6

Multiple Time Scales and Subexponentiality in MPEG Video Streams

Predrag R. Jelenković, Aurel A. Lazar and Nemo Semret
Center for Telecommunications Research
Columbia University, New York, NY 10027, USA.
Telephone: (212)-854-2399. Fax: (212) 316-9068.
email: {predrag, aurel, nemo}@ctr.columbia.edu

Abstract

We develop a practical, multiple time scale model for MPEG video traffic whose accuracy and relatively low computational complexity make it well suited for real-time traffic generation experiments on broadband networks. The major feature of our approach is the decomposition of the frame size sequence into simple slow and fast time scale components. This accurately captures aspects of queueing behavior that are difficult to model otherwise. The model also exploits the existence of deterministic patterns that are due to the MPEG coding scheme.

We also present a novel modeling approach based on spatial renewal processes (SRP). This model gives *exact* matches to any desired marginal distribution and any convex non-increasing autocorrelation function. In particular, it can match subexponentially decaying autocorrelations (i.e., can capture long range dependence), something no other model of comparable complexity can do. A SRP is suited for on-line model construction, since it involves no search in parameter spaces, and matches aggregated streams as easily as single streams. The SRP approach yields an analytically tractable queueing behavior, and thus provides a basis for admission control policies that take the dependence structure of video streams into account.

The models are validated by queueing simulations.

Keywords

ATM, VBR, traffic modeling, time scales, subexponentiality, dependence structure, real-time traffic generation, admission control.

1 INTRODUCTION

The objective of this work is to gain understanding of the statistical properties of MPEG video traffic, and in particular to identify and accurately characterize those properties that have an impact on queueing behavior when one or many video sources generate traffic to be transmitted over a network. Understanding these properties plays a crucial part in, for example, admission control policies that ensure efficient utilization of network resources while providing quality of services guarantees.

In section 2, we examine our data set, a set of MPEG-I video traces. We compute the essential statistics, such as marginal distributions and autocorrelation, for full sequences as well as component subsequences.

Then in section 3, we consider the problem of modeling MPEG video on multiple time

L. Mason et al. (eds.), *Broadband Communications*

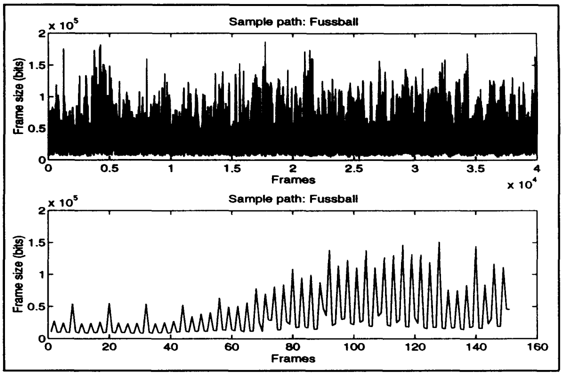

Figure 1 Sample path of Fussball sequence

scales. We develop a model which accurately matches the real data, and is well suited for real-time traffic generation. Our model consists of three deterministically interleaved independent slow time scale processes (each having positive autocorrelation and corresponding to the three frame types in MPEG), plus a single fast time scale correlated "noise" process. Through simulations we demonstrate how this separation helps account for different (temporal as well as distributional) aspects of the queue length behavior.

In section 4, we review the basic concept of subexponentiality, and introduce the Spatial Renewal Process approach to modeling. This new approach allows for matching the statistical characteristics of a broad class of processes exactly (up to second order statistics), including processes with subexponential characteristics, and can be used as a building block for more complex models. It also has the advantage of being analytical tractable, and may be very useful as a basis for implementing admission control policies.

2 DESCRIPTION OF THE MPEG TRACE DATA SET

The data set consists of sequences of MPEG-I frame sizes created at the Institute of Computer Science at the University of Würzburg (Rose (1995)). In all, 17 sequences (sportscasts, movies, music videos, newscasts, talk shows, cartoons and set top) of 40,000 frames each are available.

From the raw video frames, the MPEG coder produces three types of frames at its output: **I** frames, where only spatial redundancies are exploited; **P** frames, where motion compensation with respect to the previous I frame is used to achieve further compression; and **B** frames, where both the previous and the next I or P frames are used to minimize temporal as well as spatial redundancies. The frame types occur in a fixed periodic pattern. In this data set, the period is 12 frames, and the pattern is: IBBPBBPBBPBB. Such a segment is called a group of pictures (GOP).

Figure 1 shows the trace of frame sizes for one of the 17 sequences in our data set, a soccer game. The close-up view of a portion of the sequence clearly shows the deterministic pattern induced by the GOP pattern IBBP...: a large peak, followed by two small values, then one medium peak, etc. It also reveals that the sequence seems modulated by slower variations: a small envelope for the first 60 frames, followed by a larger one.

In Figure 2 the marginal distributions of the different types of frames show the expected differences between the three types of frames. On average, P frames are approximately twice the size of B frames, and one third the size of I frames. The autocorrelation function (ACF) of the whole sequence exhibits a complex pseudo-periodic structure. However,

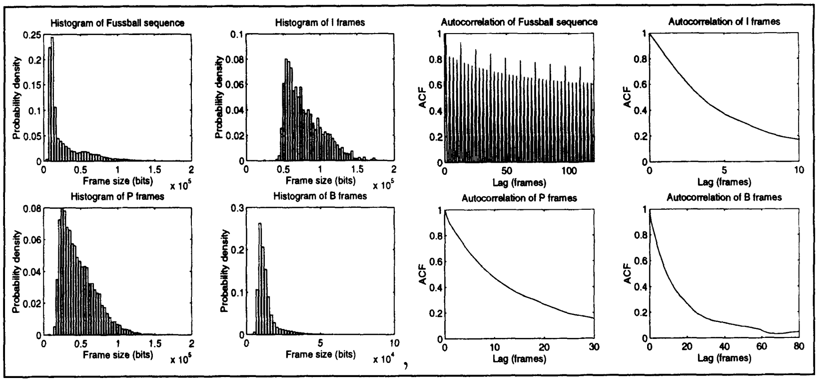

Figure 2 Statistics of Fussball MPEG sequence frame sizes

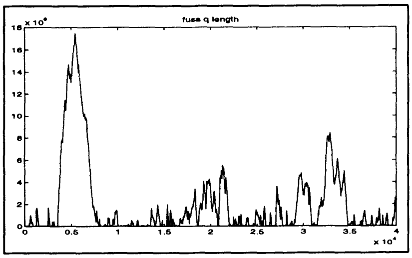

Figure 3 Queue length as a function of time: FCFS queue, utilization = 0.9

focusing on each type of frame separately, we see that the dependence there is simple: a monotonically decaying autocorrelation. Moreover, as can be seen from Figure 2, where the ACFs are plotted on the same time scale (indeed in 120 frames, there are 10 I frames, 30 P frames and 80 B frames), the time range of dependence is similar for the three types of frames. Thus we see that the complex and very heavy dependence seen in the ACF of the whole sequence (autocorrelation of 0.6 for a lag of 120 frames, i.e., 5 seconds) is to a large extent deterministic. It is due to the GOP pattern of the MPEG coder which forces every 12th frame to be of roughly the same size when compared to the overall average. The stochastic dependence that is inherent in the source (characterized by the autocorrelation function of the I, P and B frames) is much simpler: monotonically decaying in time.

Figure 3 shows the queue length (buffer size) as a function of time when our MPEG stream is fed to a FCFS single server queue with constant service rate. The utilization is 0.9. Figure 3 will be used as a basis of comparison for validating our models.

3 MULTIPLE TIME SCALE MODELING OF MPEG VIDEO

The two important features of our modeling approach are the separation of multiple time scales, and the recognition of the deterministic MPEG frame pattern.

In the literature, the existence of multiple time scale statistics in multimedia traffic has been observed from different perspectives. Li and Hwang (1993) argue, from the frequency domain point of view, that the low frequency band of the autocorrelation's Fourier transform (long term correlation) has the most significant impact on queueing. The notion of long range dependence, which has been the subject of much recent work

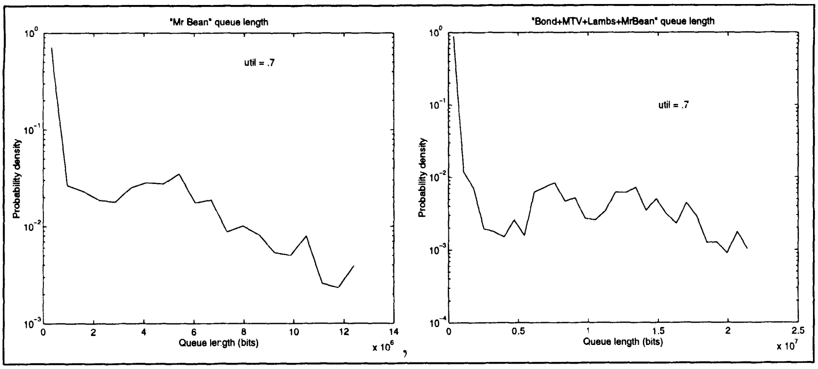

Figure 4 Queue length distribution of MPEG streams

in video traffic (see, e.g., Beran et al. (1995)), is equivalent to the existence of large time scale components. Lazar et al. (1994) develop video models for the slice and frame time scales, and show that the time scales relevant to scheduling depend on the quality of service requirements. Frost and Melamed (1994) survey a wide range of approaches to traffic modeling, several of which take multiple time scales into account, for example, self-similar or fractal models. These essentially attempt to capture an infinite number of time scales, and for that reason, they generally suffer from high run-time complexity, which makes them impractical for real-time generation. Fractal models are also relatively difficult to treat analytically in terms of queueing behavior. More recently, Landry and Stavrakakis (1995) have presented a modeling approach based on multiple time scales that is appropriate for cell and slice levels of video traffic. The types of dependency investigated (i.e., the shape of the autocorrelation), however, is restricted by the constraining *periodic* Markov chain underlying their model. Jelenkovic and Lazar (1995a) have analytically shown that when a stream with multiple time scales passes through a queue, the queue length distribution has multiple decay rates. In the current work, we find that precisely this behavior is found in the actual MPEG streams (see Figure 4). A model without multiple time scales can only capture one of the decay rates of the queue length distribution (see Figure 6 and the attendant explanation).

Other approaches that explicitly model the deterministic structure of MPEG video have been proposed recently, but they either do not capture any second order or time dependence properties at all (see Krunz et al. (1995)), or if they do, do not accurately model the cross-correlation between different frame types (see Ismail et al. (1995)).

3.1 Constructing the model

If we decompose the video stream into two time scales, the GOP pattern is naturally associated with the slow time scale, via three mutually independent (but correlated) slow time scale processes corresponding to the I, P and B substreams. Indeed, I frames are coded independently of all other frames, and therefore, their size depends only on the "complexity" of the individual images. On the other hand, B frames exploit temporal redundancy as much as possible, and therefore, their size depends more on the "action" in the sequence. Thus, it can be argued that these aspects should be modeled separately, because in general the level of action may be independent of the degree of complexity*.

*Imagine filming a painting hanging on a wall with a fixed camera: the images are complex and difficult to compress, so I frames will be large. But there is no motion at all, so B frames will be very small. At

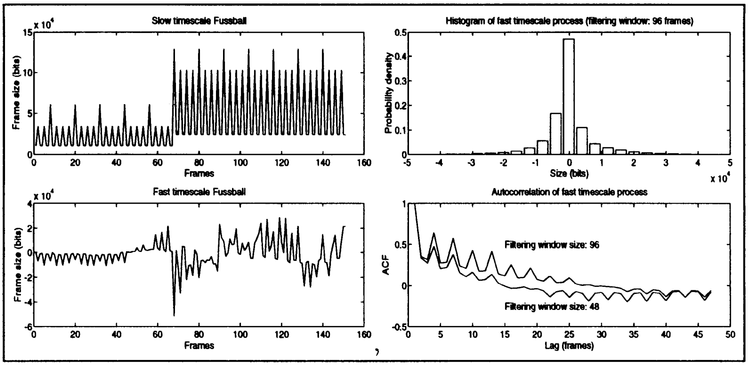

Figure 5 Decomposition into slow and fast time scale processes – averaging window of 96 frames

Also it is natural to adjust parameters that are directly tied to a visible feature of the video stream. On the fast time scale, since we would like to capture the very short term correlation (for lags as small as 1 frame) and the deterministic GOP pattern has been captured in the slow time scale, having three independent I, P and B fast time scale processes is not desirable. Instead, a single correlated fast time scale process is chosen.

We begin the development of our model by analyzing the multiple time scale statistics of an arbitrary video stream. To filter out the fast time scale process, we average the frame size sequence over non-overlapping windows. Note that we average the three types of frames separately. The averaged process gives us the slow time scale component. Figure 5 shows, for a window size of 96 frames (8 GOPs), the resulting decomposition into slow and fast time scale processes, the latter being obtained by subtracting the former from the original sequence.

To model the slow time scale process, we proceed as follows. At the beginning of a window of, say, 96 frames, we generate one frame of each type. The three frames are then repeated following the GOP pattern. The frame size sequences are generated by three independent TES (transform-expand-sample) processes, each generating one sample every 96 frames. A complete description of the TES method can be found in Melamed et al. (1991, 1992), and a detailed description of how we used it in Semret (1995). Briefly, TES methods can match the marginal distribution exactly, and the autocorrelation is approximated by tuning some model parameters. We generate samples via a fast algorithm using integer-only arithmetic operations. For the slow time scale I frames, we created a TES process, with histogram and ACF matching those of the subsequence formed by every 96th frame of the averaged process, starting from the first I frame. The TES processes for B and P frames are created similarly.

The statistics of the fast time scale process are shown in Figure 5. From the autocorrelation, it is apparent that the dependence due to the GOP pattern is still present. The fast time scale shows the variations from the long term average, which is why the histogram is symmetric about zero. Note that I frames are not only bigger than the P and B frames, they also have a larger variance. This explains the residue of the pseudo-periodic structure in the ACF. As shown in the figure, this can be reduced by averaging over windows of smaller size. In fact, how to choose the window size needs further investigation. More

the other extreme, a music video with a rapid fire sequence of simple but unrelated pictures would cause relatively small I frames, but large B frames.

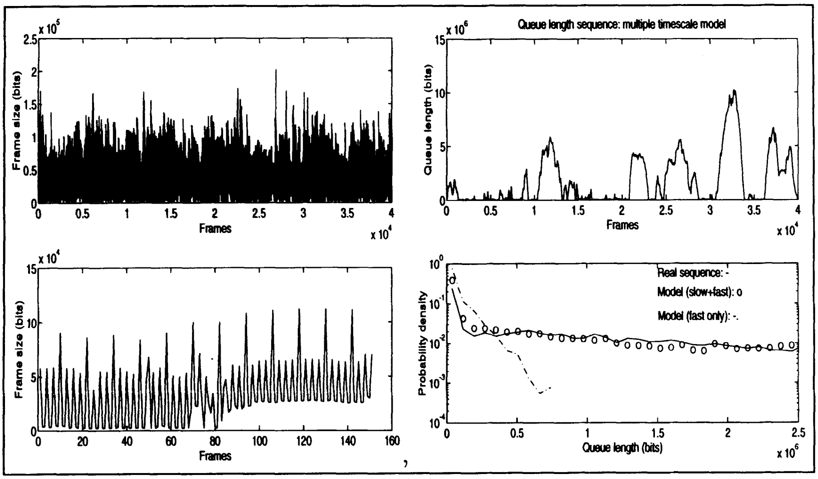

Figure 6 Sample path and queue length: multiple time scale model

analysis would have to be done to determine the optimal separation between "slow" and "fast" time scales.

In any case, compared to the slow process, the fast time scale process has a rapidly decaying ACF, so the small pseudo-periodic bumps may be ignored. Thus we choose to model the fast time scale as a simple TES process, with smoothly decaying ACF. The fast time scale process can be viewed as a small perturbation or noise with short term correlation which is added on to all frames (no distinction is made between I, P and B).

Extending the model to multiple sources: This can be done without increasing the run-time complexity of the model. Due to space limitations the reader is referred to Jelenkovic et al. (1995).

Experiments: Figure 6 (left hand side) shows a sample path generated by our model. The two time scales as well as the GOP pattern are clearly discernible. To validate our model, we conducted some queueing simulations (right hand side of Figure 6). The model **very accurately matches the real sequence in terms of the queue length distribution.** Also, the queue length sample path (top right of Figure 6) is realistic both in height and width of bursts (compare with Figure 3).

3.2 The advantage of separating time scales

In addition to its accuracy, the advantage of our approach is that the explicit separation of time scales helps account for different aspects of queueing behavior. By plotting the queue length distribution for the fast time scale alone†, we get a distribution with the same slope at small queue lengths (bottom right of Figure 6), but not at the tail. We conclude that the good match of the tail of the queue length density is due to the slow time scale component of the model.

Now, focusing on a time period where the queue is small, we see that the real video stream causes small queue buildups (top of Figure 7, around frame numbers 50 and 250),

†By fast time scale alone, we mean the fast time scale TES process plus a constant equal to the overall average of the real sequence. Thus the overall average of the frame sizes and the fast dynamics remain the same, but the slow variations are removed.

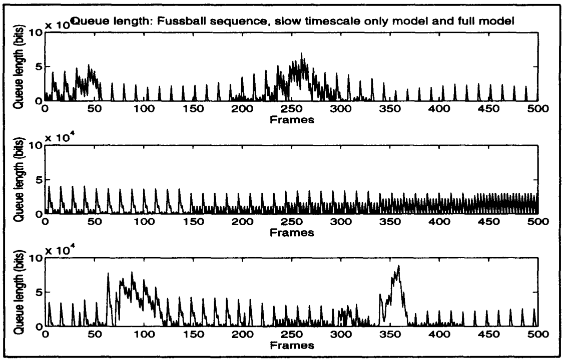

Figure 7 Queue length: Fussball MPEG sequences (top), slow time scale model only (middle), multiple time scale model (bottom). Slow time scale alone is not good enough!

which cannot be reproduced by feeding the model with slow time scale components only through the queue (middle of Figure 7). But the full model does capture this behavior (bottom), thanks to the short term correlation of the fast time scale component.

Thus either the slow or the fast time scale components (or both) may be necessary for capturing the range of queueing behavior one is interested in. That depends on the quality of service requirements. Consider a video source with mean frame size of 26 Kbits, and a rate of 24 frames per second; the utilization is 0.9 (i.e., a server with a capacity of $24 \times 26000/0.9 = 700$Kbits/sec). These were the values used in the simulations shown in Figures 6 and 7. If the maximum acceptable delay is 70ms, then the maximum allowed queue length is about 50 Kbits. In this buffer range the dominant effect on the queue is that of the fast time scale, assuming that the large bursts (which are due to the slow time scale) occur rarely. Thus, as can be seen in Figure 7, the fast time scale model is the relevant one. On the other hand if the maximum delay is ten times larger, then we are only interested in the situations where the queue length approaches 0.5 Mbits, i.e., in the big bursts. Then the fast time scale process becomes negligible, and the slow time scale should be modeled accurately.

Thus, our multiple time scale approach provides a framework for making the correct trade-offs first in what kind of data to gather, and second, what model parameters to carefully tune.

4 SUBEXPONENTIALITY OF THE VIDEO TRAFFIC

The class of subexponential distributions was first introduced by Chistakov (1964). The basic property of this class is that if $X_1, X_2, \ldots, X_n$ are independent, identically and subexponentially distributed random variables,

$$P(X_1 + X_2 + \ldots + X_n > x) \sim nP(X_1 > x), \tag{1}$$

as $x \to \infty$ (see Jelenkovic and Lazar (1995b)). In words, the big peaks tend to be isolated. Some examples of subexponential distribution functions are the Pareto family, the lognormal distribution and the Weibull distribution.

In the case of video traffic, subexponentiality can manifest itself in the marginal distributions or its time-dependent structure. In Jelenkovic and Lazar (1995b), it is shown that when either the marginal or the ACF of the input processes are subexponential, the tail

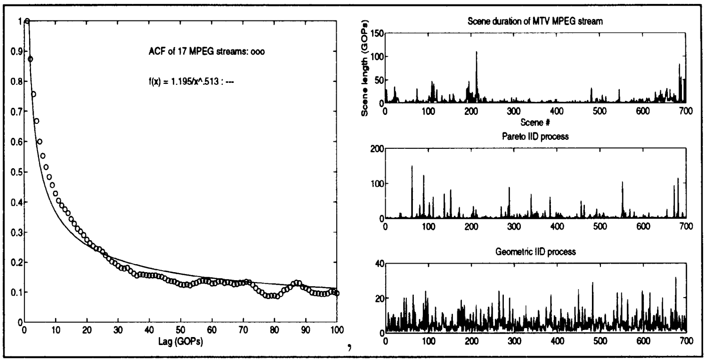

Figure 8 ACF and scene duration of MPEG streams is subexponential (matches Pareto function)

of the queue length distribution is proportional to, respectively, the tail of the marginal distribution, or the ACF. Thus, if real time video traffic exhibits subexponentiality in either form, it is crucial to capture it in modeling. As will be made clear by (4) below, the ACF is directly related to the distribution of the duration of "scenes" or "plateaus" in the video stream.

Figure 8, on the left, shows that the ACF of an MPEG video source (17 streams multiplexed) matches the (subexponential) Pareto function $f(t) = \frac{1.195}{t^{0.513}}$. As mentioned above, another way of showing the same intrinsic subexponential character is by looking at "scene durations". The right hand side of Figure 8 shows a sequence of scene durations‡ for the MTV MPEG stream (top), and for rough comparison, the sample paths generated by i.i.d. processes with Pareto (middle) and geometric distributions (bottom). Clearly, the scene durations have a subexponential character, as does the Pareto process, where the large peaks tend to be isolated in time, as suggested by (1). Note that this is unlike the geometrically distributed process depicted in Figure 8.

In the sequel, we introduce a model called the *spatial renewal process* (SRP) with arbitrary marginal distribution. We show that its autocorrelation function is equal to the integrated tail of the renewal time distribution. This result allows constructing an SRP with a given marginal distribution function and *any* convex decreasing autocorrelation function. (In particular, the model can very simply match subexponential ACFs.) The model is validated by comparing sample paths with MPEG streams, and by comparing the results of simple queueing simulations. We begin by introducing the spatial renewal process.

4.1 Constructing a Spatial Renewal Process

Consider a point process $T = \{T_0 \leq 0, T_n, n \geq 1\}$ such that $T_n - T_{n-1}, n \geq 1$, are i.i.d. with distribution function F, and mean $m = \int_0^\infty uF(u)du = \int_0^\infty [1 - F(u)]\, du$. Further, let $X = \{X_n, n \geq 0\}$, be a sequence of i.i.d. random variables with a distribution $\mathbb{P}[X_n \leq x] = G(x)$. X and T are jointly independent processes.

Now we define the process $A = \{A_t\}$ by:

$$A_t = X_n \quad \text{for} \quad T_n \leq t < T_{n+1}. \tag{2}$$

‡We define a scene change to occur when there is a jump of more than 50 Kbits in the GOP size.

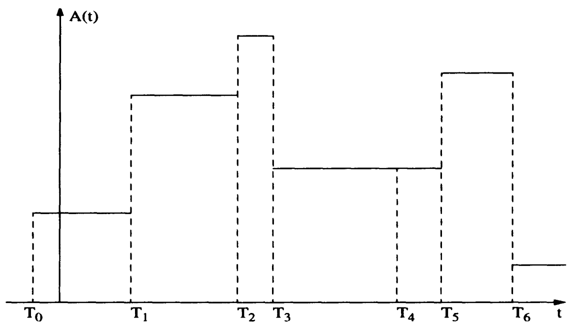

Figure 9 A possible realization of a Spatial Renewal Process.

This is called a Spatial Renewal Process (SRP). A typical sample path of such a process is given in Figure 9. In order to make the process stationary (see Cinlar (1975), section 9.3) we choose the residual time at zero until the first jump to be distributed as an integrated tail of F, i.e.,

$$F_1(t) = \mathbb{P}[\mathrm{T}_1 \le \mathrm{t}] = \mathrm{m}^{-1} \int_0^{\mathrm{t}} [1 - \mathrm{F(u)}]\,\mathrm{du}. \tag{3}$$

Now it is easy to prove (see Jelenkovic et al. (1995)) that the autocorrelation function $R(t) \stackrel{def}{=} (\mathbb{E}\mathrm{A}_0\mathrm{A}_\mathrm{t} - (\mathbb{E}\mathrm{A}_0))^2/\sigma(\mathrm{A}_0)^2$ of the SRP process satisfies the relation

$$R(t) \;=\; 1 - F_1(t). \tag{4}$$

4.2 SRP Video Traffic Model

Using the model of section 4.1, we can generate a process A_t that matches *exactly* a pair (G, R) of an arbitrary marginal distribution and an ACF which is convex, non-increasing and tends to zero for infinite lags. This can be achieved by simply generating two i.i.d. sequences τ_n and X_n, with distributions F and G respectively, corresponding to the widths and heights of the rectangles in Figure 9.

The distribution F is obtained from R. For the purpose of generation, we assume R is given as a sequence of discrete time values $R(0), R(1), \ldots$, and we seek the values $F(1), F(2), \ldots$, which will constitute the histogram of the distribution of τ_n. From (3) and (4), we have,

$$1 - R(t) = m^{-1} \sum_{i=0}^{t-1} [1 - F(i)]\,.$$

Setting $t = 1$ and requiring that $F(0) = 0$, we get

$$m = \frac{1}{1 - R(1)}, \tag{5}$$

and $[1 - R(t+1)] - [1 - R(t)] = m^{-1}\,[1 - F(t)]$ yields

$$F(t) = 1 - m\,[R(t) - R(t+1)]\,. \tag{6}$$

Conditions on R: For F to be a probability distribution, it must be non-decreasing,

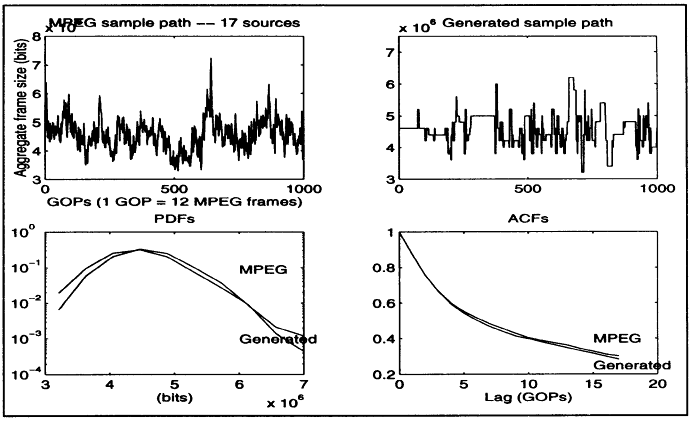

Figure 10 Matching MPEG with SRP model

which implies that R must be **convex**, and $F \leq 1$ implies that R must be **non-increasing**. Finally, consistency calls for $m = \sum_{t=0}^{\infty} [1 - F(t)] = \sum_{t=0}^{\infty} m\,[R(t) - R(t+1)]$. Since $R(0) = 1$, this implies that we must have $\lim_{t \leftarrow \infty} R(t) = 0$.

Modeling aggregate sources: This can be done by simply matching the aggregate (G, R); the aggregate marginal distribution function is given as a convolution of the individual marginal distribution functions, and the aggregate ACF is (due to independence) the weighted average of the individual ACFs (see Jelenkovic et al. (1995)).

Strengths of the SRP modeling approach: The main strengths of the SRP modeling approach are the following. First, it provides an exact match of the first and second order statistics, a property which is not achieved by any other known approach. In addition, a very broad and realistic class of autocorrelation functions can be matched: for example, Markovian models such as the Markov Modulated Poisson Process of Skelly et al. (1992), or the TES-based models, can only match exponentially decreasing ACFs. An SRP model can match *subexponential* ACFs, and as such it is a uniquely efficient and simple model capturing long range dependence. The run-time complexity is exceedingly low; the main operations involved are evaluations of F^{-1} and G^{-1}, which can be essentially done by a sequence of comparisons. The number of comparisons to generate one sample is $O(N_G + N_R)$, where N_G is the number of bins on the input histogram for the marginal G, and N_R the number of points on the graph of the ACF R. Second, matching an SRP model involves no tuning, i.e., no heuristic or algorithmic search in the parameter space. Thus, computationally, the model is very inexpensive to construct. Finally, the process has a structure which is analytically tractable, and thus constitutes a useful traffic characterization for queueing analysis and admission control. This is further explored in section 5.

4.3 SRP Traffic Model: Experiments

In this section, we use the SRP model to generate traffic matching that of 17 MPEG sources multiplexed (summed) into one stream. For clarity, we consider the stream at the GOP level, where the measured ACF exhibits the desired (convex, decreasing) properties. Of course, the SRP can also serve as a building block for composite models, capturing more complex dependence structures, exactly as the TES method was used for multiple time scale models in section 3.

Figure 10 shows portions of sample paths for the real and generated traffic, and how

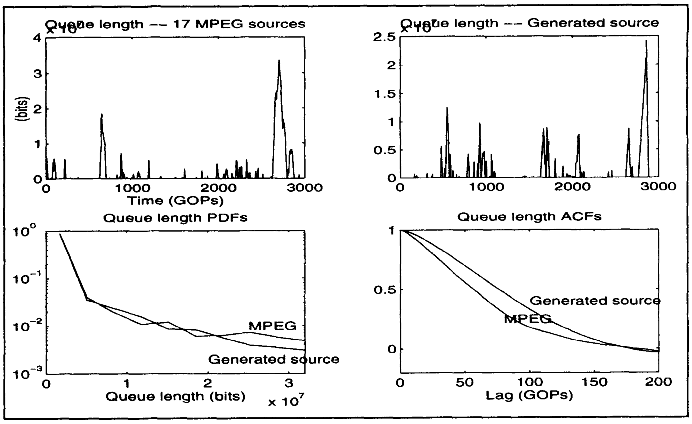

Figure 11 Queueing Simulations with MPEG and SRP model

well the statistics match. More important validation is obtained when both streams (real and generated) are run through queueing simulations, as shown in Figure 11.

The cases shown here are for a utilization of 0.9. As can be seen from the figures, the model also provides an excellent match of queueing behavior: the queue length process in time looks similar, and the queue length distribution is accurately matched.

5 SINGLE NODE ADMISSION CONTROL USING THE SRP CHARACTERIZATION

We assume that the quality of service (QOS) parameters of interest, and hence the admission control decision, can all be directly derived from the queue tail distribution $P(Q > x)^{\S}$. Note that since $P(Q > x)$ is for the aggregate stream, we can only guarantee QOS in the aggregate and not on the individual streams. Whether an individual stream may find its QOS violated depends on several other factors, including the scheduling policy.

Admission control policy: *(i) given a set of input streams, evaluate the aggregate (G, R), (ii) construct the corresponding SRP model, (iii) compute the queue length distribution, and (iv) determine if QOS is violated.*

Thanks to the simple form of the SRP, several theoretical tools can be used to calculate the queue length distribution. The queue length process at the renewal times T_n follows Lindley's recursion for the GI/GI/1 queue. When the queue increment process, $Y_n \stackrel{def}{=} (T_{n+1} - T_n)(X_n - C)$, is bounded we can numerically obtain the queue length distribution by well known z-transform techniques. However, for sources with heavy tailed autocorrelations (i.e., long range dependence) or marginals, Y_n can be large, and these techniques break down. In such cases, specifically when the distribution of Y_n is **subexponential** asymptotic results from Jelenkovic and Lazar (1995b) can be used for approximating the queue length distribution. (Again, due to space limitation, the reader is referred to Jelenkovic et al. 1995.)

§Indeed, since the service rate is constant, cell delay and cell loss probabilities can be directly derived from $P(Q > x)$.

6 CONCLUSION

We have presented two approaches to traffic modeling. The first emphasizes the multiple time scales inherent to video streams, and is geared toward real-time traffic generation. The second, the SRP model, has the advantage of being able to exactly match first and second order statistics, and of being analytically tractable and thus useful for admission control. It should be noted that the SRP model is equally efficient for real-time generation, and can replace TES in the multiple time scale model.

The admission control based on SRP as we have presented it here is complete for the single node case. As such, it is directly applicable to admission control for, e.g., a video on demand (VOD) server. Currently, we are conducting further research into the use of SRP models for admission control, in particular into extensions to the network case.

REFERENCES

J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger (1995). Long-range dependence in variable bit-rate video traffic. *IEEE Trans. Commun.*, 43:1566–1579.

V. P. Chistakov (1964). A theorem on sums of independent positive random variables and its application to branching random processes. *Theor. Probab. Appl.*, 9:640–648.

E. Cinlar (1975). *Introduction to Stochastic Processes.* Prentice-Hall.

V. S. Frost and B. Melamed (1994). Traffic modeling for telecommunication networks. *IEEE Communications Magazine*, pages 70–81, March 1994.

M. R. Ismail, I. Lambdaris, M. Devetsikiotis, and A. R. Kaye (1995). Modeling prioritized MPEG video using tes and a frame spreading strategy for transmission in ATM networks. In *Proc. IEEE Infocomm.*, pages 762–769, April 1995.

P. Jelenkovic, A.A. Lazar, and N. Semret (1995). Multiple time scales and subexponentiality in MPEG video streams. Technical Report CU/CTR/TR 430-95-36, Columbia University. `http://www.ctr.columbia.edu/comet/publications`.

P. R. Jelenković and A. A. Lazar (1995a). On the dependence of the queue tail distribution on multiple time scales of ATM multiplexers. In *Conference on Information Sciences and Systems*, pages 435–440, Baltimore, MD, March 1995. (`http: //www.ctr.columbia.edu/comet/publications`).

P. R. Jelenković and A. A. Lazar (1995b). Subexponential asymptotics of a markov-modulated G/G/1 queue. *Submitted to Journal of Appl. Prob.*.

M. Krunz, R. Sass, and H. Hughes 1995. Statistical characteristics and multiplexing of MPEG streams. In *Proc. IEEE Infocomm.*, pages 455–462, April 1995.

R. Landry and I. Stavrakakis (1995). Multiplexing ATM traffic streams with time-scale-dependent arrival processes. Preprint.

A. A. Lazar, G. Pacifici, and D. E. Pendarakis (1994). Modeling video sources for real-time scheduling. *Multimedia Systems*, 1(6):253–266.

S. Qi Li and C.-L. Hwang (1993). Queue response to input correlation functions: Discrete spectral analysis. *IEEE/ACM Trans. Networking*, 1(5):317–329.

B. Melamed (1991). TES: A class of methods for generating autocorrelated uniform variates. *ORSA Journal on Computing*, 3(4):317–329.

B. Melamed, D. Raychaudhuri, B. Sengupta, and J. Zdepski (1992). TES-based traffic modeling for performance evaluation of integrated networks. In *Proc. IEEE Infocomm.*, May 1992.

O. Rose (1995). Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems. Technical Report 101, Institute of Computer Science, University of Würzburg.

N. Semret (1995). Characterization and modeling of MPEG video traffic on multiple timescales. `http://www.ctr.columbia.edu/~nemo/mmn.ps`.

P. Skelly, S. Dixit, and M. Schwartz (1992). A histogram based model for video traffic behaviour in an ATM network node with an application to congestion control. In *Proc. IEEE Infocomm.*, May 1992.

Biographies of the authors can be found at `http://www.ctr.columbia.edu/~{predrag, aurel, nemo}`.