Multiplexing Generalized Periodic Markovian Sources with an Application to the Study of VBR Video*

Randall Landry[†] Ioannis Stavrakakis[‡]
Department of Electrical Engineering and Computer Science
University of Vermont
Burlington, VT 05405

Abstract

This paper presents a stochastic traffic model which is capable of capturing the time-scale dynamics associated with the potentially large disparity between source and output-link speeds in packet networks such as ATM. The model is applied to the study of Variable Bit Rate (VBR) video sources in which cells are delivered at a sub-frame level referred to as a slice. A tractable analytical 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.

1 Introduction

The desire to accommodate multi-media based applications 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 has been essential in determining the achievable effectiveness of statistical multiplexing.

Since the quality of service (QOS) of many applications is often expressed in terms of very low loss probabilities, performance evaluation of multiplexers in the ATM environment using standard simulation techniques is becoming a less formidable option. Consequently, it becomes increasingly important that studies be based on tractable analytical approaches under accurate traffic modeling. 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 represent time-intervals (in network time slots) at which the network perceives events of interest, such as frame arrivals from a video source or cell arrivals from an active voice source, to occur. The magnitude of these time-scales, which increases with the link capacity C, can significantly affect the impact that a traffic source has on the network.

Consider a Variable Bit Rate (VBR) video coder which pre-buffers and uniformly distributes cells in each frame, [1]-[3]. In general, such a source produces a cell stream in which the number of cells generated in consecutive frames is highly correlated. However, as far as network performance is concerned, the impact of inter-frame autocorrelations may become less significant as the link speed, and consequently the frame length and inter-cell spacing (in slots), increases. In this case, a traffic model such as that in [3], which does not account for inter-frame autocorrelations, may be sufficiently accurate. If, however, the pre-buffering scheme mentioned above is not feasible, due to the induced frame delay, pre-buffering on a shorter time-scale (referred to as a slice) may be necessary. In this case, autocorrelations that are present at this new time-scale may significantly impact the multiplexer's performance. The latter has been observed in [4] where *intra-frame* autocorrelations at the slice level are shown through simulations to have a significant negative effect on the multiplexer, while inter-frame correlations were shown to have a negligible impact.

In this paper, a discrete-time traffic model is proposed which captures 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 are made every $T_s \geq 1$ 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.

---

*Research supported in part by the Advanced Research Project Agency (ARPA) under Grant F4962-93-1-0564.
†email : rlandry@emba.uvm.edu
‡email : stavraka@emba.uvm.edu

2 Source Characterization and Queueing Analysis

In this section, the proposed traffic model is described in terms of a generalized periodic Markovian source and the generated traffic is characterized. The superposition of N such traffic sources is also considered, and a finite-capacity multiplexer fed by N generalized periodic Markovian sources is studied. For the remainder of the paper, 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.

2.1 Description of the Proposed Traffic Model

Let $\{m^k, T_s^k, f_j^k(\cdot) : 1 \leq j \leq m^k\}$ denote a set of parameters associated with the k^{th} traffic source, $1 \leq k \leq N$, where N denotes the total number of traffic sources to be multiplexed (see section 2.2) and the parameters are defined as follows:

- m^k: The number of states in an underlying periodic Markov chain associated with the k^{th} source.
- T_s^k: An integer constant (called sub-period) which represents the amount of time (in slots) spent in each state of the periodic Markov chain. In other words, a transition occurs every T_s^k time slots, and the k^{th} 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 k^{th} source when that source is in state j, $1 \leq j \leq m^k$. λ_j^k 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 $\boldsymbol{Z}^0$, where $\boldsymbol{Z}^0$ denotes the set of non-negative integers.

A traffic source described in terms of the above parameters will be referred to here as a Generalized Periodic Markov (GPM) source. The mean arrival rate for source k is simply $\lambda^k = \frac{1}{m^k}\sum_{j=1}^{m^k} \lambda_j^k$.

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 τ^k (referred to as the initialization time) denote the time instant, defined at sub-period boundaries, at which the k^{th} source was activated (from state 1), assuming that $1 \leq \tau^k \leq m^k$. Let γ_n^k denote the mean arrival rate for source k over sub-period n, $n \geq 1$, (interval $[(n-1)T_s, nT_s]$). Arrivals are assumed to be declared at the end of the time slot over which they occur and the mean arrival rate over the n^{th} sub-period is given by,

$$\gamma_n^k = \lambda_j^k \quad \text{where } j = n \bmod (m^k+1) - \tau^k + 1 \quad . \tag{1}$$

Note that since the amount of time (T_s) that the source spends in each state can be made 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 [5], where the first and second moments of the queue length process are determined for an infinite-capacity queue in discrete time.

The input rate autocorrelation function for source k, $R^k(n)$, where the lag n is indexed at sub-period boundaries, is defined as $R^k(n) = E\{\gamma_m^k \gamma_{m+n}^k\}$. It is easy to show that, for the underlying periodic Markov chain considered here, $R^k(n)$ is a periodic function with period m^k, and is given by

$$R^k(n) = \frac{1}{m^k}\sum_{i=1}^{m^k} \lambda_i^k \lambda_{i+n'}^k \ , \quad n' = n - \lfloor \frac{i+n-1}{m^k} \rfloor m^k. \tag{2}$$

The superposition of N of the underlying GPM chains is also a periodic chain with some period M. Note that since the initialization time τ^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 $\tau^1 = \ldots = \tau^N$.

It is easy to show that the period, M, of the GPM chain associated with the aggregate source is given by, $M = LCM\{m^1, m^2, \ldots, m^N\}$, 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, $1 \leq j \leq M$, is easily evaluated through the N-fold convolution

$$\tilde{f}_j(\cdot) = f_{c^1}^1(\cdot) * f_{c^2}^2(\cdot) * \cdots * f_{c^N}^N(\cdot) \quad . \tag{3}$$

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

$$c^k = \begin{cases} j + \tau^k - 1 & \text{if } 1 \leq j < \tau^k \\ j - \tau^k + 1 & \text{if } \tau^k \leq j < \tau^k + m^k \\ j - m^k - \tau^k + 1 & \text{if } \tau^k + m^k \leq j \leq M \end{cases} \quad . \tag{4}$$

2.2 Multiplexer Performance Analysis

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

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 $\{(i,j) : 0 \le i \le B, 1 \le j \le M\}$, where i (j) denotes the current state of q_k (x_k). The probability that the Markov chain $\{q_k, x_k\}_{k \ge 1}$ 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 $\{x_k\}_{k \ge 1}$, and, in fact, a fraction of $\frac{M-1}{M}$ of the total number of elements are zero.

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

$$\mathbf{P} = \begin{bmatrix} 0 & \mathbf{P}_1 & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{P}_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{P}_{M-1} \\ \mathbf{P}_M & 0 & 0 & \cdots & 0 \end{bmatrix} \quad . \tag{5}$$

It is easy to show 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 $1 \le j \le M$. The one-step transition probabilities $p_j(i,i')$, $0 \le i, i' \le B$, are given below. For $0 \le i \le B$ and $max(0, i-1) \le i' \le B-1$,

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

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. For $0 \le i \le B$ and $i' = B$,

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

where N_j denotes the maximum number of arrivals, as determined by (3), when the source is in state j. With the quantities $p_j(i,i')$ in hand, it is now easy to compute the elements of $\mathbf{P}_j$, which are simply the T_s-step transition probabilities for the finite queue when the aggregate GPM source is in state j.

The following theorem provides for the exact stationary solution of $\{q_k, x_k\}_{k \ge 1}$ by solving M Markov chains with transition matrices of order $(B+1)$. Its proof can be found in [6].

Theorem 1 *The finite irreducible periodic Markov chain, with transition matrix given by (5), has a unique stationary probability vector,* $\boldsymbol{\pi} = (\pi_{0,1} \cdots \pi_{B,M})'$, *given by*

$$\boldsymbol{\pi} = (1/M)(\boldsymbol{\pi}_1' \boldsymbol{\pi}_2' \cdots \boldsymbol{\pi}_M')' \quad , \tag{6}$$

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

The above theorem significantly reduces the numerical complexity involved in calculating the stationary probability vector of the transition matrix in (5), since only matrices of order $(B+1)$ are operated on.

Upon solving for $\boldsymbol{\pi} = (\pi_{0,1} \cdots \pi_{B,M})'$ using Theorem 1, the queue length distribution at sub-period boundaries is given by $p_q(k) = \sum_{j=1}^{M} \pi_{k,j}$, for $k = 0, \ldots, B$. Note that the queue length distribution at arbitrary time instants (slot boundaries) can be easily determined through the use of a renewal/regenerative theory-based approach (see [7] for a derivation similar to that needed here).

3 Application to the Study of VBR Sources

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. In [3] a traffic model for VBR video traffic is proposed based on an 8-bin histogram of the average bit rates over a frame. Cell arrivals over a frame are then considered to be Poisson with a constant intensity selected from the 8-bin histogram. This traffic model implies pre-buffering of the VBR traffic at the frame level as in [1] and [2]. When a number of VBR video streams are multiplexed, a cumulative average bit rate over a frame is derived through the convolution of the individual histograms and a Poisson model for the cell arrivals over a frame is considered with intensity selected from the cumulative histogram. The authors assume that the length of a video frame (in slots), is sufficiently large to drive the multiplexer to a steady-state. Therefore, it is assumed that the stationary queue length distribution $q(k)$, $0 \le k \le B$, can be accurately predicted through the *quasi-static* approximation

$$q(k) = \sum_i q(k \mid i) \cdot p_i \quad , \tag{7}$$

where $q(k \mid i)$ denotes the queue length distribution given that the aggregate source is generating cells according to the arrival process specified by bin i, and p_i denotes the probability of occurrence associated with bin i. For the constant-intensity Poisson process assumed in [3], the quantities $q(k \mid i)$ were easily computed by solving the M/D/1/B problem for each bin i in the aggregate arrival rate histogram.

In [4], 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. 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. Note that 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, which is in good agreement with experimental data [8]. 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 [3] can be used to study the multiplexing of VBR traffic which is pre-buffered at the slice level as described above.

Such a slice-level cell arrival process can be 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 slots in a slice; T_s varies with the link speed, the source rate (in slices/sec) and the slot capacity. Assuming a frame rate of 24 per second, as in [4] and [8], and a slot capacity of 53 bytes for ATM,

$$T_s = \frac{C}{305,280} \quad \text{slots} \quad , \tag{8}$$

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 ($1 \leq j \leq 30$) will be Bernoulli with rate λ_j^k.

The slice-level autocorrelation function for a VBR video source has been shown [4],[8] to possess a periodic structure with a period of 30 slices. The GPM traffic model also has a periodic sub-period-level autocorrelation function $R^k(n)$, which can be used to approximate the slice-level autocorrelation function for the k^{th} video source through an appropriate selection of the arrival rates λ_j^k, $1 \leq j \leq 30$. The following simple distribution of the rates λ_j^k will be used as a first approximation of the autocorrelation function found in [4].

Given the frame rate, which is now the mean arrival rate λ^k for the GPM source, the arrival rates λ_j^k, $1 \leq j \leq 30$ will be given by

$$\lambda_i^k = \lambda_{(31-i)}^k = ae^{-(i-1)b} \qquad i = 1, 2, \ldots, 15 \quad .$$

Upon choosing a value of b to produce the desired autocorrelation function, the constant a is calculated using the expression $\lambda^k = \frac{a}{15} \sum_{i=1}^{15} e^{-(i-1)b}$, derived 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 rate λ^k for the source. Fig. 1 shows the resulting slice-level autocorrelation functions for different values of b.

Finally, a frame-level arrival rate histogram can be generated by quantizing the frame-level arrival rates of an actual video sequence [3]. Upon defining the slice-level autocorrelation function $R^k(n)$ for source k as above, each bin in the frame-level rate histogram will completely specify a unique GPM arrival process denoted by $\{30, T_s, f_j^k(\cdot) : 1 \leq j \leq 30\}$. The aggregate rate histogram which arises when modeling the superposition of N such video sources is determined through the convolution of the individual histograms as in [3], and the corresponding GPM arrival process for each bin in the aggregate histogram is easily determined as described in section 2.1.

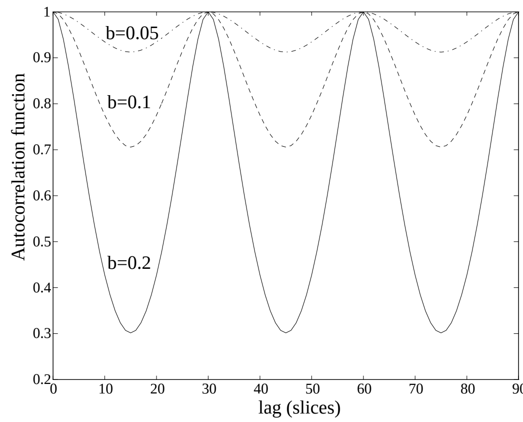

Fig. 1: *Slice-level autocorrelation function for GPM model.*

The quasi-static approximation in (7) can now be applied, and the quantities $q(k \mid i)$ can be computed by using the technique presented in section 2.2 for each i (bin in the aggregate histogram). Since the objective here is to simply motivate the adoption of the GPM source model to study the multiplexing of VBR video at slice level, only results for the conditional distribution $q(k \mid i_o)$, for a given frame rate i_o, will be considered. Note that this scenario provides queueing results for a single bin in the aggregate frame-level arrival rate histogram.

4 Numerical Results and Discussion

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 results presented here are for queue length tail probability distributions at sub-period boundaries.

Figures 2 and 3 represent the queue length distributions for a buffer capacity of $B = 70$, an output link capacity of 100 Mbps, and an input of 6 identical VBR sources each transmitting (on the slice level) at their peak frame rates of 15 Mbps. This scenario can be modeled with GPM sources through the following choice of parameters: $T_s = 327$ (see eq. (8)), $\lambda^1 = \cdots = \lambda^6 = 0.15$ and $m^1 = \cdots = m^6 = 30$.

In Fig. 2, the frame boundaries of all 6 sources are synchronized ($\tau^1 = \cdots = \tau^6$). When $b = 0$, the source behavior remains probabilistically the same from slice to slice and the queue becomes M/D/1/B with binomially distributed batch arrivals. As b increases and the slice-level bit rates become more highly variable, the queue response deteriorates dramatically. This is due to the fact that the high-rate slices are synchronized, and the queue is expected to find itself at or near capacity more frequently.

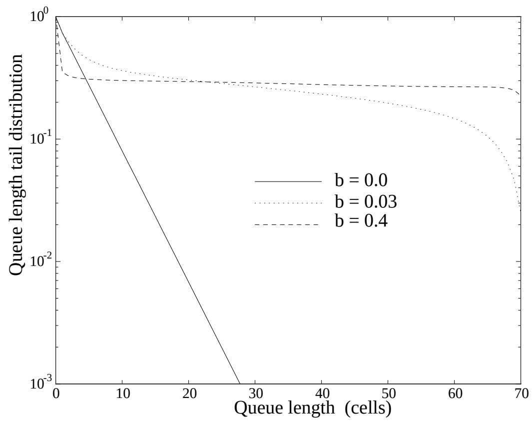

Fig. 2: *Queue length tail distributions for synchronized frame boundaries.*

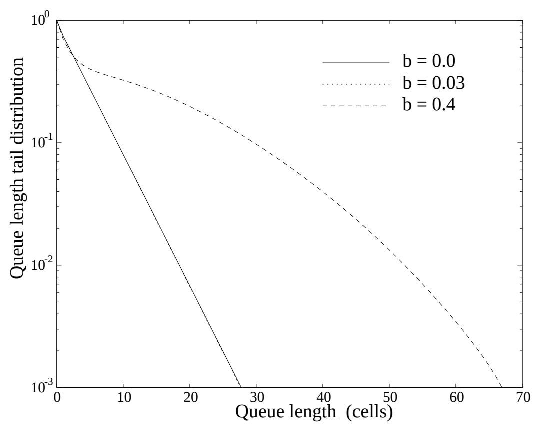

Fig. 3: *Queue length tail distributions for uniformly distributed frame boundaries.*

The impact of frame-boundary synchronization can be minimized by uniformly distributing the frame starts (boundaries) at intervals of $\lfloor \frac{M}{N} \rfloor$, as in [4]. Since any arbitrary distribution of frame starts can be considered through proper setting of the parameters τ^k $(1 \leq k \leq 30)$, the proposed model and analytical technique allow for the consideration of such a case. This is addressed in Fig. 3, where under the same multiplexing scenario as Fig. 2 (except for frame synchronization), the curves for $b = 0$ and $b = 0.03$ now coincide, and even the case of $b = 0.4$ is greatly improved. These results suggest that, if at all possible, the source frame boundaries should be uniformly distributed, or at least distributed so as to avoid complete synchronization.

Fig. 4 illustrates that the degree to which the slice-level autocorrelation function impacts on the queueing performance is heavily dependent upon the offered load. When the load is 0.9 the autocorrelation function has a major impact on the queue response, whereas for a load of 0.3 (each source transmitting at 5 Mbps), an M/D/1/B approximation may describe the queue response sufficiently well.

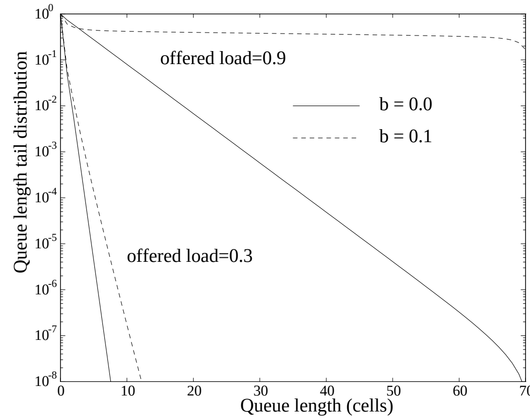

Fig. 4: *The effect of slice-level autocorrelations on queue response for low and high loads.*

Recall that the results presented here are for one possible combination of frame rates (one bin in the aggregate histogram), and that under the quasi-static approximation, all combinations of the individual frame rates would be considered. Extension of the VBR source modeling application and related queueing analysis is currently under investigation.

References

[1] B. Maglaris, D. Anastassiou, S. Prodip, G. Karlsson, J. Robbins, “Performance Models of Statistical Multiplexing in Packet Video Communications”, *IEEE Trans. Commun.*, Vol. 36, July 1988.

[2] C. Blondia, O. Casals, “Performance Analysis of Statistical Multiplexing of VBR Sources”, IEEE INFOCOM’92, Florence, Italy, 1992.

[3] P. Skelly, S. Dixit, M. Schwartz, “A Histogram-Based Model for Video Traffic Behavior in an ATM Network Node with an Application to Congestion Control”, IEEE INFOCOM’92, Florence, Italy, 1992.

[4] A. Lazar, G. Pacifici, D. Pendarakis, “Modeling Video Sources for Real-Time Scheduling”, IEEE GLOBECOM’93, Houston, TX, Dec. 1993.

[5] S.Q. Li, C.L. Hwang, “Queue Response to Input Correlation Functions: Discrete Spectral Analysis”, IEEE INFOCOM’92, Florence, Italy, 1992.

[6] J. Hunter, *Mathematical Techniques of Applied Probability*, vol. 2, *Discrete Time Models: Techniques and Applications*, Academic Press, 1983.

[7] S. Tsakiridou, I. Stavrakakis, “A Flexible Class of Performance-Driven Service Strategies”, Research Report CSEE/93/07-04, Dept. of CS/EE, University of Vermont, July 1993.

[8] M. W. Garrett, M. Vetterli, “Congestion Control Strategies for Packet Video”, Fourth International Workshop on Packet Video, Kyoto, Japan, August 1991.