Analysis, Modeling and Generation of Self-Similar VBR Video Traffic

Mark W. Garrett Walter Willinger
Bellcore
445 South St.
Morristown NJ 07960

Abstract

We present a detailed statistical analysis of a 2-hour long empirical sample of VBR video. The sample was obtained by applying a simple intraframe video compression code to an action movie. The main findings of our analysis are (1) the tail behavior of the marginal bandwidth distribution can be accurately described using "heavy-tailed" distributions (e.g., Pareto); (2) the autocorrelation of the VBR video sequence decays hyperbolically (equivalent to *long-range dependence*) and can be modeled using self-similar processes. We combine our findings in a new (non-Markovian) source model for VBR video and present an algorithm for generating synthetic traffic. Trace-driven simulations show that statistical multiplexing results in significant bandwidth efficiency even when long-range dependence is present. Simulations of our source model show long-range dependence and heavy-tailed marginals to be important components which are not accounted for in currently used VBR video traffic models.

1 Introduction

Packet switched communications technology has advanced significantly in the past decade, most notably in local area networks, the Internet and ATM technology. There are two great advantages to packet switching. One is that the bandwidth of a circuit is not restricted to a small set of allowed rates. The other is the support of variable bit rate (VBR) connections, which permits efficient statistical multiplexing of bursty traffic such as computer data. Video coders are also variable rate sources; however, the traffic is generally shaped and coded to accommodate the constant bit rate (CBR) channel of circuit-switched networks. Forcing the transmission rate to be constant results in delay, wasted bandwidth, and modulation of the video quality [ORTE93]. The availability of packet networks raises the question of whether real-time services (especially video due to its high bandwidth) can be improved in both efficiency and quality through variable rate transport.

To ensure a consistent and desirable quality of service (QOS) for a VBR video connection, the network must correctly allocate and regulate bandwidth assigned to the service. Short periods of congestion may be handled gracefully through the use of prioritization (layered coding) [PVW91], congestion notification [GARR93], and intelligent scheduling algorithms at the switch [CLAR92]. However, good design and analysis of a network requires an understanding of the traffic itself. This paper presents a long trace of VBR video with statistical analyses and simulation results which address issues of both source modeling and network performance.

Modeling of VBR video traffic is difficult, due to both the complexity of the video bandwidth trace as a stochastic process, and the problem of obtaining empirical data. To generate the present trace required 6 weeks of CPU time (in late 1990), which was arguably on the edge of practical computability. Traditionally, video coding algorithms have been designed and tested using short video sequences of 5–20 seconds that represent difficult scenes. As processor speed improves, we hope to see the use of long test sequences become standard practice for all video work.

In this study, we wish to develop an intuition about the bandwidth process of a VBR video coder. The precise quality level, picture format, and code details are not important. The distinction between intraframe and interframe coding is significant, however. Greater compression, burstiness and much stronger dependence on motion result from interframe coding, i.e., coding frame differences or use of motion prediction/compensation. Our main results do seem to extend to interframe (MPEG) video as well [GARR93a]. (See also [PANC94] for analysis of an MPEG VBR video trace.)

The next section describes the coding method and some interesting characteristics of the time series. Section 3 describes the basic statistics of the trace such as bandwidth distribution, mean, variance, burstiness, autocorrelation, and Fourier spectrum. These measures are common for traditional models. We examine the marginal distribution of the video bandwidth in detail, and suggest a model with a hyperbolically decaying tail. We measure the degree of long-range dependence, which is evident in the trace, but is not captured by standard source models. This allows us to construct a novel and accurate source model for VBR video in Section 4. In Section 5, we use the trace directly to drive network simulations, and explore the issue of resource allocation for statistically multiplexed video sources. Simulations using the source model are compared to those using the trace itself.

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association of Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
SIGCOMM 94 -8/94 London England UK
© 1994 ACM 0-89791-682-4/94/0008..$3.50

2 Video Code and Time Series Description

Coding algorithms	DCT, Run-length, Huffman
Duration	2 hours
Video frames	171,000
Frame dimensions	480 lines × 504 pels
Pel resolution	8 bits/pel (monochrome)
Frame rate	24 per second
"Slice" rate	30 per frame
Avg. bandwidth	5.34 Mb/s
Avg. compression ratio	8.70

Table 1: Parameters for generating VBR video trace.

Two hours of video material were coded using the movie "Star Wars" as the source. This represents a realistic full-length sample of entertainment video with a diverse mixture of material ranging from low complexity/motion scenes to those with very high action. The time series (or *trace*) of the bandwidth was measured at both the frame and slice time resolutions. Table 1 summarizes various statistics of the trace.

To generate such a long trace, we chose a relatively simple code. Only the luminance component is used, resulting in monochrome video. The frame is partitioned into blocks of 8×8 pels, on which a Discrete Cosine Transform (DCT) is computed. The DCT coefficients are uniformly quantized into 8 bits and compressed using run-length and Huffman coding. These algorithms comprise essentially the same coding as the JPEG standard [WALL91]. The quality of our coding is reasonable, except that block boundaries are noticeable in some cases. We assume that the fine tuning necessary for excellent visual quality would not change the character of the bandwidth statistics (for intraframe coding), other than slight shifting or scaling.

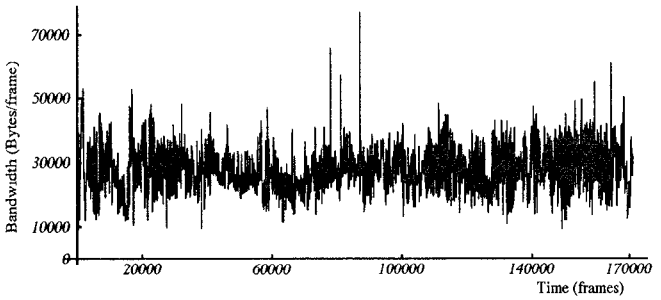

Figure 1: Time series of entire two-hour VBR video sequence.

The complete trace is shown in Fig. 1. The visible features here include three unusually high peaks near the center. These are due to visual special effects containing strong components of high spatial frequency. The scenes corresponding to these peaks are the "jump to hyperspace", a planet explosion, and the "jump from hyperspace". Two significant and unusually wide

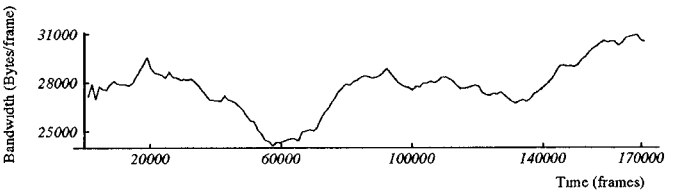

Figure 2: Low frequency content of VBR video process.

peaks occur at the very beginning and five minutes from the end, corresponding to an opening text sequence (42 seconds in duration), and the "Death Star" explosion (10 seconds). The low frequency content of the time series, shown in Fig. 2, is obtained using a moving average filter with window size of 20,000 frames (about 14 minutes). Note that the apparent variance of the trace in Fig. 1 follows a pattern similar to the moving average. This pattern also relates to the story line of the film: The action is intense in the introduction, then quite placid in the second quarter as the protagonist is developed. The pace picks up as the conflict progresses, pauses slightly and then builds to a climactic (and bandwidth-rich) finale. Such strong low frequency content is dramatic and accessible evidence of long-range dependence.

3 Statistical Analysis

Several basic statistics for the trace are given in Table 2. These describe distributional properties of the amount of information generated per frame and per slice. An important traffic descriptor is the burstiness, expressed here as the peak to mean bandwidth ratio. This measure bounds the statistical multiplexing gain (SMG) because, while a single source must be allocated approximately peak bandwidth, the allocation approaches the mean bandwidth as the number of combined sources increases. The remaining analysis divides into two categories: the marginal (or *stationary*) distribution of the process, and the time-correlation structure.

Measured by:	Frame	Slice	
Time unit, ΔT	41.67	1.389	msec
Mean bandwidth, μ	27791	926.4	bytes/ΔT
Standard deviation, σ	6254	289.5	bytes/ΔT
Coef. of variation, σ/μ	0.23	0.31	
Maximum bandwidth	78459	3668	bytes/ΔT
Minimum bandwidth	8622	257	bytes/ΔT
Peak/mean bandwidth	2.82	3.96	

Table 2: Statistics of VBR video trace.

3.1 Distributional Properties: Heavy Tails

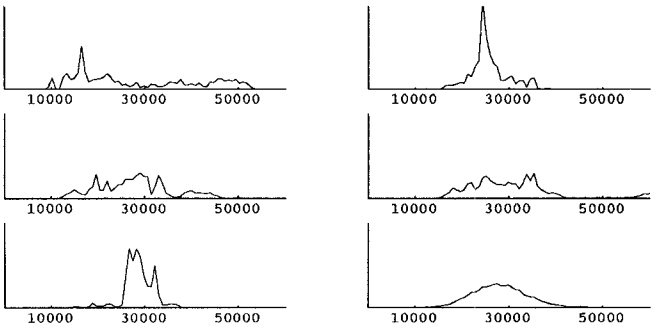

Figure 3: Bandwidth distribution for five two-minute sequences and for the complete trace (bottom right).

Figure 3 compares the distribution of bandwidth per frame for several two-minute segments of the movie, and the entire trace. Note that two minutes is a short duration compared to the complete trace, but very long compared to the amount of data in the coder or network at one time. Clearly, for periods that may seem long with respect to a queueing process, the behavior deviates significantly from the long-term statistical characterization.

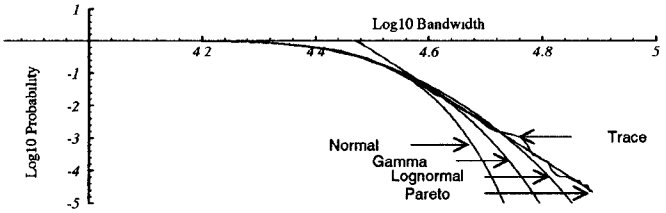

Figure 4: Log-log graph of complementary cumulative distribution compared to several common models.

To study the tail behavior in detail, it is useful to plot the complementary cumulative distribution function on a log-log scale. We compare the data to several standard distributions [LAW91, JOHN70]. While the Normal, Gamma and Lognormal distributions all have generally bell-shaped density curves and match the main body of the empirical distribution function well, Fig. 4 shows that their (right) tails decay more rapidly than the corresponding tail of the empirical distribution. We notice that the Gamma curve matches the data best—except in the extreme tail, that the Normal distribution falls off too quickly, and that the Lognormal (chosen because it has a "heavier" tail) is too heavy at first, and then falls off too rapidly. The "heavy-tailed" Pareto distribution (which decays as a power function rather than some form of exponential) yields a straight line when plotted on log-log coordinates and matches the tail behavior of the measured data very well.

We check the left tail behavior (which is not symmetrical to the right tail) in Fig. 5 and find that the Gamma distribution

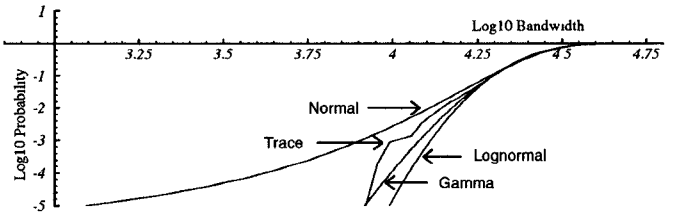

Figure 5: Log-log graph of cumulative distribution (left tail) of data compared to several common models.

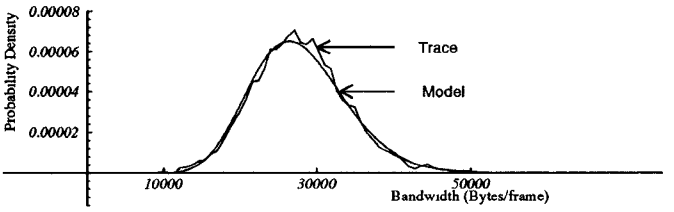

Figure 6: Probability density of trace data compared to Gamma/Pareto model.

provides an adequate fit for the lower end of the empirical distribution function. As a result, we can closely match the empirical distribution as a whole with a hybrid Gamma/Pareto distribution (denoted $F_{\Gamma/P}$). The probability densities are compared in Figure 6. Formulas for the Gamma and Pareto distributions and further details are given below in Section 4.

3.2 Time Correlation Structure and Long-Range Dependence

We now examine the time-dependent properties of the trace. We find strong evidence of *long-range dependence* (LRD), which is not accounted for in any of the commonly used stochastic models for VBR video traffic, and yet has been found to be ubiquitous in VBR video traces [BERA93], and may have important effects on performance. See [GARR93a] for an extensive literature survey of VBR video statistical measurements and models.

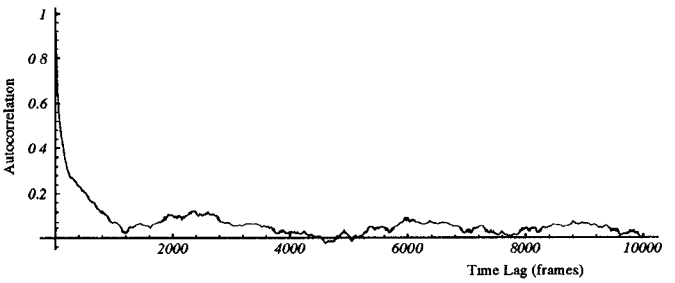

Figure 7: Autocorrelation function for video trace.

For the frame data, the empirical autocorrelation function $r(n)$ is shown in Fig. 7, with lag n ranging from 0 to 10,000 frames (about seven minutes). Notice that the initial part of the curve can be accurately matched to an exponentially decaying function, but only up to about 100–300 lags. Beyond that $r(n)$ decreases slower than exponentially, up to approximately lag 1200. It then adopts a quite erratic behavior with apparent oscillations on all scales of time. The curve does decay toward zero, but it does so extremely slowly. An autocorrelation function generally becomes inaccurate for lag values approaching the data set size, because the number of data points separated by that lag becomes small. This is *not* the cause of the erratic behavior observed in Fig. 7; there are still 161,000 observations contributing to the value of $r(10000)$. The very slowly decaying autocorrelations are indicative of LRD.

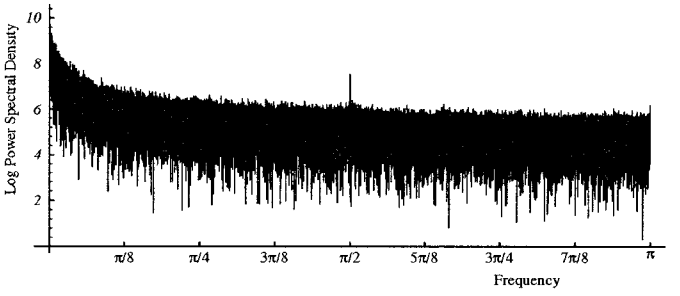

Figure 8: Frequency spectrum (periodogram) of frame data on log-linear coordinates.

Figure 8 displays the empirically measured power spectral density (also called *periodogram*) for the frame data on log-linear scale. This illustrates the frequency-domain interpretation of the observed slowly decaying autocorrelations: for low frequencies, the frequency spectrum does not seem to approach zero or a finite limit as is implied by exponentially decaying autocorrelation coefficients. Instead, the frequency spectrum observed in Fig. 8 exhibits a power law of the form $\omega^{-\alpha}$ for low frequencies which is one definition of long-range dependence.

3.2.1 Long-Range Dependence: Definition and Implications

Intuitively, long-range dependence, also known as "persistence" or the "Hurst effect," is the phenomenon of observations of an empirical record being significantly correlated to observations that are far removed in time. Formally, LRD may be captured by two essentially equivalent definitions: (i) The sum over all lags n of the autocorrelation coefficients $r(n)$ is infinite, meaning that the autocorrelations $r(n)$ ultimately decay as a hyperbolic function (i.e., as $n^{-\beta}$, as $n \to \infty, 0 < \beta < 1$) rather than a negative exponential (i.e., as ρ^n, as $n \to \infty, 0 < \rho < 1$). (ii) The periodogram, or power spectrum, behaves like $\omega^{-\alpha}$ for low frequencies, i.e., it increases without bound as the frequency $\omega \to 0$. When plotting a time series, LRD manifests itself in the presence of strong low frequency components, a property which is clearly visible for our data (see Fig. 1). An LRD process plotted on a graph of any time scale, appears to have a dominant periodicity with a few cycles fitting on the plot. If more data is plotted, new dominant modes appear [MAND69a].

The increase of apparent amplitude with increased time scale can be understood intuitively for video, especially movies. Within each scene there is random movement and variation of bandwidth. Changes of camera angle can alter the general level of complexity more than the changes within the scene, and occur on a longer time scale. Scenes themselves occur in clusters of similar type as the plot evolves (recall Fig. 2). On an even longer time scale there are different movies, and different genres of movies. For each object defined on a larger time scale, there is a larger variation of behavior.

LRD is quantified by a single parameter, H, after H. E. Hurst who studied long-term storage in water reservoirs [HURS51]. H is related to the rate of decay β of the autocorrelation coefficients and, equivalently, to the parameter α that characterizes the power law behavior of the spectral density around the origin. We will employ several techniques for estimating H for our data set.

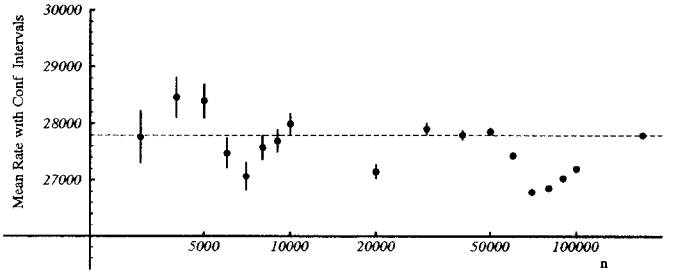

Figure 9: Estimation of mean bit rate from partial observations. 95% confidence intervals are shown for mean rate measured on first n observations.

From a statistical point of view, LRD can have unexpected and perhaps serious consequences. For example, the accuracy of a statistical measurement generally depends on having a large enough sample for the statistic to converge meaningfully. Confidence intervals (CI) are used widely in performance analysis to gauge the accuracy of parameter estimates. The conventional CI calculation not only assumes that measurement errors are Normally distributed, but also that they are i.i.d. For short-range dependent (SRD) processes (i.e., having an exponentially decreasing autocorrelation function), the correlations become negligible after a finite and usually small lag, and confidence intervals are reasonably accurate. For LRD processes, however this is not the case. Consider the estimate of the mean rate of the VBR video trace, taken on the first n observations. These estimates are shown in Fig. 9 for several values of n, together with the corresponding 95% CIs. Although the estimates gradually converge, the confidence intervals (derived under the assumption of i.i.d. or SRD) converge much more quickly than is warranted, and for most cases, the final mean value at $n = 171000$ is not even contained in the interval.

This disturbing feature will disappear when taking LRD into account, because the resulting 95% CI will be wider and will converge at a much slower rate.

3.2.2 Stationarity and Self-Similarity

It has often been claimed that VBR video is a non-stationary process. For theoretical stochastic processes the notion of stationarity is well defined. However for empirical processes, non-stationarity may mean that one has simply not yet found a satisfactory description of the process. For example, a process can be truly non-stationary if it has an explicit time-dependent trend. This can be treated by subtracting the trend and then characterizing the remaining stationary random process. However, this procedure only works if the underlying cause of the trend can be identified. It is not useful to remove the very low frequency component of our data (and model the remainder as a SRD process), because it is not deterministic, and will occur differently in another sample. Long-range dependent processes provide a convenient theory within the framework of stationarity that accounts for the observed low-frequency modulation of the statistics.

The Hurst parameter H implies a certain relationship of autocorrelations over all time scales. (If it were not the same relation for all time scales, we would need more than one parameter to describe it!) Thus, the "ideal" LRD process (the kind that comes out of LRD models having H as the only time-correlation parameter), is a *(second-order) exactly self-similar* process [COX84]. A covariance stationary process X is said to be (second-order) exactly self-similar if the corresponding "aggregated" processes $X^{(m)}$ have the same autocorrelation function as X, for all $m \geq 1$, where the processes $X^{(m)}$ are obtained by averaging the original process X over successive non-overlapping blocks of size m. (For more formal definitions of long-range dependence and self-similarity see [LELA93, BERA93].)

A process which has a hyperbolically decaying autocorrelations can always be *approximated* using SRD models. This is equivalent to, and has the problems associated with, approximating a power function by a sum of exponentials. For a specific time scale this may be done accurately, but there is always a longer time scale where the model breaks down. To model the behavior over a wide range of time scales at once requires a large number of parameters.

The present VBR video trace appears to be self-similar over a large range of time scales. In Fig. 10 we demonstrate this by comparing three processes formed by aggregating frames over blocks of size 100, 500 and 1000 frames. For an SRD process (with a reasonable number of parameters) such aggregation would result in uncorrelated white noise. The graphs in Fig. 10 not only retain significant correlations, but are quite similar in appearance. At either extreme in time scale we expect this self-similarity to break down. For short time scales ($\leq$ 200 frames $\approx$ 10 sec) the behavior is different from that of the ideally

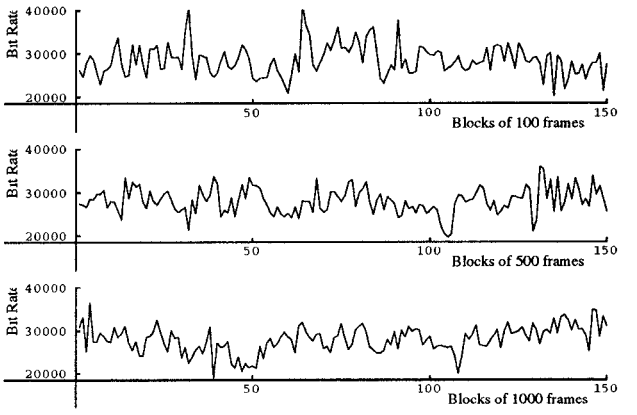

Figure 10: Self-similarity of VBR video.

self-similar process, and may be captured by augmenting the LRD model with SRD techniques such as ARIMA processes, Markov chains, etc. For this reason, our measurement of H is taken from approx. 200 frames to the longest time scale for which we can accurately estimate the behavior. At much larger time scales we may expect LRD to break down. However, since it holds over such a wide range of relevant time scales (i.e., 10 seconds to hours) it can be claimed as appropriate for traffic modeling. The assumption that self-similarity continues to hold as $\omega \to 0$, then becomes an approximation of the model.

3.2.3 Methods for Estimating H

For an i.i.d. process the variance of the sum of m observations increases in proportion to the number of points added, i.e., $\mathrm{Var} \sum^m X_i = m \, \mathrm{Var} X = m\sigma_X^2$. In terms of the aggregated processes $X^{(m)}$ introduced above, we have $\mathrm{Var}(X^{(m)}) = m^{-1}\sigma_X^2$. Asymptotically, this property still holds for SRD processes, i.e., $\mathrm{Var}(X^{(m)}) \approx m^{-1}\sigma_X^2$ (for m large). However, for LRD processes the variance of the aggregated series behaves for large m like [COX84],

$$\mathrm{Var}(X^{(m)}) \approx m^{-\beta}\sigma_X^2 \qquad (1)$$

with $0 < \beta < 1$. The "variance-time plot" is a graphical method for distinguishing between SRD ($\beta = 1$) and LRD ($0 < \beta < 1$) in a given empirical record. To estimate β, which is related to H by $\beta = 2 - 2H$, we plot the normalized variance of the aggregated series $\mathrm{Var}(X^{(m)})/\sigma_X^2$ as a function of block size m, on log-log coordinates, yielding β as the limiting slope as $m \to \infty$. Figure 11 shows that β can be consistently measured over a substantial range of m, yielding an estimate of H of about 0.78. The slope of the dotted line in Fig. 11 is $\beta = -1.0$, corresponding to an H-value of 0.5.

Another graphical method for estimating the Hurst parameter from a given empirical record is called the "R/S analysis". This method has been used to model a wide variety of geophysical phenomena [MAND69b] and is based on the

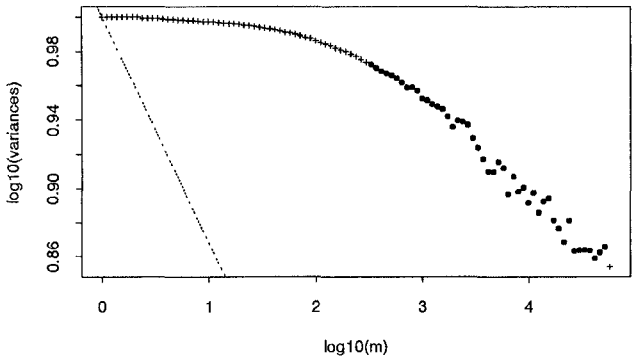

Figure 11: Variance-time plot for VBR video trace.

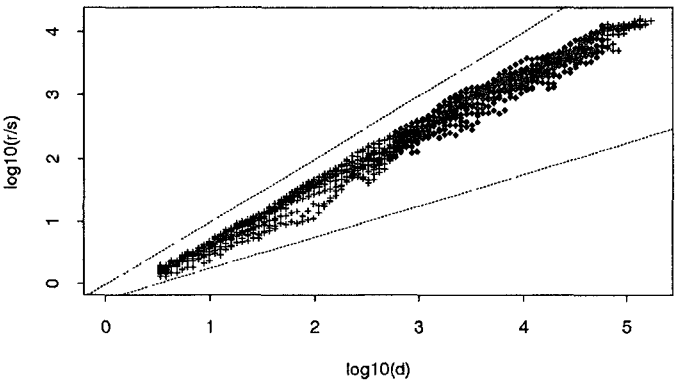

Figure 12: Pox diagram of R/S for VBR video trace.

rescaled adjusted range statistics R/S, originally introduced by H. E. Hurst [HURS51]. The heuristic behind the R/S statistic is to capture the fluctuations in a given time series in order to size, for example, a dam so that it neither underflows nor overflows given a (finite) empirical record. Formally, it is calculated as follows. From the sequence of observations, $\{X_1, X_2, ...X_n\}$, we define a sequence of *adjusted partial sums*, $W_j = (X_1 + X_2 + \cdots + X_j) - j\bar{X}(n)$, $j = 1, 2, 3, \ldots, n$, where $\bar{X}(n)$ denotes the arithmetic mean of the first n observations. Normalizing the adjusted range $R(n) = \max(0, W_1, W_2, ...W_n) - \min(0, W_1, W_2, ...W_n)$ by the sample standard deviation, $S(n)$ of the observations $X_1, X_2, \ldots, X_n$, we obtain the *rescaled adjusted range statistic*, $R(n)/S(n)$. Hurst showed empirically that for many naturally occurring time series the expected value of R/S asymptotically follows a power law [HURS51], i.e.,

$$\mathrm{E}\left[R(n)/S(n)\right] \approx n^H, \quad \text{as } n \to \infty, \tag{2}$$

with typical H-values of around 0.7. Others have since shown that (2) holds for short-range dependent processes with $H = 0.5$ [FELL51, MAND68], and for increment processes of self-similar models, where $0.5 < H < 1$ [MAND79]. When working with empirical data, a practical implementation of the R/S analysis has been proposed by [MAND69a] (see also [MAND79]) and consists of plotting $R(n)/S(n)$ versus n, for different lags n and for different partitions of the observations.

For our video trace, the resulting data is plotted in a "Pox diagram of R/S" in Fig. 12. H can be measured ideally as the asymptotic slope of a straight line. Using the points highlighted in the figure, the slope is estimated by a simple least squares regression as $H \approx 0.83$. In order to avoid possible distortions of the R/S analysis due to the presence of a particular short-range dependence structure, we also perform the R/S analysis on a number of aggregated processes $X^{(m)}$. Moreover, the size of the current data set allows us to investigate the robustness of these Hurst parameter estimates with respect to different partitions of the observations (i.e., density of points in vertical direction) and with respect to the number of lags (i.e., density of points in the horizontal direction). We find that our estimates are indeed very robust against these variations in the estimation procedure. (See Table 3.)

While variance-time plots and R/S analysis are graphical methods that produce a single point estimate of H, techniques for estimating H based on the asymptotic properties of the periodogram are known and yield confidence intervals for the estimated value of H. In particular, for Gaussian processes Whittle's approximate maximum likelihood estimator (MLE) has been studied extensively and has been shown to have many desirable statistical properties (see [BERA93, GARR93a] and references therein). In the case at hand, we consider the transformed series, $\{\log(X_i)\}$, which typically results in approximately Normal looking distributions (especially for the corresponding aggregated processes) and exhibits the same H-value as the original series. In addition, to filter out the influence of the high frequency components (since in practice, the empirical process is not exactly self-similar over all time scales), we combine the Whittle estimator with the method of aggregation and plot (not shown here) the Whittle estimator $\hat{H}^{(m)}$ with the the corresponding 95% confidence intervals $\hat{H}^{(m)} \pm 1.96\hat{\sigma}_{\hat{H}^{(m)}}$ (where $\hat{\sigma}^2_{\hat{H}^{(m)}}$ is given by a well known central limit result for the Whittle estimator) against m. This procedure suggests a Hurst parameter estimate of $\hat{H} = 0.8 \pm 0.088$, taken at $m \approx 700$.

As shown in Table 3, the estimates of H from the different methods mentioned above all fall well within the confidence intervals provided by Whittle's method. We note that other types of video generally have different values of H, and it appears that H can be used as a rough indication of scene activity. For video conferencing, for example, H tends to be smaller, typically between 0.60 - 0.75 [BERA93]. Computer traffic can be much more active, with measured H-values often close to unity [LELA93].

4 VBR Video Model Construction and Traffic Generation

We can now construct a source model for use in computer simulations that captures two important aspects from our analysis of the VBR video trace in the previous section: a precise marginal distribution with a heavy tail, and an autocorrelation

Method	H
Variance-Time	0.78
R/S Analysis	0.83
R/S Aggregated	0.78
R/S with n, M varied	0.81–0.83
Whittle estimate	0.8±0.088

Table 3: Estimates of H from all methods.

function with long-range dependence. Without both of these features, the occurrence of and persistence of "bad states" in a realization will be under-represented. The SRD structure is by default self-similar to the long-term structure. An additional set of short-term correlation parameters may be included by combining this model with an ARMA filter or modulating it with the state of a Markov chain. The accurate and meaningful modeling of short-term effects, however, is a complicated problem that we must leave for future work. Even without explicit SRD components, this model is still quite useful, and may be sufficiently accurate to represent VBR video traffic.

4.1 Generation of Fractional Noise

Realizations exhibiting long-range dependence may be obtained through a process called *fractional differencing*, which we will motivate briefly. Some stochastic processes, such as a discrete random walk or Brownian motion, tend to wander far from their origins with high probability. It is usually easier to describe these processes indirectly in terms of their increments. The incremental process may be written using a *differencing* operator, $\nabla X_k = (X_k - X_{k-1})$, which may be iterated as, $\nabla^2 X_k = (X_k - X_{k-1}) - (X_{k-1} - X_{k-2}) = X_k - 2X_{k-1} + X_{k-2}$, etc. In general,

$$\nabla^n X_k = \sum_{i=0}^{n} \binom{n}{i} (-1)^i X_{k-i}, \qquad n = 1, 2, 3, \ldots. \tag{3}$$

Just as the factorial function $n!$ is generalized to non-integer arguments by the Gamma function, we can similarly generalize ∇^n as a *fractional* differencing operator,

$$\nabla^d X_k = \sum_{i=0}^{\infty} \binom{d}{i} (-1)^i X_{k-i} \qquad -1/2 < d < 1/2 \tag{4}$$

where we now interpret the fractional binomial coefficient as,

$$\binom{d}{i}(-1)^i = \frac{\Gamma(-d+i)}{\Gamma(-d)\Gamma(i+1)}. \tag{5}$$

This generalization from integer to real functions, gives rise to the term *fractional* noise processes. Similarly, the term *fractal* is used in the context of self-similar geometric objects which can be characterized using the notion of a *fractional* dimension [MAND83].

Hosking provides an algorithm for generating a long-range dependent process called fractional ARIMA(0, d, 0), where the zeros indicate there are no autoregressive (AR) and moving average (MA) parameters specified. The basic equations for Hosking's algorithm are as follows (adapted from [HOSK84]).

The process X_k has Gaussian marginals with zero mean and variance v_0, and fractional differencing parameter $d = H - 1/2$. The autocorrelation function has an asymptotically hyperbolic shape, and is determined from d as

$$\rho_k = \frac{d(1+d)\cdots(k-1+d)}{(1-d)(2-d)\cdots(k-d)}. \tag{6}$$

X_0 is chosen from the Normal distribution $N(0, v_0)$. Set $N_0 = 0$ and $D_0 = 1$. Then generate n points by iterating the following for $k = 1 \ldots n$:

$$N_k = \rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\, \rho_{k-j} \tag{7}$$

$$D_k = D_{k-1} - N_{k-1}^2 / D_{k-1} \tag{8}$$

$$\phi_{kk} = N_k / D_k \tag{9}$$

$$\phi_{kj} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j} \qquad j = 1, \ldots, k-1 \tag{10}$$

$$m_k = \sum_{j=1}^{k} \phi_{kj} X_{k-j} \tag{11}$$

$$v_k = (1 - \phi_{kk}^2) v_{k-1} \tag{12}$$

Choose each X_k from $N(m_k, v_k)$. Since each point depends on every previous point, this algorithm requires $o(n^2)$ computation time. We found that 171,000 points could be generated in about 10 hours on an current engineering workstation.

4.2 Marginal Distribution Model

Given a realization of the fractional ARIMA(0, d, 0) process $\{X_k\}$, we can transform the marginal distribution by mapping each point as,

$$Y_k = F_{\Gamma/P}^{-1}(F_N(X_k)) \qquad k > 0 \tag{13}$$

where F_N is the cumulative probability function of the Normal distribution, and $F_{\Gamma/P}^{-1}$ is the inverse cumulative probability function of our Gamma/Pareto model. (A similar technique for distorting the marginals is used where the original process is distributed Uniformly rather than Normally [JAGE92].)

$F_{\Gamma/P}$ requires only three parameters, and is constructed as follows: The Gamma distribution has the probability density function,

$$f_\Gamma(x) = e^{-\lambda x} \frac{\lambda(\lambda x)^{s-1}}{\Gamma(s)}. \tag{14}$$

The *shape* and *scale* parameters, s and λ respectively, may be determined conveniently from the mean and variance.

The Pareto probability density function [JOHN70] is given by,

$$f_P(x) = \frac{ak^a}{x^{a+1}} \qquad x > k \tag{15}$$

with a convenient closed-form expression for the cumulative distribution,

$$F_P(x) = 1 - \left(\frac{k}{x}\right)^a \tag{16}$$

The two parameters are easily interpreted graphically. The first, k, is the minimum allowed value of x, and the second, a, is the slope of the tail on a log-log graph (see Fig. 4). We can eliminate one of these in the hybrid distribution by matching the slope and position of the two functions. The (constant) slope of the Pareto tail in Fig. 4, and the (varying) slope of the Gamma distribution match at a threshold point denoted x_{th}.

The complete Gamma/Pareto distribution model can be determined from three parameters, μ_Γ, σ_Γ, m_T, that are estimated from the empirical trace. μ_Γ and σ_Γ, are the *equivalent* mean and standard deviation of the Gamma portion of the distribution. For the present trace, it is sufficiently accurate to take the sample mean and standard deviation, because the heavy tail contains only 3% of the data. Where the tail is more significant, (as with MPEG coding [GARR93a]), these can be estimated either by graphically matching the distribution curves, or with some more elaborate estimation procedure, such as a least squares regression taken only on the Gamma part of the distribution. The value of m_T is found as the slope of the straight-line that best fits the Pareto tail. (See [GARR93a] for further details.)

We have designed and implemented a model for variable rate video with only four parameters (μ_Γ, σ_Γ, and m_T for the marginal distribution, and H for the time correlation). The realizations were tested and found to agree with the model parameters, both in marginal distribution and the value of H. There is an issue here with regard to how well the empirical distribution will converge to the model. For an LRD process the convergence is slower than for an SRD process, and this will certainly affect simulation results. (See discussion below in Section 5.2.)

To simulate the aggregation of multiple sources, we implemented a convolution of the Gamma/Pareto distribution using a table of 10,000 points to describe the distributions. Simulation results using this model are given below in Section 5.2. The empirical autocorrelation of the realization (not shown) has a slow hyperbolic decay, although it does not (nor should it) exhibit the erratic behavior evident in Fig. 7. The measured value of H is not affected by the distortion of the marginal distribution, as expected.

The (intraframe) trace data exhibits a wide variety of short-range behaviors, including periods with practically constant level. This is due to the "scene" structure of the movie, where the camera shows a scene with little change for a time, and then switches to another one. It is also common for the camera to switch between two scenes (e.g., two faces) many times, resulting in an long period of simple alternation between two levels. We have not attempted to explicitly model such scene-dependent structure, and it remains an open question whether this is necessary, and if so, how to measure and represent the scenes.

5 Trace-Driven and Model-Based Simulations

The most important reason for studying the characteristics of VBR video traffic is to determine what network resources are necessary to transport the service reliably. Also, any comparison of different methods for providing the service (such as CBR vs. VBR, or different types of VBR, etc.) must include an assessment of the resource allocation needed to provide a given quality of service.

5.1 Trace Driven Simulation

In this section we use a trace driven simulation to examine the relation between network resource allocation and performance for intraframe coded video. This is a reasonable alternative to simulation from a source model because the VBR video trace is extensive, diverse, and (hopefully) representative. The simulation also provides a useful tool for evaluating the accuracy of our source model by directly comparing its behavior to that of the trace.

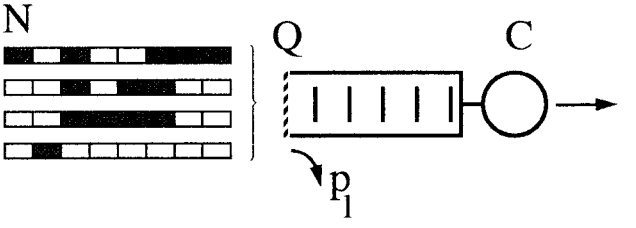

Figure 13: System modeled in trace driven simulation.

The simulation measures the performance of a single FIFO queue with a finite buffer of size Q, and fixed channel capacity, C (Fig. 13). A number of sources, N, are multiplexed to form the incoming traffic. This is implemented by combining several copies of the VBR trace that are offset by a random number of frames. Upon reaching the end of the trace, each source wraps around to the beginning, so all 171,000 frames are used once for each. The lag for each copy is chosen to be at least 1000 frames apart from each of the others. Long-range dependence implies that the cross-correlation between sources may be significant even for such long lags. For $N > 2$ therefore, we choose six different random lag combinations of the N sources and average the resulting loss rates.

In the simulation, the aggregate traffic is run through the queue, and the performance is measured as either the overall cell loss rate (P_l), or the cell loss rate in the worst errored

second (P_{l-WES}). The simple loss probability is a very general measure of quality. However, a viewer's perception will be sensitive to loss events that are localized in time, and are not apparent in the long-term average. The use of P_{l-WES} introduces a more sensitive measure.

For our results, we fix a target level of performance and measure the tradeoff between resources (Q and C) necessary to achieve the desired loss rate. Our goal is to understand how the resources are related rather than the absolute quantity of the resources needed. Given a more accurate or appropriate performance measure, this approach can be extended quite naturally.

In the longer presentation of this work [GARR93a], simulations are compared using both slice and frame data, and using uniform and random spacing of cells within the slice or frame. Note that in no case do all the cells of a frame arrive together, as is sometimes assumed. The instantaneous arrival of a whole frame would imply that the data is collected in the coder before being released to the network. This introduces unnecessary delay. We would expect real coders to be pipelined, producing cells as soon as they are ready, subject to some very small buffering, such as a DCT block, macroblock, or a row of blocks.

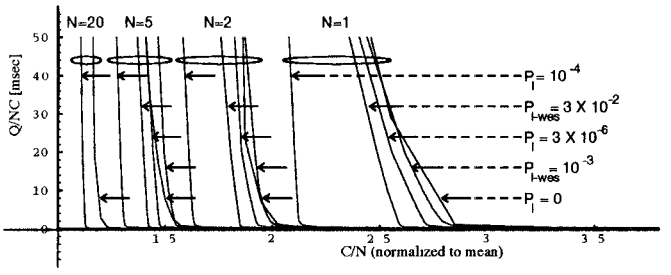

Figure 14: Behavior of statistically multiplexed video sources. Queueing delay vs. allocated bandwidth per source for several cases of multiplexed sources and target loss rate.

Figure 14 shows the basic simulation results. The maximum buffer delay, $T_{max} = Q/(NC)$ is plotted against the allocated bandwidth per source, C/N. These are meaningfully normalized measures of Q against C (so we will refer to this plot as a "Q-C curve"). The number of sources is chosen as $N = 1, 2, 5$, and 20. Curves are given for loss rates, $P_l = 0$, 10^{-4}, 3×10^{-6} and $P_{l-WES} = 10^{-3}$, 3×10^{-2}. As shown here, the bandwidth requirement is quite insensitive to the buffer size until the buffer delay is decreased to a few milliseconds. The ability to trade-off bandwidth for delay improves as the curve becomes less steep for smaller allowed loss rates. The zero-loss curves have a relatively shallow slope, and thus a better tradeoff between Q and C over a wider range. The difference in C between the curves for $P_l = 0$ and $P_l = 10^{-4}$ is substantial, especially for a single source. This means that if the video coding and transport system is designed to tolerate moderate loss, it will require much less resources than one that relies on engineering an effectively loss-free channel.

Although the QOS is specified in two ways (worst-errored-second and overall loss rates), all of the curves fall into the same general family. Note that the two types of curves lie in the same order for all values of N and C/N. This test allows us to infer that the overall loss is a good predictor of worst-errored-second loss, and is an equivalent specification of QOS, in the sense that there is a uniform, monotonic mapping from one to the other. It may often be the case that the overall packet loss rate hides information and is *not* a good indicator of user-perceived performance. By comparing Q-C curves for two performance measures we can develop a reasonably rigorous understanding of the equivalence of performance measures. This is useful when the value of one measure is in question, but a preferable measure is more difficult to implement (as is the case here).

All of these Q-C curves demonstrate a very strong "knee". This transition represents a natural operating point for the system, because at all other points, one of the resources is very sensitive to a slight change in the other. The problem of pinpointing the knee in the curve can be deceptive, and is analysed in detail in [GARR93a]. Given this point for each curve, however, we can examine C as a function of N.

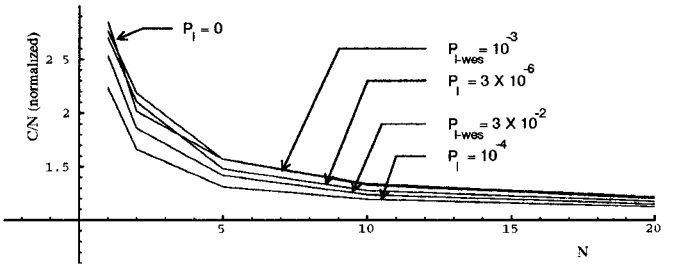

Figure 15: Required capacity allocation against number of sources multiplexed. Buffers are allocated for $T_{max} = 2$ msec.

In Fig. 15 we show the statistical multiplexing gain (SMG) that is achievable for this type of VBR video coding. The allocated bandwidth per source is shown against the number of sources for several values of acceptable loss rate (including zero). The capacity is very close to the peak rate for one source, and drops to very close to the mean rate for 20 sources. With 5 sources we have realized 72% of the possible gain, which is the difference between the peak and mean rates. (This is the average over the five curves, which are all within 4% of this figure.)

For very small loss rates in Figs. 15, 14, we find two or three curves which are very close and sometimes cross, e.g. for $P_l = 0$, $N = 2$. These represent an operating region on the threshold between zero and non-zero observed loss. Here, the resources are sufficient to avoid loss, or suffer loss only in the most extreme peaks in traffic. This indicates that it may be impossible in real traffic situations, to distinguish between the cases of $P_l = 0$, 10^{-12}, 10^{-9}, or 10^{-6}, because the differences in allocated resources are negligible.

5.2 Simulation of VBR Video Source Model

This simulator was used to evaluate the VBR video source model developed in Section 4. We compare the full model to the trace driven simulation results for $N = 1, 2, 5, 20$. We also check two variations which include only one of the two important features (the heavy-tailed distribution and long-range dependence). The Q-C curve provides a sort of "engineering test" of the model because the results are directly related to the network quality of service and resource allocation.

The Q-C curves shown Fig. 16 compare the trace to the models. Although we see the same general shape, there is a significant offset in capacity between the trace and model driven simulations. The full model performs consistently better than the two variations, indicating that both the Pareto tail and the high value of H are important components. As N increases, the marginals in all cases become more Gaussian and the special short-range time correlation effects (which account for some of the difference) are randomized. The agreement improves with N, however the distinction between the three models also diminishes.

We did find an indication that the model could, in fact, be much more accurate than these curves indicate. A comparison of the marginal distribution of the realizations show that the model does not hold the Pareto tail, but that it decays too rapidly for very high values of frame bandwidth. By slightly perturbing the parameters of the Gaussian to Gamma/Pareto mapping table, we were able to get somewhat better distribution agreement. This resulted in better agreement for $T_{max} \leq 2$ msec ($N = 1$). This illustrates an important open problem for LRD processes. Without a good theory of confidence intervals, it is impossible to know how well the extreme tail of an empirically generated trace will reflect the modeled tail shape. This discrepancy can be exaggerated when measuring simulation results which may be very sensitive to rare occurrences.

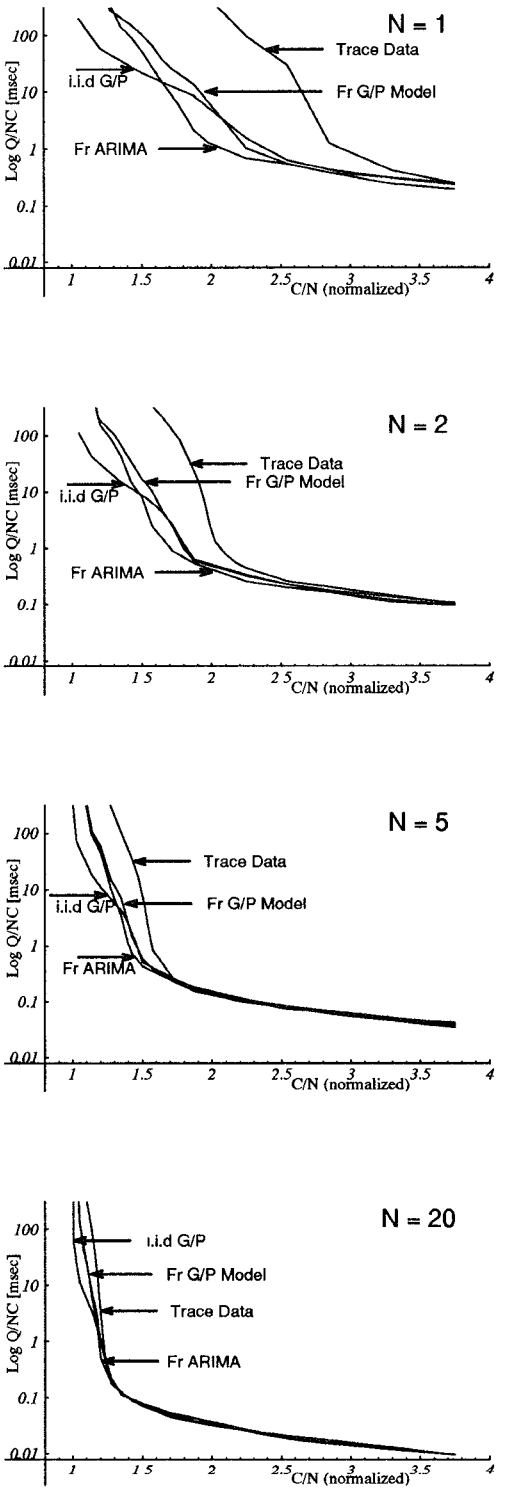

Figure 16: Comparison of simulations using VBR video trace data, fractional ARIMA model (with Gaussian marginals), fractional ARIMA model with transformed Gamma/Pareto marginal distribution, and an i.i.d. process with Gamma/Pareto marginals. ($P_l = 0$ for all cases.)

5.3 Quality of Service

The choice of a performance measure is central to system design, as well as modeling and performance evaluation. If we expect to make and keep guarantees of quality of service, the specification of the QOS has to be meaningfully associated with the traffic description, resource allocation, etc. The discrepancy between the QOS indicated by objective measures such as P_l and that perceived by the end user represents an important weakness in our ability to produce well-designed systems.

The overall expected loss rate captures no information about the correlation of losses, and since we have found P_{l-WES} to follow P_l, it appears not to be an improvement. To illustrate the problem of correlated loss, consider the two traces of Fig. 17. Here a window of 1000 frames is used to measure the running-average packet lost rate for one and twenty sources. The buffer delay is chosen at $T_{max} = 2$ msec for both, and the

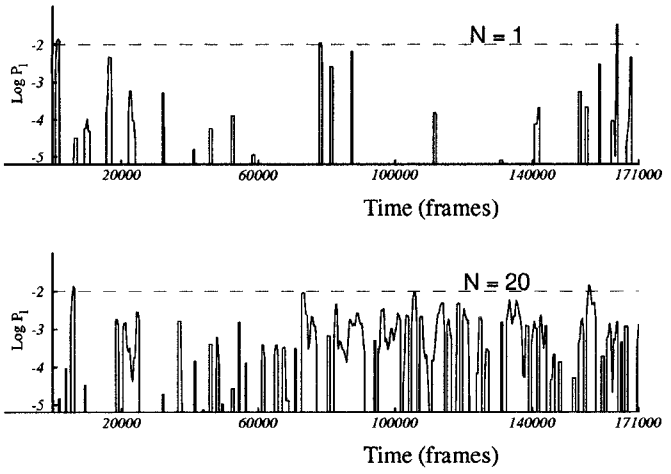

Figure 17: Error processes resulting from simulation of VBR video trace over the full two-hour interval for $N = 1$, $N = 20$ with $P_l = 10^{-3}$ in each case.

channel capacity is adjusted to give each an overall loss rate of $P_l = 10^{-3}$. (Since these are given on log scale, the apparent area under each curve is not meaningful.) Using P_l as a QOS measure, both cases have the same quality.

The question of which error process is better cannot be determined *a priori* even from the time-dependent loss rate shown. A viewer's perception will have a non-linear *threshold* effect, which is not captured by an additive measure like P_l. A large number of error events may be tolerable if they are always below the threshold of perception (or some measure of annoyance). Conversely, once a error crosses the threshold, it may not matter how severe it is. It seems reasonable to assume that these two traces would be perceived differently by a viewer, indicating that P_l does not sufficiently capture the QOS. Also, if packet loss degradations were concealed by using "layered" coding with a priority queueing discipline, then the QOS measure would have to account for this appropriately.

6 Conclusions

We have investigated the statistical properties of the stochastic traffic process generated by applying an intraframe variable rate compression code to a full-length movie. The interesting characteristics, which are not well captured by common analytic source models include a long-range dependent time correlation structure, and a heavy-tailed marginal distribution of the information content per time interval.

The trace itself can be used as a source model for this type of VBR video traffic through trace-driven simulation. Although the present dataset was generated by fixing the quantizer step size, it is probably very close to what a more sophisticated intraframe coder would produce. A video coder designed and optimized for VBR packet transport would vary the quantizer, not so much to avoid buffer overflow, but to approximate a constant quality picture. This results in a slightly higher degree of burstiness, compared to constant quantization [ORTE93]. A few extremely high peaks exist in the data, which are problematic for the network. We recommend that a realistic VBR coder should clip such peaks, rather than send them into the network. It will be much better trade-off for the coder to optimize its use of the available bandwidth and degrade the quality slightly, than for the network to accommodate such exceptional bursts. A realistic packet video coder will use layered coding[GARR93], which is not included here. The present trace, however should be a good prediction of the total information content, since the layering overhead is small.

The demonstration of long-range dependence in this trace, as well as in computer communications traffic [LELA93], has strong ramifications for the theory and practice of traffic modeling and performance analysis. The use of SRD models when inappropriate, will result in overly optimistic estimates of performance, insufficient allocation of resources and difficulty in achieving the quality of service expected by network users. Although long-range dependence exists with a significantly high value of H, the statistical treatment of the traffic and resource allocation is still possible. The statistics *do* converge, albeit slower than for i.i.d. data. Multiplexed sources are statistically better behaved than single sources, although the heavy tail of the marginals will converge to Normality only very slowly. The value of H is not reduced with traffic aggregation (due to the self-similar nature of the traffic). LRD is a relation of the frequency components of the process, not the distribution of bandwidth requirements. If the marginal distribution is (relatively) compressed because $C_v = \sigma/\mu \to 0$ as $N \to \infty$, then the traffic, for high N, is confined within narrower (statistical) bounds. The behavior *within* those bounds continues to be long-range dependent with parameter H. In the range where $\sigma/\sqrt{N} \ll \mu$, or for heavy tails, when the quantile of interest is close to the mean, then the traffic is, for all purposes, quite smooth regardless of H. Thus, H is necessary for characterizing burstiness, but not sufficient. All four parameters of our model (at least) are necessary to describe traffic which has a heavy marginal tail and LRD. (Note, that these remarks are only valid where the central limit theorem holds, i.e., σ is finite. For some cases, where $\sigma = \infty$ is a reasonable model [LELA93], the tail behavior *never* converges to Normality.)

An obvious extension of this work will be to analyse more movies of the same and different types to determine the consistency and generality of these results. Many more details and related ideas are provided in a longer version of this work [GARR93a]. This VBR dataset is available via anonymous ftp. The ftp site is thumper.bellcore.com, under directory, vbr.video.trace.

7 Acknowledgements

We would like to thank Martin Vetterli for his constructive comments on this work. Daniel V. Wilson built and maintained the video capture hardware used to collect the data set. Al Broscius provided us with a small but invaluable piece of software: a private file server that allowed our simulations to run politely on other people's machines.

References

[BERA93] J. Beran, R. Sherman, M. S. Taqqu and W. Willinger, "Variable Bit Rate Video Traffic and Long Range Dependence", *IEEE Trans. Commun.*, 1994, Accepted for publication, subject to revisions.

[CLAR92] D. D. Clark, S. Shenker and L. Zhang, "Supporting Real-Time Applications in an Integrated Services Packet Network: Architecture and Mechanism", In *Proc. ACM SIGComm Symp.*, pp. 14–26, Baltimore MD, August 1992.

[COX84] D. R. Cox, "Long-Range Dependence: A Review", In H. A. David and H. T. David, editors, *Statistics: An Appraisal*, Ames, Iowa, 1984, pp. 55–74, Iowa State Univ. Press.

[FELL51] W. Feller, "The Asymptotic Distribution of the Range of Sums of Independent Random Variables", *Ann. Math. Statists.*, Vol. 22, pp. 427–32, 1951.

[GARR93] M. W. Garrett and M. Vetterli, "Joint Source/Channel Coding of Statistically Multiplexed Real Time Services on Packet Networks", *IEEE/ACM Trans. Networking*, Vol. 1, No. 1, pp. 71–80, February 1993.

[GARR93a] M. W. Garrett, "Contributions Toward Real-Time Services on Packet-Switched Networks", Ph.D. Dissertation CU/CTR/TR 340-93-20, Columbia University, New York, N.Y., May 1993, see Chapter 4: "Statistical Analysis of a Long Trace of Variable Bit Rate Coded Video".

[HOSK84] J. R. M. Hosking, "Modeling Persistence in Hydrological Time Series Using Fractional Differencing", *Water Resources Res.*, Vol. 20, No. 12, pp. 1898–1908, 1984.

[HURS51] H. E. Hurst, "Long-Term Storage Capacity of Reservoirs", *Trans. Amer. Soc. Civil Eng.*, Vol. 116, pp. 770–799, 1951.

[JAGE92] D. L. Jagerman and B. Melamed, "The Transition and Autocorrelation Structure of TES Processes Part I: General Theory", *Stochastic Models*, Vol. 8, No. 2, pp. 193–219, 1992.

[JOHN70] N. L. Johnson and S. Kotz, *Continuous Univariate Distributions—1*, Houghton Mifflin, Boston, 1970.

[LAW91] A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*, McGraw-Hill, New York, 2nd edition, 1991.

[LELA93] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic", In *Proc. ACM SIGComm*, pp. 183–193, San Francisco, Calif., September 1993.

[MAND68] B. B. Mandelbrot and J. W. Van Ness, "Fractional Brownian Motions, Fractional Noises and Applications", *SIAM Review*, Vol. 10, pp. 422–37, 1968.

[MAND69a] B. B. Mandelbrot and J. R . Wallis, "Computer Experiments with Fractional Gaussian Noises", *Water Resources Res.*, Vol. 5, pp. 228–267, 1969.

[MAND69b] B. B. Mandelbrot and J. R . Wallis, "Some Long-Run Properties of Geophysical Records", *Water Resources Res.*, Vol. 5, pp. 321–40, 1969.

[MAND79] B. B. Mandelbrot and M. S. Taqqu, "Robust R/S Analysis of Long Run Serial Correlation", In *Proc. 42nd Session ISI*, Vol. XLVIII, Book 2, pp. 69–99, 1979.

[MAND83] B. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, New York, 1983.

[ORTE93] A. Ortega, M. W. Garrett and M. Vetterli, "Toward Joint Optimization of VBR Video Coding and Packet Network Traffic Control", In *Proc. Fifth International Workshop on Packet Video*, Berlin, Germany, March 1993.

[PANC94] P. Pancha and M. El Zarki, "MPEG Coding for Variable Bit Rate Video Transmission", *IEEE Commun. Mag.*, Vol. 32, No. 5, pp. 54–66, May 1994.

[PVW91] *Fourth International Workshop on Packet Video*, Kyoto, Japan, August 1991.

[WALL91] G. K. Wallace, "The JPEG Still Picture Compression Standard", *Commun. of the ACM*, Vol. 34, No. 4, pp. 31–44, April 1991.