ELEC 599 Project Report
Queuing Analysis of Long-Range Dependent Traffic

Vinay J. Ribeiro
Department of Electrical and Computer Engineering
Rice University
Email: vinay@rice.edu

February 24, 1999

Abstract

Network traffic has been empirically shown to possess long-range dependence (LRD). LRD has important implications on the design, control and analysis of networks. This is mainly due to the queuing behavior of LRD traffic which is drastically different to that predicted by classical traffic models. Modeling of LRD traffic assumes great importance in this light. One such model, the multifractal wavelet model (MWM) has been tested for its validity as a traffic model in this project. Results indicate that it matches the queuing behavior of traffic well.

Experiments have been carried out to shed light on whether or not a single parameter H (the Hurst parameter) is sufficient to describe the queuing performance of LRD traffic. Results indicate the negative.

LRD is only one property of traffic that can affect queuing. Queuing experiments to study the effect of marginals on queuing reveal that heavy-tailed marginals give rise to more packet losses which leads us to conclude that second-order statistics alone cannot suffice to describe the queuing behavior of traffic.

1 Introduction

Ethernet data traffic has been convincingly shown to possess fractal or self-similar properties (structural similarities over a wide range of scales) in the seminal paper by Leland et. al. [1]. Other empirical studies have provided concrete evidence that such traffic exists in wide-area networks also [2]. Intuitively, the critical characteristic of self-similar traffic is that there is no natural length of a *burst*. Due to this *burstiness* the tail queue probability of a queue with self-similar traffic as input decays much slower than would normally be expected from classical queuing analysis and this has serious implications for the design, control, and analysis of high-speed networks [3]. For example, Norros showed that in the case of fractional Brownian motion (fBm) (a self-similar process) input to a single-server queue, a substantial increase in utilization for the same loss probability would require a tremendous increase in buffer size [4]. This peculiar queuing behavior is associated with the long-range dependence (LRD) property (or Joseph effect[1]) of self-similar processes. Thus, admission control and other congestion control algorithms that strive to guarantee certain quality of service (QoS) requirements of connections cannot afford to ignore LRD.

Self-similar processes are mono-fractal and the degree of self-similarity or measure of burstiness is given by a single parameter H (the Hurst parameter). Self-similar traffic models like fBm, attempt to capture the LRD inherent in data traffic. However, network traffic cannot always be expected to be self-similar and a question of significant importance that needs to be answered is whether or not the single parameter H can accurately describe the queuing behavior of LRD network traffic. Queuing experiments that we conducted for this purpose indicate the negative.

LRD is only one feature of traffic with fractal properties. As a first step in understanding the effect of other traffic characteristics on queuing, we have carried out experiments to unravel the effect of the marginal density on queuing. The results suggest that traces with heavy-tailed marginals give rise to larger buffer overflow probabilities.

We used traffic models (developed at Rice by Matthew Crouse) in the simulation experiments that can synthesize traces given the required correlation structure. We performed queuing experiments to ascertain the validity of one of these models called the multifractal wavelet model (βMWM) [5] and it was shown to match the queuing behavior of real data well.

The rest of this report is organized as follows. Sections 2 to 6 give a brief overview of the vast field of LRD and its impact on networking. After defining LRD and second-order self-similarity in Section 2, two models for LRD traffic are introduced in Section 3, one of which is fBm. We then study the immense impact of LRD on queuing in Section 4. Of the myriad methods for estimating the intensity of LRD, we introduce three methods in Section 5. We also introduce wavelets in Section 5.3 which are used in one of the three estimators and also in the multifractal wavelet model (MWM) that is described in Section 7. The issue of non-stationarity vs. LRD is touched upon in Section 6. Queuing experiments to validate the MWM as a traffic model are described in Section 8 and further tests with the aim of understanding the queuing behavior of LRD traffic better are presented in Section 9. Finally, we end with the conclusions in Section 10 and plot a future course of action in Section 11.

2 Long-Range Dependence (LRD) and Second-Order Self-Similarity

2.1 Definition of LRD

The presentation below follows a review of LRD by Cox [6]. Consider a discrete-time random process $\{X_t, \quad t \in \mathbb{Z}\}$, that is second-order stationary with auto-covariance function

$$r_X[k] = \mathrm{cov}(X_t, X_{t+k}), \tag{1}$$

[1] A reference to Joseph, a Biblical figure who foretold of the 'seven fat years and the seven lean years' that ancient Egypt was to experience.

and power spectral density function

$$S_X(f) = \sum_{k=-\infty}^{\infty} r_X[k]\ e^{-j2\pi fk}. \tag{2}$$

Consider a new series X_t^m formed by averaging the original series X_t in non-overlapping blocks of length m, replacing each block by its mean. That is,

$$X_t^m \ = \ \frac{X_{tm-m+1} + ... + X_{tm}}{m}. \tag{3}$$

Let the auto-covariance of X_t^m be given by $r_X^m[k]$. Processes that satisfy the following essentially equivalent conditions are termed *short-range dependent* (SRD).

1. $\sum_{k=0}^{\infty} r_X[k]$ is convergent.

2. $S_X(0)$ is finite.

3. $r_X^m[0]$ is for large m asymptotically of the form $\frac{v'}{m}$ where v' is a constant.

For such processes, the averaged process X_t^m tends to second order pure noise as $m \ \to \ \infty$. The above conditions are satisfied by all stationary autoregressive-moving average processes (ARMA) of finite order, i.e., processes with rational power spectral densities. By contrast *LRD* processes are those which satisfy the following equivalent conditions.

1. $\sum_{k=0}^{\infty} r_X[k]$ is divergent.

2. $S_X(f)$ is singular near $f \ = \ 0$.

3. $m\ r_X^m[0] \ \to \ \infty$ as $m \ \to \ \infty$.

From a statistical point of view, the salient feature of LRD processes is that the variance of a mean decreases more slowly than the reciprocal of sample size. From an intuitive point, possibly the most enlightening property is that the averaged process X_t^m takes a non-degenerate correlation structure as $m \ \to \ \infty$.

2.2 Second-Order Self-Similarity

A process X_t is called exactly second-order self-similar if it has an identical covariance structure as the averaged process X_t^m, i.e.,

$$\frac{r_X^m[k]}{r_X^m[0]} = \frac{r_X[k]}{r_X[0]},\ \ \forall m, k \in \mathbb{N}. \tag{4}$$

This similarity on different time-scales (aggregation levels) gives the processes their name. The only solution to equation (4) is

$$\frac{r_X[k]}{r_X[0]} = \frac{1}{2}\delta^2(k^{2H}), \quad \forall k \in \mathbb{N}, \ \ 0 \ < \ H \ < \ 1, \tag{5}$$

where $\delta^2(f)$ denotes the second central difference operator applied to a function f, i.e., $\delta^2(f(k)) = f(k+1) - 2f(k) + f(k-1)$. Thus, for exactly second-order self-similar processes, the power spectral density is of the form

$$S_X(f) \sim \frac{c}{f^{2H-1}}, \tag{6}$$

for f near 0, where c is a constant. Due to this property, self-similar processes are often loosely called $\frac{1}{f}$ processes. One can roughly say that the closer the value of H is to 1, the more bursty the process is. When $H = 1/2$, the process is white noise. For $0 < H < 1/2$ the auto-covariance goes negative. Processes with $1/2 < H < 1$ exhibit LRD and we will be concerned with only this case.

The process X_t is called asymptotically second-order self-similar with parameter H (Hurst parameter) if

$$\frac{r_X^m[1]}{r_X^m[0]} \rightarrow 2^{2H-1} - 1, \quad \text{as } m \rightarrow \infty, \tag{7}$$

and

$$\frac{r_X^m[k]}{r_X^m[0]} \rightarrow \frac{1}{2}\delta^2(k^{2H}), \text{as } m \rightarrow \infty \text{ and } k = 2, 3, ... \tag{8}$$

Thus, exactly second-order self-similar processes have correlation structure that is determined by a single parameter H. The same goes for asymptotically second-order self-similar processes for large m.

3 Stochastic Modeling of LRD traffic

In order to better understand the impact of LRD traffic on various aspects of networks (design of networks, queuing etc.), LRD traffic models are essential. In this section, two such models are described.

3.1 Superposition of ON/OFF sources

The superposition of ON/OFF sources with strictly alternating ON and OFF periods and whose ON and OFF periods exhibit the *Noah effect* [2] (i.e. have high variability or infinite variance) produces aggregate network traffic that exhibits LRD [7]. The Noah effect guarantees that each ON/OFF source individually exhibits characteristics that cover a wide range of time scales (or aggregation levels: see Section 2.2). This is because the source can be in the ON state or OFF state for a very large time with non-negligible probability, which gives rise to strong correlations over large lags.

Consider a single source and focus on the stationary binary time series $\{W(t),\ t \geq 0\}$ it generates. $W(t) = 1$ means that there is a packet at time t and $W(t) = 0$ means that there is no packet. The length of the ON-periods are i.i.d., those of the OFF-periods are i.i.d., and the lengths of ON and OFF-periods are independent. An OFF period always follows an

[2] A reference to Noah, a Biblical figure waited 40 days and nights for a mighty flood to stop and several more for the rain water to disappear and leave dry land.

ON-period. Suppose now that there are M i.i.d. sources. Since each source sends its own sequence of packet trains (traffic) it has its own reward sequence $\{W^{(m)}(t), t \geq 0\}$ where m identifies the source. Rescaling time by a factor T, consider

$$W_M(Tt) = \int_0^{Tt} \left(\sum_{m=1}^{M} W^{(m)}(u) \right) \mathrm{d}u, \tag{9}$$

the aggregated cumulative packet counts in the interval $[0, Tt]$. With an appropriate choice for the distributions of the ON and OFF periods, $\{W_M(Tt), t \geq 0\}$ adequately normalized is $\{\sigma_L B_H(t),\ t \geq 0\}$, where σ_L is a finite positive constant and B_H is fractional Brownian motion which is described later in Section 3.2. For example, if the complementary (or tail) distribution of the ON and OFF periods are

$$F_{ONc}(x) \ = \ K^{\alpha_{\mathrm{ON}}} x^{-\alpha_{\mathrm{ON}}}, \tag{10}$$

$$F_{OFFc}(x) \ = \ K^{\alpha_{\mathrm{OFF}}} x^{-\alpha_{\mathrm{OFF}}}, \tag{11}$$

for $x \geq K > 0$, $1 < \alpha_{\mathrm{ON}}, \alpha_{\mathrm{OFF}} < 2$ and equal 0 for $x < K$, then $\{W_M(Tt), t \geq 0\}$ behaves statistically like fBm with Hurst parameter $H = \frac{3-\alpha_{\min}}{2}$. The parameter $\alpha_{\min}$ is the smaller of α_{ON} and α_{OFF}. Analysis of Ethernet traffic [7] showed that individual source-destination pairs behaved like the above mentioned ON/OFF sources and this could be one reason for the LRD of Ethernet traffic.

3.2 Fractional Brownian Motion and Fractional Gaussian Noise

Fractional Brownian Motion (fBm) is a moving average of $\mathrm{d}B(t)$, where $B(t),\ t \in \mathbb{R}$ is ordinary Brownian motion, in which past increments of $B(t)$ are weighted by the kernel $(t-s)^{H-1/2}$ for $0 < H < 1$ [8]. Brownian motion, $B(t)$, is real random function with independent Gaussian increments such that $B(t_1) - B(t_2)$ has mean zero and variance $|t_2 - t_1|$, and such that $B(t_1) - B(t_2)$ is independent of $B(t_4) - B(t_3)$ if the intervals (t_1, t_2) and (t_3, t_4) do not overlap. Reduced fBm for $0 < H < 1$ and starting value b_0 at time 0 is defined by

$$B_H(0) = b_0, \tag{12}$$

$$B_H(t) - B_H(0) = \frac{1}{\Gamma(H+\frac{1}{2})} \int_{-\infty}^{0} [(t-s)^{H-\frac{1}{2}} - (-s)^{H-\frac{1}{2}}] \, \mathrm{d}B(s) + \int_0^t (t-s)^{H-\frac{1}{2}} \mathrm{d}B(s), \tag{13}$$

for $t > 0$. Intuitively, for every value of t, $B_H(t)$ depends on the entire past values of $B(s)$ and thus fBm has strong correlations over large lags.

A formal definition of self-similarity (not to be confused with 2^{nd} order self-similarity of Section 2.2) is as follows. The increments of a random function $\{Y(t), t \in \mathbb{R}\}$ are said to be self-similar with parameter H ($H \geq 0$) if for any $h > 0$ and any t_0,

$$Y(t_0 + \tau) - Y(t_0) \ \stackrel{\mathrm{fd}}{=} \ h^{-H} \left[Y(t_0 + h\tau) - Y(t_0) \right], \tag{14}$$

where $\stackrel{\mathrm{fd}}{=}$ indicates an equality in finite-dimensional distributions. The increments of fBm are stationary and self-similar with parameter H (called H-sssi). The only non constant

Gaussian random function that is H-sssi and mean square continuous[3] is fBm. However, fBm is non-stationary. This can easily be seen from its autocorrelation function which is given by

$$r_{B_H}(t,s) = \frac{\sigma^2}{2}\left[|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right] \tag{15}$$

where σ^2 is the variance of $B_H(1)$. Its power spectral density is given by

$$S_{B_H}(f) = \frac{c}{|f|^{2H+1}}, \tag{16}$$

where c is a constant. Also, fBm does not have a derivative and one works with sampled continuous-time fBm for practical computations. The discrete-time process thus formed is called fractional Gaussian noise (fGn) and is given by

$$B_G[n] = B_H(n\Delta t) - B_H((n-1)\Delta t), \quad n \in \mathbb{Z} \tag{17}$$

and has autocorrelation structure

$$r_{B_G}[k] = \frac{\sigma^2}{2}|\Delta t|^{2H}[|k+1|^{2H} + |k-1|^{2H} - 2|k|^{2H}], \quad k \in \mathbb{N} \tag{18}$$

and power spectral density

$$S_{B_G}(f) \ \sim \ \frac{c}{|f|^{2H-1}}. \tag{19}$$

near $f = 0$. In fact, fGn is an exactly second-order self similar process. Clearly, since fGn is a stationary Gaussian process, it is defined completely by its mean and autocovariance (which is specified by its variance and H). Thus, fGn can capture the LRD of real world phenomena but is incapable of imbibing the short-range dependence (SRD) structure of signals due to its rigid correlation structure. However, there exist other traffic models like the fractional autoregressive integrated moving average (F-ARIMA) model [6][10], which can capture both LRD and SRD[1].

4 Effect of LRD on Queuing

The LRD property of traffic has not only statistical significance in traffic measurements (for e.g., the sample mean has high variance if the data points are few) but also has a considerable impact on queuing performance, and is a dominant characteristic for a number of packet traffic engineering problems.

4.1 Experimental evidence

The effect of LRD on queuing was empirically demonstrated in [3] with the help of actual Ethernet traffic traces. The correlation of the traces at large lags (LRD) was destroyed through shuffling of packet inter-arrival times, with the purpose of approximately keeping the

[3]For defn. of m.s. continuity see pg. 336 of [9].

marginal distribution the same. The resulting trace exhibited much smaller average queuing delays than the original trace for link utilization above 0.4. This was for a single-server infinite buffer queue with deterministic service. This result suggests that the burstiness of traffic cannot simply be explained by renewal arrival processes with heavy tailed distributions for the inter-arrival times (not to be confused with aggregation of such processes as in Section 3.1). This follows from the fact that in the experiment the marginal distribution of inter-arrival times was not changed.

In [3], LRD was demonstrated to be the primary cause for the larger queuing delays and not the SRD. This was done by dividing the sequence of inter-arrival times into blocks of length m where $m \ll N$ (N = the total number of data points of the trace). The data within each block was then randomly shuffled and the position of each block in the trace left unchanged, with the purpose of destroying only the SRD. The resulting trace displayed queuing behavior strikingly similar to that of the original trace. When the blocks were shuffled keeping the data within each block unchanged,with the purpose of destroying only the SRD, the resulting trace had a markedly different average queue delay than the original trace.

4.2 Theoretical tail queue probability

Consider an unlimited buffer with constant leak rate (or link rate) $c > 0$ to which a LRD input process is fed [4]. The cumulative arrivals process (continuous time) is given by

$$A_t = mt \; + \; (am)^{1/2} Z_t, \quad t \in \mathbb{R} \tag{20}$$

where Z_t is normalized fBm with Hurst parameter H, m is the mean rate and a is the index of dispersion at time unit ($t = 1$). The stationary storage process (buffer occupancy) is defined as

$$V_t^{(m,a,H,c)} = \sup_{s \le t} \left(A_t - A_s - c(t-s) \right), \quad t \in \mathbb{R} \tag{21}$$

We also denote

$$V_t^{(b)} = \sup_{s \le t} \left(Z_t - Z_s - b(t-s) \right), \quad t \in \mathbb{R} \tag{22}$$

Using the self-similarity of Z, it is shown in [4] that it is sufficient to study the process $V^{(1)}$ in order to obtain $V^{(m,a,H,c)}$. This can be stated as

$$\mathrm{P}(V^{(m,a,H,c)} > x) \; = \; \mathrm{P}\left(V^{(1)} > \frac{x}{(am)^{1/2}} \left(\frac{(c-m)}{(am)^{1/2}} \right)^{\frac{H}{1-H}} \right). \tag{23}$$

From the theory of large deviations the following approximate stationary distribution of $V^{(1)}$ is obtained.

$$\mathrm{P}(V^{(1)} > x) \; \approx \; e^{\frac{-x^{2-2H}}{2\kappa(H)^2}}, \tag{24}$$

where $\kappa(H) = H^H(1-H)^{1-H}$. The Weibull tail distribution for the storage process given by equation(24) is in stark contrast to the exponential distributions obtained for processes with only SRD like finite state Markovian processes, given by equation(25).

$$\mathrm{P}(V^{(1)} > x) \;\approx\; e^{-\delta x}, \tag{25}$$

where δ depends on the service rate at the queue and the statistical properties of the arrivals process.

This peculiar queuing behavior of LRD processes clearly shows that ignoring LRD will give rise to very optimistic expectations of queuing performance which can have dire consequences in a network that makes important quality of service (QoS) guarantees (e.g. delay bound, delay jitter etc.) to connections. Neglecting LRD in the design of networks also takes its toll. For example, it was observed [1] that overall packet loss decreases very slowly with increasing buffer capacity, in sharp contrast to Poisson-based models where losses decrease exponentially as suggested by equation(25). Also, packet delay was found to always increase with buffer capacity, in contrast to classical models where delay does not exceed a fixed limit regardless of buffer size.

5 Statistical methods for testing for self-similarity

As mentioned in Section 2.2 the parameter H roughly gives an estimate of the 'intensity' of the LRD of a process. Thus an estimate of H can prove very useful in problems that require decisions or predictions to be made based on the LRD of a process. Not surprisingly, estimation of H is an important research area by itself. A few methods for estimating H are presented here. Details of a plethora of techniques for the same can be found in [11].

In Section 2.1 we saw that the LRD of a process is manifest in

1. the auto-covariance structure,
2. the power spectral density and
3. in the variances that decay slowly with aggregation level.

Thus, the problem of estimating H can be approached from these three angles. We use the same notation here as that used in Section 2.

5.1 Variance-time plot method

This method makes use of the third property mentioned above. For a second-order self-similar process $\{X_t, \;\; t \in \mathbb{Z}\}$, the variance of the aggregated process $\{X_t^m\}$, behaves as follows.

$$\frac{r_X^m[0]}{r_X[0]} \;\sim\; m^{2H-2}, \;\; m \;\rightarrow\; \infty. \tag{26}$$

For exactly second-order self-similar process the approximation is exact. Thus, we can obtain an estimate of $2H-2$ by plotting the logarithm of variance of the process at different scales against the logarithm of scale $\log(m)$ (called the variance-time plot (VTP)). In actual practice we have only a single sample function of the process X_t and so we compute estimates of $r_X^m[0]$ as follows.

$$r_X^m[0] \;=\; \frac{m}{N}\sum_{k=1}^{\frac{N}{m}}\left(X^{(m)}(k)\right)^2 \;-\; \left(\frac{m}{N}\sum_{k=1}^{\frac{N}{m}}X^{(m)}(k)\right)^2 \tag{27}$$

where N is the total number of data points of X_t (here assumed to be a multiple of m). The values of m are chosen such that they are equidistant on a log-scale. In practice, we estimate the slope by fitting a line to the points obtained from the VTP. The drawback of this method is that due to strong correlations in the data, the estimates may not be accurate. However, if N and m are large and $N \gg m$, the estimates will be fairly accurate.

The VTP contains all the auto-covariance information of a wide-sense stationary process as given by equation(28).

$$r_X[k] = \frac{1}{2}\delta^2(h^2 r_X^m[0]), \tag{28}$$

where δ^2 denotes the central second difference operator defined in Section 2.2.

5.2 The R/S method

For a process $\{X_t,\quad t \in \mathbb{N}\}$, with partial sum $Y(n) \;=\; \sum_{i=1}^n X_i$ and sample variance $S^2(n) := (1/n)\sum_{i=1}^n X_i^2 \;-\; (1/n)^2 Y(n)^2$, the R/S statistic, or the *rescaled adjusted range*, is given by:

$$\frac{R}{S}(n) \;:=\; \frac{1}{S(n)}\left[\max_{0\le t\le n}\left(Y(t)-\frac{t}{n}Y(n)\right) - \min_{0\le t\le n}\left(Y(t)-\frac{t}{n}Y(n)\right)\right] \tag{29}$$

For fractional Gaussian noise (fGn),

$$\mathbb{E}\left[\frac{R}{S}(n)\right] \;\sim\; C_H n^H,\quad n\to\infty, \tag{30}$$

where C_H is a positive, finite constant not dependent on n.

The procedure for determining H using the R/S statistic is spelled out below. Subdivide a given time series of length N into K blocks, each of size N/K. For each lag n, compute $R(k_i,n)/S(k_i,n)$ starting at points $ki = iN/K \;+\; 1,\; i \in \mathbb{N}$ such that $k_i \;+\; n \le N$. For values of n smaller than N/K, we get K different estimates of $R(n)/S(n)$. For values on n approaching N, we get fewer values, as few a 1 when $n \ge N - N/K$. Choosing logarithmically spaced values of n, $\log[R(k_i,n)/S(k_i,n)]$ is plotted vs. $\log n$. The resulting plot is called a *pox plot* for the R/S statistic. We estimate the parameter H by fitting a line to the points in the pox plot. We usually do not use the low and high ends of the plot (i.e. low and high values of n) in estimating H because the low end is affected by the short-term correlations and the high end will have too few points for each n.

5.3 A wavelet based method

This method makes use of the power spectral density of the process. First, we begin with a short introduction to wavelets. Wavelets will also be used later in this report in Section 7.

5.3.1 Wavelets

The discrete wavelet transform is an atomic decomposition that represents a one-dimensional signal $x(t)$ in terms of shifted and dilated versions of a prototype bandpass wavelet function $\psi(t)$, and shifted versions of a low-pass scaling function $\phi(t)$ [12]. For special choices of the wavelet and scaling functions, the atoms

$$\psi_{j,k}(t) \;:=\; 2^{-j/2}\,\psi\Big(2^{-j}t-k\Big), \qquad \phi_{j,k}(t) \;:=\; 2^{-j/2}\,\phi\Big(2^{-j}t-k\Big), \qquad j,k\in\mathbb{Z} \tag{31}$$

form an orthonormal basis, and we have the signal representation

$$x(t) \;=\; \sum_k u_k\,\phi_{J_0,k}(t) \;+\; \sum_{j=-\infty}^{J_0}\sum_k w_{j,k}\,\psi_{j,k}(t), \tag{32}$$

with $w_{j,k} := \int x(t)\,\psi^*_{j,k}(t)\,dt$, and $u_{j,k} := \int x(t)\,\phi^*_{j,k}(t)\,dt$. For a wavelet $\psi(t)$ centered at time zero and frequency f_0, the *wavelet coefficient* $w_{j,k}$ measures the signal content around time 2^jk and frequency $2^{-j}f_0$. The *scaling* coefficient $u_{j,k}$ measures the local mean around time 2^jk. In the wavelet transform, j indexes the *scale* or resolution of analysis: J_0 indicates the coarsest scale or lowest resolution of analysis, and smaller j corresponds to higher resolution analysis. The wavelet transform synthesis equation(32) employs scaling coefficients only at scale J_0; scaling coefficients at scales $j<J_0$ represent higher resolution approximations to the signal. Any filter bank or lifting wavelet transform implementation produces all of the scaling coefficients $u_{j,k}$, $j<J_0$, as a natural byproduct [12].

5.3.2 A Wavelet-based estimate of H

As stated in Section 2.2, the power spectral density of a second-order self-similar process is given by

$$S_X(f) \;\sim\; \frac{c}{f^{2H-1}}, \tag{33}$$

for f near 0 where c is a constant. As stated in Section(5.3.1), the wavelet coefficient $w_{j,k}$ measures the signal content around time 2^jk and frequency $2^{-j}f_0$. Thus, a useful spectral estimator can be designed by performing a time average of the $|w_{j,k}|^2$ at a given scale, i.e.,

$$\hat{S}_X(2^{-j}f_0) \;=\; \frac{1}{n_j}\sum_k |w_{j,k}|^2, \tag{34}$$

where n_j is the number of wavelet coefficients at octave j. $\hat{S}_X(f)$ is therefore a measure of the amount of energy that lies within a given bandwidth around the frequency f and can therefore be regarded as a statistical estimator for the spectrum $S_X(f)$ [13]. When the process X_t is wide sense stationary, the expectation of the spectral estimator is

$$\mathbb{E}\left(\hat{S}_X(2^{-j}f_0)\right) \;=\; \int S_X(f)2^j|\Psi(2^jf)|^2\,\mathrm{d}f, \tag{35}$$

where Ψ denotes the Fourier transform of the analyzing wavelet $\psi(t)$. Equation(35) reveals that $\hat{S}_X$ suffers from a convolutive bias, that is, the spectrum to be estimated is mixed within a frequency range corresponding to the frequency width of the analyzing window at scale j. The crucial point here is that for second-order self-similar signals this bias reduces naturally to a simple form, enabling an unbiased estimation of H. This can be seen from equations (33) and (35) which give

$$\begin{aligned} \mathbb{E}\left(\hat{S}_X(2^{-j}f_0)\right) &= S_X((2^{-j}f_0)|f_0|^{(2H-1)} \textstyle\int |f|^{(1-2H)}|\Psi(f)|^2 \, \mathrm{d}f && (36) \\ &= c\; 2^{j(2H-1)} \textstyle\int |f|^{(1-2H)}|\Psi(f)|^2 \, \mathrm{d}f && (37) \end{aligned}$$

It is therefore possible to design an estimator $\hat{H}$ for the parameter H from a simple linear regression of $\log_2 \hat{S}_X(2^{-j}f_0)$ on j, that is,

$$\log_2 \hat{S}_X(2^{-j}f_0) = \log_2\left(\frac{1}{n_j}\sum_k |w_{j,k}|^2\right) = (2\hat{H}-1)j + \hat{c_f} \tag{38}$$

where c_f estimates

$$\log_2\left(c \textstyle\int |f|^{(1-2H)}|\Psi(f)|^2 \, \mathrm{d}f\right)$$

provided that the integral

$$\int |f|^{(1-2H)}|\Psi(f)|^2 \, \mathrm{d}f \tag{39}$$

converges. Performing a weighted least squares fit between the scales j_1 and j_2 yields the following explicit formula for the estimator of H:

$$\hat{H}(j_1,j_2) \;:=\; \frac{1}{2}\left[\frac{\sum_{j=j_1}^{j_2} S_j j \eta_j - \sum_{j=j_1}^{j_2} S_j j \sum_{j=j_1}^{j_2} S_j \eta_j}{\sum_{j=j_1}^{j_2} S_j \sum_{j=j_1}^{j_2} S_j j^2 - (\sum_{j=j_1}^{j_2} S_j j)^2} + 1\right] \tag{40}$$

where

$$\eta_j = \log_2(\tfrac{1}{n_j}\textstyle\sum_k |w_{j,k}|^2)$$

and the weight $S_j = \frac{n \ln^2 2}{2^{j+1}}$ is the inverse of the theoretical asymptotic variance of η_j (n=number of data points). For the integral (39) to converge, the number of vanishing wavelet moments N_ψ must be greater than $H-1$, i.e.,

$$\int t^k \psi(t) \, \mathrm{d}t = 0 \tag{41}$$

for $k = 0, 1, 2, ..., N_\psi - 1$. This condition is trivial because by definition $\int \psi(t)\, dt = 0$ and $H < 1$.

It is interesting to note that the estimator $\hat{H}$ is asymptotically unbiased and has very low bias even for short data sets. Another fact of great importance [14] is that for a second-order self-similar process with parameter H,

$$\mathbb{E}(w_{j,k}w_{j,k'}) = O(|k = k'|^{2H-2-2N_\psi}), \quad |k = k'| \to \infty. \tag{42}$$

Thus strongly correlated data gets decorrelated in the wavelet domain. Thus, the sample variance of the wavelet coefficients is much more accurate when compared to the sample variance of the data.

6 Non-stationarity vs. LRD

One of the typical features of stationary long-memory processes is that there appear to be local trends and cycles [10], which are, however, spurious and disappear after some time. This property can make it rather difficult to distinguish a stationary LRD process from a non-stationary process. This problem becomes more severe if the observed time-series is short. Additional non-numerical information about the process can help in detecting non-stationarities. Also, simple persisting trends can be distinguished from stationarity, if we wait long enough. For instance, we can distinguish a linear trend from a stationary process, if the sample size is large. There is a vast range of possible non-stationarities and it is therefore not possible to decide for sure whether the underlying process in indeed stationary. Ultimately, this is a rather philosophical question that might not be of primary practical interest. From the practical point of view, the assumption of stationarity provides a simple framework for modeling traffic. Unless there is evidence against it, it tends to be more useful than complicated non-stationary models. Subtle trends like slowly decaying polynomials can be confused with LRD [10]. These can be eliminated by the wavelet based estimator of H of Section 5.3 (due to the zero wavelet moments) [13]. For other estimators that are not affected by such trends, the interested reader is referred to [10].

7 The Multifractal Wavelet Model (MWM)

For modeling LRD process, fractional Brownian motion (fBm) has played a central role in many fields. Although fBm is a tractable signal model, its strict self-similarity is too restrictive to adequately characterize many types of signals. For example: (1) many signals exhibit a self-similar scaling behavior over one range of resolutions, but then a totally different scaling behavior over another range; (2) many signals obey different types of scaling behavior in different regions of the signal; (3) many signals possess significant long-range dependence, but display short-term correlations inconsistent with strict self-similarity; (4) in many signals, different moments will scale with different exponents; and (5) many signals have increments that are inherently positive and hence non-Gaussian. Thus, self-similar models are not always rich enough to describe complex real-world signals accurately.

The *multifractal wavelet model* (MWM) [5] is a multifractal signal model especially suited to positive LRD data. The MWM has many attractive features. Some of these are mentioned below.

1. It provides a guaranteed positive output.

2. The LRD is easily tunable (including, but not limited to, a strict power-law correlation decay).

3. Computations involving the MWM are extremely efficient — synthesis of a trace of N sample points requires only $O(N)$ computations.

4. The parameters of the MWM are simple enough to be easily inferred from observed data or chosen a priori.

7.1 Wavelet construction of positive process

Since wavelets are self-similar by their very construction [12], they are well-suited for analysis and synthesis of random fractals. As mentioned in Section 5.3, highly-correlated, LRD signals become nearly uncorrelated in the wavelet domain. If the de-correlation were exact, then specifying the correct variances of the wavelet coefficients would fully capture the correlation structure of the signal. Thus by modeling in the wavelet domain, we can approximate a process simply by characterizing how its wavelet coefficient energy varies as a function of scale. The notation used in this section is the same as that of Section 5.3.1 where an introduction to wavelets was given.

The basic idea behind the MWM is simple: The Haar wavelet transform combined with a positivity constraint on the data naturally results in a binomial multifractal [15]. In a Haar wavelet transform, the scaling and wavelet coefficients can be recursively computed using

$$u_{j+1,k} = 2^{-1/2}(u_{j,2k} + u_{j,2k+1}) \tag{43}$$

and

$$w_{j+1,k} = 2^{-1/2}(u_{j,2k} - u_{j,2k+1}). \tag{44}$$

Capturing non-negativity is straightforward for the Haar wavelet transform. The dependency between the wavelet coefficient $W_{j,k}$[4] and its corresponding scaling coefficient $U_{j,k}$ is given by

$$U_{j,2k} = 2^{-1/2}(U_{j+1,k} + W_{j+1,k}) \tag{45}$$

and

$$U_{j,2k+1} = 2^{-1/2}(U_{j+1,k} - W_{j+1,k}). \tag{46}$$

For non-negative signals, $U_{j,k} \geq 0$, implying that

$$|W_{j,k}| \leq U_{j,k}. \tag{47}$$

In addition, since each wavelet coefficient is obtained via a bandpass filtering of a stationary process, we observe that $\mathbb{E}[W_{j,k}] = 0$ and that $W_{j,k}$ will be distributed symmetrically around 0.

These constraints lead to a very simple multiplicative signal model for positive processes. Let $A_{j,k}$ be symmetric random variables supported on the interval $[-1, 1]$ and define the wavelet coefficients recursively by

$$W_{j,k} = A_{j,k}\, U_{j,k}. \tag{48}$$

This ensures that $|W_{j,k}| \leq U_{j,k}$ and, hence, that the corresponding process is non-negative. By specifying the densities for the $A_{j,k}$'s and the coarsest scaling coefficients $U_{J_0,k}$ we, thus, specify the joint density function for the entire wavelet transform.

[4] Capital letters are used since the underlying signal X is random.

7.2 Synthesis of traffic via matching

We synthesize traffic using the MWM by making use of the statistics of a given trace. First, we take a wavelet transform of the given trace, and from this we estimate the variance of wavelet coefficients at different scales. The multipliers $A_{j,k}$'s are required to be symmetric about 0 for stationary processes and we obtain their variances easily from those of the $W_{j,k}$'s [5]. Since β distributions are used for the $A_{j,k}$'s the resulting model is christened the βMWM. We then generate synthetic traces by the process outlined in Section 7.1 which can be visualized as a course-to-fine synthesis. The coarsest scaling coefficients $U_{J_0,k}$ are modeled using multivariate densities with strictly non-negative support or by selecting their values deterministically to exactly match the overall mean. Once we know the $U_{J_0,k}$, we form the coarsest wavelet coefficients $W_{J_0,k}$ via equation(48). Through equations (45) and (46) we combine $U_{J_0,k}$ and $W_{J_0,k}$ to form $U_{J_0-1,k}$, the scaling coefficients at level J_0-1. We then proceed to form $W_{J_0-1,k}$, the wavelet coefficients at level J_0-1, via (48). We repeat this process until the finest scale ($j=J_1$) wavelet coefficients and scaling coefficients (the signal $X[k]=U_{J_1,k}$) are generated. We then invert the synthesized wavelet transform to obtain the required data trace.

7.3 Multifractal properties

Many real-world phenomena (including network traffic [16]) are *multifractal.* Roughly speaking, in a multifractal, the scaling H in equation(14) changes erratically with each t_0. Multifractal analysis aims to describe this inherent, ubiquitous occurrence of irregularities that govern the shape and complexity of multifractal signals and hence can be used to find useful information about network traffic.

The multiplicative structure of the construction of the MWM is very typical of a multifractal, for example, the random binomial [15]. Consequently, the model exhibits multifractal properties and can match non-Gaussian processes beyond the second-order statistics matched by fBm.

8 Validation of the MWM: Queuing Experiments

A key test to assess the accuracy of the βMWM is to compare the queuing behavior of simulated traffic traces against real data. Standard tests were performed in order to evaluate the accuracy of the βMWM. The experiments were aimed at comparing the queuing performance of the traces generated by the βMWM with real network traffic. The tests made use of the August 1989 Bellcore Ethernet trace *pAug* of one-million inter-arrival times as measured by Leland et al. [1]. Inter-arrival times were modeled, because they do not take quantized values like packets per time and are thus better suited for the βMWM. From the variance-time plot, it was observed that the trace exhibits LRD with an average H of approximately 0.79. For simplicity all packets were assumed to be of equal length. Traces were synthesized using the method described in Section 7.2 In all the simulation experiments, the performance of an infinite length single server queue with a single trace as input was considered. A constant service rate of 500 packets/sec was assumed.

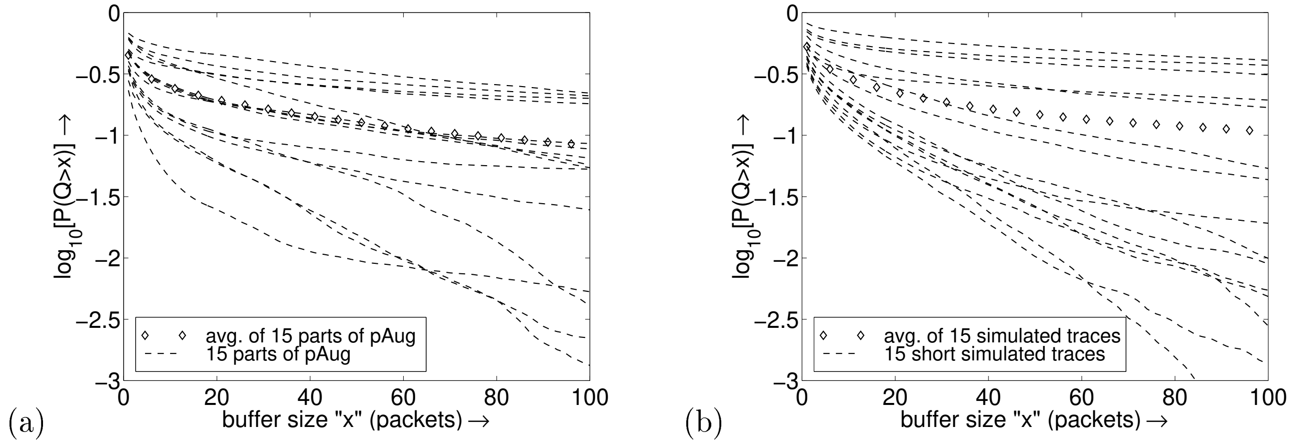


Figure 1: *Here the pAug trace was partitioned into 15 sub-traces of equal number of packets and their queueing behavior was compared with that of 15 synthesized traces of the same length. In (a), observe that the sub-traces have a wide variation in tail queue distributions. In (b), observe that the synthesized traces display a similar variation in tail queue distributions as the sub-traces do.*

To learn about the queuing behavior of the actual data trace that was modeled, we partitioned the actual trace into 15 sub-traces of equal length (2^{16} packets). We assume each sub-trace to be a different realization of the underlying process and expect their queuing behavior to resemble that of different sample functions of the process. We then used the βMWM to synthesize 15 traces each of 2^{16} packet inter-arrival times. The performance of the queue with the actual and βMWM sub-traces as input are shown in Figure 1. We observes that the sub-traces of the real data have widely varying queue distributions, a phenomenon that is well captured by the simulation traces.

The next experiment was conducted to compare the queuing performance of the entire real data trace with that of 20 traces of the same length, generated using the βMWM (see Figure 2). We observe that the simulated traces in Figure 2(a) exhibit a wide variation in tail queue distributions. The results of the previous experiment indicate that this is to be expected. We also observe that the average tail queue distribution of the simulated traces matches that of the real trace exceedingly well (see Figure 2(b)). However, as the previous experiment suggests, the real data can be viewed as just one possible realization of our model, and so it cannot be expected to always exhibit the same queuing behavior as the average of several simulated traces.

When the same variables as those in Figure 2(b) were plotted in loglog versus log coordinates, i.e. $\log(\log(\mathrm{P}(Q > x)))$ was plotted against $\log(x)$, a Weibull law was observed that is characteristic of LRD traffic (Section 4.2). This indicates that the simulated traces do exhibit queuing behavior consistent with that expected from modeling with fBm. However, the exponent that was found (≈ 0.2) was smaller than $2 - 2H \approx 0.4$ (see equation(24)).

In summary, the queuing experiments demonstrate that the synthesized traffic traces not only match real traffic in its various statistical properties but also in its queuing behavior.

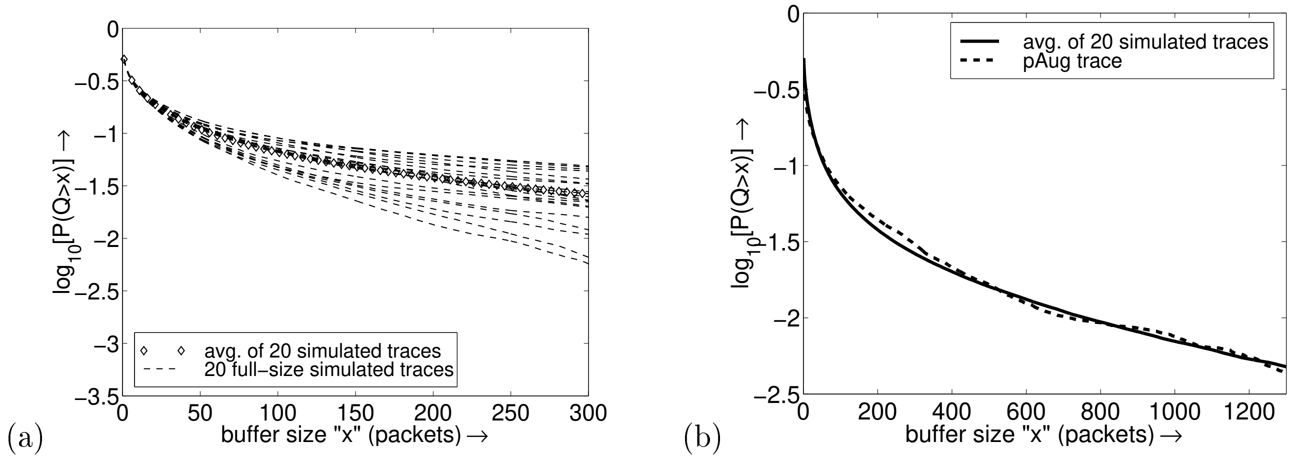


Figure 2: *Comparison of the queuing behavior of pAug with 20 full size synthesized traces. Displayed are the tail probabilities of buffer occupancy against buffer size. In (a), observe the variability of the queue performance of the synthesized trace. In (b), observe that the average queue performance of simulated traces and that of the real trace match closely.*

9 Further Queuing Experiments on LRD traffic

As seen in previous sections, a single parameter H has often been used to characterize LRD traffic. However, in practice we cannot always expect traffic to be strictly self-similar. Also, analytical results for queuing behavior of fBm assume Gaussian marginals, while traffic that is inherently positive cannot be Gaussian. Experiments have been carried out to throw more light on the queuing behavior of traffic when these assumptions do not hold, the details of which are given below.

In all the experiments described in this section, a traffic model [17] created by Matthew Crouse at Rice university, has been used. This model (let us name it FFTGEN) synthesizes traces with Gaussian marginals given the required autocovariance structure. All experiments involved calculation of the probability of overflow of single-server finite length buffers with lengths varying from 1 to 200 packets. A deterministic service rate of 50 packets/sec was assumed for each queue. Every trace used in the simulation experiments was input to all the queues and the ratio of number of packets dropped to the total number of packets input to the queue was obtained (P(buffer overflow)). The traces generated by FFTGEN were that of packets/unit time.

9.1 Effect of marginals on queuing

In each of these experiments, we held the mean and autocovariance structure of traces constant, while changing the marginals between Gaussian, lognormal and uniform densities. An fGn autocovariance structure for a given H was used. Though the traffic model can generate traces with Gaussian marginals only, it is possible to transform the Gaussian data to obtain other marginals. This is done as outlined below [17].

1. Given a mean, autocovariance structure and a required marginal density, modify the autocovariance and mean appropriately (pre-warping step).

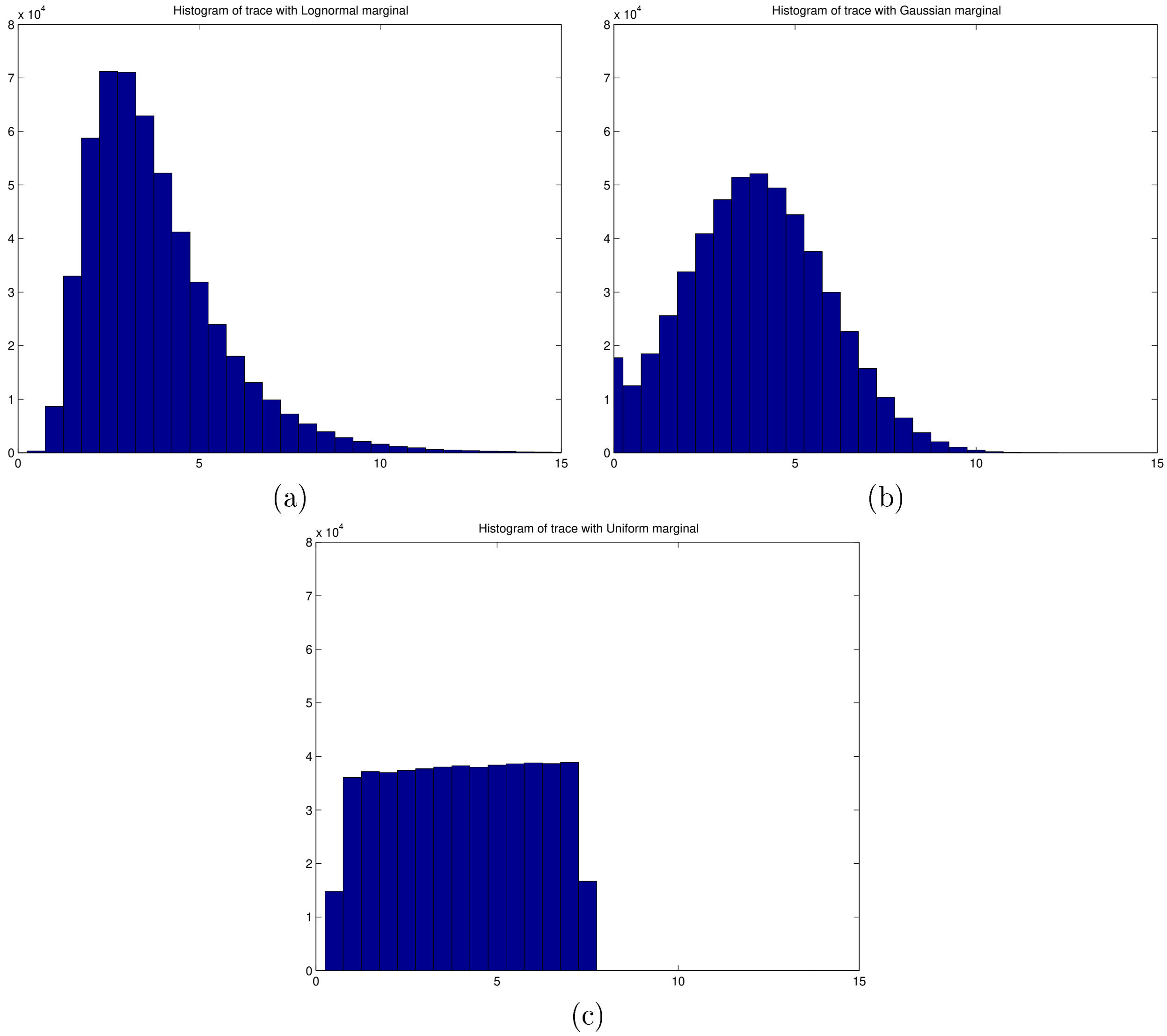


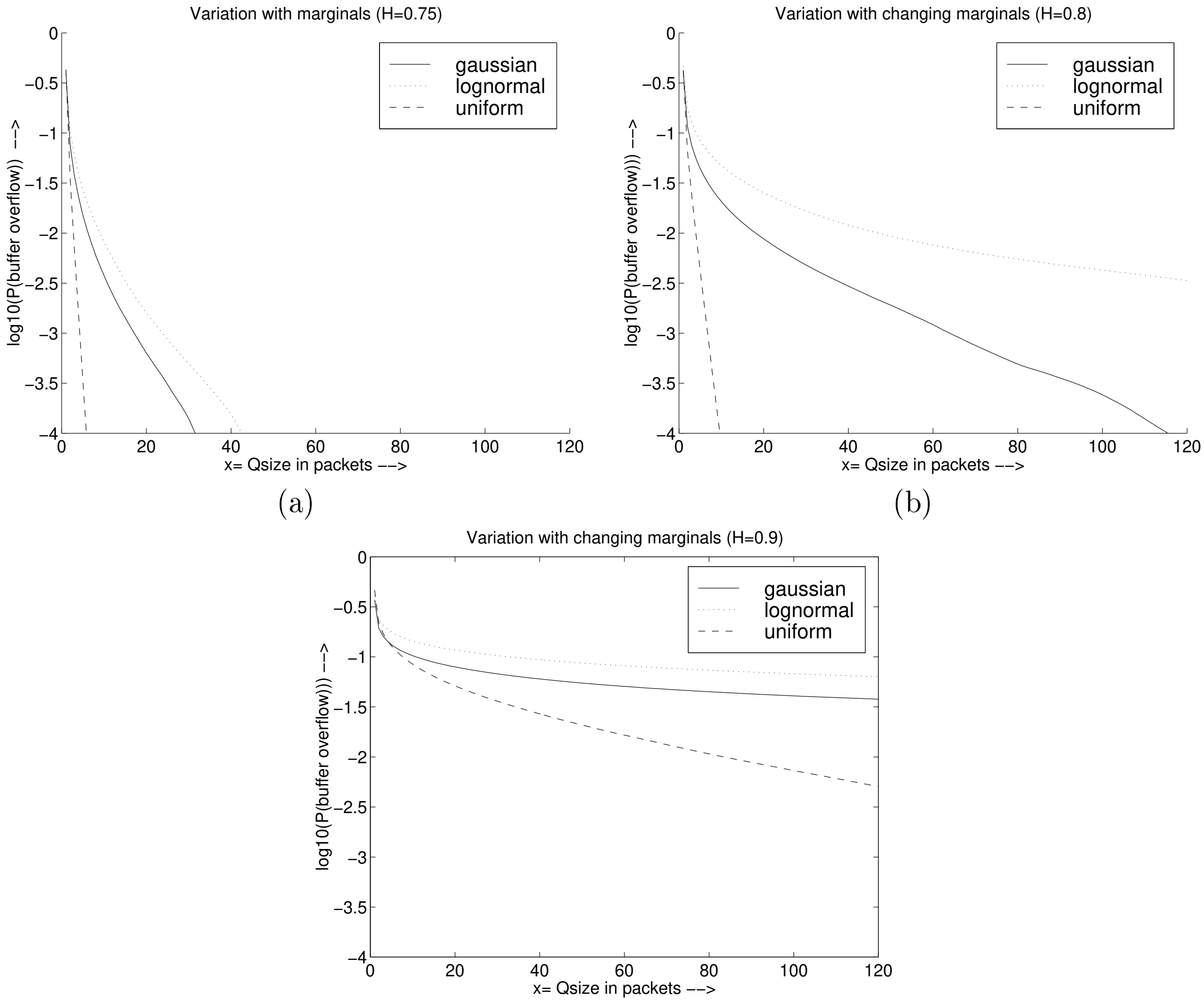
Figure 3: *Histogram plots of the marginal density of traces used to test the effect of marginals on queuing. Each plot corresponds to one trace. Trace with (a) Lognormal marginal, (b) Gaussian marginal, and (c) Uniform marginal.*

2. Generate a synthetic trace with Gaussian marginal using the pre-warped autocovariance and the FFTGEN .

3. Transform the trace to obtain the required marginal (warping).

We generated traces with lognormal and uniform marginal densities in this way. Figure 3 shows the marginal density of the sample traces. The traces with Gaussian marginals have negative values and these were made zero in the experiments. This leads to a bump in the histogram plot of Figure 3(b) near zero.

We used a mean packet input rate of 40 pkts/unit time, variance of 400 (pkts/unit time)^2. We observed that roughly one in every 100 data points was negative for the traces with Gaussian marginal. We varied the value of H in the experiment and investigated the effect



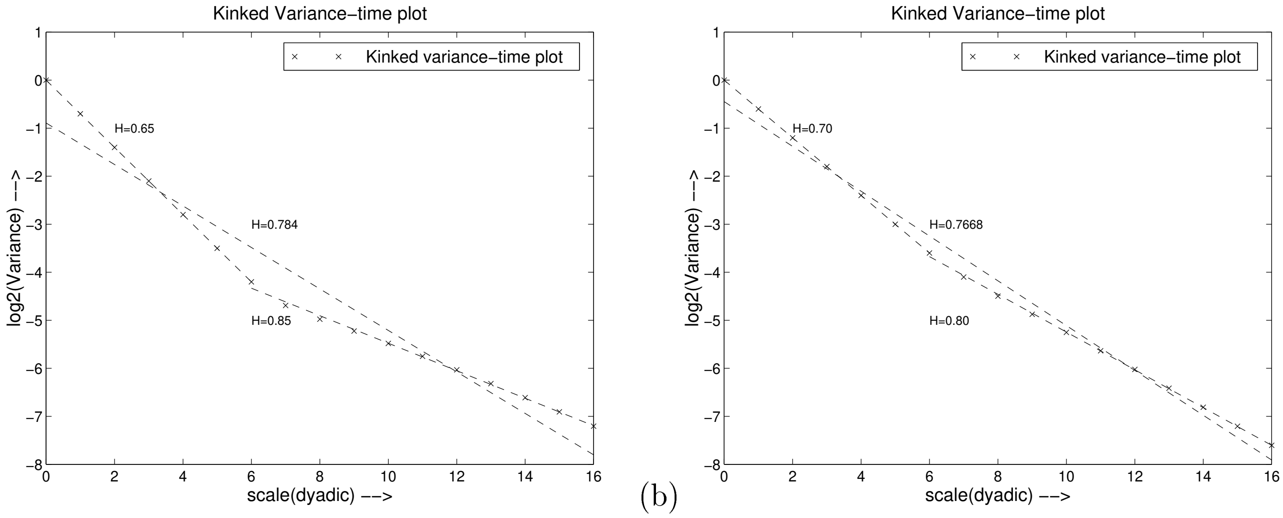
(c)

Figure 4: *For different values of H, traces with different marginal were input to short queues of different lengths. We observe that in general the traces with lognormal marginal density have higher loss probabilities than those with Gaussian marginals which in turn display higher loss probabilities than the traces with uniform marginals.*

of marginals on queuing. We generated 30 traces for each set of statistics (mean, H and marginal) and obtained the average probability of buffer overflow.

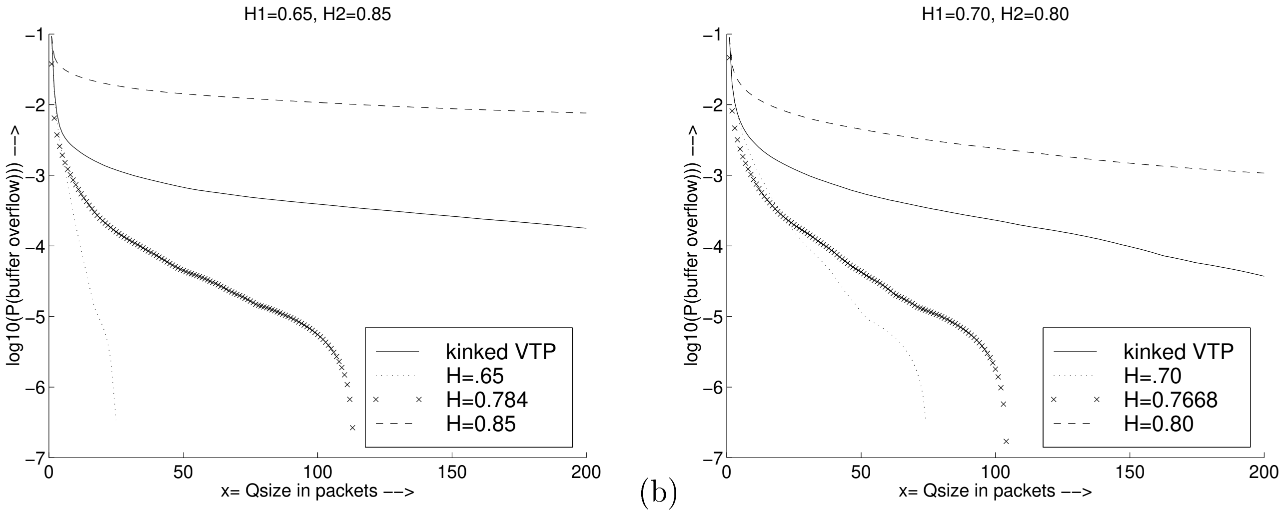
We observe from Figure 4 that traces with uniform marginals have much smaller buffer overflow probability than that of traces with Gaussian and lognormal marginals. Traces with lognormal marginals show greater buffer overflow probabilities than those with Gaussian marginals. This indicates that marginals with *heavier tails* display worse queuing behavior (more losses). With increasing values of H, the effect of the marginal appears to decrease for any particular finite buffer length, as seen from Figure 4. The results of this experiment suggest that queuing behavior cannot always be completely described by the second-order statistics of traffic.

9.2 Experiments with kinked variance-time plots



(a) (b)

Figure 5: *Kinked variance-time plots used for generating traces. In a) the kink is more pronounced than in b). For (a), $H_1 = 0.65$, $H_2 = 0.85$ and $H_3 = 0.784$. For (b),$H_1 = 0.70$, $H_2 = 0.80$ and $H_3 = 0.7668$.*



(a) (b)

Figure 6: *Buffer overflow probability of traces with kinked VTP and self-similar traces with Hurst parameter H1, H2 and H3 are shown. None of the self-similar traces match the queuing behavior of the traces with kinked VTPs.*

The VTP of an exactly second-order self-similar process is a straight line with slope $2H - 2$. However, real data traces like MPEG video have show *kinks* in their variance-time plots, that is, different slopes in different regions of the variance-time plot. We conducted the experiments described below to find out if a single parameter H is sufficient to describe such traffic.

We generated traces with lognormal densities corresponding to the two kinked VTPs shown in Figure 5 using the FFTGEN. We also generated self-similar traces with values of

H (H_1 and H_2) corresponding to the slopes of the variance-time plots on either side of the kink (at dyadic scale 6) and H (say H_3) corresponding to slope of the least square fit to the kinked variance time plot. We then compared the queuing performance of the traces. For the traces with kinked VTP, the mean rate of packet arrivals was 47 pkts/unit time, the variance of packet arrivals 200 (pkts./unit time)^2, and link capacity 50 pkts./unit time. Traces with Hurst parameters H_1 and H_2 had the same mean and variance as the kinked traces. The traces with $H = H_3$ were chosen to have variance corresponding to the Y-intercept of the *least square fit* line of the kinked VTP.

We observe from Figure 6 that all the self-similar traces fail to match the queuing behavior of the traces with kinked VTPs, irrespective of the magnitude of the kink ($H_2 - H_1$). These results suggest that incorporating the entire autocovariance structure in a traffic model can improve its matching of the queuing performance of real data substantially.

10 Conclusions

Experiments conducted to validate the MWM reveal that it captures the queuing behavior of real traffic well. Additional experiments indicate that the correlation matching property of the MWM can help significantly in matching the queuing performance of real traces. The experiment to study the effect of the marginal density on queuing reveals that second-order statistics do not suffice in describing the queuing behavior of LRD processes.

11 Future Work

The results of this project indicate that an in-depth study is required into the effects of various traffic characteristics on queuing. Among these, multifractal properties of traffic could play a very important role in queuing analysis as they describe the local irregularities of signals that intuitively would seem to affect short finite length buffers. Analytical methods must be used to study queuing behavior of traffic to supplement simulation results. Also, it is necessary to substantiate the results of this project with exhaustive tests.

References

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

[2] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson Modeling," *IEEE/ACM Transactions on Networking*, vol. 3, pp. 226–244, 1995.

[3] A. Erramilli, O. Narayan, and W. Willinger, "Experimental queueing analysis with long-range dependent traffic," *IEEE/ACM Transactions on Networking*, vol. 4, pp. 209–223, April 1996.

[4] I. Norros, “Four approaches to the fractional Brownian storage,” *Fractals in Engineering*, pp. 154–169, 1997.

[5] R. R. Riedi, M. Crouse, V. J. Ribeiro, and R. G. Baraniuk, “A Multifractal Wavelet Model with Application to TCP Network Traffic,” *submitted to IEEE Transaction on Information Theory*, February 1998.

[6] D. R. Cox, “Long-range dependence : A review,” *Statistics : An Appraisal*, pp. 55–74, 1984.

[7] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, “Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level (extended version),” *IEEE/ACM Transactions on Networking*.

[8] B. B. Mandelbrot and J. W. V. Ness, “Fractional Brownian Motions, Fractional Noises and Applications,” *SIAM Review*, vol. 10, pp. 422–437, October 1968.

[9] A. Papoulis, *Probability, Random Variables, and Stochastic Processes.* McGraw Hill Book Co., 1991.

[10] J. Beran, *Statistics for long-memory processes.* Chapman and Hall, 1994.

[11] M. S. Taqqu, V. Teverovsky, and W. Willinger, “Estimators for long-range dependence: An empirical study,” *Fractals*, vol. 3, no. 4, pp. 785–798, 1995. World Scientific Publishing Company.

[12] C.S.Burrus, R. A. Gopinath, and H. Guo, *Introduction to wavelets and wavelet transforms.* Prentice Hall, 1998.

[13] P. Abry and D. Veitch, “Wavelet analysis of long-range-dependent traffic,” *IEEE transactions on information theory*, vol. 44, pp. 2–15, January 1998.

[14] A. H. Tewfik and M. Kim, “Correlation structure of the discrete wavelet coefficients of fractional Brownian motion,” *IEEE transactions on information theory*, vol. 38, pp. 904–909, March 1992.

[15] R. R. Riedi, “An introduction to multifractals,” *available at: http://www-ece.rice.edu/~riedi/publications.html.*

[16] R. H. Riedi and J. L. Véhel, “TCP traffic is multifractal: A numerical study,” *submitted to IEEE Transactions of Networking*, October 1997.

[17] M. Crouse and R. G. Baraniuk, “Fast, exact synthesis of Gaussian and non-Gaussian long-range dependent processes,” *to be submitted to IEEE Transactions on Information Theory*, 1998.