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

Marwan Krunz
Department of Electrical & Computer Engineering
University of Arizona
Tucson, AZ 85721
krunz@ece.arizona.edu

Armand Makowski
Department of Electrical Engineering
University of Maryland
College Park, MD 20742
armand@eng.umd.edu

LAST REVISED: May 12th, 1997

Abstract

Statistical evidence suggests that the autocorrelation function of a compressed-video sequence is better captured by $e^{-\beta\sqrt{k}}$ than by $k^{-\beta} = e^{-\beta \log k}$ (long-range dependence) or $e^{-\beta k}$ (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-servers 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 $M/G/\infty$ processes a good candidate for modeling many types of correlated traffic in computer networks. We derive the appropriate G that gives the desired correlation function $e^{-\beta\sqrt{k}}$. Though not Markovian, the model is shown to exhibit short-range dependence. Poisson variates of the $M/G/\infty$ model are appropriately transformed to capture the frame size distribution of a video sequence. Using the performance of a 'real' video stream as a reference, we study via simulations the queueing performance for two models: the $M/G/\infty$ model and the fractional ARIMA model [5] (which exhibits LRD). Our results indicate that the $M/G/\infty$ model is much more accurate in predicting the actual queueing performance. Only $\mathcal{O}(n)$ computations are required to generate an $M/G/\infty$ trace, compared to $\mathcal{O}(n^2)$ for a F-ARIMA trace (n being the trace length). Finally, we reexamine the variance-time test, which is typically used to detect LRD in empirical data. By applying it to synthetic realizations of the $M/G/\infty$ model, we show that this test can give misleading indications on the dependence structure of the examined data.

keywords: traffic modeling, VBR video, $M/G/\infty$ process, fitting, long-range dependence.

1 Introduction

One of the most controversial issues in current teletraffic research is the relevance of traffic correlations to the dimensioning of the network. While researchers agree on the importance of correlations in general, they widely disagree on how much of these correlations should be captured in a traffic model. Earlier traffic models are Markovian in nature, with an autocorrelation function that drops off exponentially. Examples of such models include autoregressive and Markov modulated processes (see [4] for a survey). Markovian models exhibit *short-range dependence* (SRD), where the sum of correlations at all lags is finite (note, however, that a SRD model is not necessarily Markovian). Recently, a number of studies indicated the existence of persistent correlations in several types of traffic, including Ethernet LAN [3, 14], WAN [19], and variable-bite-rate (VBR) video traffic [1, 5]. It was argued that such persistence can be adequately captured by models with *long-range dependence* (LRD), whose autocorrelation functions decay slowly (often as $k^{-\beta}$ with $0 < \beta < 1$), so that the correlations have an infinite sum. While the LRD phenomenon was long observed by researchers in other domains such as hydrolics and economics (see [1] and the references therein), recent indications of its presence in computer traffic spurred considerable debate on its relevance to network analysis. On one side, advocates of LRD argue that such a phenomenon has significant impact on network performance, and thus must be incorporated when modeling the traffic. On the other side, advocates of Markovian modeling, though acknowledging the presence of such a phenomenon, argue that for networks with finite buffers it is sufficient to incorporate the correlations up to some finite lag that is proportional to the buffer size [20, 8, 6].

In this paper, we investigate the modeling of VBR compressed video traffic. The literature abounds in video models with Markovian correlation behavior. Long-range dependence in video traffic was first indicated by Beran et al. [1]. Foremost, the statistical evidence supporting LRD was based on the estimated value of the Hurst parameter, which is between 0.5 and 1 for LRD models. Following that study, a video model for the intra-coded *Star Wars* movie was proposed [5] based on the fractional ARIMA (F-ARIMA) process. Huang et al. [12] provided a hybrid model with a correlation structure that exhibits an exponential behavior at small lags and LRD at large lags. The fitting was applied to an MPEG-compressed movie with the three types of compressed frames (I, P, and B).

Putting aside the issue of the relevance of long-term correlations to the dimensioning of network resources, we revisit the statistical evidence that motivated the introduction of LRD models. While the correlation structure of a real video sequence clearly indicates a non-Markovian behavior, this structure is not well captured by LRD models either. In particular, the form $k^{-\beta}$ significantly underestimates the empirical correlations at small and intermediate lags. We found that the empirical autocorrelation is better characterized by $e^{-\beta\sqrt{k}}$ (k is the frame lag). Such a non-Markovian form, though short-range dependent, can be produced by the so-called $M/G/\infty$ input processes. An $M/G/\infty$ process is a correlated process that models the number of busy servers in a discrete-time $M/G/\infty$ queue. One of its attractive features is that its correlation structure is dictated by the service distribution G. Diverse correlation structures, including those with LRD behavior, can be produced using different service distributions [17] (In fact, the $M/G/\infty$ process was first mentioned by Cox [2] as an example of a process that exhibits LRD, which occurs when G is a Pareto distribution).

With the strong evidence that $e^{-\beta\sqrt{k}}$ is more accurate than $k^{-\beta}$ in capturing the autocorrelations of a real video sequence, we present an $M/G/\infty$-based video model with an appropriate G. The resulting Weibull-like form of G renders the model SRD. To capture the frame size distribution of a real video sequence, the Poisson marginal distribution of the $M/G/\infty$ process is transformed into a hybrid Gamma/Pareto distribution, which was earlier suggested by Garrett and Willinger [5]. The nonlinear transformation is shown to have negligible impact on the original correlation structure.

We then study via simulations the queueing performance of the proposed $M/G/\infty$ model and that of the F-ARIMA model of Garrett and Willinger [5]. Two measures of performance are examined: The cell loss rate due to buffer overflow and the frame error rate. The 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. In addition, the computational complexity for generating synthetic $M/G/\infty$ traces is $\mathcal{O}(n)$, compared to $\mathcal{O}(n^2)$ for the F-ARIMA traces, where n is the trace length. The accuracy and low complexity of the $M/G/\infty$ model makes it quite useful in dimensioning network resources. Finally, we examine the variance-time test which is typically used to detect LRD in an empirical sequence. By applying it to $M/G/\infty$ traces, we show that this test can lead to incorrect conclusions about the dependency structure of the underlying data.

The paper is structured as follows. In Section 2 we examine the appropriateness of LRD to the modeling of video traffic. $M/G/\infty$ input processes are discussed in Section 3. A traffic model for VBR video based on such processes is presented in Section 4. For the purpose of validating the model, simulation results of the queueing performance under both the $M/G/\infty$ and the F-ARIMA models are provided in Section 5. In Section 6 we examine the validity of inference from the variance-time plots. Section 7 concludes the paper.

2 Relevance of Long-Range Dependence to Video Traffic

LRD modeling of VBR video was first suggested by Beran et al. [1], who argued that the non-Markovian behavior observed in many empirical video traces can be captured by LRD models. The statistical tests which were used to infer the existence of the LRD phenomenon are mainly based on the estimated value of the Hurst parameter. Although the examined traces under these tests pass for LRD, we show in Section 6 that such tests can be misleading. And even if video data exhibit asymptotic LRD behavior, capturing the correlations at large lags only, as is the case with LRD models, is not enough as evident by the discrepancy in the queueing performance between a LRD model and a real video stream (Section 5). Before discussing the indications that led us to question the appropriateness of LRD modeling of video, we briefly review the basic notion of LRD and self-similarity. The reader is referred to [2, 1] for more details.

2.1 LRD and Self-Similarity

Consider a second-order stationary process $\{X_n : n = 1, 2, \ldots\}$ with mean $\overline{X}$ and variance v. Let $C_k \triangleq \text{cov}(X_n, X_{n+k}) = E\left[(X_n - \overline{X})(X_{n+k} - \overline{X})\right]$. The autocorrelation function is given by $\rho_k \triangleq C_k/v$, for $k = 0, 1, \ldots$, which can equivalently be represented by its power spectral density $g(\omega) \triangleq (1/2\pi)\sum_{k=-\infty}^{\infty} \rho_k e^{-ik\omega}$. For $m = 1, 2, \ldots$, let

$$X_n^{(m)} \triangleq \frac{\sum_{i=nm-m+1}^{nm} X_i}{m}, \quad n = 1, 2, \ldots \tag{1}$$

so that $\{X_n^{(m)}\}$ is an averaged version of $\{X_n\}$, with the averaging being taken over non-overlapping blocks of length m. The variance of the new time series is given by:

$$v_m \triangleq \text{var}(X_n^{(m)}) = \frac{v}{m} + \frac{2}{m^2}\sum_{p=1}^{m-1}\sum_{q=1}^{p} C_q \tag{2}$$

According to Cox [2], the process $\{X_n\}$ is LRD if it satisfies any of the following, virtually equivalent, conditions: (i) $\sum_{k=0}^{\infty} \rho_k = \infty$, (ii) $g(w) \to \infty$ as $w \to 0$, and (iii) $mv_m \to \infty$ as $m \to \infty$. In contrast, if $\sum_{k=0}^{\infty} \rho_k$ is finite, $g(0)$ is finite, or $lim_{j\to\infty} mv_m =$ constant, then $\{X_n\}$ is short-range dependent. Various forms of LRD are possible according to the above definition. One particular form which received much attention is when:

$$\rho_k \sim ck^{-\beta} \text{ as } k \to \infty, \tag{3}$$

where $0 < \beta < 1$ and c is a constant. An equivalent statement to (3) is that:

$$\lim_{m\to\infty} \rho_k^{(m)} = (1/2)\delta^2(k^{2-\beta}) \approx \acute{c}k^{-\beta} \text{ (for large } k) \tag{4}$$

where $\rho_k^{(m)}$ is the autocorrelation in $\{X_n^{(m)}\}$ at lag k, $\delta^2(.)$ is the central second difference operator, and $\acute{c}$ is a constant. A process that satisfies (4) is said to exhibit *asymptotic second-order self-similarity* [2]. When (4) is satisfied for all k, the process is said to exhibit *exact second-order self-similarity*, since in this case, $\rho_k^{(m)} = \rho_k$ for all k and m. For SRD processes, $\rho_k^{(m)} \to 0$ as $m \to \infty$, for all k, i.e., the process tends to white noise.

2.2 How Appropriate is LRD Modeling of Video Traffic?

While a model is expected to capture certain aspects of the underlying physical phenomenon, its goodness is ultimately determined according to its ability to achieve the goal it is designed for. In traffic modeling, the goal of a model is to predict the network performance, and subsequently be used in dimensioning network resources. Thus, the queueing performance is the crucial factor that determines the appropriateness of a traffic model. However, because of the significant impact of traffic correlations on the queueing behavior, one could obtain a first evaluation of a model by examining its correlation structure. We provide two preliminary indications from the correlation structure of a video sequence that point to the inappropriateness of LRD modeling. Such indications are confirmed after examining the queueing performance. For brevity, we mainly present the results from the fitting of one real trace, namely the intra-coded version of the *Star Wars* movie (see [5]). This trace consists of 171,000 frames, and is the longest trace available in public domain. We also examined other shorter MPEG traces (with I, P, and B frames being modeled separately), and reached similar conclusions. Throughout the paper, a trace (or a sequence) refers to the sequence of frames sizes of a real or a synthetic VBR stream, given in 48-byte ATM cells.

First, we examine the autocorrelations in a video sequence. Correlations, after all, were the motivation for considering LRD models in the first place. The autocorrelation functions (ACFs) of the *Star Wars* trace and of three fits are shown in Figure 1. The fits are: (1) $e^{-\beta k}$ (Markovian), (2) $k^{-\beta}$ (LRD), and (3) $e^{-\beta\sqrt{k}}$ ($M/G/\infty$ model). For fits (1) and (3), β is obtained by usual least-square fitting, which resulted in $\beta \approx 0.0075$ and $\beta \approx 0.076$, respectively. For the LRD fit, $\beta = 0.4$ is obtained from the estimated value of the Hurst parameter ($H = 1 - \beta/2 \approx 0.8$), which was reported in [5]. 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, $e^{-\beta\sqrt{k}}$ appears to be a much better fit than the other two. Note that using a larger value for H will 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. Similar observations can be made from the fitting of another shorter trace (Figure 2), which was generated by an MPEG-2 encoder. This trace consists of 12600 frames from *The Wizard of Oz* movie (see [13]). In this case, least-square fitting resulted in $\beta \approx 0.003$ for the first fit, and $\beta \approx 0.055$ for the third fit. For the $k^{-\beta}$ fit, $\beta \approx 0.238$ was estimated from the asymptotic slope of the variance-time plot. Once again, the real ACF is best captured by $e^{-\beta\sqrt{k}}$.

Next, we examine the self-similarity in video data. An exact second-order self-similar process must satisfy $\rho_k^{(m)} = \rho_k$ for all k and m. Of course, empirical traces can only be expected to exhibit asymptotic self-similarity, if any. The ACFs of the aggregated sequences of the *Star Wars* trace are shown in Figure 3 using blocks of sizes 1, 10, and 100. It is obvious that the correlation structure varies with the block size, tending to that of white noise as block size increases. No indication of self-similarity is observed. Note, however, that as m increases, the ACFs of the aggregated sequence tend to look similar to each other, but also similar to the ACF of white noise. This should not be interpreted as second-order self-similarity since $\{\rho_k^{(m)}\}$ behaves as such even when the underlying process is SRD.

Both indications raise doubts on whether LRD is an appropriate technique for modeling video traffic. The statistical tests that led to suggesting the presence of LRD in video data (for example, the variance-time plots) must also be reexamined. We come back to this issue in Section 6.

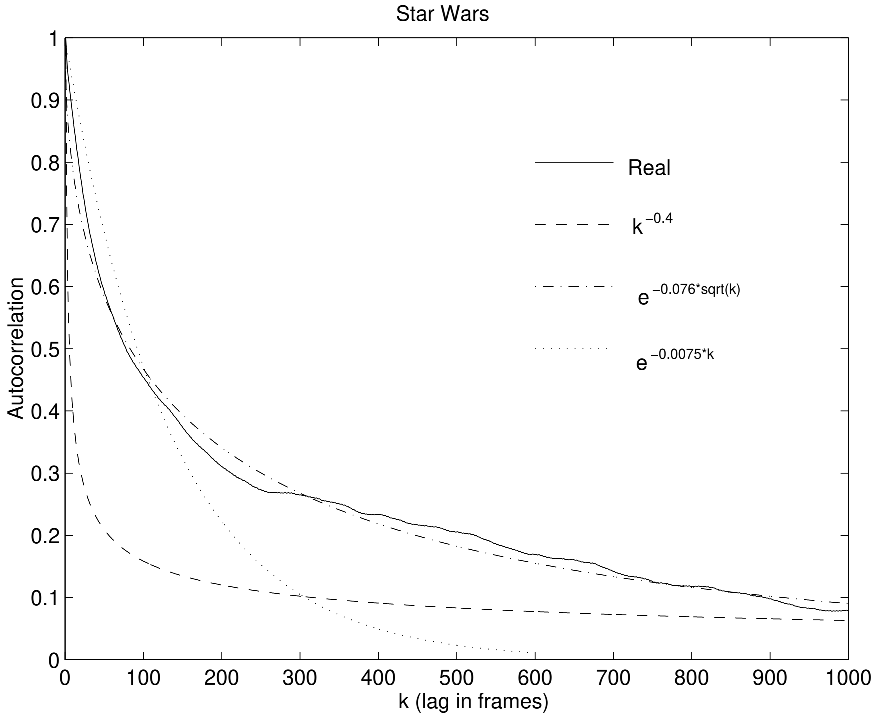

Figure 1: Correlation structure of an intra-coded video sequence.

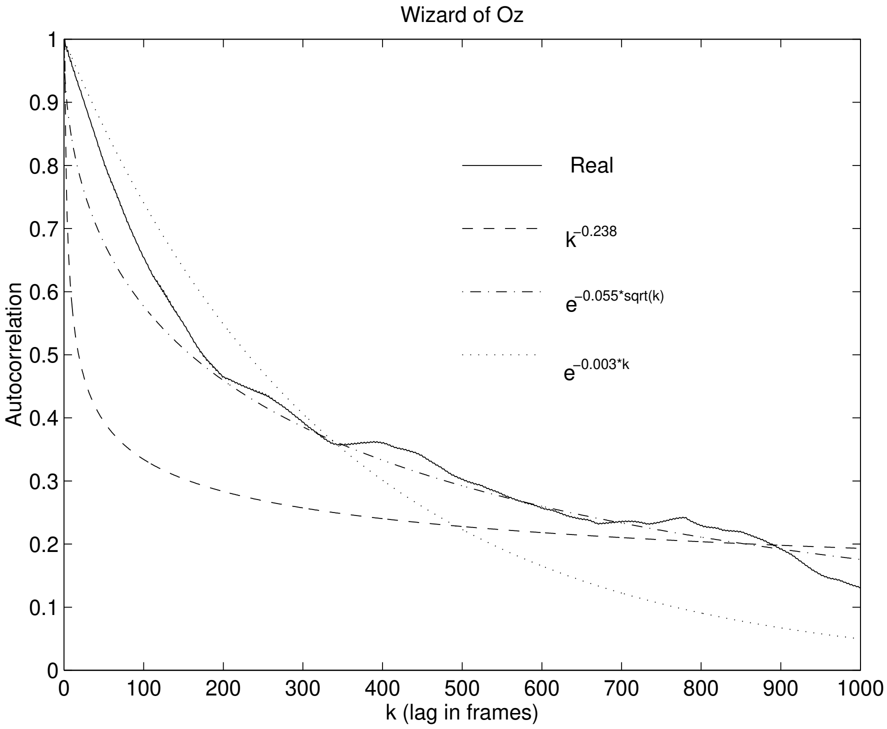

Figure 2: Correlation structure of an MPEG-2 coded video sequence.

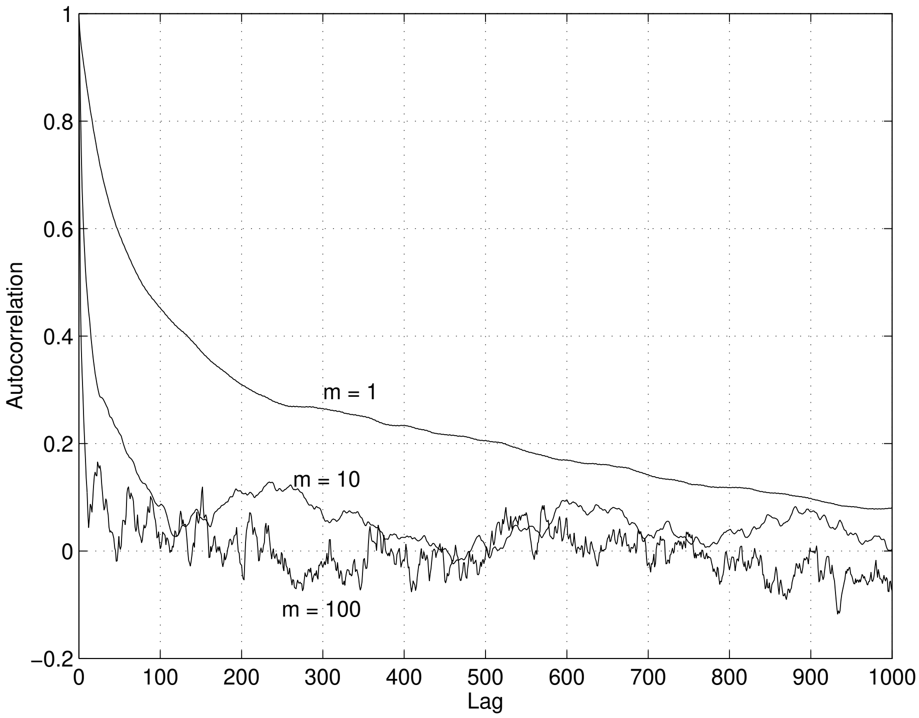

Figure 3: ACFs of aggregated versions of *Star Wars* sequence (block size $= 1, 10$, and 100).

3 $M/G/\infty$ Input Processes

As indicated in Figures 1 and 2, the ACF of a video sequence is adequately captured by $e^{-\beta\sqrt{k}}$. Neither Markovian nor LRD models exhibit such a correlation structure. A model with ACF of the form $e^{-\beta\sqrt{k}}$ can be constructed using $M/G/\infty$ input processes. This versatile class of processes is capable of displaying many forms of correlations including LRD. Our interest here is in the use of $M/G/\infty$ processes as a traffic model for video. In this section, we summarize some of the important properties of $M/G/\infty$ processes, as they relate to our modeling effort. Further details can be found in [18, 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 $\sigma_{n+1,1}, \ldots, \sigma_{n+1,\xi_{n+1}}$ be the integer-valued service times for customers $1, 2, \ldots, \xi_{n+1}$ 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 $\sigma_{0,1}, \ldots, \sigma_{0,b_{0^-}}$, which are mutually independent.

Let b_n be the number of busy servers (i.e., remaining customers) at time n^+, $n = 0, 1, \ldots$, (after counting arrivals and departures at the start of slot $[n, n+1)$). The process $\{b_n : n = 0, 1, \ldots\}$ is known as the $M/G/\infty$ input process. Let $b^{(n)}$ be the $(\mathbb{N}+1)-$valued vector $(b_0, b_1, \ldots, b_n)$. The fact that $\{b_n\}$ exhibits some form of dependency 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 $f, g : \mathbb{N}^{n+1} \to \mathbb{R}$,*

$$E[f(b^n)g(b^n)] \geq E[f(b^n)]\,E[g(b^n)] \tag{5}$$

provided the expectations exist and are finite.

From (5), it is easy to see that:

$$\mathrm{cov}[b_n, b_{n+k}] \geq 0, \quad \forall n, k = 0, 1, \ldots \tag{6}$$

It has been shown that $\{b_n\}$ can exhibit various forms of positive autocorrelations, the extent of which is controlled by the tail behavior of G [17]. In fact, it can even exhibit LRD when G is a Pareto distribution [2]. Likhanov et al. [15] showed that the superposition of a large number of On/Off sources with Pareto distributed activity periods tends in the limit to the $M/G/\infty$ model. In [19], the $M/G/\infty$ model with G lognormal was found appropriate for some WAN traffic.

To use the $M/G/\infty$ process in modeling correlated traffic, it is helpful to establish the relationship between the correlation structure of this process and the distribution G. Then, G can be appropriately chosen to produce a given correlation structure. In general, $\{b_n\}$ is not stationary, but it admits a stationary and ergodic version $\{b_n^*\}$ [16]:

Proposition 2 *There exists a stationary and ergodic process $\{b_n^* : n = 0, 1, \ldots\}$ such that*

$$\{b_{n+k},\ n = 0, 1, \ldots\} \xrightarrow{d} \{b_n^*,\ n = 0, 1, \ldots\} \quad \textit{as } k \to \infty \tag{7}$$

for any initial distribution of the rv b_{0^-} and any corresponding initial service times. Moreover, for $n = 0, 1, \ldots$, b_n^* is a Poisson rv with parameter $\lambda E[\sigma]$.

We use the process $\{b_n^*\}$ as the basis for our video source model. The covariance structure for such a process is given by the following proposition, the proof of which is provided in the appendix.

Proposition 3 *The covariance structure for the process* $\{b_n^*\}$ *is given by*

$$\Gamma(k) \triangleq \mathrm{cov}[b_n^*, b_{n+k}^*] = \lambda E\left[(\sigma - k)^+\right], \quad \textit{for all } n, k = 0, 1, \ldots \tag{8}$$

where $(x)^+ = x$ *if* $x \geq 0$, *and* 0 *otherwise.*

Although Equation (8) relates the covariance structure to the service time, it does not provide a recipe for selecting an appropriate G that results in a desired correlation structure. A more explicit relationship is needed. Let $\widetilde{\sigma}$ be the forward recurrence associated with the service time σ, and let $f_{\widetilde{\sigma}}$ be its *pmf*:

$$f_{\widetilde{\sigma}}(i) \triangleq \mathrm{P}\left[\widetilde{\sigma} = i\right] = \frac{\mathrm{P}\left[\sigma \geq i\right]}{\mathrm{E}\left[\sigma\right]}, \quad i = 1, 2, \ldots \tag{9}$$

Define the sequence:

$$u_n \triangleq -\ln \mathrm{P}\left[\widetilde{\sigma} > n\right], \quad n = 0, 1, \ldots \tag{10}$$

The covariance structure of $\{b_n^*\}$ can be written as follows: For a fixed k, we have

$$\begin{aligned}
\Gamma(k) &= \lambda \mathrm{E}\left[(\sigma - k)^+\right] = \lambda \sum_{i=0}^{\infty} \mathrm{P}\left[(\sigma - k)^+ > i\right] \\
&= \lambda \sum_{i=0}^{\infty} \mathrm{P}\left[(\sigma > k + i\right] = \lambda \sum_{i=k+1}^{\infty} \mathrm{P}\left[\sigma \geq i\right] \\
&= \lambda \mathrm{E}\left[\sigma\right] \sum_{i=k+1}^{\infty} \mathrm{P}\left[\widetilde{\sigma} = i\right] = \lambda \mathrm{E}\left[\sigma\right] \mathrm{P}\left[\widetilde{\sigma} > k\right] \\
&= \lambda \mathrm{E}\left[\sigma\right] e^{-u_k}
\end{aligned} \tag{11}$$

Equation 11 relates the monotonic behavior of the ACF to the service distribution G. As indicated in [17], the sequence $\{u_k,\ k = 0, 1, \ldots\}$ plays a major role in the determination of the buffer asymptotics for an $M/G/\infty$ input traffic. Using large deviation results, it was shown [17] that $\{u_k\}$ is the appropriate *scaling sequence* needed to obtain large deviation estimates of the buffer asymptotics for an infinite single server queue with deterministic service time and $M/G/\infty$ input traffic. Using (11), it can be shown [17] that

$$\sum_{k=0}^{\infty} \Gamma(k) = \lambda \mathrm{E}\left[\sigma\right] \mathrm{E}\left[\widetilde{\sigma}\right] = \frac{\lambda}{2} \mathrm{E}\left[\sigma(\sigma + 1)\right], \tag{12}$$

so the process $\{b_n^*\}$ is LRD if and only if $\mathrm{E}\left[\sigma^2\right]$ is infinite, as when G is a Pareto distribution.

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

A common practice in traffic studies is to construct a model that captures the first and second-order statistics of the empirical sequence, including the mean, the variance, and the correlation structure. The parameters of the $M/G/\infty$ process that can be used in the fitting are G and λ. While G can be chosen to provide a given autocorrelation structure (using (11)), λ can only be fitted to one moment (the mean or the variance). To provide a fitting for the complete marginal distribution (including the mean and variance) as well as the correlation structure, we first choose G of the $M/G/\infty$ model that results in Poisson variates with the autocorrelation structure $e^{-\beta\sqrt{k}}$. Then, we apply an appropriate transformation on these variates that captures the empirical distribution of the data.

4.1 Modeling the Correlation Structure

For a correlation structure of the form $e^{-\beta\sqrt{k}}$, the corresponding service distribution can be obtained using (11). From (9) and (10), we have:

$$\begin{aligned} \mathrm{P}[\sigma \geq k] &= \mathrm{E}[\sigma]\,\mathrm{P}[\tilde{\sigma} = k], \quad k = 1, 2, \ldots \\ &= \mathrm{E}[\sigma]\left(\mathrm{P}[\tilde{\sigma} > k-1] - \mathrm{P}[\tilde{\sigma} > k]\right) \\ &= \mathrm{E}[\sigma]\left(e^{-u_{k-1}} - e^{-u_k}\right) = \mathrm{E}[\sigma]\left(\rho_{k-1} - \rho_k\right) \end{aligned} \tag{13}$$

Thus,

$$\mathrm{P}[\sigma = k] = \mathrm{E}[\sigma]\left(\rho_{k-1} + \rho_{k+1} - 2\rho_k\right), \quad k = 1, 2, \ldots \tag{14}$$

which gives the *pmf* of σ in terms of the ACF. One could suggest using the empirical ACF instead of the fit $e^{-\beta\sqrt{k}}$. However, the sample autocorrelations are not guaranteed to be always monotone, which is needed to ensure that $\mathrm{P}[\sigma \geq k]$ in (13) is positive for $k = 1, 2, \ldots$. By setting $k = 1$ in (13), we obtain the mean service time:

$$\mathrm{E}[\sigma] = \frac{\mathrm{P}[\sigma \geq 1]}{\rho_0 - \rho_1} = \frac{1}{1 - e^{-\beta}} \tag{15}$$

For *Star Wars*, $\beta \approx 0.076$. Thus, $\mathrm{E}[\sigma] = 13.65$.

It can be easily shown that $\mathrm{P}[\sigma = k] > 0$ for $k = 1, 2, \ldots$. This follows from the fact that $e^{-\beta\sqrt{t}}$ (in the continuous domain) is concaved up. Thus,

$$\left(e^{-\beta\sqrt{k-1}} - e^{-\beta\sqrt{k}}\right) > \left(e^{-\beta\sqrt{k}} - e^{-\beta\sqrt{k+1}}\right) \quad \text{for } k = 1, 2, \ldots \tag{16}$$

and the result follows immediately from (14).

With a correlation structure of the form $e^{-\beta\sqrt{k}}$, *the $M/G/\infty$ input model is short-range dependent, since*

$$\sum_{k=0}^{\infty} \rho_k = 1 + \sum_{k=1}^{\infty} \rho_k \leq 1 + \int_0^{\infty} e^{-\beta\sqrt{t}}\,dt = 1 + \frac{2}{\beta^2} < \infty \tag{17}$$

Short-range dependence can also be shown by the asymptotic behavior of the variance of the aggregated sequences. From (2), we have:

$$\begin{aligned} v_m &= \frac{v}{m} + \frac{2v}{m^2}\sum_{p=1}^{m-1}\sum_{q=1}^{p}\rho_q \quad \text{for } m = 1, 2, \ldots \\ &= \frac{v}{m} + \frac{2v}{m^2}\sum_{p=1}^{m}(m-p)\rho_p = \frac{v}{m} + \frac{2v}{m^2} g_m \end{aligned} \tag{18}$$

where $g_m \triangleq \sum_{p=1}^{m}(m-p)\rho_p$. We only need to show that $g_m \sim \mathcal{O}(m)$ as $m \to \infty$.

$$\begin{aligned} g_m &= \sum_{j=1}^{m}(m-j)\rho_j = m\sum_{j=1}^{m}\rho_j - \sum_{j=1}^{m} j\rho_j \\ &= m\left(\sum_{j=1}^{\infty}\rho_j - \sum_{j=m+1}^{\infty}\rho_j\right) - \left(\sum_{j=1}^{\infty} j\rho_j - \sum_{j=m+1}^{\infty} j\rho_j\right) \end{aligned} \tag{19}$$

As $m \to \infty$, $\sum_{j=m+1}^{\infty}\rho_j \to 0$. Also, since $j\rho_j = je^{-\beta\sqrt{j}}$ is *asymptotically* monotonically decreasing, $\sum_{j=m+1}^{\infty} j\rho_j \to 0$ as $m \to \infty$. Thus,

$$g_m \sim m\sum_{j=1}^{\infty}\rho_j - \sum_{j=1}^{\infty} j\rho_j \tag{20}$$

for large m. We showed earlier that $\sum \rho_j$ is convergent. We only need to check $\sum j\rho_j$. In the continuous domain, $te^{-\beta\sqrt{t}}$ attains its maximum at $t^* = 4/\beta^2$, and decreases monotonically after that. Let

$$A \triangleq \sum_{j=1}^{\lceil t^* \rceil} j\rho_j < \infty \tag{21}$$

Then,

$$\sum_{j=1}^{\infty} j\rho_j = A + \sum_{j=\lceil t^* \rceil+1}^{\infty} j\rho_j < A + \int_{\lceil t^* \rceil}^{\infty} te^{-\beta\sqrt{t}}dt < A + \int_{0}^{\infty} te^{-\beta\sqrt{t}}dt \tag{22}$$

But $\int_0^{\infty} te^{-\beta\sqrt{t}}dt = 12/\beta^4$. Thus, $\sum j\rho_j$ is convergent, and $v_m \sim \mathcal{O}(1/m)$ as $m \to \infty$. Thus, the model is SRD.

4.2 Modeling the Marginal Distribution

The $M/G/\infty$ model produces correlated variates with Poisson marginal distribution. However, the tail of a Poisson CDF is quite different from the tail of the empirical distribution of a real video sequence. This is illustrated in Figure 4 where the $M/G/\infty$ fit is obtained by matching the means (sample mean $= \lambda E[\sigma]$), and setting λ accordingly. Several theoretical fits have been suggested for the frame-size distribution, including Gamma [10], lognormal [13, 9], and hybrid Gamma/Pareto distributions [5]. The last fit was found quite appropriate for *Star Wars* data. Accordingly, we use it to characterize the frame-size distribution. As explained in [5], 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 be the cumulative probability functions for Gamma and Pareto distributions, respectively. F_P has no closed-form expression. Its derivative is given by

$$f_\Gamma(x) = e^{-\lambda x} \frac{\lambda(\lambda x)^{s-1}}{\Gamma(s)} \tag{23}$$

where $\Gamma(.)$ is the Gamma function. The Pareto distribution has the explicit form: $F_P(x) = 1 - (\acute{x}/x)^a$, for $x \geq \acute{x}$; and zero, otherwise. Both $\acute{x}$ and a are determined from the fitting.The hybrid Gamma/Pareto distribution is given by:

$$F_{\Gamma/P}(x) = \begin{cases} F_\Gamma(x), & \text{if } x \leq x^* \\ F_P(x), & \text{if } x > x^* \end{cases} \tag{24}$$

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 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 (see Figure 4). Using $F_\Gamma(x^*) = F_P(x^*)$ along with least-square fitting of the Pareto tail, estimates of $\acute{x}$ and a can be obtained. Table 1 gives the estimated parameters for *Star Wars* trace (with frame sizes in 48-byte cells).

Sample Mean (cells)	Std. Deviation (cells)	λ (1/cells)	s	x^* (cells)	$\acute{x}$ (cells)	a
579.5	130.3	3.41×10^{-2}	19.78	650	576	10.662

Table 1: Estimated values of various parameters in the hybrid Gamma/Pareto model.

The Poisson variates of the $M/G/\infty$ process can now be transformed into Gamma/Pareto variates. Let $\{X_n\}$ be a realization of the $M/G/\infty$ process with $\lambda = 1$. Then, $\{X_n\}$ is transformed into a new sequence $\{Y_n\}$ using:

$$Y_n = F_{\Gamma/P}^{-1}(F_O(X_n)) \quad \text{for } n = 1, 2, \ldots \tag{25}$$

where F_O is a Poisson CDF with parameter $\lambda E[\sigma] = E[\sigma]$, and

$$F_{\Gamma/P}^{-1}(y) = \begin{cases} F_P^{-1}(y) = k/(1-y)^{1/a}, & \text{if } y > 1 - (k/x^*)^a \\ F_\Gamma^{-1}(y), & \text{otherwise} \end{cases} \tag{26}$$

with F_Γ^{-1} obtained numerically.

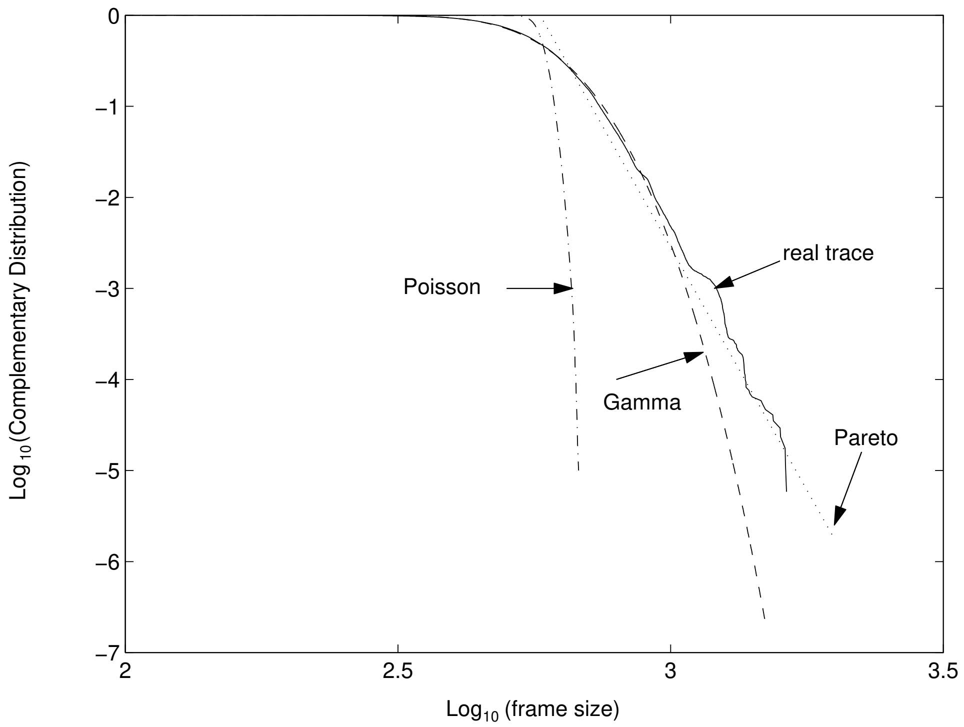

Figure 4: Complementary frame-size distribution of a real video stream and of several fits.

When applying the above transformation, a legitimate concern arises naturally regarding the impact of this transformation on the original correlation structure. In general, a nonlinear transformation is expected to affect the correlation structure. However, we observed that this effect is quite negligible. 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 5. The average ACF is almost indistinguishable from the theoretical ACF of the non-transformed $M/G/\infty$ process.

5 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 model. The queueing system consists of a single-server FIFO queue with capacity B (in cells) and constant service rate C. First, we consider the performance for a single stream (i.e., no multiplexing), since discrepancies in the buffer overflow behavior among different models are most apparent in the single-stream case. We use the queueing performance for the original *Star Wars* stream as a reference for contrasting both models. In all experiments, we assume that cells in each frame are evenly distributed over the frame duration (e.g., 1/30 sec). 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.

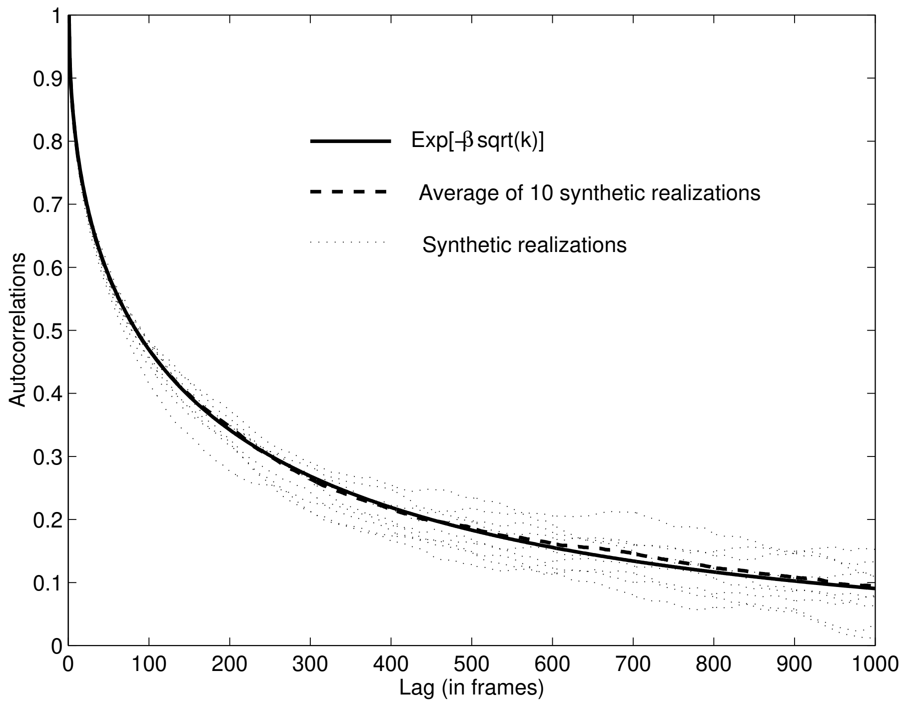

Figure 5: ACFs of ten $M/G/\infty$ realizations after transformation, along with their sample average.

5.1 Generating Synthetic Traces

Synthetic realizations of the $M/G/\infty$ and F-ARIMA models are used to simulate the queueing performance. 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. Such long traces are needed to obtain meaningful results under small cell loss probabilities. In fact, we first tried using shorter traces of length 100,000, and found that at realistic loss rates, losses occur in only few frames (for example, in a given experiment with $U = 60\%$ and $B = 2500$, a loss rate of 8.3E–6 (484 cells) came from only five errored frames). 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, in order to display the maximum value of the *Star Wars* trace, an $M/G/\infty$ realization (before transformation) must include a point with value 33. For a sequence of N positively correlated rvs, $X_1, \ldots, X_N$, with common CDF $F_X(x)$, it is easy to see that

$$\begin{aligned}
\Pr[\max_i X_i > x] &= 1 - \prod_{i=2}^{N} \Pr\left[X_i < x / X_{i-1} < x, \ldots, X_1 < x\right] \Pr\left[X_1 < x\right] \\
&< 1 - \prod_{i=1}^{N} \Pr\left[X_i < x\right] = 1 - \left(F_X(x)\right)^N
\end{aligned}$$

Thus, for $N = 100,000$ and a Poisson distribution, $\Pr[\max X_i > 32] < 1 - (F_X(32))^{100000} = 0.4745$ (less than 50% chance that the 100,000-long realization reaches the real maximum frame size).

Note that 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. It requires $\mathcal{O}(n^2)$ computations to generate a F-ARIMA trace (before transformation) using Hosking algorithm [11], where n is the trace length. In contrast, an $M/G/\infty$ trace requires only $\mathcal{O}(n)$ computations. Even to generate a 100,000-long F-ARIMA trace, Hosking algorithm requires about 3 days of execution on a Sparc-10 workstation. Instead, we used an approximation due to Haslett and Raftery [7]. With this approximation, it takes about 2 days to obtain a F-ARIMA trace of 500,000 points. In contrast, it takes less than a minute to obtain an $M/G/\infty$ trace of 1,000,000 points.

An example of the sample ACF of a synthetic F-ARIMA realization (before and after transformation) is shown in Figure 6. It should be noted that the theoretical ACF of a F-ARIMA process, which is given by

$$\rho_k = \frac{d(1+d)\cdots(k-1+d)}{(1-d)(2-d)\cdots(k-d)}, \quad k = 1, 2, \ldots, \text{ where } 0 < d < 0.5 \tag{27}$$

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 $k^{-\beta}$. As with the $M/G/\infty$ model, transforming the normal variates of a F-ARIMA realization into Gamma/Pareto variates is not observed to have any noticeable impact on the correlations.

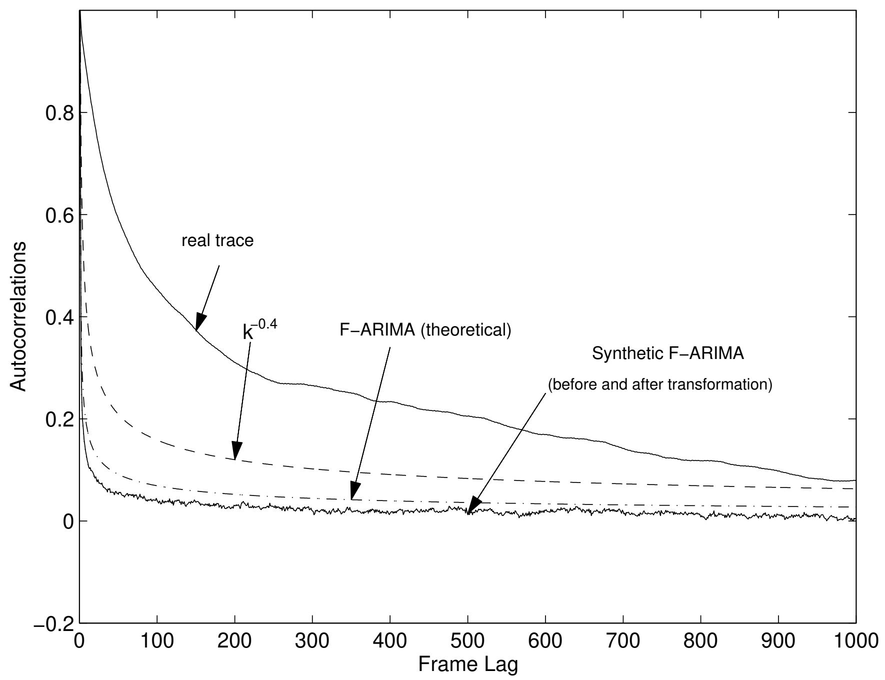

Figure 6: ACFs of a F-ARIMA process and of synthetic F-ARIMA traces.

A Csim program 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 cells is generated according to a Poisson distribution with $\lambda = 1$. Each arriving cell is kept for a random time σ whose *pmf* is given by (14). A synthetic $M/G/\infty$ trace is obtained from the number of remaining cells at the beginning of each time slot. This trace is then transformed into one with a Gamma/Pareto marginal distribution.

A summary statistic of the VBR sequence in the real stream and in both models is given in Table 2. For each model, the statistics are averaged over ten independent realizations. The 95% confidence intervals for the last three statistics are given in parenthesis as percentages of their average values. For the mean and standard deviation, the confidence intervals (not shown) are sufficiently tight. Both models provide acceptable approximation of the first-order statistics, with the standard deviation of the F-ARIMA traces being closer to the real one. Also, the $M/G/\infty$ traces slightly underestimate the tail behavior, whereas the F-ARIMA traces overestimate it. As shown in the next section, these differences do not have a noticeable impact on the queueing behavior.

Video Source	Frame-Size Statistics (in cells)				
	Mean	Std. Deviation	% frames > 900	% frames > 1200	% frames > 1500
Real data	579.5	130.3	1.66E–2	1.12E–3	4.68E–5
$M/G/\infty$ model	590.4	121.5	1.34E–2 (± 8%)	4.11E–4 (± 20%)	2.69E–5 (± 42%)
F-ARIMA model	573.6	132.2	1.40E–2 (± 6%)	9.92E–4 (± 7%)	1.22E–4 (± 16%)

Table 2: Summary statistic of frames sizes in real and synthetic video data. The statistics for each of the two models are based on 10 independent realizations.

5.2 Queueing Performance for a Single Stream

Simulations of a finite-capacity FIFO queue were conducted using synthetic and real data (the *Star Wars* trace). The cell loss and frame error rates are examined at three loads: $U = 80\%$, 60%, and 40%. A summary of the simulations results (to two significant digits) is given in Table 3. For the $M/G/\infty$ and F-ARIMA models, the results are based on the average of ten independent runs. For values above 1.0E–5, the confidence intervals (not shown) are within 25% of the mean values.

For $U = 80\%$ and 60%, the buffer size is varied from 100 to 2500 cells. As expected, the cell loss and frame error rates of 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 cell loss performance by about an order of magnitude. The buffer size seems to have a bigger impact on the frame error rate than on the cell loss rate. For both $U = 80\%$ and $U = 60\%$, the frame error rate of the real stream decreases by about 50% when B is increased from 100 to 2500 cells. *Comparing the performance of the two models to that of the real stream, it is clear that the $M/G/\infty$ model provides much accurate predictions of both cell loss and frame error rates.* 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 cell loss and frame error rates 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.

Interestingly, the $M/G/\infty$ model is quite accurate at extremely small loss rates, 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 cell loss and frame error rates, and eventually underestimates them as B increases to 1000 cells.

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 real trace exhibit stronger statistical correlations than the rest of the trace. We speculate that both ends correspond to the compressed frames in the credits (the portion that contains the names of actors, acknowledgements, etc.). By filtering the first and last 1000 frames of the trace, the "nonstationarity" in the real data was significantly reduced, and consequently, the queueing performance was observed to be much closer to that of the $M/G/\infty$ model.

5.3 Queueing Performance for Multiplexed Streams

Despite the fact that video models in general do not lend themselves to queueing analysis, they provide a practical means to generate independent but structurally similar 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. Our goal here is to contrast the behavior of the two models, and not to investigate the multiplexing gain per se. 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, which becomes almost independent of the number of multiplexed streams. When simulating the performance for N multiplexed streams, we first obtain an *aggregate* trace from the pointwise sum of the N traces. If during a time slot, one can show that buffer overflow is impossible, then the aggregate trace can be used to update the buffer content at the end of that slot. Otherwise, the frame sizes in the individual traces are used to simulate the performance on a cell-by-cell basis. Fortunately, buffer overflow occurs during only a small percentage of time slots. Let $\{\overline{X}_j\}$ be the aggregate trace, C the service rate, and Q_k the queue length just before the start of the kth slot. It can be easily shown that each of the following two conditions guarantees that the buffer does not overflow during the kth slot: (1) $\overline{X}_k \leq C$ and $Q_k \leq B - N$; and (2) $\overline{X}_k > C$, $\overline{X}_k - C < B - Q_k$, and $Q_k \leq B - N$. The proof is simple and is skipped. With this optimization, the simulation time is $\mathcal{O}(n + \alpha nWN)$, 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 during buffer overflow. Typically, $\alpha W \ll 1$, making the complexity much less than $\mathcal{O}(Nn)$.

Average cell loss and frame error rates under both models are given in Table 4 for $N = 5$ and

Buffer Size (cells)	Cell Loss Rate			Frame Error Rate		
	Real	$M/G/\infty$	F-ARIMA	Real	$M/G/\infty$	F-ARIMA
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

(a) $U = 80\%$

Buffer Size (cells)	Cell Loss Rate			Frame Error Rate		
	Real	$M/G/\infty$	F-ARIMA	Real	$M/G/\infty$	F-ARIMA
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

(b) $U = 60\%$

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

(c) $U = 40\%$

Table 3: Average cell loss and frame error rates at three different loads. Ten independent replications are used to obtain the values for $M/G/\infty$ and F-ARIMA models.

$N = 10$ at 80% load. The depicted values were averaged over five independent runs, where each run provided the mean cell loss and frame error rates of N multiplexed streams. The 95% confidence intervals are shown in parenthesis as a percentage of the average. To give an idea about the efficiency of the simulation approach, we report the average simulation times for the $M/G/\infty$ traces (1,000,000 points each). When $N = 5$, it took from 120 min (with $B = 200$) to 75 min (with $B = 800$) to execute one simulation using an Alpha workstation. When $N = 10$ the simulation time was further reduced to 20 min and 13 min for $B = 100$ and $B = 400$, respectively. Intuitively, the simulation time decreases with B. From Table 4 it can be observed that, as in the case of $N = 1$, the performance of F-ARIMA model is more sensitive to the buffer size than the $M/G/\infty$ model.

Buffer	$M/G/\infty$ Model		F-ARIMA Model	
Size (cells)	Cell Loss Rate	Frame Error Rate	Cell Loss Rate	Frame Error Rate
200	2.3E–4 (± 14%)	3.7E–3 (± 8%)	2.1E–4 (± 7%)	4.8E–3 (± 5%)
400	2.0E–4 (± 16%)	2.9E–3 (± 11%)	6.9E–5 (± 12%)	1.5E–3 (± 11%)
600	1.8E–4 (± 18%)	2.5E–3 (± 13%)	2.6E–5 (± 16%)	5.5E–4 (± 13%)
800	1.7E–4 (± 21%)	2.2E–3 (± 13%)	1.1E–5 (± 23%)	2.2E–4 (± 19%)

(a) $N = 5$

Buffer	$M/G/\infty$ Model		F-ARIMA Model	
Size (cells)	Cell Loss Rate	Frame Error Rate	Cell Loss Rate	Frame Error Rate
100	7.4E–6 (± 30%)	2.2E–4 (± 22%)	2.1E–5 (± 24%)	7.9E–4 (± 17%)
200	6.3E–6 (± 37%)	1.7E–4 (± 30%)	1.3E–5 (± 27%)	4.6E–4 (± 21%)
300	5.6E–6 (± 42%)	1.5E–4 (± 34%)	7.8E–6 (± 36%)	3.0E–4 (± 24%)
400	4.8E–6 (± 44%)	1.2E–4 (± 39%)	4.7E–6 (± 39%)	1.8E–4 (± 27%)

(a) $N = 10$

Table 4: Average cell loss and frame error rates for N multiplexed streams ($U = 80\%$).

6 On the Validity of Inference from Variance-Time Plots

Researchers often rely on the variance-time (V-T) plots to detect LRD in empirical data and estimate the corresponding Hurst parameter. According to condition (iii) of LRD, a process for which $v_m \sim \mathcal{O}(m^{-\beta})$ as $m \to \infty$, with $0 < \beta < 1$, is LRD. In the V-T test, $\log v_m$ is plotted as a function of $\log m$, and the asymptotic slope of the plot is taken as an estimate of $-\beta$ in $k^{-\beta}$. If $\beta < 1$, the empirical sequence is believed to exhibit LRD. Although when applied to real video sequences, the V-T test suggests an asymptotic value of $\beta < 1$, we now show that this test can lead to misleading conclusions.

The V-T plot for the original *Star Wars* trace is shown in Figure 7. Its asymptotic slope, ignoring aggregation levels smaller than 100, is estimated by least-square method to be 0.43, roughly in agreement with the numbers in [1] and [5]. Taking one of the 1,000,000-long $M/G/\infty$ traces, we obtain its V-T plot before and after the Gamma/Pareto transformation, as shown in Figures 8 and 9, respectively (both versions are examined to show that the transformation does not influence the outcome of the V-T test). Both figures indicate an asymptotic slope of about 0.59 (starting with aggregation level of $10^{2.5} = 317$), which according to the V-T test should suggest that the underlying data exhibit LRD. However, the $M/G/\infty$ model is SRD. Despite the length of the $M/G/\infty$ trace (longer than any of the empirical video records ever tested), the V-T test can wrongly suggest the presence of LRD in a synthetic realization of a SRD model. Similar conclusions were reached for other synthetic $M/G/\infty$ traces. Figure 10 depicts the negative of the slope of the V-T plot as a function of the first aggregation level (S) that is used in computing the slope (i.e., only variances with aggregation level $m \geq S$ are used in fitting the V-T plot). Note that in the case of the $M/G/\infty$ trace, the slope increases very slowly with S, and expectedly tends to -1 as $S \to \infty$. However, the convergence is quite slow, that even for traces 1,000,000-long, the V-T plot could wrongly indicate the presence of LRD.

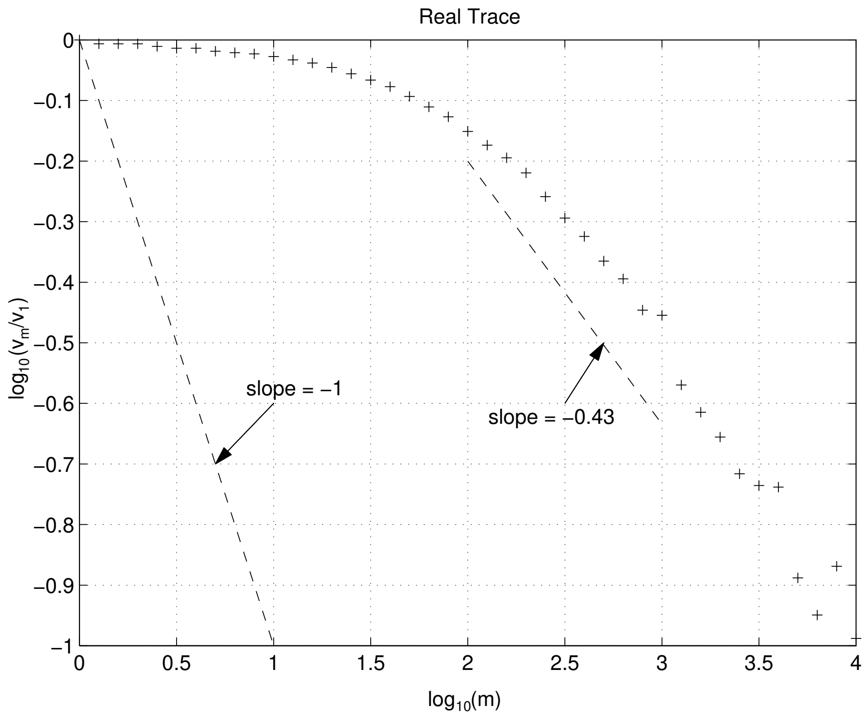

Figure 7: Variance-time plot for the real *Star Wars* trace.

7 Concluding Remarks

In this paper, we investigated the appropriateness of LRD modeling of video data. A compelling statistical evidence based on the autocorrelations of the original and the aggregated video sequences suggests that LRD models only capture the extreme tail of the ACF, and fail to represent the correlations at small and intermediate lags. Neither do Markovian models, which have an exponential tail, succeed in properly capturing the real ACF. We found that an ACF of the form $e^{-\beta\sqrt{k}}$ is a better match to the real ACF than both $e^{-\beta k}$ (Markovian) and $e^{-\beta \log k}$ (LRD). Accordingly, we introduced a new video model based on the $M/G/\infty$ input processes, whose ACF is given by $e^{-\beta\sqrt{k}}$. The $M/G/\infty$

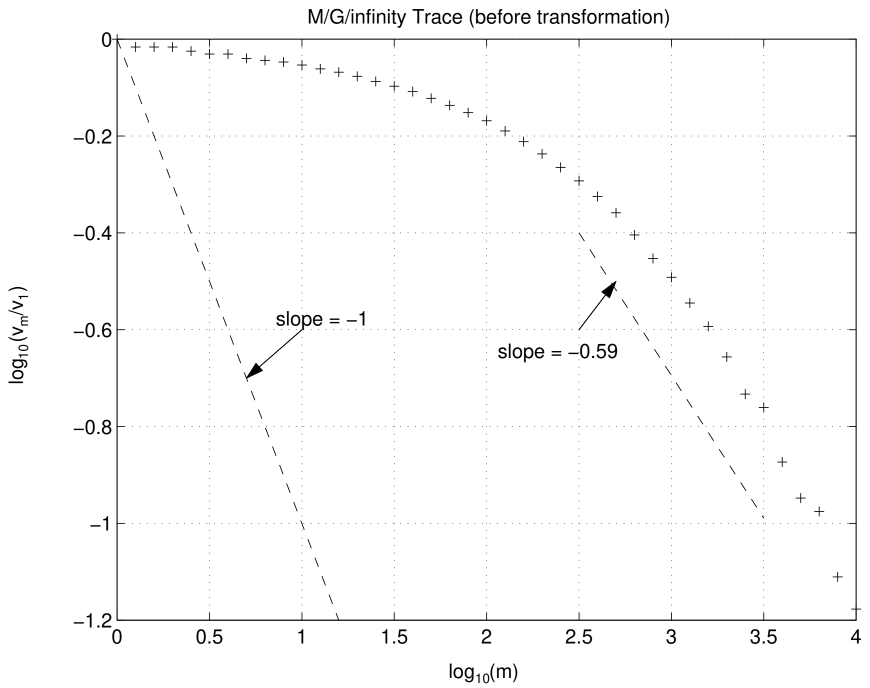

Figure 8: Variance-time plot for an $M/G/\infty$ trace before transformation.

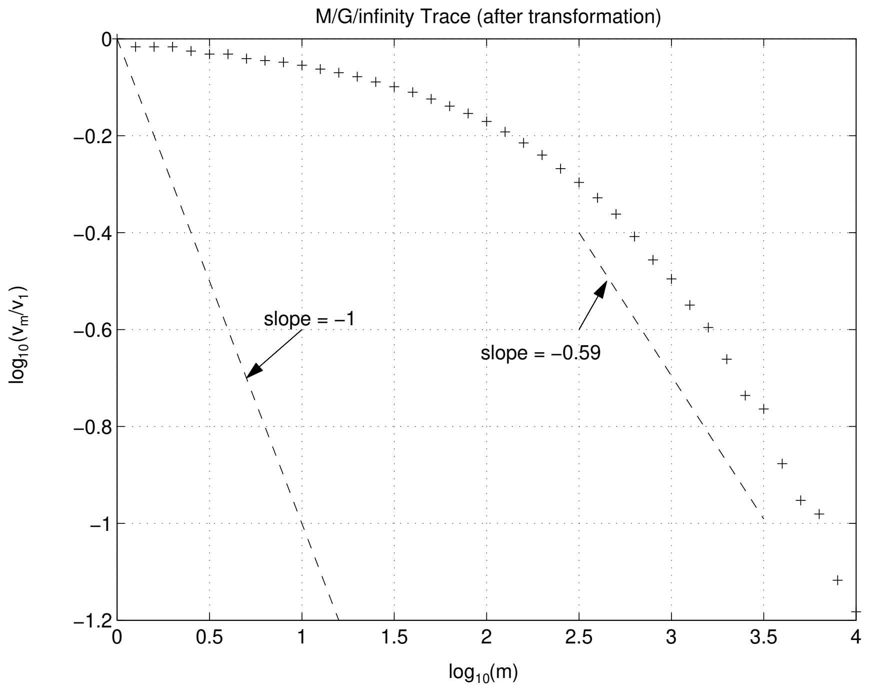

Figure 9: Variance-time plot for an $M/G/\infty$ trace after transformation.

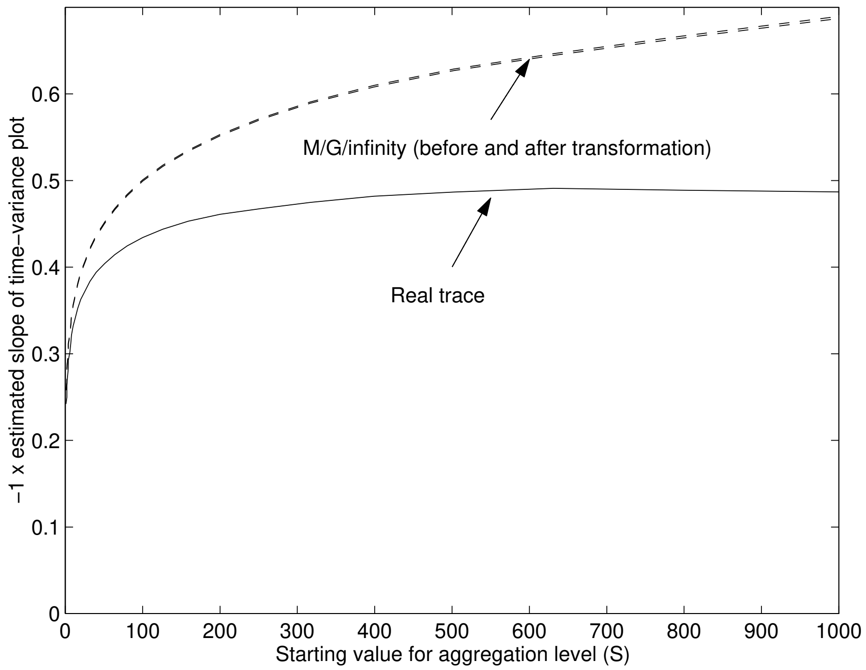

Figure 10: Slope of variance-time plots as a function of the first aggregation level, S, used in computing the slope (aggregation levels smaller than S are not used).

video model is shown to be short-range dependent (though not Markovian). With the performance of a real stream taken as a reference, the $M/G/\infty$ model was shown to provide better predictions of the queueing performance (in terms of the cell loss and frame error rates) than the F-ARIMA model. In particular, the performance of the F-ARIMA model is much more sensitive to changes in the buffer size, making it underestimate the real performance at large buffers by orders of magnitude. In contrast, the $M/G/\infty$ model exhibits similar buffer size/performance trend to the real traffic, and provides sufficiently accurate predictions. We provide a case study of the multiplexing performance under both models. The study indicates, once again, the overly sensitive behavior of the F-ARIMA model to changes in buffer size. LRD in video data is typically shown by means of variance-time plot. We reexamine such a statistical test and show that, even for SRD $M/G/\infty$ traces of substantial length, the variance-time plots can give misleading indications regarding the dependency structure of the examined traces.

Appendix

Proof of Proposition 3

Consider an $M/G/\infty$ input process $\{b_n : n = 0, 1, \ldots\}$. Then, b_n^* can be written as

$$b_n = b_n^{(0)} + b_n^{(a)}, \quad n = 0, 1, \ldots \tag{28}$$

where $b_n^{(0)}$ is the contribution of the customers that are initially present in the system at $n = 0$:

$$b_n^{(0)} = \sum_{i=1}^{b_{0^-}} \mathbf{1}[\sigma_{0,i} > n], \quad n = 0, 1, \dots \tag{29}$$

and $b_n^{(a)}$ is the contribution of all other customers, which can be written as

$$b_n^{(a)} = \sum_{s=1}^{n} \sum_{i=1}^{\xi_s} \mathbf{1}[\sigma_{s,i} > n - s], \quad n = 0, 1, \dots \tag{30}$$

It has been shown [16] that for the stationary version $\{b_n^*\}$, b_n^* can be represented by (28) with $b_n^{(a)}$ as in (30) and $b_n^{(0)}$ given by

$$b_n^{(0)} = \sum_{i=1}^{b_{0^-}} \mathbf{1}[\widetilde{\sigma}_i > n] \tag{31}$$

where $\{\widetilde{\sigma}_i : i = 1, \dots, b_{0^-}\}$ are *i.i.d.* rvs distributed according to the forward recurrence time $\widetilde{\sigma}$ (this distribution was given in (9)), and b_{0^-} is a Poisson rv with mean $\lambda \mathrm{E}[\sigma]$. In the above, $\mathbf{1}[\ ]$ is the indicator function. Before proceeding further with the proof, let us first introduce the following result:

Lemma 1 *Let* $\{X_1, X_2, \dots\}$ *be* $\mathbb{R}$*-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*

$$\begin{aligned} \mathrm{cov}\left[\sum_{j=1}^{\nu} f(X_j), \sum_{i=1}^{\nu} g(X_i)\right] &= \mathrm{E}[\nu]\,\mathrm{E}[f(X_1)g(X_1)] + \left(\mathrm{E}[\nu(\nu-1)] - (\mathrm{E}[\nu])^2\right)\mathrm{E}[f(X_1)]\,\mathrm{E}[g(X_1)] \\ &= \mathrm{E}[\nu]\,\mathrm{cov}[f(X_1), g(X_1)] + \mathrm{var}(\nu)\mathrm{E}[f(X_1)]\,\mathrm{E}[g(X_1)] \end{aligned} \tag{32}$$

Proof:

$$\mathrm{cov}\left[\sum_{j=1}^{\nu} f(X_j), \sum_{i=1}^{\nu} g(X_i)\right] = \mathrm{E}\left[\sum_{j=1}^{\nu} f(X_j) \cdot \sum_{i=1}^{\nu} g(X_i)\right] - \mathrm{E}\left[\sum_{j=1}^{\nu} f(X_j)\right]\mathrm{E}\left[\sum_{i=1}^{\nu} g(X_i)\right]$$

$$= \mathrm{E}\left[\sum_{i=1}^{\nu} f(X_i)g(X_i) + \sum_{1 \le i,j \le \nu,\ i \ne j} f(X_j)g(X_i)\right] - \mathrm{E}\left[\sum_{j=1}^{\nu} f(X_j)\right]\mathrm{E}\left[\sum_{i=1}^{\nu} g(X_i)\right]$$

$$= \mathrm{E}[\nu]\,\mathrm{E}[f(X_1)g(X_1)] + \mathrm{E}[\nu(\nu-1)]\,\mathrm{E}[f(X_1)]\,\mathrm{E}[g(X_1)] - \mathrm{E}[\nu]^2\mathrm{E}[f(X_1)]\,\mathrm{E}[g(X_1)]$$

Under the enforced independence assumptions, we have

$$\boldsymbol{\Gamma(h)} \stackrel{\triangle}{=} \operatorname{cov}(\boldsymbol{b}_{\boldsymbol{n}}^{*}, \boldsymbol{b}_{\boldsymbol{n+h}}^{*}) = \operatorname{cov}(\boldsymbol{b}_{\boldsymbol{n}}^{(0)}, \boldsymbol{b}_{\boldsymbol{n+h}}^{(0)}) + \operatorname{cov}(\boldsymbol{b}_{\boldsymbol{n}}^{(a)}, \boldsymbol{b}_{\boldsymbol{n+h}}^{(a)}), \quad \boldsymbol{n = 0, 1, \ldots; \; h = 1, 2, \ldots} \tag{33}$$

First, consider $\operatorname{cov}(\boldsymbol{b}_{\boldsymbol{n}}^{(a)}, \boldsymbol{b}_{\boldsymbol{n+h}}^{(a)})$:

$$\begin{aligned}
\operatorname{cov}(\boldsymbol{b}_{\boldsymbol{n}}^{(a)}, \boldsymbol{b}_{\boldsymbol{n+h}}^{(a)}) &= \operatorname{cov}\left(\boldsymbol{b}_{\boldsymbol{n}}^{(a)}, \sum_{s=1}^{n}\sum_{i=1}^{\xi_s} \mathbf{1}[\sigma_{s,i} > n+h-s]\right) \\
&\quad + \operatorname{cov}\left(\boldsymbol{b}_{\boldsymbol{n}}^{(a)}, \sum_{s=n+1}^{n+h}\sum_{i=1}^{\xi_s} \mathbf{1}[\sigma_{s,i} > n+h-s]\right) \\
&= \operatorname{cov}\left(\boldsymbol{b}_{\boldsymbol{n}}^{(a)}, \sum_{s=1}^{n}\sum_{i=1}^{\xi_s} \mathbf{1}[\sigma_{s,i} > n+h-s]\right) \quad \text{(by independence)} \\
&= \sum_{r=1}^{n}\sum_{s=1}^{n} \operatorname{cov}\left(\sum_{j=1}^{\xi_r} \mathbf{1}[\sigma_{r,j} > n-r], \sum_{i=1}^{\xi_s} \mathbf{1}[\sigma_{s,i} > n+h-s]\right) \\
&= \sum_{s=1}^{n} \operatorname{cov}\left(\sum_{s=1}^{\xi_s} \mathbf{1}[\sigma_{s,j} > n-s], \sum_{i=1}^{\xi_s} \mathbf{1}[\sigma_{s,i} > n+h-s]\right) \\
&= \sum_{s=1}^{n} \{\mathrm{E}[\xi_s]\, \mathrm{E}[\mathbf{1}[\sigma_{s,1} > n-s] \cdot \mathbf{1}[\sigma_{s,1} > n+h-s]] \\
&\quad + \left(\mathrm{E}[\xi_s(\xi_s - 1)] - \mathrm{E}[\xi_s]^2\right) \mathrm{P}[\sigma_{s,1} > n-s]\, \mathrm{P}[\sigma_{s,1} > n+h-s]\Big\}
\end{aligned} \tag{34}$$

where the last equality is obtained using Lemma 1. Let $\boldsymbol{\sigma}$, $\widetilde{\boldsymbol{\sigma}}$, and $\boldsymbol{\xi}$ denote generic rvs with the same distributions as $\boldsymbol{\sigma_{n,i}}$, $\widetilde{\boldsymbol{\sigma}}_{\boldsymbol{n}}$, and $\boldsymbol{\xi_i}$, respectively. Then (34) reduces to

$$\begin{aligned}
\operatorname{cov}(\boldsymbol{b}_{\boldsymbol{n}}^{(a)}, \boldsymbol{b}_{\boldsymbol{n+h}}^{(a)}) &= \mathrm{E}[\xi] \sum_{s=1}^{n} \mathrm{P}[\sigma > n+h-s] \\
&\quad + \left(\mathrm{E}[\xi(\xi-1)] - \mathrm{E}[\xi]^2\right) \cdot \sum_{s=1}^{n} \mathrm{P}[\sigma > n-s]\, \mathrm{P}[\sigma > n+h-s] && (35) \\
&= \mathrm{E}[\xi] \sum_{r=1}^{n} \mathrm{P}[\sigma > r+h-1] && (36) \\
&\quad + \left(\mathrm{E}[\xi(\xi-1)] - \mathrm{E}[\xi]^2\right) \sum_{r=1}^{n} \mathrm{P}[\sigma > r-1]\, \mathrm{P}[\sigma > r+h-1] && (37) \\
&= \lambda \sum_{r=1}^{n} \mathrm{P}[\sigma > r+h-1] && (38)
\end{aligned}$$

Next, consider $\mathrm{cov}(b_n^{(0)}, b_{n+h}^{(0)})$:

$$
\begin{aligned}
\mathrm{cov}(b_n^{(0)}, b_{n+h}^{(0)}) &= \mathrm{cov}\left(\sum_{i=1}^{b_{0-}} \mathbf{1}[\tilde{\sigma}_i > n], \sum_{j=1}^{b_{0-}} \mathbf{1}[\tilde{\sigma}_j > n+h]\right) &\quad (39)\\
&= \mathrm{E}\left[b_{0-}\right] \mathrm{P}\left[\tilde{\sigma} > n+h\right] \\
&\quad + \left(\mathrm{E}\left[b_{0-}(b_{0-}-1)\right] - (\mathrm{E}\left[b_{0-}\right])^2\right) \mathrm{P}\left[\tilde{\sigma} > n\right] \mathrm{P}\left[\tilde{\sigma} > n+h\right] &\quad (40)\\
&= \lambda \mathrm{E}\left[\sigma\right] \mathrm{P}\left[\tilde{\sigma} > n+h\right] &\quad (41)
\end{aligned}
$$

(since $b_{0-} \sim$ Poisson with mean $\lambda \mathrm{E}[\sigma]$). Combining (38) and (41), we have

$$
\begin{aligned}
\mathrm{cov}(b_n^*, b_{n+h}^*) &= \lambda \mathrm{E}\left[\sigma\right] \mathrm{P}\left[\tilde{\sigma} > n+h\right] + \lambda \sum_{r=1}^{n} \mathrm{P}\left[\sigma > r+h-1\right] \\
&= \lambda \mathrm{E}\left[\sigma\right] \sum_{r=1}^{\infty} \mathrm{P}\left[\tilde{\sigma} = n+h+r\right] + \lambda \sum_{r=1}^{n} \mathrm{P}\left[\sigma > r+h-1\right] \\
&= \lambda \sum_{r=1}^{\infty} \mathrm{P}\left[\sigma \geq n+h+r\right] + \lambda \sum_{r=1}^{n} \mathrm{P}\left[\sigma > r+h-1\right] \\
&= \lambda \sum_{r=1}^{\infty} \mathrm{P}\left[\sigma \geq n+h+r\right] + \lambda \sum_{r=1}^{n} \mathrm{P}\left[\sigma \geq r+h\right] \\
&= \lambda \sum_{r=1}^{\infty} \mathrm{P}\left[\sigma \geq h+r\right] = \lambda \sum_{r=1}^{\infty} \mathrm{P}\left[\sigma - h \geq r\right] \\
&= \lambda \sum_{r=1}^{\infty} \mathrm{P}\left[(\sigma-h)^+ \geq r\right] = \lambda \mathrm{E}\left[(\sigma-h)^+\right] &\quad (42)
\end{aligned}
$$

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] 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.

[3] 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.

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

[5] M. W. Garrett and W. Willinger. Analysis, modeling, and generation of self-similar VBR video traffic. In *Proceedings of the SIGCOMM'94 Conference (Computer Communications Review)*, 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] 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.

[8] D. Heyman and T. Lakshman. What are the implications of long-range dependence for VBR video traffic engineering? Technical report, Bellcore, 1995.

[9] 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.

[10] 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.

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

[12] 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 SIGCOMM '95*, 1995.

[13] 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.

[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 *Proceedings of IEEE INFOCOM '95*, pages 985–992, Boston, MA, Apr. 1995.

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

[17] M. Parulekar and A. M. Makowski. Tail probabilities for $M/G/\infty$ input processes (i): Preliminary asymptotics. Submitted for publication.

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

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

[20] B. Rye 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*, Aug. 1996.