Modeling Video Traffic Using $M|G|\infty$ Input Processes: A Compromise Between Markovian and LRD Models

Marwan M. Krunz*
Department of ECE
University of Arizona
Tucson, AZ 85721
krunz@ece.arizona.edu
Tel: (520) 621-8731
Fax: (520) 621-3862

Armand M. Makowski
Department of EE
University of Maryland
College Park, MD 20742
armand@isr.umd.edu
Tel: (301) 405-6844
Fax: (301) 314-9281

Abstract

Statistical evidence suggests that the autocorrelation function $\rho(k)$ ($k = 0, 1, \ldots$) 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. In essence, the $M|G|\infty$ process is a stationary version of the busy-server process of a discrete-time $M|G|\infty$ queue. By varying G, many forms of time dependence can be displayed, which makes the class of $M|G|\infty$ input models a good candidate for modeling many types of correlated traffic in computer networks. For video traffic, we derive the appropriate G that gives the desired correlation function $\rho(k) = e^{-\beta\sqrt{k}}$. Though not Markovian, this model is shown to exhibit short-range dependence. Poisson variates of the $M|G|\infty$ model are appropriately transformed to capture the marginal distribution of a video sequence. Using the performance of a real video stream as a reference, we study via simulations the queueing performance under three video models: our $M|G|\infty$ model, the fractional ARIMA model [9] (which exhibits LRD), and the DAR(1) model (which exhibits a Markovian structure). Our results indicate that only the $M|G|\infty$ model is capable of *consistently* providing acceptable predictions of the actual queueing performance. 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.

keywords: Traffic modeling, VBR video, $M|G|\infty$ process, fitting, correlated variates.

[*] Correspondence Author.

1 Introduction

Recent indications of persistent correlations in various types of network traffic (including Ethernet LAN [7, 19], WAN [28], and variable-bite-rate (VBR) video traffic [2, 9]) have spurred an ongoing debate on the relevance of these correlations to the dimensioning of network resources. While there is general agreement on the importance of traffic correlations, researchers tend to disagree on how much of them should be incorporated in a traffic model. Conventional traffic models are Markovian in nature, with an autocorrelation function (ACF) that drops off exponentially. They include many familiar models such as autoregressive models, Markov Arrival processes (MAP), and Markov modulated processes (cf. [1, 8, 23] for surveys). Markovian models exhibit *short-range dependence* (SRD), in that the ACF $\rho(k)$ $(k = 1, 2, \ldots)$ is summable, i.e., $\sum_k \rho(k) < \infty$, implying a rapid decay of the ACF for large lags k. Note, however, that a SRD model is not necessarily Markovian. The persistence of traffic correlations and their presence at multiple time scales have prompted some researchers to consider instead *long-range dependent* (LRD) models. The ACF in LRD models drops off slowly (typically as a power function) to the extent that the correlations now have an infinite sum; $\sum_k \rho(k) = \infty$. The LRD phenomenon has long been observed in other domains such as hydraulics and economics (see [2] and the references therein). In teletraffic studies, advocates of LRD argue that such a phenomenon has significant impact on network performance, and thus must be taken into account when dimensioning network resources. On the other hand, 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 [12, 10, 29].

As indicated above, the key difference between these two modeling approaches lies in asymptotic behavior of the ACF: Markovian models give rise to an ACF of the form $\rho(k) \sim e^{-\beta k}$ $(\beta > 0)$, whereas in LRD models we find $\rho(k) \sim k^{-\beta} = e^{-\beta \log k}$ $(\beta > 0)$, which drops off much slower than its Markovian counterpart. These ACFs represent two extremes, between which other forms can be envisioned, at least in principle. More generally, the ACF can have the general representation $\rho(k) \sim e^{-f(k)}$, for some monotone function $f : \mathbb{N} \to \mathbb{R}_+$ which increases no slower than $\log k$ but no faster than k.

The challenge for the traffic modeler is to identify a class of stochastic processes that can display forms of correlations as diverse as possible. One such class, which is considered here, is the class of $M|G|\infty$ input processes, which are obtained from the (correlated) busy-server process of a discrete-time $M|G|\infty$ queue. The viability of $M|G|\infty$ processes for modeling network traffic can be attributed to several factors [25]. 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; in fact, the $M|G|\infty$ process was first mentioned by Cox [3] as an example of a process exhibiting LRD (which occurs when G is a Pareto distribution). Secondly, the $M|G|\infty$ model arises naturally in teletraffic as the limiting case for the aggregation of on/off sources [20]. Thirdly, queueing performance for these processes is sometimes feasible, as demonstrated in [4, 27, 26, 21]. Finally, when their queueing analysis in not tractable (as in the case of the video model presented in this paper), the computational complexity for generating synthetic $M|G|\infty$ traces is only $\mathcal{O}(n)$, with n being the trace length. This low complexity allows for fast generation of these traces to be used in network simulations.

In this paper, we investigate the use of $M|G|\infty$ processes in modeling VBR compressed video streams. We start by re-examining the empirical ACF of four VBR video sequences, which were generated by JPEG and MPEG-2 encoders. Statistical evidence suggests that the empirical ACF is better captured by $\rho(k) \sim e^{-\beta\sqrt{k}}$ than by $\rho(k) \sim k^{-\beta} = e^{-\beta \log k}$ (LRD) 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 $\rho(k) \sim e^{-\beta\sqrt{k}}$. We determine the appropriate G that provides such an ACF. Though non-Markovian, this model is SRD. The variates in the basic $M|G|\infty$ process are Poisson distributed. To capture the frame-size distribution of a real video sequence, the Poisson marginal distribution is transformed into a hybrid Gamma/Pareto distribution, in line with the findings in [9]. This nonlinear transformation is shown to have negligible impact on the original correlation structure.

As a means of validating the appropriateness of our $M|G|\infty$ model, we study its queueing performance via simulations and contrast it to two previously proposed video models: the F-ARIMA model [9] (a LRD model) and the discrete autoregressive of order one model (DAR(1)) [14] (a Markovian model). Using the queueing performance for the real video streams as a reference point, we evaluate the performance for the three models with respect to two measures: the cell loss rate due to buffer overflow and the frame error rate. The main conclusions drawn from our study are that (i) the $M|G|\infty$ model provides acceptable performance predictions over a wide spectrum of traffic loads; (ii) the performance of the F-ARIMA model is overly sensitive to the size of the buffer, which makes it in certain cases underestimate the actual performance by several orders of magnitude; and (iii) the DAR(1) model provides very good performance predictions at heavy loads, but performs poorly at light loads. The adequacy of the $M|G|\infty$ video model is justified by the fact that it attempts to capture both short-term and long-term correlations, hence combining the goodness of Markovian models at small lags with that of LRD models at large lags. It is a compromise that incorporates the benefits of the two competing paradigms.

The rest 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 three video models. Section 7 concludes the paper.

2 $M|G|\infty$ Input Processes

In this section, we formally introduce the class of $M|G|\infty$ processes, and summarize some of their properties as they relate to our modeling efforts; additional information can be found in [24, 26].

2.1 Stationary $M|G|\infty$ Input Processes

Consider a discrete-time system with an infinite number of servers. During time slot $[n, n+1)$ ($n = 0, 1, \ldots$), ξ_{n+1} new customers arrive into the system. Customer j, $j = 1, \ldots, \xi_{n+1}$, is presented to its own server, which begins its service by the start of slot $[n+1, n+2)$, with a service time $\sigma_{n+1,j}$ (in number of slots). Let b_n denote the number of busy servers, or equivalently, the number of customers present in the system at the beginning of time slot $[n, n+1)$, with b_0 being the initial number of customers present in the system. It is assumed that the $\mathbb{N}$–valued random variables (rvs) b_0, $\{\xi_{n+1},\ n = 0, 1, \ldots\}$, $\{\sigma_{n,j},\ n = 1, 2, \ldots;\ j = 1, 2, \ldots\}$ and $\{\sigma_{0,j},\ j = 1, 2, \ldots\}$ satisfy the following assumptions: (i) they are mutually independent; (ii) $\{\xi_{n+1},\ n = 0, 1, \ldots\}$ are *i.i.d.* Poisson rvs with parameter $\lambda > 0$; (iii) $\{\sigma_{n,j},\ n = 1, \ldots;\ j = 1, 2, \ldots\}$ are *i.i.d.* rvs with common pmf G on $\{1, 2, \ldots\}$. Let σ be a generic $\mathbb{N}$–valued rv distributed according to the pmf G; assume that $\mathbf{E}[\sigma] < \infty$. Then, the $M|G|\infty$ input process is simply the busy-server process $\{b_n,\ n = 0, 1, \ldots\}$.

For $n = 0, 1, \ldots$, let b^n denote the $\mathbb{N}^{n+1}$-valued rv $(b_0, b_1, \ldots, b_n)$. The fact that the $M|G|\infty$ process $\{b_n,\ n = 0, 1, \ldots\}$ exhibits some form of positive dependence is indicated by the following result [27]:

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 $f, g : \mathbb{N}^{n+1} \to \mathbb{R}$, it holds that*

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

provided the expectations exist and are finite.

From (1), we already conclude that

$$\text{cov}[b_n, b_{n+k}] \geq 0, \quad n, k = 0, 1, \ldots \tag{2}$$

The notion of *association* used above was introduced in [5], and has been found useful in many contexts when formalizing the idea of *positive* dependence.

Thus far, no additional assumptions are made on the rvs $\{\sigma_{0,i},\ i = 1, 2, \ldots\}$, which represent the service durations of the b_0 customers initially present in the system. Various scenarios can, in principle, be accommodated: If the initial customers start their service at time $n = 0$, then it is appropriate to assume that the rvs $\{\sigma_{0,j},\ j = 1, 2, \ldots\}$ are also *i.i.d.* rvs with common pmf G. On the other hand, if we take the viewpoint that the system has been in operation for some time, then these rvs $\{\sigma_{0,j},\ j = 1, 2, \ldots\}$ may be interpreted as the residual work (expressed in time slots) that the b_0 "initial" customers require from their respective servers before service is completed. In general, the statistics of the rvs $\{\sigma_{0,j},\ j = 1, 2, \ldots\}$ cannot be specified in any meaningful way, except for the situation when the system is in steady state.

Although the busy server process $\{b_n,\ n = 0, 1, \ldots\}$ is in general *not* a (strictly) stationary process, it does admit a stationary and ergodic version. The existence of this stationary regime emerges very naturally through the following proposition. We use $\Longrightarrow$ to indicate weak convergence.

Proposition 2 *There exists a stationary and ergodic* $\mathbb{N}$*–valued process* $\{b^\star_n,\ n = 0, 1, \ldots\}$ *such that*

$$\{b_{n+k},\ n = 0, 1, \ldots\} \Longrightarrow \{b^\star_n,\ n = 0, 1, \ldots\} \quad \text{as } k \to \infty \tag{3}$$

for any choice of the initial condition rv b_0 *and of the service times* $\{\sigma_{0,i},\ i = 1, 2, \ldots\}$.

This stationary version $\{b^\star_n,\ n = 0, 1, \ldots\}$ admits an explicit construction, which corresponds to taking (i) b_0 to be Poisson distributed with parameter $\lambda \mathbf{E}[\sigma]$; (ii) $\{\sigma_{0,j},\ j = 1, 2, \ldots\}$ to be *i.i.d.* rvs distributed according to the *forward recurrence time* $\hat{\sigma}$ associated with σ. The pmf of $\hat{\sigma}$ is given by

$$\mathbf{P}[\hat{\sigma} = r] \overset{\triangle}{=} \frac{\mathbf{P}[\sigma \geq r]}{\mathbf{E}[\sigma]}, \quad r = 1, 2, \ldots \tag{4}$$

Based on the above construction, several useful properties of the stationary version $\{b^\star_n,\ n = 0, 1, \ldots\}$ are readily obtained [24]:

Proposition 3 *The stationary and ergodic version* $\{b^\star_n,\ n = 0, 1, \ldots\}$ *of the busy-server process has the following properties:*

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

2. *It holds that*

$$\lim_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} b^\star_k = \lambda \mathbf{E}[\sigma] \quad a.s. \tag{5}$$

3. *The covariance structure of* $\{b^\star_n,\ n = 0, 1, \ldots\}$ *is given by*

$$\Gamma(k) \equiv \mathrm{cov}[b^\star_n, b^\star_{n+k}] = \lambda \mathbf{E}\left[(\sigma - k)^+\right], \quad n, k = 0, 1, \ldots \tag{6}$$

Henceforth, by an $M|G|\infty$ input process we mean its *stationary* version $\{b^\star_n,\ n = 0, 1, \ldots\}$, as described above. This stationary process, which is fully characterized by the pair (λ, G), will be used here as the basis for traffic modeling.

2.2 Correlation Properties of $M|G|\infty$ Input Processes

We note from (6) that

$$\Gamma(k) \;=\; \lambda \sum_{i=0}^{\infty} \mathbf{P}\left[(\sigma - k)^+ > i\right]$$

$$
\begin{aligned}
&= \lambda \sum_{i=0}^{\infty} \mathbf{P}\left[\sigma > k+i\right] \\
&= \lambda \sum_{i=k+1}^{\infty} \mathbf{P}\left[\sigma \geq i\right] \\
&= \lambda \mathbf{E}\left[\sigma\right] \sum_{i=k+1}^{\infty} \mathbf{P}\left[\hat{\sigma} = i\right] \\
&= \lambda \mathbf{E}\left[\sigma\right] \mathbf{P}\left[\hat{\sigma} > k\right], \quad k = 0, 1, \ldots \qquad (7)
\end{aligned}
$$

Thus, the ACF for an $M|G|\infty$ process is given by

$$
\rho(k) \overset{\triangle}{=} \frac{\Gamma(k)}{\Gamma(0)} = \mathbf{P}\left[\hat{\sigma} > k\right], \quad k = 0, 1, \ldots \qquad (8)
$$

since $\Gamma(0) = \lambda \mathbf{E}\left[\sigma\right]$ by (6). 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.

To close this section, we point out that the process $\{b_n^*,\ n = 0, 1, \ldots\}$ can induce *both* SRD and LRD behaviors: From (8), it follows readily [27] that

$$
\sum_{k=0}^{\infty} \Gamma(k) = \lambda \mathbf{E}\left[\sigma\right] \mathbf{E}\left[\hat{\sigma}\right] = \frac{\lambda}{2} \mathbf{E}\left[\sigma(\sigma+1)\right], \qquad (9)
$$

whence

$$
\sum_{k=0}^{\infty} \rho(k) = \mathbf{E}\left[\hat{\sigma}\right] = \frac{1}{2} + \frac{\mathbf{E}\left[\sigma^2\right]}{2\mathbf{E}\left[\sigma\right]}. \qquad (10)
$$

Consequently, the process $\{b_n^*,\ n = 0, 1, \ldots\}$ is LRD (resp. SRD) *if and only if* $\mathbf{E}\left[\sigma^2\right]$ is infinite (resp. finite). In particular, the $M|G|\infty$ input traffic will be LRD when G is Pareto, with a shape parameter in the interval $(1, 2)$ [3].

3 Correlation Structure of VBR Video Sources

In our study, we examined four public-domain VBR video traces (Table 1). These traces were generated using three different encoding mechanisms (see references for further details). Each trace represents an integer-valued sequence of number of cells per frame for a given movie.

Movie	Source	Trace Length (frames)	Compression Scheme
Star Wars	M. Garrett [9]	174,000	DCT (intra-coding)
Beauty and the Beast	W. Feng [6]	143,442	JPEG
Crocodile Dundee	W. Feng [6]	168,565	JPEG
Wizard of Oz	M. Krunz [18]	12,600	MPEG-2 (I sequence)

Table 1: Summary of the four VBR traces used in the study.

While a model is expected to capture some statistical properties of the underlying empirical data, its goodness is ultimately determined based on its ability to achieve the goal it was designed for. In teletraffic studies, the goal of a model is to predict accurately the network performance for the purpose of dimensioning network resources. Thus, the queueing performance is the crucial factor that determines the appropriateness of a traffic model. 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 was fitted by three functions: (a) $\rho(k) = e^{-\beta k}$ (Markovian), (b) $\rho(k) = k^{-\beta}$ (LRD), and (c) $\rho(k) = e^{-\beta\sqrt{k}}$. The last fit was chosen because its drop-off behavior is similar to that of the empirical ACF (but other forms are also possible). For fits (a) and (c), 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 [9]. For the other traces, the Hurst parameter was estimated by several methods, including variance-time plots, R/S analysis, and Whittle's approximation (see [2, 31] for discussion of these tests). In the interest of 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, the LRD fit becomes acceptable. In contrast, the choice $\rho(k) = e^{-\beta\sqrt{k}}$ provides a very good fit at both small and large lags, particularly for the first three traces. Note that 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.

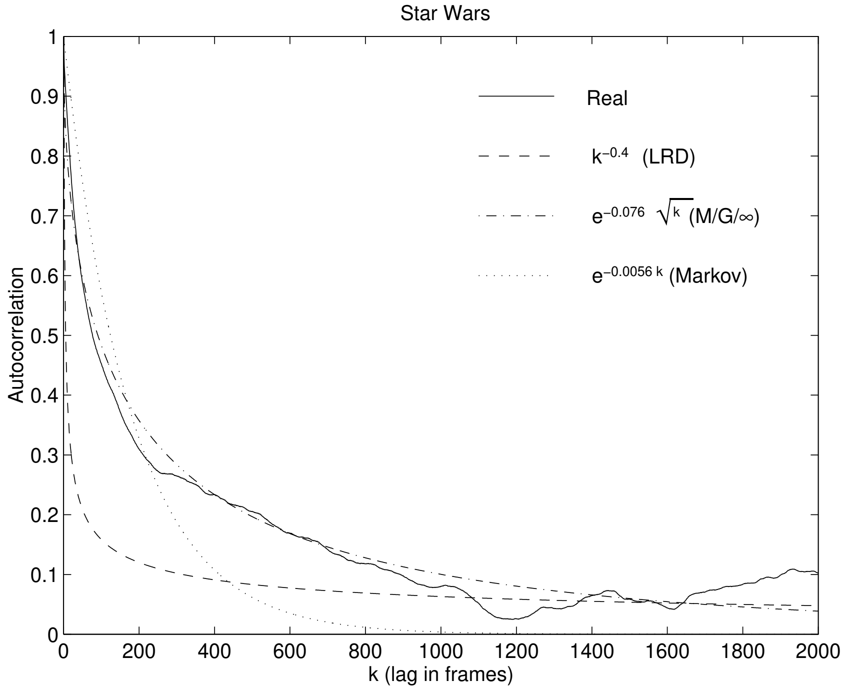
4 $M|G|\infty$-Based Model for Video Traffic

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

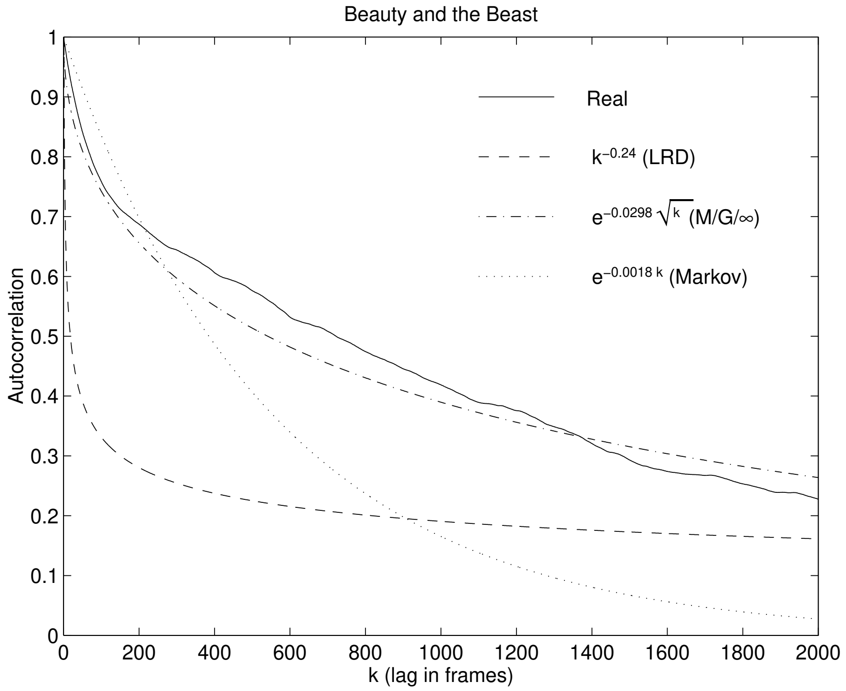
$$\rho(k) = e^{-\beta\sqrt{k}}, \quad k = 0, 1, 2, \ldots \tag{11}$$

for some constant $\beta > 0$. A model with such an ACF can be constructed using $M|G|\infty$ input processes. In teletraffic modeling studies, a common practice is to try to capture the first two moments, the autocorrelation structure, and the general shape of the marginal distribution. More recently, researchers have realized the importance of capturing the tail of the marginal distribution (e.g., [9, 10, 22]), which is especially important for computing the buffer overflow probability at a multiplexer.

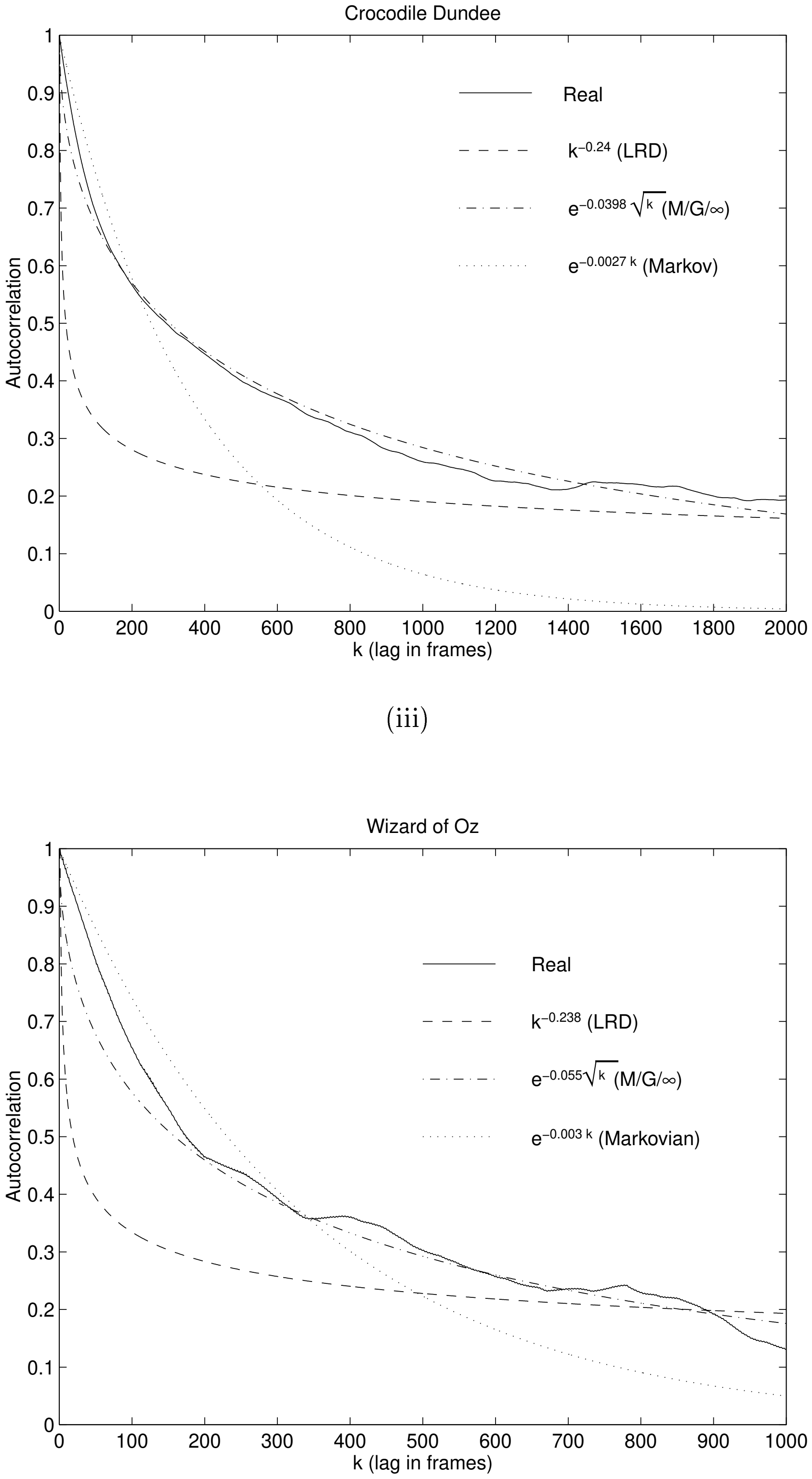
The parameters of the $M|G|\infty$ process that can be used in the fitting are the service distribution G and the arrival rate λ. While G can be chosen to provide a given autocorrelation structure (via (6)), the arrival rate λ can only be fitted to one moment (mean or variance). To capture the complete marginal


Star Wars
Real
$k^{-0.4}$ (LRD)
$e^{-0.076\sqrt{k}}$ (M/G/∞)
$e^{-0.0056k}$ (Markov)
Autocorrelation
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
200
400
600
800
1000
1200
1400
1600
1800
2000
k (lag in frames)

(i)


Beauty and the Beast
Real
$k^{-0.24}$ (LRD)
$e^{-0.0298\sqrt{k}}$ (M/G/∞)
$e^{-0.0018k}$ (Markov)
Autocorrelation
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
200
400
600
800
1000
1200
1400
1600
1800
2000
k (lag in frames)

(ii)



(iii)

(iv)

Figure 1: Fitting the autocorrelation function of VBR video sequences.

distribution (including the mean and variance) as well as the correlation structure, we proceed in two steps. First, we choose G in the $M|G|\infty$ model that provides the target ACF. Then, we identify a pointwise transformation that transforms the Poisson marginal distribution of the original $M|G|\infty$ process into a more appropriate distribution. These steps are described next.

4.1 Modeling the Correlation Structure

We seek the pmf G which is responsible for a correlation sequence of the form (11). To that end, we note from Proposition 3 and (8) that the correlation structure of the stationary $M|G|\infty$ input process (which is parameterized by λ and G) is completely determined by the pmf of $\hat{\sigma}$ (thus of σ). It turns out that the inverse is also true, as we now show.

Indeed, if $\rho(k),\ k = 0, 1, \ldots$, is the ACF of the stationary $M|G|\infty$ input process (λ, σ), then (4) and (8) together imply

$$\begin{aligned} \rho(k) - \rho(k+1) &= \mathbf{P}[\hat{\sigma} > k] - \mathbf{P}[\hat{\sigma} > k+1] \\ &= \mathbf{P}[\hat{\sigma} = k+1] \\ &= \frac{1}{\mathbf{E}[\sigma]}\mathbf{P}[\sigma > k], \quad k = 0, 1, \ldots \end{aligned} \tag{12}$$

so that the mapping $k \to \rho(k)$ is necessarily decreasing and integer–convex. Taking into account the facts $\rho(0) = 1$ and $\mathbf{P}[\sigma > 0] = 1$, we conclude from (12) (with $k = 0$) that

$$\mathbf{E}[\sigma]^{-1} = 1 - \rho(1) \tag{13}$$

with $\rho(1) < 1$ necessarily by the finiteness of $\mathbf{E}[\sigma]$. Combining (12) and (13) we find that

$$\mathbf{P}[\sigma > k] = \frac{\rho(k) - \rho(k+1)}{1 - \rho(1)}, \quad k = 0, 1, \ldots \tag{14}$$

Note also from (14) that

$$\mathbf{E}[\sigma] = \sum_{k=0}^{\infty} \mathbf{P}[\sigma > k] = \frac{1 - \lim_{k\to\infty} \rho(k)}{1 - \rho(1)} \tag{15}$$

and (13) then imposes $\lim_{k\to\infty} \rho(k) = 0$. A moment of reflection readily yields the following invertibility result.

Proposition 4 *An $\mathbb{R}_+$–valued sequence $\{\rho(k),\ k = 0, 1, \ldots\}$ is the autocorrelation function of the stationary $M|G|\infty$ process with integrable σ if and only the corresponding mapping $k \to \rho(k)$ is decreasing and integer–convex with $\rho(0) = 1 > \rho(1)$ and $\lim_{k\to\infty} \rho(k) = 0$, in which case the pmf G of σ is given by (14).*

Differencing (14) yields the pmf of σ:

$$\mathbf{P}[\sigma = k] = \frac{\rho(k-1) - 2\rho(k) + \rho(k+1)}{1-\rho(1)}, \quad k = 1, 2, \ldots \tag{16}$$

The mapping $x \to e^{-\sqrt{x}}$ is decreasing and convex on $\mathbb{R}_+$, so that the sequence $k \to e^{-\beta\sqrt{k}}$ is automatically decreasing and integer-convex on $\mathbb{N}$. Proposition 4 can thus be applied to the correlation sequence (11). Upon substitution into (13) and (16), we find that the desired pmf for σ is simply

$$\mathbf{P}[\sigma = k] = \frac{e^{-\beta\sqrt{k-1}} - 2e^{-\beta\sqrt{k}} + e^{-\beta\sqrt{k+1}}}{1-e^{-\beta}}, \quad k = 1, 2, \ldots \tag{17}$$

and its mean service time is given by

$$\mathbf{E}[\sigma] = (1 - e^{-\beta})^{-1}. \tag{18}$$

The value of β used in (17) and (18) is obtained by fitting the empirical ACF. It might be suggested that in determining the pmf of σ, the empirical ACF be used directly in (16) instead of an analytical fit just performed. However, the empirical ACF is *not* always monotone, and thus there is no *a priori* guarantee that $\mathbf{P}[\sigma \geq k] \geq 0$ in (14) for all $k = 1, 2, \ldots$.

To conclude, we observe by an elementary comparison that

$$\sum_{k=0}^{\infty} \rho(k) = 1 + \sum_{k=1}^{\infty} e^{-\beta\sqrt{k}} \leq 1 + \int_0^{\infty} e^{-\beta\sqrt{t}} dt = 1 + \frac{2}{\beta^2} < \infty, \tag{19}$$

and the correlation structure (11) indeed gives rise to an SRD model.

4.2 Modeling the Marginal Distribution

By Proposition 3, the $M|G|\infty$ model produces correlated variates with a Poisson marginal distribution $F_{Poisson}$, whose tail drops faster than that of the empirical distribution of a real video sequence. This is illustrated in Figure 2 for the *Star Wars* sequence where the parameter of the Poisson distribution (of the $M|G|\infty$ fit) is obtained by matching the sample mean to $\lambda\mathbf{E}[\sigma]$, and setting λ accordingly ($\mathbf{E}[\sigma]$ is estimated from the empirical ACF via (18)). Indeed, the sample mean provides a natural estimate of $\lambda\mathbf{E}[\sigma]$ owing to the ergodic property (5) of $M|G|\infty$ processes. The tail of the marginal distribution plays an important role in determining the buffer overflow probability at a multiplexer [10]. Hence, we need to provide a better fit to the empirical tail than the Poisson fit. To do that, we transform the Poisson distribution of the $M|G|\infty$ process into a more appropriate distribution. The key idea here resides in the following well-known observation: For a frame-size distribution F, a transformation $T : \mathbb{R} \to \mathbb{R}$ can always be constructed so that if the $\mathbb{R}$-valued rv X is distributed according to some distribution H, then the $\mathbb{R}$-valued rv $Y = T(X)$ is distributed according to F.

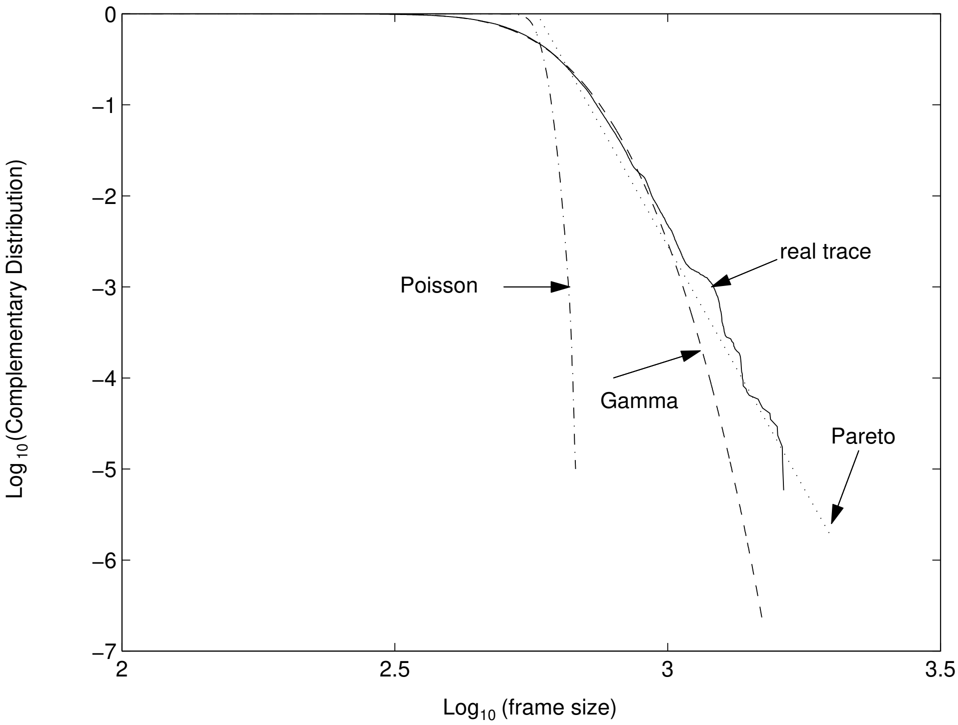


Figure 2: Complementary frame-size distribution for the *Star Wars* trace along with several fits.

Indeed, it suffices to take

$$T(x) \stackrel{\triangle}{=} F^{-1}(H(x)), \quad x \in \mathbb{R} \tag{20}$$

where F^{-1} denotes the (generalized) inverse of F.

The program is now clear: Consider a (stationary) $M|G|\infty$ process $\{b^\star_n,\ n = 0, 1, \ldots\}$ characterized by the pair (λ, G). The common distribution H of these variates is Poisson with parameter $\lambda \mathbf{E}[\sigma]$. For any frame-size distribution F, define the transformed process $\{a_n,\ n = 0, 1, \ldots\}$ as

$$a_n \stackrel{\triangle}{=} T(b^\star_n) = F^{-1}(H(b^\star_n)) = F^{-1}(F_{Poisson}(b^\star_n)), \quad n = 1, 2, \ldots \tag{21}$$

For each $n = 0, 1, \ldots$, the rv a_n will be distributed according to F. In fact, the transformed process $\{a_n,\ n = 0, 1, \ldots\}$ is still stationary and ergodic. In general, the covariance structures of the two processes will *not* be exactly the same. The best one may hope for is that these covariance structures are approximately equal, i.e.,

$$\text{cov}[a_{n+k}, a_n] \simeq \text{cov}[b^\star_{n+k}, b^\star_n], \quad n, k = 0, 1, \ldots \tag{22}$$

Next, we need to select an appropriate distribution F. Several theoretical fits have been suggested for the frame-size distribution of a video sequence, including Gamma [14], lognormal [13, 17], and hybrid Gamma/Pareto distributions [9]. The last fit was found quite appropriate for *Star Wars* data.

Accordingly, we use it here to model the frame-size distribution. As explained in [9], the Gamma distribution is used to capture the general shape of the empirical distribution, whereas the Pareto distribution is used to capture the tail of the empirical distribution. Let F_Γ and F_P denote the cumulative probability functions for the Gamma and Pareto distributions, respectively. Although F_Γ has no closed-form expression, its derivative is given simply by

$$f_\Gamma(x) = \frac{\omega^s}{\Gamma(s)} x^{s-1} e^{-\omega x}, \quad x \geq 0 \tag{23}$$

where the parameters $s > 0$ and $\omega > 0$ are the shape and scale parameters, respectively, and the standard Gamma function $\Gamma(s)$ is given by

$$\Gamma(s) \triangleq \int_0^\infty x^{s-1} e^{-x} dx, \quad s > 0. \tag{24}$$

The Pareto distribution we use has the explicit form

$$F_P(x) = \begin{cases} 1 - (\frac{a}{x})^\alpha & \text{if } x \geq a \\ 0 & \text{if } x < a \end{cases} \tag{25}$$

with parameters $\alpha > 0$ and $a > 0$ which are both determined by fitting.

The hybrid Gamma/Pareto distribution $F_{\Gamma/P}$ is then given by

$$F_{\Gamma/P}(x) = \begin{cases} F_\Gamma(x) & \text{if } x \leq x^* \\ F_P(x) & \text{if } x > x^* \end{cases} \tag{26}$$

for some $x^* > 0$. As in [9], 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 (Figure 2). Using the continuity condition $F_\Gamma(x^*) = F_P(x^*)$ along with least-square fitting of the Pareto tail, estimates of a and α can be obtained. Table 2 gives the estimated parameters for three traces (frame sizes are in 48-byte cells). Since the fourth video trace is relatively short, accurate fitting of its extreme tail is not possible.

Trace	Mean	Std. Dev.	ω	s	x^*	a	α
	(cells)	(cells)	(1/cells)		(cells)	(cells)	
Star Wars	579.5	130.3	3.41E–2	19.78	650	576	10.7
Beauty & Beast	264.3	74.6	4.75E–2	12.55	398	215	5.31
Crocodile Dundee	225.0	48.7	9.50E–2	21.4	355	224	10.1

Table 2: Estimated values of various parameters in the hybrid Gamma/Pareto model.

Thus, we select $F = F_{\Gamma/P}$, and the Poisson variates of the $M|G|\infty$ process can now be transformed into Gamma/Pareto variates. Let $\{b_n^\star,\ n = 0, 1, \ldots\}$ denote the $M|G|\infty$ process with $\lambda = 1$ and service time distribution (17), so that its correlation structure is given by (11). The sequence $\{b_n^\star,\ n = 0, 1, \ldots\}$ is transformed into a new sequence $\{a_n,\ n = 0, 1, \ldots\}$ through the transformation

$$a_n = F_{\Gamma/P}^{-1}(F_{Poisson}(b_n^\star)), \quad n = 1, 2, \ldots \tag{27}$$

where $F_{Poisson}$ is the cumulative probability function of a Poisson rv with parameter $\mathbf{E}\left[\sigma\right]$ (given by (18)) and

$$F_{\Gamma/P}^{-1}(y) = \begin{cases} F_P^{-1}(y) = a/(1-y)^{1/\alpha} & \text{if } y > F_P(x^*) = 1 - (a/x^*)^\alpha \\ F_\Gamma^{-1}(y) & \text{otherwise} \end{cases} \tag{28}$$

with F_Γ^{-1} obtained numerically.

Since only the Gamma part is used in fitting the mean and variance, the mean and variance of a_n will be slightly different from their empirical counterparts. For example, the mean frame size in a synthetic trace is given by

$$\mathbf{E}\left[a_n\right] = \int_0^{x^*} x f_\Gamma(x)dx + \int_{x^*}^{\infty} f_P(x)dx \tag{29}$$

while the empirical mean is fitted to $\int_0^\infty x f_\Gamma(x)dx$. However, this slight discrepancy is of no significance.

As pointed out above, this transformation does not, in general, preserve the original correlation structure. However, in all our experiments, the effect of transformation was barely noticeable. An example of the sample ACFs of several transformed $M|G|\infty$ traces, along with their average (i.e., the average of the sample ACFs) is shown in Figure 3 based on *Star Wars* fitting. The average ACF is almost indistinguishable from the theoretical ACF of the non-transformed $M|G|\infty$ process.

5 Synthetic Trace Generation and Computational Issues

Ideally, we would like to analytically determine the queueing performance for a traffic model so that control decisions related to call admission and resource allocation can be done on-line. However, there is a natural tradeoff between the complexity of a model and the relative accuracy of its queueing predictions. A detailed video model, such as the one considered in this paper, does not easily lend itself to queueing analysis, but can be used to drive network simulations. Performance evaluation by means of simulations is useful in off-line dimensioning problems (e.g., buffer sizing under a fixed quality of service). The simulation time can sometimes be reduced by employing certain problem-specific techniques (some of which are discussed in the next section). Separating the issue of model construction from that of queueing tractability allows highly accurate models to be developed. It should also be mentioned that models with analytically tractable performance are not always usable in on-line traffic control problems, particularly when extensive numerical computations are needed to

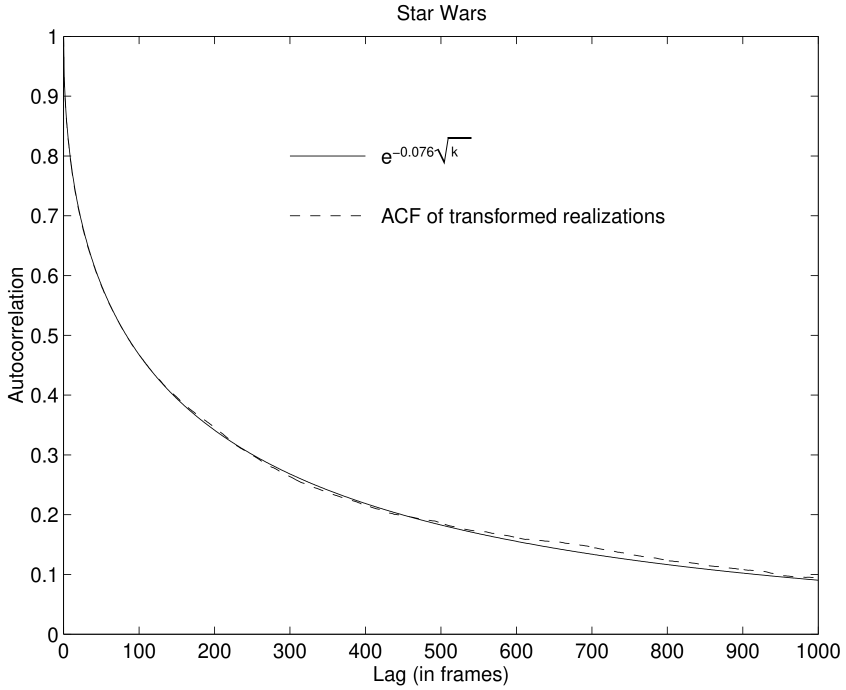


Figure 3: Impact of transformation on the autocorrelation structure.

obtain the results. While network simulations can be driven by "real" data, such data sets are often not available or very difficult to obtain. A stochastic model, on the other hand, encompasses many realizations that represent independent yet structurally" similar (i.e., homogeneous) streams, which are ideal for statistical multiplexing studies.

5.1 Simulation Models

To verify the appropriateness of the $M|G|\infty$-based model, we investigate its queueing performance and contrast it with the performance of two popular video models: the F-ARIMA model [9] (which exhibits LRD), and the DAR(1) model [14] (which exhibits a Markovian structure). By a suitable transformation, we ensure that all models share the same hybrid Gamma/Pareto marginal distribution, thereby eliminating the impact of the marginal frame-size distribution. In all three models, the hybrid Gamma/Pareto distribution is discretized to obtain integer-valued frame sizes.

Synthetic realizations from the three video models were generated and used in the queueing simulations described in the next section. Each of the $M|G|\infty$ and DAR(1) traces consists of $1,000,000$ data points, while each F-ARIMA trace consists of $500,000$ data points (a data point corresponds to a frame size measured in cells). The F-ARIMA traces are shorter than their $M|G|\infty$ and DAR(1) counterparts since generating F-ARIMA traces of length $1,000,000$ is computationally prohibitive. More specifically, it requires $\mathcal{O}(n^2)$ computations to generate a F-ARIMA trace of length n using Hosking's algorithm [15] (before transformation). In contrast, both the $M|G|\infty$ and DAR(1) models require only $\mathcal{O}(n)$ computations per trace. To generate a F-ARIMA trace of only $100,000$ points using Hosking's algorithm, it took about three days of execution on a Sparc-10 workstation. To generate $500,000$-long F-ARIMA traces, we used an approximation due to Haslett and Raftery [11], which was incorporated in the S-Plus package. Even with this approximation, it took about two days to obtain one $500,000$-long F-ARIMA trace, compared to less than a minute for a $1,000,000$-long $M|G|\infty$ or DAR(1) trace. Extensive simulations based on the three models were conducted. For brevity, we show the results for one real trace (the *Star Wars*) and its corresponding models.

5.2 F-ARIMA and DAR(1) Models

The F-ARIMA model [9] used here is constructed by transforming a fractional ARIMA process with a standard normal marginal distribution into one with a hybrid Gamma/Pareto distribution. An example of the sample ACF of a synthetic F-ARIMA realization for the *Star Wars* trace is shown in Figure 4.

The theoretical ACF of a F-ARIMA process is given by

$$\rho(k) = \frac{d(1+d)\cdots(k-1+d)}{(1-d)(2-d)\cdots(k-d)}, \quad k = 1, 2, \ldots, \; 0 < d < 0.5 \tag{30}$$

which behaves as $k^{-\beta}$ only *asymptotically* ($d = H - 1/2$). In fact, the ACF of the F-ARIMA model underestimates the short-term correlations of the real data even more than $k^{-\beta}$.

We have transformed the normally distributed variates of the standard F-ARIMA model into Gamma/Pareto variates. Here, as with the $M|G|\infty$-based model, inspection of Figure 4 suggests that the transformation has almost no impact on the correlation structure of the original F-ARIMA process. This is in keeping with the work in [16] where under mild conditions, a transformed LRD Gaussian process is shown to maintain its Hurst value.

The DAR(1) model is obtained as follows [1]: Let $\{V_n, \; n = 0, 1, \ldots\}$ and $\{U_n, \; n = 0, 1, \ldots\}$ be two mutually independent processes of *i.i.d.* rvs. For $n = 0, 1, \ldots$, the rv V_n is Bernoulli with $\mathbf{P}[V_n = 1] = 1 - \mathbf{P}[V_n = 0] = r$, and the rv U_n is an $\mathbb{N}$-valued rv distributed according to the pmf $\pi(i) \overset{\triangle}{=} \mathbf{P}[U_n = i]$, $i = 0, 1, \ldots$. A DAR(1) process $\{X_n, \; n = 0, 1, \ldots\}$ is defined through the recursion

$$X_n = V_n X_{n-1} + (1 - V_n) U_n, \quad n = 1, 2, \ldots \tag{31}$$

with given X_0. The sequence $\{X_n, \; n = 0, 1, \ldots\}$ is a Markov chain with the same marginal distribution as $\pi = (\pi(0), \pi(1), \ldots,)$, i.e., $\mathbf{P}[X_n = i] = \pi(i)$, $i = 0, 1, \ldots$, and with an ACF of of the form $\rho_k = r^k$, similar to that of the familiar AR(1) process. In [14] the DAR(1) model was used to characterize video-teleconferencing streams, with the marginal distribution taken as a negative binomial distribution; the discrete analog of a Gamma distribution. Here, instead, we use a hybrid Gamma/Pareto marginal distribution, consistent with our choice for the other two models examined in the paper.

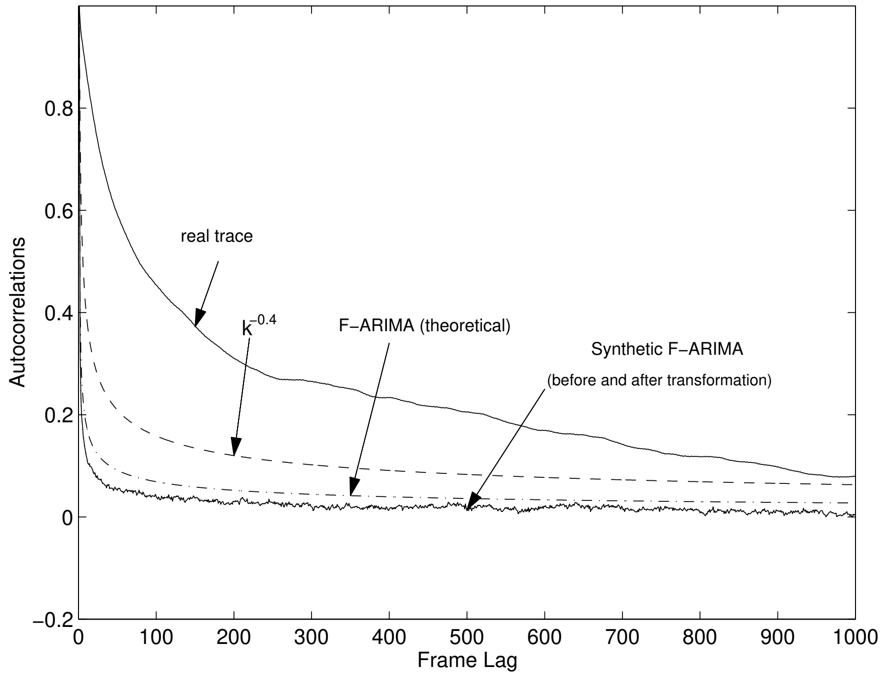


Figure 4: Autocorrelation function of F-ARIMA model.

5.3 Generation of $M|G|\infty$ traces

A CSIM program[1] was written to generate synthetic $M|G|\infty$ traces. The program simulates an $M|G|\infty$ queue with infinite servers. Time is slotted in frame periods. At the start of a time slot, a batch of arrivals is generated according to a Poisson distribution with $\lambda = 1$. Each arrival is kept for a random time σ whose pmf is given by (17). A synthetic $M|G|\infty$ trace is obtained from the number of remaining customers at the beginning of each time slot. This trace is then transformed into one with a Gamma/Pareto marginal distribution.

The computational complexity for generating an $M|G|\infty$ trace of length n (before transforming the marginals) is $\mathcal{O}(n)$. To show that, we provide a sequential version of our CSIM program, which is shown in Figure 5. There are three nested 'for' loops in this algorithm. In each iteration of the outermost loop, a batch of Poisson arrivals is generated. A service time is obtained for each customer in the batch (the second 'for' loop). Finally, the effect of the service time of a customer is incorporated in the innermost loop by incrementing the values of busy servers in future time slots during which a customer is being served. It is easy to see that the *average* complexity of the algorithm is $\mathcal{O}(n\lambda\mathbf{E}[\sigma])$. Since λ and $\mathbf{E}[\sigma]$ are fixed and independent of n, the complexity is $\mathcal{O}(n)$. The computational complexity for the generation of DAR(1) traces is also $\mathcal{O}(n)$ (see [14] for details on how to generate DAR(1) traces).

We note that due to the correlated nature of cell losses, extremely long traces are needed to obtain meaningful results under small cell loss probabilities. In fact, we first tried using shorter traces of

[1] CSIM is a C-based discrete-event simulation language [30].

```
$A \overset{\triangle}{=}$ Poisson rv with mean $\lambda = 1$ (batch size)
$\sigma \overset{\triangle}{=}$ rv with distribution $G$ given in (16)
$X[k] \overset{\triangle}{=} k$th value in the resulting trace (number of busy servers)
$n \overset{\triangle}{=}$ trace length
for $i = 1$ to $n$ do
    Set $X[i] := 0$ /* initialize counters */
end for
 for $i = 1$ to $n$ do
    generate batch size: $A \sim$ Poisson($\lambda$)
        for $j = 1$ to $A$ do
            generate a service time: $\sigma \sim G$
            for $k = 1$ to $\sigma$
                increment counter: $X[i+k] = X[i+k] + 1$
            end for
        end for
end for
```

Figure 5: Sequential algorithm for generating an $M|G|\infty$ synthetic trace.

length 100,000, and found that for realistic loss rates, losses occur in only few frames, e.g., in one particular experiment, a loss rate of 8.3E–6 (484 cells) came from five errored frames only. Intuitively, correlations make it more likely that large frames follow each other, thus causing correlated periods of buffer overflow. Moreover, 100,000-long realizations may not be long enough to display the extreme tail of the frame-size distribution, causing the loss performance to be underestimated. For example, the maximum frame size in the *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., $F^{-1}_{\Gamma/P}(F_{Poisson}(33)) = 894$. An $M/G/\infty$ trace before transformation is a realization of n identically distributed rvs $b^\star_1, \ldots, b^\star_n$, which are *associated* (Proposition 1). By the well-known properties of associated rvs [5], we have

$$\mathbf{P}\left[\max_{i=1,\ldots,n} b^\star_i > x\right] \leq 1 - \prod_{i=1}^{n} \mathbf{P}\left[b^\star_i < x\right] = 1 - F_{Poison}(x)^n, \quad x \in \mathbb{R}\,. \tag{32}$$

Thus, for $n = 100,000$, we find that

$$\mathbf{P}\left[\max_{i=1,\ldots,n} b^\star_i > 32\right] < 1 - (F_{Poisson})(32))^{100,000} = 0.4745, \tag{33}$$

i.e., there is less than 50% chance that the 100,000-long realization reaches the real maximum frame size.

6 Queueing Performance

To verify the appropriateness of the $M|G|\infty$ model, we investigate its queueing performance and compare it to the performance of the F-ARIMA and DAR(1) video models. For brevity, we show the results for one real trace (the *Star Wars*) and its corresponding video models. The queueing system consists of a single-server FIFO queue with capacity B (in cells) and constant release rate C (cells/slot). Two types of simulations were conducted. The first is for a single stream (i.e., no multiplexing), which is used to contrast the performance of the three models with reference to the performance under the real trace. It is expected that discrepancies in the buffer overflow behavior are most apparent in the single-stream case. In the second type of simulations, we investigate the performance for several, statistically multiplexed streams. Obtaining the performance for real video streams in this case raises a fundamental challenge: Since no two real traces exhibit the same statistical structure (due to differences in scene dynamics), in principle one cannot obtain the multiplexing performance for independent *and* homogeneous real video streams. Possible approximate approaches that can be used for this purpose include:

(i) Obtain multiple "real" streams from a single empirical trace by arranging the trace as a circular list, starting each stream at a random location in this list, and proceeding sequentially until the circle is completed [14]. The problem with this approach is that the resulting streams are *not* independent, particularly if the starting times are not sufficiently separated.

(ii) Multiplex traces of different movies. Since traces typically differ in their statistical properties (e.g., mean, variance, etc.), multiplexing them amounts to multiplexing heterogeneous streams. This works well if we are only interested in the heterogeneous case. However, we are also interested in the homogeneous case which gives us a better understanding of the multiplexing gain and the *average* loss performance that individual streams will experience.

While neither approach is completely satisfactory, unfortunately one has no other alternatives. We opted for a modified version of the first approach, whereby the starting times are chosen to be maximally separated (to reduce the potential dependence between the multiplexed streams). Furthermore, we limit our study of the performance under real streams to the case of five multiplexed streams, so that streams' starting points are sufficiently distanced from each other. Of course, no such restriction is necessary when studying the multiplexing performance under the various traffic models. In all experiments, we assume that cells in each frame are evenly distributed over the frame duration. Two measures of performance are considered: the cell loss rate, and the frame error rate. A frame is errored when one or more of its cells are lost. This measure is important for applications that do not implement error concealment mechanisms for recovery from partial frame losses.

6.1 Single Stream

The cell loss rate (CLR) and the frame error rate (FER) are examined at three loads: $U = 80\%$ (heavy load), 60% (moderate load), and 40% (light load). A summary of the simulations results to two significant digits is given in Table 3. The depicted results for the three models represent the averages of ten independent runs.

For $U = 80\%$ and 60%, the buffer size is varied from 100 to 2500 cells. As expected, CLR and FER for a real stream are quite high at $U = 80\%$. 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. At both $U = 80\%$ and $U = 60\%$, the FER for the real stream decreases by about 50% when B is increased from 100 to 2500 cells.

By comparing the performance for the three models with reference to the performance for the real stream, we observe the following: In the heavy-load regime, both $M/G/\infty$ and DAR(1) models provide acceptable predictions of CLR and FER, with DAR(1) being slightly more accurate. Under the F-ARIMA model, the performance is overly sensitive to the buffer size, to the extent that it underestimates the actual CLR and FER by orders of magnitude when B is large. This is clearly a consequence of not sufficiently capturing the short-term correlations. Going to the moderate-load regime, we observe that once again both $M/G/\infty$ and DAR(1) models provide significantly more accurate predictions of CLR and FER than the F-ARIMA model. In this regime, DAR(1) and $M/G/\infty$ models give comparable results (particularly, with respect to the CLR measure).

Interestingly, in the light-load regime ($U = 40\%$), the DAR(1) model is no more capable of providing acceptable performance predictions. In fact, no losses were observed in any of the DAR(1) simulations (although 10 independent simulations each with a 1,000,000-long trace were used). The $M/G/\infty$ model is quite accurate in this regime. The F-ARIMA model is still overly sensitive to the buffer size, although the gap between its performance and the real performance is now smaller (when B is small the F-ARIMA model overestimates CLR and FER, but as B increases the model starts to underestimate both performance measures). The main conclusion to be drawn from Table 3 is that *of the three examined models, only the $M/G/\infty$ model is observed to consistently provide acceptable performance predictions at various traffic loads.* The performance of the $M/G/\infty$ is always within an order of magnitude of the real performance. The capability of the $M/G/\infty$ model of providing acceptable results can be attributed to the fact that it incorporates the good aspects of Markovian and LRD models; similar to Markov models, it incorporates the short-term correlations, and similar to LRD models, it captures the slowing decaying nature of the correlation structure in a VBR video sequence.

The $M|G|\infty$ model slightly underestimates the actual queueing performance, particularly at intermediate loss rates (i.e., 1.0E–3 to 1.0E–4) and large buffer sizes. An examination of the real trace reveals that much of the discrepancy is related to some ‘nonstationarity” in the real data, which is not accounted for in the $M|G|\infty$ model. In particular, the first and last few thousand frames of the *Star Wars* trace exhibit stronger statistical correlations than the rest of the trace. We speculate these frames correspond to the compressed frames in the credits (the portion that contains the names of actors, acknowledgements, etc.).

In the above simulations, the simulation time was significantly reduced by conducting the discrete-event simulation at the frame level (rather than the cell level). The algorithm that was used for these single-stream simulations is shown in Figure 6. It exploits the fact that only the frame sizes, the service rate, the maximum buffer size, and the queue length at the beginning and end of each time slot are relevant to the computation of the CLR and FER measures.

```
$Q_j \triangleq$ length of the queue at the beginning of the $j$th slot
$C \triangleq$ service rate (cells/slot)
$B \triangleq$ buffer size
$X_j \triangleq$ $j$th value in the trace (e.g., frame size)
Initialize
  $Q_1 := 0$
  $lost\_cells := 0$
for $j = 1$ to $last\_frame$ do
  if $X_j \leq C$ then /* underflow */
    $Q_{j+1} := \max\{Q_j + X_j - C, 0\}$
  else
    $T := Q_j + X_j - C - B$
    if $T > 0$          /* cell losses */
      $lost\_cells := lost\_cells + T$
      $Q_{j+1} := B$
    else
      $Q_{j+1} := T + B$
    end if
  end if
end for
```

Figure 6: Algorithm for approximating the loss rate for a single trace.

6.2 Multiplexed Streams

In this section, we investigate the multiplexing performance for the three models for the purpose of contrasting their different behaviors. It is not our objective here to provide a thorough evaluation of the multiplexing gain and the associated resource allocation problem, which will be the topic of a

Buffer	Cell Loss Rate				Frame Error Rate			
Size (cells)	Real	$M/G/\infty$	F-ARIMA	DAR(1)	Real	$M/G/\infty$	F-ARIMA	DAR(1)
100	1.7E–2	1.1E–2	8.5E–3	1.9E–2	1.1E–1	6.7E–2	4.8E–2	1.2E–1
500	1.6E–2	9.6E–3	1.6E–3	1.7E–2	9.1E–2	5.0E–2	7.3E–3	9.9E–2
1000	1.5E–2	8.7E–3	4.6E–4	1.5E–2	7.9E–2	4.3E–2	1.9E–3	8.0E–2
1500	1.4E–2	8.0E–3	1.7E–4	1.3E–2	7.0E–2	3.8E–2	6.7E–4	6.7E–2
2000	1.3E–2	7.6E–3	7.9E–5	1.2E–2	6.3E–2	3.5E–2	2.9E–4	5.7E–2
2500	1.2E–2	7.2E–3	4.1E–5	1.1E–2	5.8E–2	3.2E–2	1.5E–4	4.9E–2

(a) $U = 80\%$

Buffer	Cell Loss Rate				Frame Error Rate			
Size (cells)	Real	$M/G/\infty$	F-ARIMA	DAR(1)	Real	$M/G/\infty$	F-ARIMA	DAR(1)
100	1.2E–3	6.6E–4	7.6E–4	5.6E–4	6.2E–3	2.9E–3	3.4E–3	4.7E–3
500	1.1E–3	4.9E–4	7.1E–5	4.8E–4	5.1E–3	1.7E–3	2.4E–4	3.4E–3
1000	1.0E–3	4.0E–4	8.9E–6	4.0E–4	4.4E–3	1.3E–3	2.3E–5	2.5E–3
1500	1.0E–3	3.4E–4	1.6E–6	3.4E–4	3.9E–3	1.1E–3	4.2E–6	2.0E–3
2000	9.6E–4	3.0E–4	5.2E–7	3.0E–4	3.6E–3	9.1E–4	6.0E–7	1.6E–3
2500	9.1E–4	2.7E–4	2.0E–7	2.6E–4	3.3E–3	7.9E–4	2.0E–7	1.3E–3

(b) $U = 60\%$

Buffer	Cell Loss Rate				Frame Error Rate			
Size (cells)	Real	$M/G/\infty$	F-ARIMA	DAR(1)	Real	$M/G/\infty$	F-ARIMA	DAR(1)
10	1.3E–5	1.2E–5	5.3E–5	0	6.4E–5	7.1E–5	1.8E–4	0
20	1.3E–5	1.1E–5	4.9E–5	0	5.8E–5	5.7E–5	1.7E–4	0
30	1.3E–5	1.1E–5	4.7E–5	0	5.8E–5	4.0E–5	1.6E–4	0
40	1.3E–5	1.1E–5	4.4E–5	0	5.8E–5	3.8E–5	1.5E–4	0
50	1.2E–5	1.0E–5	4.2E–5	0	5.8E–5	3.7E–5	1.4E–4	0
60	1.2E–5	1.0E–5	3.9E–5	0	5.8E–5	3.6E–5	1.3E–4	0
70	1.2E–5	9.7E–6	3.7E–5	0	5.3E–5	3.5E–5	1.2E–4	0
80	1.2E–5	9.5E–6	3.5E–5	0	5.3E–5	3.5E–5	1.1E–4	0
90	1.2E–5	9.2E–6	3.3E–5	0	5.3E–5	3.5E–5	1.1E–4	0
100	1.2E–5	9.0E–6	3.1E–5	0	5.3E–5	3.5E–5	9.9E–5	0
200	1.1E–5	7.2E–6	1.9E–5	0	4.7E–5	2.3E–5	5.4E–5	0
300	9.9E–6	6.2E–6	1.1E–5	0	4.1E–5	1.9E–5	3.0E–5	0
400	8.9E–6	5.6E–6	7.4E–6	0	4.1E–5	1.7E–5	1.9E–5	0
500	7.9E–6	5.0E–6	4.8E–6	0	3.5E–5	1.4E–5	1.3E–5	0
600	6.8E–6	4.5E–6	2.9E–6	0	3.5E–5	1.3E–5	8.8E–6	0
700	5.8E–6	4.2E–6	1.7E–6	0	2.9E–5	1.1E–5	5.4E–6	0
800	4.8E–6	3.9E–6	1.0E–6	0	2.9E–5	9.8E–6	2.8E–6	0
900	3.8E–6	3.6E–6	7.8E–7	0	2.3E–5	8.8E–6	8.0E–7	0
1000	2.8E–6	3.4E–6	6.7E–7	0	2.3E–5	7.9E–6	6.0E–7	0

(c) $U = 40\%$

Table 3: Average cell loss and frame error rates at three different loads (*Star Wars* trace). Ten independent replications are used to obtain the values for each model.

future study. For simplicity, we assume that frames' boundaries of multiplexed streams are aligned in time, so the time axis is slotted in frame periods. This specialization allows us to significantly reduce the simulation time using the following optimization.

Consider a simulation experiment in which N video streams (indicated by their frame-size traces) are to be multiplexed. Assume that the N streams have the same number of frames, n. Let $\{X_j^{(k)},\ j = 1, 2, \ldots, n\}$ be the frame-size sequence for the kth stream, $k = 1, 2, \ldots, N$. To obtain the CLR and FER for the multiplexed N streams, we first compute an *aggregate* trace $\{\overline{X}_j,\ j = 1, 2, \ldots, n\}$ from the pointwise sum of the N traces, i.e., $\overline{X}_j = \sum_{k=1}^{N} X_j^{(k)}$, for $j = 1, \ldots, n$. For a time slot (i.e., a frame period) in which buffer overflow cannot occur, the aggregate trace can be used to update the buffer occupancy at the end of that slot. This updating is done on a frame-by-frame basis, using an algorithm similar to the one in Figure 6. For time slots during which buffer overflow is possible (based on some sufficient conditions that will be introduced shortly), the individual traces are used to simulate the performance on a cell-by-cell basis.

Fortunately, buffer overflow occurs only in a small fraction of the total number of simulated time slots (n). Let Q_j denote the queue length at the beginning of the jth slot. It can be shown that either of the following two conditions guarantees no buffer overflow during the jth slot:

1. $(\overline{X}_j \leq C) \bigcap (Q_j \leq B - N)$.
2. $(\overline{X}_j > C) \bigcap (\overline{X}_j - C \leq B - Q_j) \bigcap (Q_j \leq B - N)$.

With this optimization, the simulation time for computing the queueing performance for N multiplexed streams is $\mathcal{O}(n + \alpha n W N)$, where n is the trace length, α is the fraction of slots for which neither of the above conditions is satisfied, and W is the average number of cells per frame per stream during buffer overflow. Typically, $\alpha W \ll 1$, making the complexity much less than $\mathcal{O}(Nn)$.

To give an idea about the efficiency of the above simulation approach, Table 4 gives an example of the simulation times for ten multiplexed $M/G/\infty$ streams with different buffer sizes (the results in the table were based on a single run). As the buffer size increases, both CLR and FER decrease, resulting in shorter simulation times. In this example, a reduction of almost an order of magnitude in the CLR resulted in an equivalent reduction of an order of magnitude in simulation time.

The multiplexing performance for the three models is shown in Table 5 for $N = 5$ and $N = 10$ at a load of $U = 80\%$. Each value in the table represents an *average* over N streams and over 5 independent simulations. In the case of $N = 5$, we have also provided results for real streams, with the five streams being derived from the original empirical trace as described before. It can be observed that the $M/G/\infty$ model provides the closest performance to the real performance. The F-ARIMA model is overly sensitive to the buffer size (i.e., the CLR and FER in the F-ARIMA model decrease with an increase in the buffer size faster than the corresponding trend seen by real video sources). Such an overly sensitive behavior (which we have seen before in the case of $N = 1$) can lead to

Buffer Size (cells)	Average CLR	Simulation Time (seconds)
100	3.2E–5	512.84
200	3.0E–5	381.95
300	2.8E–5	362.72
400	2.7E–5	328.50
⋮	⋮	⋮
3500	7.8E–6	108.13
4000	6.1E–6	95.47
4500	4.4E–6	88.39
5000	2.7E–6	63.52

Table 4: Example of the reduction in the simulation time as the CLR decreases ($N = 10$).

overallocation or underallocation of buffer and bandwidth resources. While the DAR(1) model shows acceptable sensitivity to buffer size, it is shown to overestimate both CLR and FER.

The results for $N = 10$ are provided for the purpose of contrasting the three models. But since the performance under real video streams is not available in this case, one cannot make definite conclusions about the relative accuracy of the three models (for $N = 10$, we could not use the previous trick to obtain ten streams from a single empirical trace since the extracted streams start to show some non-negligible cross-correlations). However, one can make few observations by contrasting the behavior when $N = 10$ to the previous case when $N = 5$. As in the case of $N = 5$, the performance for the F-ARIMA model when $N = 10$ is more sensitive to the buffer size than the other two models. Both DAR(1) and $M/G/\infty$ models display comparable sensitivities to buffer size. However, the CLR performance for the DAR(1) model is more than an order of magnitude higher than that of the $M/G/\infty$ model. Given the performance for the real streams when $N = 5$ and that for the DAR(1) model when $N = 10$, one could conclude that the DAR(1) is probably overestimating the CLR performance (realistically, we should expect an appreciable reduction in the CLR when going from $N = 5$ to $N = 10$). Of course, a conclusive judgment would require obtaining the performance for ten multiplexed real streams.

7 Concluding Remarks

In this paper, we investigated a new approach for characterizing VBR video streams based on $M|G|\infty$ processes. These processes enjoy several attractive features that make them a viable approach for modeling various types of network traffic. A compelling statistical evidence from four different video traces suggests that the ACF of a VBR sequence is better captured by $e^{-\beta\sqrt{k}}$ than by $e^{-\beta k}$ (Markovian) or $e^{-\beta \log k}$ (LRD). While Markovian models capture the short-term correlations and LRD models capture the long-term correlations, the fit $e^{-\beta\sqrt{k}}$ is shown to sufficiently capture the empirical correlations at all lags. To display such a correlation structure, an $M|G|\infty$-based model for video was developed, which exhibits short-range dependence (though not Markovian). The Poisson marginals of

Buffer	Cell Loss Rate				Frame Error Rate			
Size (cells)	Real	$M/G/\infty$	F-ARIMA	DAR(1)	Real	$M/G/\infty$	F-ARIMA	DAR(1)
200	2.6E–4	2.3E–4	2.1E–4	8.2E–4	6.2E–3	3.7E–3	4.8E–3	1.3E–2
400	2.4E–4	2.0E–4	6.9E–5	7.7E–4	5.7E–3	2.9E–3	1.5E–3	1.1E–2
600	2.3E–4	1.8E–4	2.6E–5	7.2E–4	5.3E–3	2.5E–3	5.5E–4	9.6E–3
800	2.2E–4	1.7E–4	1.1E–5	6.8E–4	5.1E–3	2.2E–3	2.2E–4	9.3E–3

(a) $N = 5$

Buffer	Cell Loss Rate			Frame Error Rate		
Size (cells)	$M/G/\infty$	F-ARIMA	DAR(1)	$M/G/\infty$	F-ARIMA	DAR(1)
100	7.4E–6	2.1E–5	1.4E–4	2.2E–4	7.9E–4	3.0E–3
200	6.3E–6	1.3E–5	1.3E–4	1.7E–4	4.6E–4	2.8E–3
300	5.6E–6	7.8E–6	1.3E–4	1.5E–4	3.0E–4	2.6E–3
400	4.8E–6	4.7E–6	1.2E–4	1.2E–4	1.8E–4	2.4E–3

(b) $N = 10$

Table 5: Average cell loss and frame error rates for N multiplexed streams ($U = 80\%$).

the $M|G|\infty$ process were transformed into ones with a more appropriate distribution (due to Garrett and Willinger [9]). The impact of the transformation is shown to be negligible. With the performance of a real stream taken as a reference, we examined the queueing performance under the $M|G|\infty$ model and contrasted it to the performances for two popular video models: the F-ARIMA model (LRD) and the DAR(1) (Markovian). Our simulation results indicate that the $M|G|\infty$ model consistently provides acceptable predictions of the actual cell loss and frame error rates at various traffic loads and buffer sizes. In contrast, the performance for F-ARIMA model is overly sensitive to the buffer size, to the extent that it sometimes underestimates the real performance by several orders of magnitude. The DAR(1) model, while showing acceptable trend to changes in buffer size, sometimes gives unacceptably optimistic predictions (e.g., the case of a single stream with 40% traffic intensity), and in other times pessimistic predictions (case of multiplexed streams). An additional advantage of the $M|G|\infty$ model over the F-ARIMA is that only $\mathcal{O}(n)$ computations are needed to generate a synthetic trace of size n, compared to $\mathcal{O}(n^2)$ for a F-ARIMA trace. Our future work will focus on using the $M|G|\infty$ model in on-line admission control and dynamic resource allocation. Towards this end, we have been working on analytically obtaining the queueing performance for multiplexed $M|G|\infty$ sources and using such performance to compute the effective bandwidth. Results of this research will be reported in a future work.

Acknowledgement

The authors would like to thank Mark Garrett and Wu Feng for making their video traces available. They also thank the anonymous reviewers for their careful evaluation of the paper. The work of the first author was supported through a University of Arizona grant. 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. The final version of this manuscript was prepared while the second author was visiting INRIA - Sophia Antipolis; the hospitality of Projet MISTRAL is gratefully acknowledged.

Appendix

A Proof of Equation (6)

The derivation of (6) is based on the following well-known result on random sums of *i.i.d.* rvs.

Lemma 1 *Let* $\{X, X_n,\ n = 1, 2, \ldots\}$ *be a sequence of* $\mathbb{R}$*-value i.i.d. rvs which are independent of an* $\mathbb{N}$*-valued rv* ν*. For any two functions* $f, g : \mathbb{R} \to \mathbb{R}$*, we have*

$$\begin{aligned} \mathrm{cov}\left[\sum_{j=1}^{\nu} f(X_j), \sum_{i=1}^{\nu} g(X_i)\right] &= \mathbf{E}\left[\nu\right]\mathbf{E}\left[f(X)g(X)\right] + \Big(\mathbf{E}\left[\nu(\nu-1)\right] - (\mathbf{E}\left[\nu\right])^2\Big)\mathbf{E}\left[f(X)\right]\mathbf{E}\left[g(X)\right] \\ &= \mathbf{E}\left[\nu\right]\mathrm{cov}[f(X), g(X)] + \mathrm{var}(\nu)\mathbf{E}\left[f(X)\right]\mathbf{E}\left[g(X)\right] \qquad (34) \end{aligned}$$

provided the expectations exist. □

Consider the $M|G|\infty$ input process $\{b_n, n = 0, 1, \ldots\}$. For each $n = 0, 1, \ldots$, we note that

$$b_n = b_n^{(0)} + b_n^{(a)} \qquad (35)$$

where the rvs $b_n^{(0)}$ and $b_n^{(a)}$ describe the contributions to the number of customers in the system at the beginning of slot $[n, n+1)$ from those initially present (at $n = 0$) and from the new arrivals, respectively. Under the enforced operational assumptions, we readily have

$$b_n^{(a)} = \sum_{s=1}^{t} \sum_{i=1}^{\xi_s} \mathbf{1}\left[\sigma_{s,i} > n - s\right] \qquad (36)$$

and

$$b_n^{(0)} = \sum_{i=1}^{b_0} \mathbf{1}\left[\sigma_{0,i} > n\right]. \qquad (37)$$

The stationary version $\{b_n^\star,\ n = 0, 1, \ldots\}$ is obtained by assuming that (i) the rv b_0 is a Poisson rv with parameter $\lambda\mathbf{E}\left[\sigma\right]$; (ii) the rvs $\{\sigma_{0,j},\ j = 1, 2, \ldots\}$ are *i.i.d.* rvs distributed according to the pmf (4) of the forward recurrence time associated with σ.

Fix $n = 0, 1, \ldots$ and $k = 1, 2, \ldots$. By independence, we have

$$\Gamma(k) \stackrel{\triangle}{=} \mathrm{cov}[b_n, b_{n+k}] = \mathrm{cov}[b_n^{(0)}, b_{n+k}^{(0)}] + \mathrm{cov}[b_n^{(a)}, b_{n+k}^{(a)}]. \qquad (38)$$

First we consider the term $\mathrm{cov}[b_n^{(a)}, b_{n+k}^{(a)}]$: Under the enforced independence assumptions,

$$\mathrm{cov}\left[b_n^{(a)}, \sum_{s=n+1}^{n+k} \sum_{i=1}^{\xi_s} \mathbf{1}\left[\sigma_{s,i} > n+k-s\right]\right] = 0 \tag{39}$$

so that

$$\begin{aligned}
\mathrm{cov}[b_n^{(a)}, b_{n+k}^{(a)}] &= \mathrm{cov}\left[b_n^{(a)}, \sum_{s=1}^{n} \sum_{i=1}^{\xi_s} \mathbf{1}\left[\sigma_{s,i} > n+k-s\right]\right] \\
&= \mathrm{cov}\left[b_n^{(a)}, \sum_{s=1}^{n} \sum_{i=1}^{\xi_s} \mathbf{1}\left[\sigma_{s,i} > n+k-s\right]\right] \quad \text{(by independence)} \\
&= \sum_{r=1}^{n} \sum_{s=1}^{n} \mathrm{cov}\left[\sum_{j=1}^{\xi_r} \mathbf{1}\left[\sigma_{r,j} > n-r\right], \sum_{i=1}^{\xi_s} \mathbf{1}\left[\sigma_{s,i} > n+k-s\right]\right] \\
&= \sum_{s=1}^{n} \mathrm{cov}\left[\sum_{j=1}^{\xi_s} \mathbf{1}\left[\sigma_{s,j} > n-s\right], \sum_{i=1}^{\xi_s} \mathbf{1}\left[\sigma_{s,i} > n+k-s\right]\right] \\
&= \sum_{s=1}^{n} \mathbf{E}\left[\xi_s\right] \mathbf{E}\left[\mathbf{1}\left[\sigma_{s,1} > n-s\right] \cdot \mathbf{1}\left[\sigma_{s,1} > n+k-s\right]\right] \\
&\quad + \sum_{s=1}^{n} \left(\mathbf{E}\left[\xi_s(\xi_s-1)\right] - \mathbf{E}\left[\xi_s\right]^2\right) \mathbf{P}\left[\sigma_{s,1} > n-s\right] \mathbf{P}\left[\sigma_{s,1} > n+k-s\right]
\end{aligned} \tag{40}$$

where the last equality follows by Lemma 1. Making use of the fact that the *i.i.d.* rvs $\{\xi_{n+1},\ n = 0, 1, \ldots\}$ are Poisson rvs with parameter λ, we see that (40) reduces to

$$\mathrm{cov}[b_n^{(a)}, b_{n+k}^{(a)}] = \lambda \sum_{r=1}^{n} \mathbf{P}\left[\sigma > r+k-1\right]. \tag{41}$$

Next, we consider $\mathrm{cov}[(b_n^{(0)}, b_{n+k}^{(0)}]$. Again, making use of Lemma 1 under the enforced independence assumptions, we conclude that

$$\begin{aligned}
\mathrm{cov}[b_n^{(0)}, b_{n+k}^{(0)}] &= \mathrm{cov}\left[\sum_{i=1}^{b_0} \mathbf{1}\left[\sigma_{0,i} > n\right], \sum_{j=1}^{b_0} \mathbf{1}\left[\sigma_{0,j} > n+k\right]\right] \\
&= \mathbf{E}\left[b_0\right] \mathbf{P}\left[\hat{\sigma} > n+k\right] \\
&\quad + \left(\mathbf{E}\left[b_0(b_0-1)\right] - (\mathbf{E}\left[b_0\right])^2\right) \mathbf{P}\left[\hat{\sigma} > n\right] \mathbf{P}\left[\hat{\sigma} > n+k\right] \\
&= \lambda \mathbf{E}\left[\sigma\right] \mathbf{P}\left[\hat{\sigma} > n+k\right]
\end{aligned} \tag{42}$$

since b_0 is a Poisson rv with mean $\lambda \mathbf{E}\left[\sigma\right]$. Combining (41) and (42), we have

$$\mathrm{cov}[b_n^*, b_{n+k}^*] = \lambda \mathbf{E}\left[\sigma\right] \mathbf{P}\left[\hat{\sigma} > n+k\right] + \lambda \sum_{r=1}^{n} \mathbf{P}\left[\sigma > r+k-1\right]$$

$$
\begin{aligned}
&= \lambda \mathbf{E}[\sigma] \sum_{r=1}^{\infty} \mathbf{P}[\hat{\sigma} = n+k+r] + \lambda \sum_{r=1}^{n} \mathbf{P}[\sigma > r+k-1] \\
&= \lambda \sum_{r=1}^{\infty} \mathbf{P}[\sigma \geq n+k+r] + \lambda \sum_{r=1}^{n} \mathbf{P}[\sigma \geq r+k] \\
&= \lambda \sum_{r=1}^{\infty} \mathbf{P}[\sigma \geq k+r] \\
&= \lambda \sum_{r=1}^{\infty} \mathbf{P}\left[(\sigma - k)^{+} \geq r\right] \qquad (43)
\end{aligned}
$$

and the proof of (6) is now completed.

References

[1] A. Adas. Traffic models in broadband networks. *IEEE Communications Magazine*, 35(7):82–89, July 1997.

[2] 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.

[3] D. R. Cox. Long-range dependence: A review. In H. A. David and H. T. David, editors, *Statistics: An Appraisal*, pages 55–74. The Iowa State University Press, Ames, Iowa, 1984.

[4] N. G. Duffield and N. O'Connell. Large deviations and overflow probabilities for the general single server queue, with applications. In *Proc. of the Cambridge Philosophical Society*, pages 363–374, 1995.

[5] J. Esary, F. Proschan, and D. Walkup. Association of random variables, with applications. *Annals of Mathematical Statistics*, 38:1466–1474, 1967.

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

[7] 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.

[8] V. S. Frost and B. Melamed. Traffic modeling for telecommunications networks. *IEEE Communications Magazine*, 32(3):70–81, Mar. 1994.

[9] 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.

[10] 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.

[11] J. Haslett and A. E. Raftery. Space-time modeling with long-memory dependence: assessing ireland's wind power resource (with discussion). *Applied Statistics*, 38:1–50, 1989.

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] M. Krunz and S. K. Tripathi. On the characterization of of VBR MPEG streams. In *Proceedings of SIGMETRICS '97 Conference)*, pages 192–202, June 1997.

[19] 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.

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

[21] Z. Liu, P. Nain, D. Towsley, and Z.-L. Zhang. Asymptotic behavior of a multiplexer fed by a long-range dependent process. *Journal of Applied Probability*, 1998. (To appear).

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

[23] H. Michiel and K. Laevens. Teletraffic engineering in a broadband era. *Proceedings of the IEEE*, 85:2007–2033, 1997.

[24] M. Parulekar. *Buffer Engineering for Self-Similar Traffic.* PhD thesis, University of Maryland, College Park, Aug. 1998.

[25] M. Parulekar and A. Makowski. M/G/∞ input processes: A versatile class of models for network traffic. In *Proceedings of IEEE INFOCOM '97*, pages 1452–1459, Kobe, Japan, Apr. 1997.

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

[27] M. Parulekar and A. M. Makowski. Tail probabilities for M/G/∞ input processes (I): Preliminary asymptotics. *Queueing Systems - Theory & Applications*, 1998. (in press).

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

[29] 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, Stanford University, CA, Aug. 1996.

[30] H. Schwetman. *CSIM User's Guide*, 1991.

[31] M. S. Taqqu, V. Teverovsky, and W. Willinger. Estimators for long-range dependence: An empirical study. *Fractals*, 3(4):785–798, 1995.