a VBR video sequence.

the network with a cell stream which is as smooth as possible; this can be accomplished by simply storing data temporarily in a pre-buffer as in Fig. 7.

In [11], a flexible simulation model is developed which captures the behavior of VBR video traffic at both the frame and slice level. There are 30 equal-sized slices in every frame, and the inter-frame as well as the inter-slice autocorrelation functions are matched to those from experimental data provided by [12]. The cell generation rate is considered to be constant throughout the slice, modeling the VBR video traffic scheme delivered to the multiplexer under a slice-level pre-buffering scheme. Through a simulation study driven by a number of source models which match the characteristics of a full-motion video movie, [12], the authors show that the impact of inter-frame correlation is negligible while the impact of inter-slice correlation is significant.

These simulation results suggest that if pre-buffering at the frame level is not practical, and transmission takes place at the slice level, then an effective traffic model should capture the time-scales associated with both the frame and the slice. The former will determine the time-scale at which inter-frame correlations exist, while the latter will determine the horizon of the constant-rate cell delivery to the multiplexer (slice). Note that under pre-buffering at the frame level a single rate of the resulting traffic over a frame is generated. Under pre-buffering at the slice level, and for a selected average cell rate over a frame, 30 potentially different and correlated rates of the resulting traffic over a frame are produced; each rate remains fixed over a slice.

Due to the high positive autocorrelations in the number of cells generated in a frame, the cell arrival process (at slice level) over a frame can be considered to remain virtually unchanged over a sequence of consecutive frame horizons. This is in agreement with the experimental data from [12] (Fig. 8) in which the slice-level arrival rate is shown to have a periodic behavior (with period 30 slices), and the frame-level arrival process remains relatively constant over many frames (or periods). Assuming that the time horizon associated with changes in the frame-level arrival rate is sufficiently long to induce the steady-state behavior of the multiplexer, a quasi-static approximation similar to that in [9] can be used to study the multiplexing of VBR traffic which is pre-buffered at the slice level as described above. However, a constant-rate arrival process over a frame, such as the Poisson process considered in [9], will clearly not be applicable here to approximate the arrival process over a number of video frames. Instead, assuming a frame-level arrival rate which remains constant over many frames, a periodic cell arrival process (with a period equal to the frame length) within each frame could be considered.

4.1. GPM traffic model for the periodic slice-level arrival process

The periodic cell arrival process described above can be easily captured through the proposed GPM traffic model by setting the period of each GPM source (m^k) to be 30 and the sub-period (T_s) equal to the number of

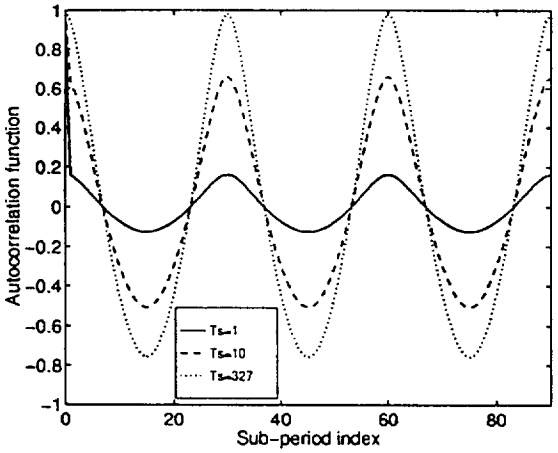

Fig. 9. Sub-period-level autocorrelation function for $\lambda^k = 0.2$, $b = 0.2$ and various values of T_s.

slots in a slice; T_s varies with the link speed, the frame rate (in frames/sec) and the slot capacity. Assuming a frame rate of 24 per second, as in [11] and [12], and a slot capacity of 53 bytes (48 bytes of payload and 5 bytes of header) for ATM,

$$T_s = \frac{C \text{ bits}}{\text{sec}} \cdot \frac{1 \text{ byte}}{8 \text{ bits}} \cdot \frac{1 \text{ slot}}{53 \text{ bytes}} \cdot \frac{1 \text{ frame}}{30 \text{ slices}} \cdot \frac{1 \text{ sec}}{24 \text{ frames}}$$

$$= \frac{C}{305{,}280} \text{ slots/slice}, \tag{18}$$

where C is the link capacity in units of bits/sec. To capture the uniform distribution of cells within a slice, the PMF $f_j^k(\cdot)$ for the number of cells generated in each slot when source k is in state j ($1 \leqslant j \leqslant 30$) will be Bernoulli with rate λ_j^k. Note that this can be seen as statistically smoothing the source at the slice level. It is believed that an assumption of statistical smoothing should provide for a conservative (or pessimistic) estimation of delivered QoS if the VBR source were deterministically smoothed (cells are delivered to the network equi-spaced within each slice). This has been shown to be the case in [23] when the VBR source is pre-buffered and smoothed on a frame by frame basis.

As was the case with the slice-level arrival rate process, the slice-level autocorrelation function for a VBR video source has been shown [10,12,13] to possess a periodic structure with a period of 30 slices. The GPM traffic model also has a periodic sub-period-level autocorrelation function $R^k(n)$, as given in Eq. (4). In fact, for a given frame-level cell arrival rate for source k, denoted by λ^k in Eq. (1), the slice-level autocorrelation function for the kth video source can be approximated by $R^k(n)$ through an appropriate selection of the arrival rates λ_j^k, $1 \leqslant j \leqslant 30$.

In order to produce a first approximation for the autocorrelation function for cells/slice which is similar to the functions found in [24], for example, the following distribution of the rates λ_j^k will be considered [3]:

$$\lambda_i^k = \lambda_{(31-i)}^k = a e^{-(i-1)b}, \quad i = 1, 2, \ldots, 15. \tag{19}$$

[3] Note that there exist many possible choices for the rate distribution function λ_i^k. The distribution given in Eq. (19) is intended only to serve as an example, and is not intended to closely match any experimental data set.

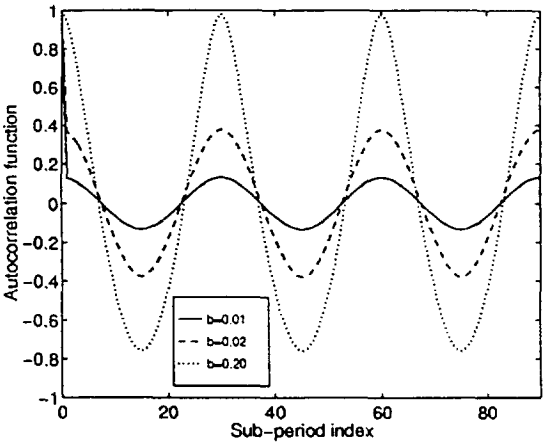

Fig. 10. Sub-period-level autocorrelation function for $\lambda^k = 0.2$, $T_s = 327$ and various values of b.

Upon choosing a value of b to produce the desired profile for the autocorrelation function, the constant a is calculated using the expression

$$\lambda^k = \frac{a}{15} \sum_{i=1}^{15} e^{-(i-1)b}, \tag{20}$$

which is derived from (1) by requiring that the mean rate over states 1 through 15 (which equals the mean rate over states 16 through 30) be equal to the frame cell rate λ^k for the source. Note that according to (4), the autocorrelation function for cells/sub-period is dependent upon both T_s and the sub-period rates λ_j^k. Fig. 9 shows the resulting sub-period-level autocorrelation functions for a source with frame cell rate $\lambda^k = 0.2$, $b = 0.2$ and different values of T_s ranging from 1 to 327. The latter corresponds to a video source being multiplexed onto a 100 Mbps ATM link (see Eq. (18)).

Similarly, Fig. 10 presents sub-period-level autocorrelation functions for the source with $T_s = 327$ and different sub-period rate distributions resulting by varying b from 0.01 to 0.20. Note that as b approaches 0, the GPM source looks more and more like an i.i.d. cell arrival process. In fact, when $b = 0$, the GPM source is simply a binomially distributed batch arrival process. It is clear from Figs. 9 and 10 that similar sub-period-level autocorrelation functions can be produced by either fixing b and varying T_s, or fixing T_s and varying b.

4.2. Numerical results and discussion for the VBR application

In this section some numerical results are presented for the analysis of the finite capacity multiplexer with an input process consisting of the superposition of N GPM traffic sources, which are adopted as explained in the previous section to model VBR video sources pre-buffered at the slice level. The stationary queue distributions computed using the techniques presented in Section 2.4 yield queueing results for video sources generating traffic according to a single average frame rate. In order to compute the queue length distributions over all possible frame rates, a frame rate histogram can be derived as described in [9], and the quasi-static approximation presented there can be applied. Application of the quasi-static approximation is beyond the scope of this paper. The primary objective here is to determine the impact of periodic slice-level autocorrelations on queue response.

To address the effectiveness of the slice-level autocorrelation function in describing a VBR input source, consider the functions pictured in Fig. 11. Although the two autocorrelation functions are almost identical, one is produced by an underlying periodic Markov chain which makes transitions every time slot ($T_s = 1$), while the

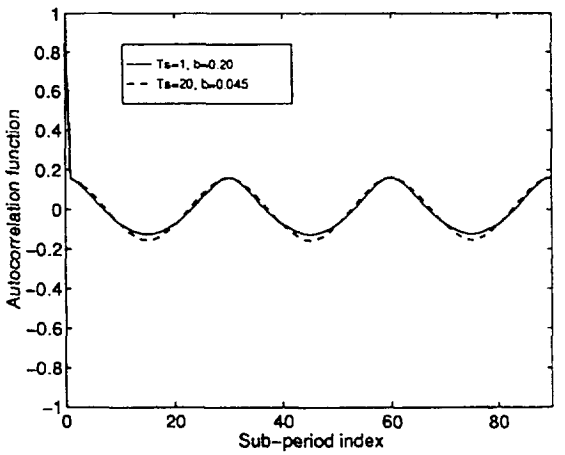

Fig. 11. Sub-period-level autocorrelation functions for sources with $\lambda^k = 0.2$, and different values of T_s.

other is produced by a source making transitions every 20 slots ($T_s = 20$). The autocorrelation functions are matched by simply varying the distribution of sub-period rates through the parameter b; the total rate for each GPM source is $\lambda^k = 0.2$. The queue response to these sources is examined by considering a multiplexer of capacity $B = 30$ cells with an input of 4 of the GPM sources. The input sources with $T_s = 20$ slots have a much more negative impact on the multiplexer than those with $T_s = 1$, which is illustrated in Fig. 12 through tail distributions for buffer occupancy as well as loss probabilities L. It is clear that simply matching the autocorrelation function for cells/slice is not sufficient to estimate multiplexer performance. The actual time-scale associated with the slice needs to be considered carefully.

In fact, by adopting different source time-scales T_s, performance studies can lead to contrasting conclusions about the feasibility of multiplexing VBR video which is not pre-buffered at the frame level. For instance, based on numerical results for input sources with periodic autocorrelations which make transitions every time slot, the author in [25] concludes that loss probabilities are higher when VBR video sources are pre-buffered and smoothed on a frame by frame basis. A similar conclusion would be drawn using GPM sources with $T_s = 1$. For

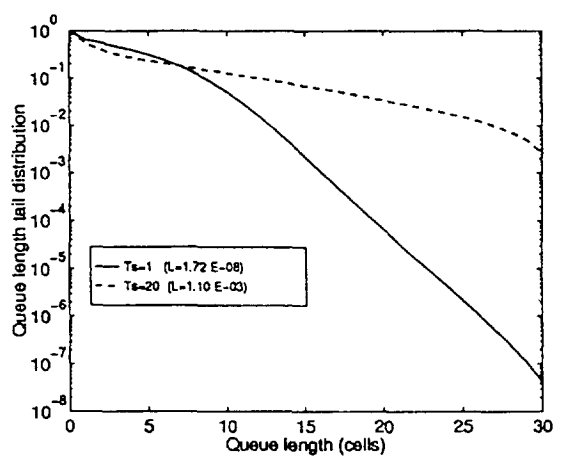

Fig. 12. Tail distributions and loss probabilities for matching autocorrelation functions with $T_s = 1$ and 20; total load is $\lambda = 0.8$ in both cases.

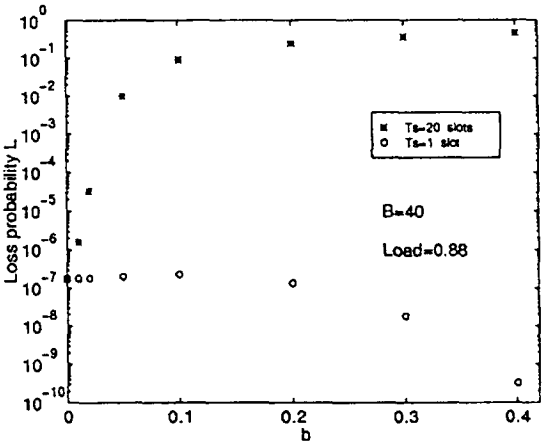

Fig. 13. Cell loss probabilities for $N = 4$ GPM sources, $B = 40$ and T_s of 1 and 20; total load is $\lambda = 0.88$.

GPM input sources, a frame-level pre-buffering scheme is modeled by setting constant arrival rates over all 30 sub-periods (or slices), which is the case when $b = 0$ under the rate distribution strategy of Eq. (19). Note that this eliminates the periodic nature of the slice-level autocorrelation function.

In Fig. 13 cell loss probabilities are plotted versus b for $N = 4$ input sources, each with rate 0.22, at a multiplexer of capacity $B = 40$ for $T_s = 1$ and $T_s = 20$. As expected, when $b = 0$ (which models frame-level pre-buffering) the rates do not change from slice to slice and the loss probability does not depend upon T_s. When $T_s = 1$, notice that loss probabilities decrease as the periodic fluctuations in arrival rate increase with b. Therefore, use of this model would lead to the false conclusion, as in [25], that the periodic autocorrelations provide for improved statistical multiplexing gain over frame-level pre-buffering. As T_s is increased to as little as 20 slots, which corresponds to an output link of capacity $C = 6.1$ Mbps (Eq. (18)) for the adopted GPM model, numerical results for L lead to a completely different conclusion. In this case it would seem that smoothing the periodic autocorrelations through frame-level pre-buffering provides greatly improved QoS. These results illustrate the importance of incorporating time-scales into traffic models in order to accurately assess the impact of the modeled traffic sources on network performance.

In order to investigate a more realistic multiplexing scenario, Table 1 depicts cell loss probabilities for a buffer capacity of $B = 50$, an output link capacity of 100 Mbps, and an input consisting of 10 identical VBR sources each transmitting at a rate of 8 Mbps. This scenario can be modeled with GPM sources through the following choice of parameters: $T_s = 327$, $\lambda^1 = \cdots = \lambda^{10} = 0.08$ and of course $m^1 = \cdots = m^{10} = 30$. As explained in Section 4.1, each source generates cells according to the appropriate Bernoulli PMF to approximate the smoothing of cells within a slice.

In Table 1, three different autocorrelation functions for cells/slice are considered. One set of results represents the frame-level pre-buffering case ($b = 0$) while the others correspond to two of the autocorrelation

Table 1
Loss probabilities for different values of b under both *synchronized* and *uniformly distributed* frame boundaries

b	L for *synchronized* frame boundaries	L for *uniformly distributed* frame boundaries
0.00	$6.57\,e^{-12}$	$6.57\,e^{-12}$
0.10	$7.43\,e^{-10}$	$6.54\,e^{-12}$
0.20	$4.67\,e^{-7}$	$6.45\,e^{-12}$

curves pictured in Fig. 10 ($b = 0.01$ and $b = 0.02$). Clearly, the periodic autocorrelation function has a negative impact on the queue, and increasing the positive and negative peaks in the autocorrelation function by increasing b results in even worse performance. Note that the worst-case scenario in terms of possible starting times among the input sources is achieved when the frame boundaries of all 10 sources are synchronized ($c^1 = \cdots = c^{10}$).

The impact of frame-boundary synchronization can be minimized by uniformly distributing the frame starts (boundaries) at intervals of $\lfloor M/N \rfloor$, as is the case in [11], which represents a best-case scenario in terms of source starting times. While ensuring such a uniform distribution of frame starts in practice is highly unlikely, the positive impact on the multiplexer is indisputable, as is seen in Table 1, where loss probabilities virtually coincide for all values of b.

Recall that the results presented here are for one possible combination of frame cell rates, and that under a quasi-static approximation, all combinations of the individual frame cell rates would be considered. However, it should be noted that the averaging effect of the quasi-static approximation may mask the extreme queueing conditions that occur over time horizons for which the offered load is high. Hence, it may be more reasonable, when applying the quasi-static approximation, to consider only the "high-rate" bins, which would yield a more detailed account of the queueing behavior during the times in which the QoS requirements for the VBR sources are most likely to be violated. It is possible that these QoS violations, which may be sufficiently rare when considered over all possible combinations of frame cell rates, are severe enough during the high-load periods to disrupt, or at least impair, service.

5. Conclusions

In this paper, the concept of source time-scales was applied to the discussion of traffic sources in broadband packet networks such as ATM. A Generalized Periodic Markovian (GPM) traffic model was introduced to capture the cell arrival dynamics associated with these time-scales, and a finite-capacity multiplexer with an input process consisting of the superposition of N GPM sources was analyzed in discrete time. An efficient numerical technique was presented for the solution of a potentially large irreducible Markov chain by solving smaller chains with state space size independent of the number of states in the input process.

As expected, numerical results suggest that the time-scales at which variations in the cell arrival process occur, impact significantly on the queue response. This fact seems to indicate that these relevant time-scales, which change with the link speed, should be considered as important traffic descriptors, and incorporated into source models in high-speed environments.

The proposed GPM traffic source model and developed analytical technique were applied to the study of Non-Deterministic Periodic (NDP) traffic sources as well as to VBR video sources which deliver cells at the slice level, as opposed to smoothing and delivering them on a frame by frame basis. Results from the NDP source study indicate that, depending upon the operating conditions at the finite-capacity multiplexer, the period of the NDP input streams may have either a positive or negative impact on queueing performance, relative to that of a M/D/1/B queue. This implies that M/D/1/B results should not be used for bounding loss probabilities when studying the queue response to NDP input processes. Furthermore, the asymptotic (large N) behavior of NDP input sources was shown to be poorly approximated by the Poisson process when the potential cell generation times occur periodically with period greater than 1 (i.e. $m > 1$).

Results from the VBR video application illustrate that the inter-slice autocorrelation function, which is periodic in nature, has a major impact on the finite-capacity queue. The video application was based on the assumption, first made in [9], that the system can be approximately analyzed by independently studying the queue response to different frame cell rates and then conditioning over all rates. Numerical results, which were derived for a single frame cell rate, indicate that frame-level pre-buffering provides improved QoS when slice-level autocorrelations are periodic.

Appendix A. Derivation of the sub-period-level autocorrelation function $R(n)$

This appendix is concerned with the derivation of $R(n)$, defined as

$$R(n) \triangleq \frac{\mathrm{E}\{A_{n'}A_{n'+n}\} - \mathrm{E}^2\{A_{n'}\}}{\mathrm{E}\{A_{n'}^2\} - \mathrm{E}^2\{A_{n'}\}}. \tag{A.1}$$

The quantity $\mathrm{E}\{A_{n'}A_{n'+n}\}$ will be derived by considering the cases $n = 0$ and $n > 0$ separately.
For $n = 0$,

$$\begin{aligned}\mathrm{E}\{A_{n'}^2\} &= \sum_k k^2 \Pr\{A_{n'} = k\} \\ &= \sum_i \sum_k k^2 \Pr\{A_{n'} = k \mid \gamma_{n'} = \lambda_i\} \Pr\{\gamma_{n'} = \lambda_i\}.\end{aligned} \tag{A.2}$$

The second moment of $A_{n'}$, given $\gamma_{n'}$, will be denoted by ξ_2, where

$$\xi_2 = \sum_k k^2 \Pr\{A_{n'} = k \mid \gamma_{n'} = \lambda_i\}. \tag{A.3}$$

This quantity can be expressed in terms of the GPM parameters defined earlier once the cell arrival PMF $f_i(\cdot)$, $1 \leqslant i \leqslant m$, is known. For instance, if $f_i(\cdot)$ is a Bernoulli PMF with rate λ_i, then $\Pr\{A_{n'} = k \mid \gamma_{n'} = \lambda_i\}$ is simply a binomial distribution given by

$$\Pr\{A_{n'} = k \mid \gamma_{n'} = \lambda_i\} = \binom{T_s}{k} (\lambda_i)^k (1 - \lambda_i)^{T_s - k}. \tag{A.4}$$

In this case, it is easy to show that ξ_2 is simply

$$\xi_2 = T_s^2\lambda_i^2 + T_s\lambda_i(1 - \lambda_i), \tag{A.5}$$

and (22) can be expressed as

$$\begin{aligned}\mathrm{E}\{A_{n'}^2\} &= \sum_i \left[T_s^2\lambda_i^2 + T_s\lambda_i(1 - \lambda_i)\right]\pi_i \\ &= T_s^2 \sum_i \lambda_i^2\pi_i + T_s\lambda - T_s\sum_i \lambda_i^2\pi_i.\end{aligned} \tag{A.6}$$

For $n > 0$,

$$\begin{aligned}\mathrm{E}\{A_{n'}A_{n'+n}\} &= \sum_{k_1}\sum_{k_2} k_1k_2 \Pr\{A_{n'} = k_1,\ A_{n'+n} = k_2\} \\ &= \sum_i \sum_j \sum_{k_1} \sum_{k_2} k_1k_2 \Pr\{A_{n'} = k_1 \mid \gamma_{n'} = \lambda_i\} \Pr\{A_{n'+n} = k_2 \mid \gamma_{n'+n} = \lambda_j\} \\ &\qquad \cdot \Pr\{\gamma_{n'+n} = \lambda_j \mid \gamma_{n'}\lambda_i\} \Pr\{\gamma_{n'} = \lambda_i\} \\ &= T_s^2 \sum_i \sum_j \lambda_i\lambda_j p_{i,j}^{(n)}\pi_i.\end{aligned} \tag{A.7}$$

From (A.6) and (A.7), the quantity $\mathrm{E}\{A_{n'}A_{n'+n}\}$, for $n \geqslant 0$, can be expressed as

$$\mathrm{E}\{A_{n'}A_{n'+n}\} = T_s^2 \sum_i \sum_j \lambda_i\lambda_j p_{i,j}^{(n)}\pi_i + T_s\left(\lambda - \sum_i \lambda_i^2\pi_i\right)\delta_n, \tag{A.8}$$

where δ_n is the Kronecker delta.

This expression can be further simplified by noting that, due to the periodic structure of the Markov chain considered here, the stationary probability π_i is simply $1/m$, where m represents the number of states in the periodic Markov chain. Also note that the one-step transition probabilities for this source are defined by

$$p_{i,j} = \begin{cases} 1 & \text{if } 1 \leqslant i \leqslant m \text{ and } j = (i) \bmod (m) + 1, \\ 0 & \text{otherwise.} \end{cases} \tag{A.9}$$

From these quantities it is an easy computation to show that the n-step transition probabilities are given by

$$p_{i,j}^{(n)} = \begin{cases} 1 & \text{if } 1 \leqslant i \leqslant m \text{ and } j = (i + n - 1) \bmod (m + 1), \\ 0 & \text{otherwise.} \end{cases} \tag{A.10}$$

Eq. (A.8) can now be rewritten as

$$\mathrm{E}\{A_{n'} A_{n'+n}\} = \frac{T_s^2}{m} \sum_{i=1}^{m} \lambda_i \lambda_j + T_s \left(\lambda - \frac{1}{m} \sum_{i=1}^{m} \lambda_i^2 \right) \delta_n, \tag{A.11}$$

where $j = (i + n - 1) \bmod(m) + 1$.

Finally, $\mathrm{E}\{A_{n'}\}$ can be derived as follows:

$$\begin{aligned} \mathrm{E}\{A_{n'}\} &= \sum_k k \Pr\{A_{n'} = k\} \\ &= \sum_i \sum_k k \Pr\{A_{n'} = k \mid \gamma_{n'} = \lambda_i\} \Pr\{\gamma_{n'} = \lambda_i\} \\ &= \sum_i T_s \lambda_i \pi_i \\ &= T_s \lambda. \end{aligned} \tag{A.12}$$

By substituting (A.11) and (A.12) into (A.1), the final expression for the autocorrelation function $R(n)$ is obtained:

$$R(n) = \frac{\dfrac{1}{m} \sum_{i=1}^{m} \lambda_i \lambda_j + \dfrac{1}{T_s} \left(\lambda - \dfrac{1}{m} \sum_{i=1}^{m} \lambda_i^2 \right) \delta_n - \lambda^2}{\dfrac{1}{m} \sum_{i=1}^{m} \lambda_i^2 + \dfrac{1}{T_s} \left(\lambda - \dfrac{1}{m} \sum_{i=1}^{m} \lambda_i^2 \right) - \lambda^2}, \tag{A.13}$$

where again $j = (i + n - 1) \bmod(m) + 1$.

Appendix B. Derivation of the quantities $\hat{\pi}_{i,j}$ and L

Let X_n denote the process $\{q_n, x_n\}_{n \geqslant 0}$, where n is indexed at *slot* boundaries. Recall that the process $\{q_k, x_k\}_{k \geqslant 1}$, which has been considered for the analysis of the GPM sources at a finite-capacity multiplexer, is indexed at sub-period boundaries; the length of a sub-period is equal to T_s slots. The Markov chain $\{q_k, x_k\}_{k \geqslant 1}$ has state-space $S = \{(0, 1), \ldots, (B, M)\}$. Note that $\{q_k, x_k\}_{k \geqslant 1}$ is embedded in the process $\{q_n, x_n\}_{n \geqslant 0}$, and $\{X_{(k-1)T_s}\}_{k \geqslant 1} = \{q_k, x_k\}_{k \geqslant 1}$.

This appendix is concerned with the computation of the stationary probabilities for the process X_n, denoted by $\hat{\pi}_{i,j}$ where

$$\hat{\pi}_{i,j} = \lim_{n \to \infty} \Pr\{q_n = i, x_n = j\},$$

as well as the cell loss probability L.

Let ϕ denote some event of interest which can be observed at slot boundaries. $\Pr\{\phi\}$ will denote the probability of this event. Let $N_{i,j}(\phi)$ denote the number of occurrences of ϕ when the Markov chain $\{q_k, x_k\}_{k\geq 1}$ is in state $(i, j) \in S$. $\bar{N}_{i,j}(\phi)$ will denote the expected value of this quantity. If $\pi_{i,j}$ denotes the stationary probability that the Markov chain $\{q_k, x_k\}_{k\geq 1}$ is in state $(i, j) \in S$, then it can be shown [26] that

$$\Pr\{\phi\} = \frac{1}{T_s} \sum_i \sum_j \bar{N}_{i,j}(\phi)\pi_{i,j}. \tag{B.1}$$

Computation of $\hat{\pi}_{i,j}$. Let ϕ represent the event that the system is found in state (i, j) at an arbitrary time instant, and let $N_{i',j'}(i, j)$ denote the number of times the system is in state (i, j) over the current sub-period, given that the sub-period began in state (i', j'). Also denote by $I^n_{i',j'}(i, j)$ an indicator function which takes the value 1 if the system is in state (i, j) after n time slots in the sub-period which began in state (i', j'); the quantity equals 0 otherwise. Clearly, $N_{i',j'}(i, j)$ can be expressed as

$$N_{i',j'}(i, j) = \sum_{n=1}^{T_s} I^n_{i',j'}(i, j).$$

The expected value of $N_{i',j'}(i, j)$ is given by

$$\begin{aligned}\bar{N}_{i',j'}(i, j) &= \sum_{n=1}^{T_s} \mathrm{E}\{I^n_{i',j'}(i, j)\} \\ &= \sum_{n=1}^{T_s} \Pr\{q_n = i,\ x_n = j \mid q_0 = i',\ x_0 = j'\} \\ &= \sum_{n=1}^{T_s} p^{(n)}_j(i', i)\mathbf{1}_{\{j=j'\}}.\end{aligned}$$

Finally, a direct application of Eq. (B.1) yields the result

$$\hat{\pi}_{i,j} = \frac{1}{T_s} \sum_{n=1}^{T_s} \sum_{i'=0}^{B} p^{(n)}_j(i', i)\pi_{i',j} \quad \text{for}(i, j) \in S. \tag{B.2}$$

Computation of L. In this case, ϕ is the event of a cell being lost in an arbitrary time slot. By definition, the loss rate, or cell loss probability L, is given by

$$L = \frac{\Pr\{\phi\}}{\lambda},$$

where λ is the total offered load, or the probability that a cell arrives in an arbitrary time slot. $N_{i'j'}$ will denote the total number of cells lost over the sub-period which began in state (i', j'); $\bar{N}_{i'j'}$ will denote its expected value. Let $I^n_{i'j'}$ denote the number of cells lost over the nth slot of this sub-period and let a_n denote the number of cell arrivals over the nth slot. It is easy to see that

$$I^n_{i'j'} = \begin{cases} k & \text{if } a_n = B + 1 - q_{n-1} + k, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,

$$N_{i'j'} = \sum_{n=1}^{T_s} I^n_{i'j'},$$

and

$$\begin{aligned}\bar{N}_{i'j} &= \sum_{n=1}^{T_s} \mathrm{E}\{I_{i'j}^n\} \\ &= \sum_{n=1}^{T_s} \sum_{k} k \Pr\{a_n = B+1-q_{n-1}+k \mid q_0 = i',\ x_0 = j'\} \\ &= \sum_{n=1}^{T_s} \sum_{i=0}^{B} \sum_{k} k \Pr\{a_n = B+1-i+k \mid q_0 = i',\ x_0 = j'\} \Pr\{q_{n-1} = i \mid q_0 = i',\ x_0 = j'\} \\ &= \sum_{n=1}^{T_s} \sum_{i=0}^{B} \sum_{k=K_1}^{K_2} k p_j^{(n-1)}(i', i) \tilde{f}_j(B+1-i+k),\end{aligned}$$

where K_1 is the minimum number of losses which can occur over the given time slot, and K_2 is the maximum number of such losses. It is easy to see that $K_1 = \max(0, i-B-1)$ and $K_2 = N-B-1+i$.

Finally, applying Eq. (B.1) yields

$$L = \frac{1}{\lambda T_s} \sum_{i'=0}^{B} \sum_{j'=1}^{M} \sum_{n=1}^{T_s} \sum_{i=I_1}^{B} \sum_{k=K_1}^{K_2} k p_j^{(n-1)}(i', i) \tilde{f}_j(B+1-i+k) \pi_{i',j}, \tag{B.3}$$

where $I_1 = \max(0, B+2-N)$.

References

[1] H. Heffes and D. Lucantoni, A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer analysis, *IEEE J. Select. Areas Commun.* (September 1986).

[2] K. Sriram and W. Whitt, Characterizing superposition arrival processes in packet multiplexers for voice and data, *IEEE J. Select Areas Commun.* (September 1986).

[3] J. Daigle, Y. Lee and N. Magalhaes, Discrete time queues with phase dependent arrivals, in: *Proc. IEEE INFOCOM'90*, San Francisco, CA, 1990.

[4] I. Stavrakakis, Analysis of a statistical multiplexer under a general input traffic model in: *Proc. IEEE INFOCOM'90*, San Francisco, CA, 1990.

[5] S.Q. Li, A general solution technique for discrete queueing analysis of multimedia traffic on ATM, *IEEE Trans. Commun.* 39 (7) (1991).

[6] K. Sohraby, On the asymptotic behavior of heterogeneous statistical multiplexer with applications, in: *Proc. IEEE INFOCOM'92*, Florence, Italy, 1992.

[7] B. Maglaris, D. Anastassiou, S. Prodip, G. Karlsson and J. Robbins, Performance models of statistical multiplexing in packet video communications, *IEEE Trans. Commun.* 36 (July 1988).

[8] C. Blondia and O. Casals, Performance analysis of statistical multiplexing of VBR sources, in: *Proc. IEEE INFOCOM'92*, Florence, Italy, 1992.

[9] 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).

[10] P. Pancha and M. El Zarki, MPEG coding for variable bit rate video transmission, *IEEE Commun. Mag.* 32 (5) (1994).

[11] A. Lazar, G. Pacifici and D. Pendarakis, Modeling video sources for real-time scheduling, in: *Proc. IEEE GLOBECOM'93*, Houston, TX, 1993.

[12] M.W. Garrett and M. Vetterli, Congestion control strategies for packet video, in: *Proc. 4th Internat. Workshop on Packet Video*, Kyoto, Japan, 1991.

[13] B. Melamed, D. Raychaudhuri, B. Sengupta and J. Zdepski, Tes-based traffic modeling for performance evaluation of integrated networks, in: *Proc. IEEE INFOCOM'92*, Florence, Italy, 1992.

[14] S.Q. Li and C.L. Hwang, Queue response to input correlation functions: Discrete spectral analysis, in: *Proc. IEEE INFOCOM'92*, Florence, Italy, 1992.

[15] J. Hunter, *Mathematical Techniques of Applied Probability, Discrete Time Models: Basic Theory*, Volume 1 (Academic Press, New York, 1983).
[16] J. Hunter, *Mathematical Techniques of Applied Probability, Discrete Time Models: Techniques and Applications*, Volume 2 (Academic Press, New York, 1983).
[17] V. Ramaswami and G. Latouche, Modeling packet arrivals from asynchronous input lines, in: *Proc. ITC 12*, 1988.
[18] T. Eliazov, V. Ramaswami, W. Willinger and G. Latouche, Performance of an ATM switch: Simulation study, in: *Proc. IEEE INFOCOM'90*, San Francisco, CA, 1990.
[19] A. Eckberg, The single server queue with periodic arrival process and deterministic service time, *IEEE Trans. Commun.* **27** (March 1979).
[20] J. Roberts and J. Virtamo, The superposition of periodic cell arrival streams in an ATM multiplexer, *IEEE Trans. Commun.* **39** (2) (1991).
[21] H. Kroner, Statistical multiplexing of sporadic sources-exact and approximate performance analysis, in: *Teletraffic and Datatraffic in a Period of Change: ITC 13* (Elsevier, Amsterdam, 1991).
[22] I. Stavrakakis, Statistical multiplexing of correlated slow traffic sources, in: *Proc. GLOBECOM'92*, Orlando, FL, 1992.
[23] N. Shroff and M. Schwartz, Video modeling within networks using deterministic smoothing at the source, in: *Proc. IEEE INFOCOM'94*, Toronto, Canada, 1994.
[24] B. DeCleene, P. Panch, M. El Zarki and H. Sorensen, Comparison of priority partition methods for VBR MPEG, in: *Proc. IEEE INFOCOM'94*, Toronto, Canada, 1994.
[25] C. Herrmann, VBR video in ATM without frame-buffering: Influence of a periodic correlation function on QoS parameters, in: *Proc. 2nd Workshop on Performance Modelling and Evaluation of ATM Networks*, Bradford, UK, 1994.
[26] H.C. Tijms, *Stochastic Modeling and Analysis: A Computational Approach* (Wiley, New York, 1986).

Randall Landry received his B.S. degree in Electrical Engineering from the University of Southern Maine in 1990 and his M.S. and Ph.D. degrees in Elecrical Engineering from the University of Vermont in 1992 and 1994.
He is currently employed in the Core Network Technology Branch of the DSP R&D Center at Texas Instruments, Inc., Dallas, Texas. His research interests include queueing theory, performance analysis and design of communication systems, and modeling of broadband integrated services traffic.
Dr. Landry is currently engaged in research efforts aimed at developing and optimizing technologies for the internetworking of data communication protocols such as Ethernet and ATM. His work in this area encompasses protocol, algorithm and hardware development. He is also involved in efforts to develop very high-speed switch architectures and traffic management protocols.

Ioannis Stavrakakis received the Diploma in Electrical Engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, 1983, and the Ph.D. degree in Electrical Engineering from the University of Virginia, 1988.
In 1988, he joined the faculty of Computer Science and Electrical Engineering at the University of Vermont as an assistant and then associate professor. Since 1994, he has been an associate professor of Electrical and Computer Engineering at Northeastern University, Boston. His research interests are in stochastic system modeling, teletraffic analysis and discrete-time queueing theory, with primary focus on the design and performance evaluation of Broadband Integrated Services Digital Networks (B-ISDN).
Dr. Stavrakakis is a senior member of IEEE and a member of the IEEE communications Society, Technical Committee on Computer Communications. He has organized and chaired sessions, and has been a technical committee member, for conference such as GLOBECOM, ICC and INFOCOM.