A Source Model for VBR Video Traffic Based on $M/G/\infty$ Input Processes

Marwan Krunz
Department of ECE
University of Arizona
Tucson, AZ 85721
krunz@ece.arizona.edu
Tel: (520) 621-8731

Armand Makowski
ISR and Department of EE
University of Maryland
College Park, MD 20742
armand@isr.umd.edu
Tel: (301) 405-6844

Abstract

Statistical evidence suggests that the autocorrelation function of a compressed-video sequence is better captured by $\rho(k) = e^{-\beta\sqrt{k}}$ than by $\rho(k) = k^{-\beta} = e^{-\beta \log k}$ (long-range dependence) or $\rho(k) = e^{-\beta k}$ (Markovian). A video model with such a correlation structure is introduced based on the so-called $M/G/\infty$ input processes. Though not Markovian, the model exhibits short-range dependence. Using the queueing performance under 'real' video traffic as a reference, we study via simulations the queueing performance under two video models: the $M/G/\infty$ model and the fractional ARIMA (F-ARIMA) model (which exhibits LRD). Our results indicate that the $M/G/\infty$ model is much more accurate in predicting the actual queueing performance than the F-ARIMA model. Furthermore, only $\mathcal{O}(n)$ computations are required to generate an $M/G/\infty$ trace of length n, compared to $\mathcal{O}(n^2)$ for a F-ARIMA trace.

1 Introduction

Several studies have recently indicated the persistence of the autocorrelations in various types of network traffic, including Ethernet LAN [4, 14], WAN [20], and variable-bite-rate (VBR) video traffic [1, 5, 11]. Such persistence have spurred an ongoing debate on its relevance to the dimensioning of network resources. While researchers generally agree on the importance of traffic correlations, they still disagree on how much of them should be incorporated in a traffic model. Earlier traffic models are Markovian in nature, with an autocorrelation function $\rho(k)$ that drops off exponentially; $\rho(k) \sim e^{-\beta k}$ for large k ($\beta > 0$). These Markovian models exhibit *short-range dependence* (SRD), in that the autocorrelation function (ACF) is summable, i.e., $\sum_k \rho(k) < \infty$. (Note that a SRD model is not necessarily Markovian.) The persistence of traffic correlations at multiple time scales has motivated some researchers to consider instead *long-range dependent* (LRD) models, for which the ACF drops off slowly (typically as $\rho(k) \sim k^{-\beta} = e^{-\beta \log k}$, $0 < \beta < 1$), to the extent that $\sum_k \rho(k) = \infty$. Advocates of LRD modeling argue that the LRD phenomenon has significant impact on network performance, and thus it must be accounted for in dimensioning network resources. On the other side, supporters of Markovian modeling, while acknowledging the presence of such a phenomenon, argue that for networks with *finite* buffers it is sufficient to incorporate correlations up to some finite lag that is proportional to the buffer size [6, 7, 21].

The two asymptotic forms of autocorrelation associated with Markovian and LRD models represent two extremes within which other forms can be envisioned. For example, one can envision a versatile class of stochastic processes in which the ACF has the generic form $\rho(k) \sim e^{-f(k)}$ for some monotone function $f : \mathbb{N} \to \mathbb{R}_+$ which increases no slower than $\log k$ but no faster than k. The challenge in traffic modeling is to identify such a diverse class and use it to display various forms of correlations. One such class, which is considered here, is the class of $M/G/\infty$ processes, which are obtained from the (correlated) busy-server process of a discrete-time $M/G/\infty$. The viability of $M/G/\infty$ processes for modeling network traffic can be attributed to several factors [18]. Firstly, they constitute a versatile class of processes, which can display various forms of time dependencies, the extent of which is governed by the service time distribution G. Secondly, the $M/G/\infty$ model arises naturally in teletraffic as the limiting case for the aggregation of on/off sources [15]. Thirdly, queueing performance for these processes is sometimes feasible as demonstrated in [18, 19]. Finally, the computational complexity for generating synthetic $M/G/\infty$ traces is only $\mathcal{O}(n)$ (n

being the length of the trace), which allows for fast trace generation in network simulations.

In this paper, we investigate the use of $M/G/\infty$ processes in modeling VBR compressed video streams. We start by examining the empirical ACF of four VBR video sequences. Statistical evidence suggests that the empirical ACF is better captured by $\rho(k) \sim e^{-\beta\sqrt{k}}$ than by $\rho(k) \sim k^{-\beta} = e^{-\beta \log k}$ (long-range dependence) or $\rho(k) \sim e^{-\beta k}$ (Markovian), where k is the lag between frames. Accordingly, we introduce an $M/G/\infty$-based video model with an ACF of the form $\rho(k) \sim e^{-\beta\sqrt{k}}$. Though non-Markovian, our model exhibits SRD. To evaluate the appropriateness of our $M/G/\infty$ model, we study its queueing performance via simulations and contrast it with the (LRD) F-ARIMA video model [5]. Our results indicate that the $M/G/\infty$ model is much more accurate than the F-ARIMA model in predicting the queueing behavior of a real video stream, with the additional advantage of requiring fewer computations for generating synthetic traces.

The remaining of the paper is structured as follows. In Section 2 we give an overview of $M/G/\infty$ input processes. In Section 3 we present the fitting results for the ACFs of four video sequences. The $M/G/\infty$-based video model is introduced in Section 4. Issues related to generating synthetic $M/G/\infty$ traces are discussed in Section 5. In Section 6 we present simulations of the queueing performance under the $M/G/\infty$ and the F-ARIMA models. The paper is concluded in Section 7.

2 $M/G/\infty$ Input Processes

In this section, we summarize some of the important properties of $M/G/\infty$ processes as they relate to our modeling effort. Additional details can be found in [17]. Consider a discrete-time $M/G/\infty$ queueing system in which customers arrive in *i.i.d.* Poisson batches of mean λ. Let ξ_{n+1} be the size of the $(n+1)$th batch, i.e., the number of arrivals during time slot $[n, n+1)$. The arrival process can be thought of as a discrete-time version of a Poisson process. Upon arriving at the system, customers are presented to an infinite group of servers. Arrivals during time slot $[n, n+1)$ are serviced at the beginning of slot $[n+1, n+2)$. Let $\sigma_{n+1,1}, \sigma_{n+1,2}, \ldots$ be the integer-valued service times for customers $1, 2, \ldots$ of the $(n+1)$th batch, respectively. It is assumed that service times are *i.i.d.* with a common distribution G. We use σ to indicate a generic rv for the service time of a customer. Initially, there are b_0 customers in the system with corresponding (residual) service times $\sigma_{0,1}, \sigma_{0,2}, \ldots$, which are mutually independent. Let b_n be the number of busy servers (i.e., remaining customers) at the end of slot $[n, n+1)$, $n = 0, 1, \ldots$. The process $\{b_n,\ n = 0, 1, \ldots\}$ is known as the $M/G/\infty$ input process.

For each $n = 0, 1, \ldots$, we introduce the $(n+1)$-tuple $(b_0, b_1, \ldots, b_n)$. The fact that the process $\{b_n, n = 0, 1, \ldots\}$ exhibits some form of positive dependence is indicated by the following result [17]:

Proposition 1 *For any choice of the initial condition rv b_0 and of the service times $\{\sigma_{0,i},\ i = 1, 2, \ldots\}$, the rvs $\{b_n,\ n = 0, 1, \ldots\}$ are associated in the following sense: For any $n = 0, 1, \ldots$ and any pair of non-decreasing mappings $f, g : \mathbb{N}^{n+1} \to \mathbb{R}$,*

$$\mathbf{E}\left[f(b^n)g(b^n)\right] \geq \mathbf{E}\left[f(b^n)\right]\mathbf{E}\left[g(b^n)\right] \tag{1}$$

provided the expectations exist and are finite.

The notion of *association* was introduced in [2], and has been found useful in many contexts when formalizing the idea of positive correlation.

Although the busy-server process is in general *not* strictly stationary, it does admit a stationary and ergodic version [17]. This corresponds to taking (i) the rv b_0 to be Poisson distributed with parameter $\lambda\mathbf{E}[\sigma]$; (ii) the rvs $\{\sigma_{0,j},\ j = 1, 2, \ldots\}$ to be *i.i.d.* rvs distributed according to the *forward recurrence time* $\hat{\sigma}$ associated with σ; this integer-valued rv $\hat{\sigma}$ is distributed according to

$$\mathbf{P}\left[\hat{\sigma} = r\right] \triangleq \frac{\mathbf{P}\left[\sigma \geq r\right]}{\mathbf{E}\left[\sigma\right]}, \quad r = 1, 2, \ldots \tag{2}$$

Several useful properties of the stationary version $\{b_n^\star,\ n = 0, 1, \ldots\}$ are readily obtained [17]:

Proposition 2 *The stationary and ergodic version $\{b_n^\star,\ n = 0, 1, \ldots\}$ of the busy server process has the following properties:*

1. *For each $n = 0, 1, \ldots$, the rv $b_n^\star$ is a Poisson rv with parameter $\lambda\mathbf{E}[\sigma]$;*

2. *Its covariance structure is given by*

$$\Gamma(k) \equiv \mathrm{cov}[b_n^\star, b_{n+k}^\star] = \lambda\mathbf{E}\left[\sigma\right]\mathbf{P}\left[\hat{\sigma} > k\right] \tag{3}$$

for all $n, k = 0, 1, \ldots$, where we use the notation $x^+ = \max(x, 0)$ for any real x.

Henceforth, by an $M/G/\infty$ input process we mean its *stationary* version $\{b_n^\star,\ n = 0, 1, \ldots\}$ which is fully characterized by the pair (λ, G) and which we use here as the basis for traffic modeling.

We note from (3) that the ACF for an $M/G/\infty$ process is given by

$$\rho(k) \triangleq \frac{\Gamma(k)}{\Gamma(0)} = \mathbf{P}\left[\hat{\sigma} > k\right], \quad k = 0, 1, \ldots. \tag{4}$$

By varying G, the process $\{b_n^*,\ n = 0, 1, \ldots\}$ can display various forms of positive autocorrelations, the extent of which is controlled by the tail behavior of G.

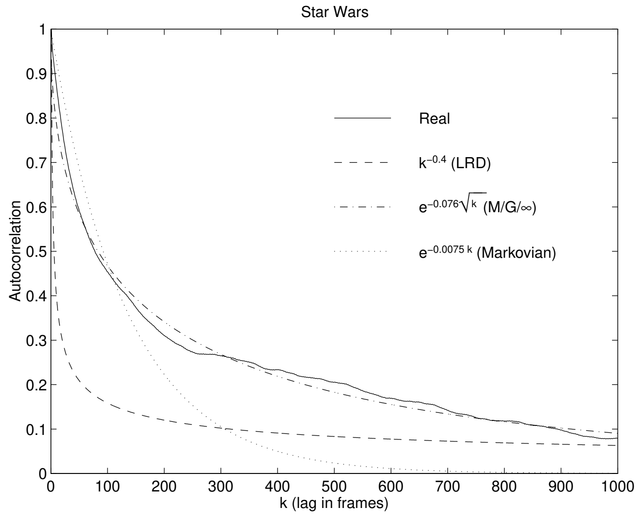
3 Autocorrelations in Video Traffic

Our modeling study is based on four video traces, which were compressed using three different compression schemes (see Table 1). These traces are available in the public domain, and the details of their compression can be found in the cited references. Frame sizes are given in 48-byte ATM cells (rounded up to give an integer number of cells per frame).

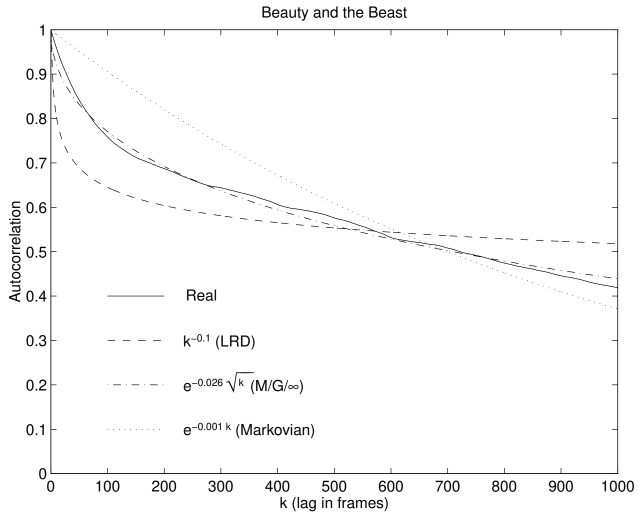
Trace	Compression Scheme
Star Wars [5]	DCT (intra-coding)
Beauty and the Beast [3]	JPEG
Crocodile Dundee [3]	JPEG
Wizard of Oz [13]	MPEG-2 (I sequence)

Table 1: VBR traces used in the study.

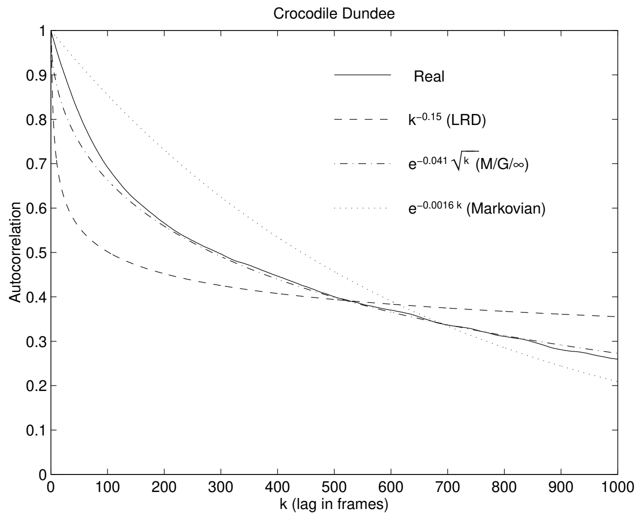
In teletraffic studies, the usefulness of a traffic model lies in its ability to predict network performance for the purpose of dimensioning network resources. Since traffic correlations are known to have a profound impact on queueing behavior, preliminary indications of the goodness of a model can be obtained by examining its correlation structure. The ACFs for the four traces are shown in Figure 1. Each empirical ACF is fitted by three functions: (1) $e^{-\beta k}$ (Markovian), (2) $k^{-\beta}$ (LRD), and (3) $e^{-\beta\sqrt{k}}$. The last fit was chosen because its drop-off behavior is similar to that of the empirical ACF (other forms are also possible). For fits (1) and (3), β is obtained by least-square fitting. For the LRD fit of *Star Wars* trace, $\beta = 0.4$ was obtained from the estimated value of the Hurst parameter ($H = 1 - \beta/2 \approx 0.8$), which was reported in [5]. For the other traces, the Hurst parameter was estimated using several tools, including variance-time plots, R/S analysis, and Whittle's approximation. For brevity, we only display the estimated values for the various parameters in Figure 1. Clearly, the Markovian fit drops off much faster than the real ACF, so it only captures the short-term correlations. The LRD fit is not adequate either since it underestimates the correlations at lags 1 through 1000, and even beyond. Only at very large lags does the LRD fit become acceptable. In contrast, $e^{-\beta\sqrt{k}}$ gives a very good fit at small and large lags, particularly for the first three traces. Using a larger value for H would not improve the LRD fit, since $k^{-\beta}$ always drops off fast and then maintains almost a flat appearance. Hence, it always underestimates the correlations up to some lag, and overestimates them beyond that lag.



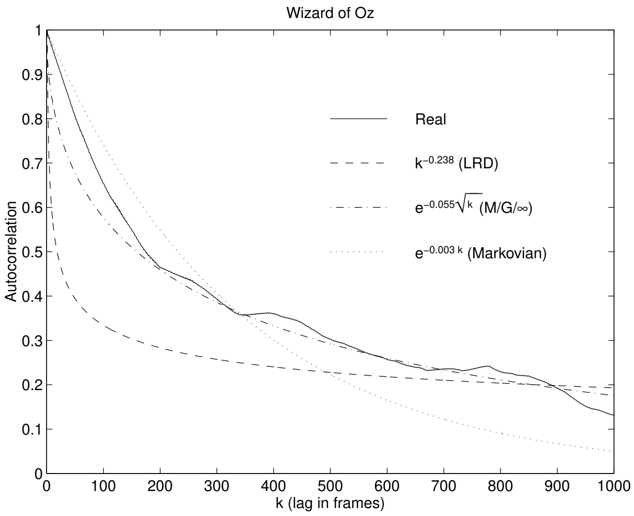
(i)



(ii)



(iii)



(iv)

Figure 1: ACFs of four video sequences along with various fits.

4 $M/G/\infty$-Based Video Model

As indicated in Figure 1, the ACF of a video sequence is adequately captured by

$$\rho(k) = e^{-\beta\sqrt{k}}, \quad k = 0, 1, \ldots \tag{5}$$

A model with such an ACF can be constructed using $M/G/\infty$ input processes. In teletraffic modeling studies, it is common practice to try to capture the first two moments, the autocorrelation structure, and the general shape of the marginal distribution (particularly, the tail of this distribution [5, 6, 16]). Of the parameters G and λ of the $M/G/\infty$ process that can be used in the fitting, G can be chosen to provide a given autocorrelation structure via (4), but λ can only be fitted to one moment (mean or variance). Thus, to capture *both* the complete marginal distribution (including mean and variance) and the correlation structure, we proceed in two steps. First, we choose G in the $M/G/\infty$ model that provides the target ACF (5). Then, we transform the Poisson marginal distribution into a more appropriate distribution.

4.1 Modeling the Correlation Structure

An ACF of the form (5) is generated by an $M/G/\infty$ process whose service distribution G is determined as follows: From (4) and (2) we have

$$\rho(k) - \rho(k+1) = \mathbf{E}[\sigma]^{-1}\,\mathbf{P}[\sigma > k] \tag{6}$$

for $k = 0, 1, \ldots$. Specializing this last relation for $k = 0$ and using the fact $\mathbf{P}[\sigma > 0] = 1$, we obtain

$$\mathbf{E}[\sigma]^{-1} = \rho(0) - \rho(1) = 1 - \rho(1) = 1 - e^{-\beta}. \tag{7}$$

Combining (6) and (7), we conclude that

$$\mathbf{P}[\sigma > k] = \frac{\rho(k) - \rho(k+1)}{1 - \rho(1)}, \quad k = 1, 2, \ldots \tag{8}$$

Hence,

$$\begin{aligned} \mathbf{P}[\sigma = k] &= \mathbf{P}[\sigma > k-1] - \mathbf{P}[\sigma > k] \\ &= \frac{\rho(k-1) - 2\rho(k) + \rho(k+1)}{1 - \rho(1)}. \end{aligned} \tag{9}$$

Substituting (5) into (9) we find

$$\mathbf{P}[\sigma = k] = \frac{e^{-\beta\sqrt{k-1}} - 2e^{-\beta\sqrt{k}} + e^{-\beta\sqrt{k+1}}}{1 - e^{-\beta}}. \tag{10}$$

With a correlation structure of the form (5), the model is SRD since

$$\sum_{k=1}^{\infty} \rho(k) \leq \int_0^{\infty} e^{-\beta\sqrt{t}}dt = \frac{2}{\beta^2} < \infty.$$

4.2 Modeling the Marginal Distribution

The $M/G/\infty$ model produces correlated variates with a Poisson marginal distribution, whose tail drops faster than that of the empirical distribution of a real video sequence. Several fits have been suggested for the frame-size distribution, including Gamma [9], lognormal [8, 12], and hybrid Gamma/Pareto distributions [5]. Of these fits, the last is particularly appropriate for the tail of the frame-size distribution, and is thus used in our modeling study. As explained in [5], the Gamma part in the hybrid Gamma/Pareto fit is used to capture the general shape of the empirical distribution, whereas the Pareto part captures its tail. Let F_Γ and F_P be the cumulative probability functions for Gamma and Pareto distributions, respectively. Although no closed-form expression exists for F_Γ, its density f_Γ has the following form:

$$f_\Gamma(x) = e^{-\omega x}\frac{\omega(\omega x)^{s-1}}{\Gamma(s)}, \quad x \geq 0 \tag{11}$$

for some parameters $\omega > 0$ and $s > 0$, with $\Gamma(.)$ being the Gamma function. The Pareto distribution has the explicit form

$$F_P(x) = 1 - \min\left(1, \frac{x}{a}\right)^{-\alpha}, \quad x \geq 0 \tag{12}$$

with parameters $a > 0$ and $\alpha > 1$ determined by fitting. The hybrid Gamma/Pareto distribution $F_{\Gamma/P}$ is given by

$$F_{\Gamma/P}(x) = \begin{cases} F_\Gamma(x) & \text{if } x \le x^* \\ F_P(x) & \text{if } x > x^*. \end{cases} \tag{13}$$

As in [5], the parameters of the Gamma distribution are obtained by matching the first and second moments of the empirical sequence to those of a Gamma rv. Once the Gamma part is fitted, x^* can be estimated graphically by inspecting the tail of the empirical distribution, and determining where it starts to deviate from the tail of the Gamma fit. Using the continuity condition $F_\Gamma(x^*) = F_P(x^*)$ along with least-square fitting of the Pareto tail, we can obtain estimates of a and α.

The $M/G/\infty$ process $\{X_n,\ n = 0, 1, \ldots\}$ with $\lambda = 1$ is transformed into a new sequence of Gamma/Pareto variates $\{Y_n,\ n = 0, 1, \ldots\}$ using

$$Y_n = F^{-1}_{\Gamma/P}(F_{Poisson}(X_n)), \quad n = 1, 2, \ldots \tag{14}$$

where $F_{Poisson}$ is a Poisson distribution of parameter $\mathbf{E}[\sigma]$, and

$$F^{-1}_{\Gamma/P}(y) = \begin{cases} F_P^{-1}(y) & \text{if } y > 1 - (k/x^*)^\alpha \\ F_\Gamma^{-1}(y) & \text{otherwise} \end{cases}$$

with F_P^{-1} derived from (12) and F_Γ^{-1} obtained numerically. In principle, the nonlinear transformation (14) could affect the original correlation structure. However, in all our experiments the effect of transformation was barely noticeable. This is illustrated in Figure 2, which depicts the *average* ACF of ten transformed $M/G/\infty$ traces along with the analytical ACF of the non-transformed $M/G/\infty$ process.

5 Synthetic Trace Generation

Synthetic realizations of traffic models are often used in trace-driven simulations of the queueing performance. To investigate the queueing performance under our model and contrast it with the (LRD) F-ARIMA model [5], we generated many synthetic realizations from the two models with the same hybrid Gamma/Pareto marginal distribution (in the F-ARIMA model, the marginal distribution is transformed from a normal distribution).

Each $M/G/\infty$ realization consists of 1,000,000 points (i.e., frame sizes), and each F-ARIMA realization consists of 500,000 points. Due to the correlated nature of cell losses, such long traces are needed to obtain meaningful results under small cell loss probabilities. Intuitively, correlations make it more likely

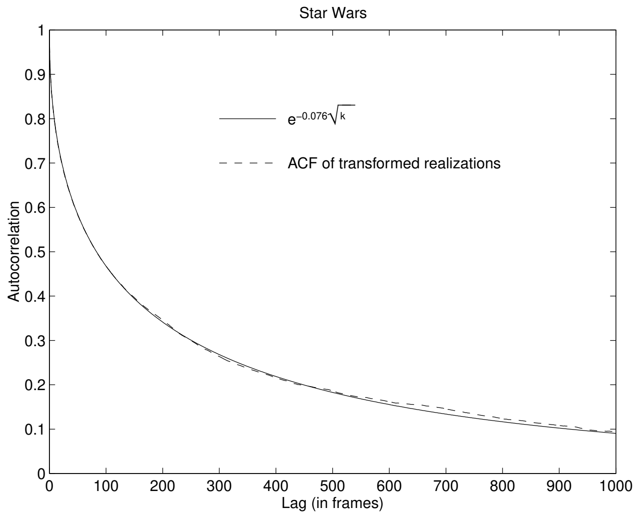


Figure 2: Impact of transformation on the ACF.

for large frames to follow each other, causing extended periods of buffer overflow. Long traces are also needed to display the extreme tail of the frame-size distribution. For example, the maximum frame size in the real *Star Wars* trace is 894 cells. In order to display this value in a transformed $M/G/\infty$ trace, the corresponding value before transformation is 33, i.e., $F^{-1}_{\Gamma/P}(F_{Poisson}(33)) = 894$. The $M/G/\infty$ trace is a realization of n identically distributed Poisson rvs $X_1, \ldots, X_n$ which are associated (Proposition 1). Thus, by well-known properties of associated rvs [2], we have

$$\mathbf{P}\left[\max_{i=1,\ldots,n} X_i > x\right] \le 1 - F_{Poisson}(x)^n, \tag{15}$$

for each x in $\mathbb{R}$. Taking $n = 100,000$ in (15), we get $\mathbf{P}[\max X_i > 32] \le 1 - (F_{Poisson}(32))^{100000} = 0.4745$, i.e., there is more than 50% chance that an 100,000-long realization will never reach the maximum frame-size of the real trace.

The F-ARIMA traces are not as long as the $M/G/\infty$ traces since generating F-ARIMA traces of length 1,000,000 is computationally prohibitive. Hosking's algorithm [10] requires $\mathcal{O}(n^2)$ computations to generate a F-ARIMA trace of length n (before transformation). In contrast, the generation of an $M/G/\infty$ trace requires only $\mathcal{O}(n)$ computations. It takes about two days of execution to generate a 500,000-long F-ARIMA trace using the S-Plus package, (running on a Sparc-10 workstation), compared to less than a minute for a 1,000,000-long $M/G/\infty$ trace (the $M/G/\infty$ traces were generated using a simple C program that simulates an $M/G/\infty$ queueing system).

6 Queueing Performance

To verify the appropriateness of the $M/G/\infty$ video model, we investigate its queueing performance and compare it to the performance of the F-ARIMA model. For brevity, we show the results based for one real trace (the *Star Wars*) and its corresponding $M/G/\infty$ and F-ARIMA models. The queueing system consists of a single-server FIFO queue with capacity B (in cells) and constant service rate C (in cells per time slot). Two types of simulations are conducted: (1) single-stream performance (i.e., no multiplexing), and (2) multiplexing performance. In all experiments, we assume that cells in each frame are evenly distributed over the frame duration (e.g., 1/30 sec). The two performance measures considered here are the cell loss rate (CLR), and the frame error rate (FER). A frame is in error when one or more of its cells are lost. FER is a useful measure for applications that do not implement error concealment mechanisms for recovery from partial frame losses.

6.1 Single Stream

A summary of the simulation results (to two significant digits) is given in Table 2 for three different loads: $U = 80\%$, 60%, and 40%. For the $M/G/\infty$ and F-ARIMA models, the depicted results represent the averages of ten independent runs.

At $U = 80\%$ and under 'real' traffic, both the CLR and the FER are expectedly high. Adding extra buffer barely provides any improvement in performance. In contrast, reducing the load from 80% to 60% (i.e., increasing bandwidth by 33%) improves the CLR by about an order of magnitude. The buffer size seems to have a bigger impact on the FER than on the CLR. For both $U = 80\%$ and $U = 60\%$, the FER for real traffic decreases by about 50% when B is increased from 100 to 2500 cells. *Comparing the performance under the two models to that under the real stream, it is clear that the $M/G/\infty$ model provides much accurate predictions of both CLR and FER.* Under the F-ARIMA model, the performance is much more sensitive to the buffer size. When $B = 2500$ and $U = 80\%$ or 60%, the F-ARIMA model underestimates the CLR and FER by several orders of magnitude. The $M/G/\infty$ slightly underestimates the performance by no more than an order of magnitude, and often less than that.

The $M/G/\infty$ model is quite accurate at extremely small CLRs, as is the case when $U = 40\%$. In this regime, errors are mainly due to very large frames. The F-ARIMA predictions improves as well at this load. However, as observed at other loads, the performance of the F-ARIMA model is more sensitive to the buffer size than the real traffic and the $M/G/\infty$ model. In fact, when $U = 40\%$ and B is small, the F-ARIMA model overestimates the CLR and FER, and eventually underestimates them as B increases to 1000 cells.

6.2 Multiplexed Streams

While many video models do not lend themselves to queueing analysis, they still provide the means to generate independent, homogeneous synthetic streams, which are ideal for statistical multiplexing studies. As a proof-of-concept study, we evaluate the multiplexing performance under both the $M/G/\infty$ and F-ARIMA models. For simplicity, we assume that frames boundaries of the multiplexed streams are aligned, so that the time axis is slotted in frame periods. This restriction allows us to significantly reduce the simulation time. To multiplex K streams, we first obtain an *aggregate* trace $\{\overline{X}_j : j = 1, \ldots, n\}$ from the pointwise sum of the K traces. Thus, if $X_j^{(k)}$ indicates the size of the jth frame in the kth trace, then $\overline{X}_j = \sum_{k=1}^{K} X_j^{(k)}$. For a frame period (slot) in which buffer overflow cannot occur, the aggregate trace can be used to update the buffer occupancy at the end of that period. Otherwise, the individual traces are used to simulate the performance on a cell-by-cell basis.

Fortunately, buffer overflow occurs in a small fraction of frame periods. Let Q_k denote the queue length at the beginning of the kth slot. It can be shown that either of the following two conditions guarantees no buffer overflow during the kth slot:

1. $(\overline{X}_k \leq C) \bigcap (Q_k \leq B - K)$.
2. $(\overline{X}_k > C) \bigcap (\overline{X}_k - C < B - Q_k) \bigcap (Q_k \leq B - K)$.

In the first condition, the total arrival rate over a time slot is less than the service rate. However, this alone is not sufficient to prevent buffer overflow, which could be caused by simultaneous cell arrivals from several streams. For this reason, we also require that $Q_k \leq B - K$. In the second condition, the total arrival rate is greater than the service rate over a time slot, but the difference is not large enough to overflow the buffer during that slot.

Based on these facts, the number of computations needed to simulate the queueing performance for K multiplexed streams is $\mathcal{O}(n + \alpha n W K)$, where α is the fraction of slots for which neither of the above two conditions is satisfied and W is the average number of cells per frame during buffer overflow. Typically, $\alpha W \ll 1$, making the complexity much less than $\mathcal{O}(Kn)$. More importantly, α is proportional to the cell loss rate. Thus, the simulations are even faster at smaller cell loss rates.

Load	Buffer Size (cells)	Cell Loss Rate			Frame Error Rate		
		Real	$M/G/\infty$	F-ARIMA	Real	$M/G/\infty$	F-ARIMA
80%	100	1.7E–2	1.1E–2	8.5E–3	1.1E–1	6.7E–2	4.8E–2
	500	1.6E–2	9.6E–3	1.6E–3	9.1E–2	5.0E–2	7.3E–3
	1000	1.5E–2	8.7E–3	4.6E–4	7.9E–2	4.3E–2	1.9E–3
	1500	1.4E–2	8.0E–3	1.7E–4	7.0E–2	3.8E–2	6.7E–4
	2000	1.3E–2	7.6E–3	7.9E–5	6.3E–2	3.5E–2	2.9E–4
	2500	1.2E–2	7.2E–3	4.1E–5	5.8E–2	3.2E–2	1.5E–4
60%	100	1.2E–3	6.6E–4	7.6E–4	6.2E–3	2.9E–3	3.4E–3
	500	1.1E–3	4.9E–4	7.1E–5	5.1E–3	1.7E–3	2.4E–4
	1000	1.0E–3	4.0E–4	8.9E–6	4.4E–3	1.3E–3	2.3E–5
	1500	1.0E–3	3.4E–4	1.6E–6	3.9E–3	1.1E–3	4.2E–6
	2000	9.6E–4	3.0E–4	5.2E–7	3.6E–3	9.1E–4	6.0E–7
	2500	9.1E–4	2.7E–4	2.0E–7	3.3E–3	7.9E–4	2.0E–7
40%	10	1.3E–5	1.2E–5	5.3E–5	6.4E–5	7.1E–5	1.8E–4
	100	1.2E–5	9.0E–6	3.1E–5	5.3E–5	3.5E–5	9.9E–5
	200	1.1E–5	7.2E–6	1.9E–5	4.7E–5	2.3E–5	5.4E–5
	400	8.9E–6	5.6E–6	7.4E–6	4.1E–5	1.7E–5	1.9E–5
	600	6.8E–6	4.5E–6	2.9E–6	3.5E–5	1.3E–5	8.8E–6
	800	4.8E–6	3.9E–6	1.0E–6	2.9E–5	9.8E–6	2.8E–6
	1000	2.8E–6	3.4E–6	6.7E–7	2.3E–5	7.9E–6	6.0E–7

Table 2: Average cell loss and frame error rates for a single video stream at three different loads. Ten independent replications are used to obtain the values for $M/G/\infty$ and F-ARIMA models.

The average CLR and FER under both models are given in Table 3 for $K = 5$ at 80% load. The performance under 'real' traffic is obtained by arranging the real *Star Wars* trace as a circular list, and deriving the five streams from the *same* list but with different starting points. To reduce the dependency among the five 'real' streams, the starting points are taken to be maximally distanced from each other. However, for large K this approach is unreliable due to the potential cross-correlation between the 'real' streams. For this reason, we only depict the performance for a small value of $K = 5$. As shown in Table 3, the performance under the $M/G/\infty$ model is very close to the real performance. In contrast, the F-ARIMA model is more sensitive to changes in the buffer size than both the real traffic and the $M/G/\infty$ model.

7 Concluding Remarks

In this paper, we investigated the modeling of VBR video streams using $M/G/\infty$ processes. Statistical evidence from four different video traces suggests that the ACF of a VBR video sequence is adequately captured by $\rho(k) = e^{-\beta\sqrt{k}}$. To display such a correlation structure, an $M/G/\infty$-based video model was introduced. To further capture the frame-size distribution of a real video sequence, the Poisson marginal distribution of the $M/G/\infty$ process was transformed into a hybrid Gamma/Pareto distribution, as suggested earlier in [5]. With the performance under real video traffic taken as a reference, we examined the queueing performance under the $M/G/\infty$ model and contrasted it with that of the F-ARIMA model. The performance was investigated in terms of the cell loss rate and the frame error rate, for single and multiplexed streams. In all cases, it was observed that the $M/G/\infty$ model is much more accurate than the F-ARIMA model in predicting the real queueing performance.

Acknowledgement

The authors would like to thank Mark Garrett and Wu Feng for making their video traces available. The work of the second author was supported partially through NASA Grant NAGW77S and the Army Research Laboratory under Cooperative Agreement No. DAAL01-96-2-0002.

References

[1] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger. Long-range dependence in variable bit-rate video traffic. *IEEE Transaction on Communications*, 43:1566–1579, 1995.

[2] J. Esary, F. Proschan, and D. Walkup. Association of random variables, with applications. *Annals of Mathematical Statistics*, 38:166–1474, 1967.

Buffer Size (cells)	Cell Loss Rate			Frame Error Rate		
	Real	$M/G/\infty$	F-ARIMA	Real	$M/G/\infty$	F-ARIMA
200	2.6E–4	2.3E–4	2.1E–4	6.2E–3	3.7E–3	4.8E–3
400	2.4E–4	2.0E–4	6.9E–5	5.7E–3	2.9E–3	1.5E–3
600	2.3E–4	1.8E–4	2.6E–5	5.3E–3	2.5E–3	5.5E–4
800	2.2E–4	1.7E–4	1.1E–5	5.1E–3	2.2E–3	2.2E–4

Table 3: Average cell loss and frame error rates for five multiplexed streams ($U = 80\%$).

[3] W. Feng. *Video-on-Demand Services: Efficient Transportation and Decompression of Variable Bit Rate Video.* PhD thesis, University of Michigan, Apr. 1996.

[4] H. J. Fowler and W. E. Leland. Local area network traffic characteristics, with implications for broadband network congestion management. *IEEE Journal on Selected Areas in Communications*, 9:1139–1149, 1991.

[5] M. W. Garrett and W. Willinger. Analysis, modeling, and generation of self-similar VBR video traffic. In *Proc. of the SIGCOMM'94 Conference*, pages 269–280, Sept. 1994.

[6] M. Grossglauser and J.-C. Bolot. On the relevance of long-range dependence in network traffic. In *Proceedings of the ACM SIGCOMM '96 Conference*, 1996.

[7] D. Heyman and T. Lakshman. What are the implications of long-range dependence for VBR video traffic engineering? *IEEE/ACM Transactions on Networking*, 4:301–317, June 1996.

[8] D. Heyman, E. Tabatabai, and T. Lakshman. Statistical analysis of MPEG2-coded VBR video traffic. In *Proc. of the Sixth International Workshop on Packet Video*, 1994.

[9] D. P. Heyman, A. Tabatabai, and T. V. Lakshman. Statistical analysis and simulation study of video teleconferencing traffic in ATM networks. *IEEE Trans. on Circuits and Systems for Video Technology*, 2(1):49–59, Mar. 1992.

[10] J. R. M. Hosking. Modeling persistence in hydrological time series using fractional differencing. *Water Resources Res.*, 20(12):1898–1908, 1984.

[11] C. Huang, M. Devetsikiotis, I. Lambadaris, and A. Kaye. Modeling and simulation of self-similar variable bit rate compressed video: A unified approach. In *Proceedings of the SIGCOMM '95 Conference*, pages 114–125, 1995.

[12] M. Krunz, R. Sass, and H. Hughes. Statistical characteristics and multiplexing of MPEG streams. In *Proc. of the IEEE INFOCOM '95 Conference*, pages 455–462, Boston, Apr. 1995.

[13] M. Krunz and S. K. Tripathi. On the characterization of of VBR MPEG streams. In *Performance Evaluation Review (Proceedings of SIGMETRICS '97 Conference)*, volume 25, pages 192–202, June 1997.

[14] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, Feb. 1994.

[15] N. Likhanov, B. Tsybakov, and N. Georganas. Analysis of an ATM buffer with self-similar fractal input traffic. In *Proc. of the IEEE INFOCOM '95 Conference*, pages 985–992, Apr. 1995.

[16] B. Melamed and D. Pendarakis. A TES-based model for compressed Star Wars video. In *Proceedings of IEEE GLOBCOM '94*, pages 120–126, 1994.

[17] M. Parulekar. *Buffer Engineering for Self-Similar Traffic.* PhD thesis, University of Maryland, College Park, 1997.

[18] M. Parulekar and A. Makowski. M/G/∞ input processes: A versatile class of models for network traffic. In *Proceedings of IEEE INFOCOM '97*, Apr. 1997.

[19] M. Parulekar and A. M. Makowski. Tail probabilities for a multiplexer with self-similar traffic. In *Proceedings of the IEEE INFOCOM '96 Conference*, pages 1452–1459, 1996.

[20] V. Paxson and S. Floyd. Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3:226–244, 1993.

[21] B. Ryu and A. Elwalid. The importance of long-range dependence of VBR video traffic in ATM traffic engineering: Myths and realities. In *Proceedings of the ACM SIGCOMM '96 Conference*, pages 3–14, Aug. 1996.