

Performance of Communication Networks Fielding Bursty Data Traffic

G. R. Dattatreya¹ and Sarvesh S. Kulkarni²

¹*Department of Computer Science, University of Texas at Dallas, Richardson, TX
75083-0688, Email: datta@utdallas.edu*

²*Department of Electrical and Computer Engineering, Villanova University,
Villanova, PA 19085, Email: sarvesh.kulkarni@villanova.edu*

Abstract

This paper develops a traffic model and an efficient traffic generation method for limited scale burstiness. A review of existing models of burstiness and their properties lays the background for such development. The effects of various models of bursty traffic on single queues are studied, including analytical derivation, where possible. The effects of bursty data traffic (based on the model developed here) in multihop networks are studied through extensive simulation experiments. These experiments are conducted with two types of network operations: nonadaptive and adaptive. In adaptive operation, network nodes continuously react to the ill effects of bursty traffic. The simulation results are discussed comparatively with corresponding results for Poisson traffic and self-similar traffic reported elsewhere. Results convincingly show that adaptive operation shows a significant reduction in the performance degradation caused by bursty traffic.

1 Introduction

The intensity of data traffic arriving at a point can be specified in various ways such as bits per second, packets per unit time, ordered pair of *ON* time for a packet and *OFF* time between successive packets, or, the total number of bits or packets received over a given large time interval. Unless otherwise specified, there is the implicit notion that these intensity figures (which are average quantities) do not change from one time interval to the next, and that the quantities in successive intervals are independent, over meaningful ranges of times. The observed traffic in data networks appears to violate this property. Packets may arrive in clusters, with occasional packets present between clusters separated by appreciable time intervals. These occasional packets may not allow us to model the

incoming traffic with intermittent time periods of arrivals. Furthermore, the variance in the amount of traffic may not decrease rapidly as the length of the time interval used for averaging. Therefore, averaging the arrivals over a long time interval to include several packet clusters in the interval does not help us in realizing stable values for intensities. Traffic exhibiting such characteristics is referred to as *bursty* traffic. The two causes of traffic burstiness are (1) unbounded variance and (2) long range dependence (LRD) in inter-arrival times. Many researchers study self-similarity and the related LRD in different ways and from different starting points. Therefore, the first objective of this paper is to present a tutorial development of many properties of burstiness, LRD, and self-similarity, starting from first principles. This serves the additional purpose of providing the background and motivation for the development of our bursty traffic model. The second objective of the paper is the development of the N^{th} order auto-regressive (AR) model for bursty traffic and an efficient technique to generate such traffic. This model possesses limited scale burstiness. The third objective is to examine the behavior of single queues and multi-hop networks of queues with data traffic generated from smooth and different bursty models. The fourth and final objective is to demonstrate that the ill effects of bursty traffic in multi-hop networks can be significantly reduced with the help of adaptive routing.

This paper is organized as follows. The popular models of bursty traffic and the techniques for generation of bursty traffic are surveyed in Section 2. A clear distinction between smooth and bursty traffic is brought out in Section 3. The properties of self-similarity are developed and a particular self-similar process, the Fractional Brownian Motion (FBM) process is studied in Section 4. The driver process of FBM is the Fractional Gaussian Noise (FGN) process. Problems in generating such a process are identified. An N^{th} order AR model to generate bursty traffic with limited scale burstiness is proposed in Section 5. A classical algorithm called Durbin's recursive procedure, is used to determine the parameters of this model, for any degree of burstiness. The rest of the paper studies the effects of bursty traffic on single queues and feed-forward data networks. Single queues with traffic based on the AR model are simulated first. The effect of bursty arrivals into a single queue, based on Pareto distributed inter-arrival times, is analyzed in Section 6. Section 7 discusses the details of a known technique for the synthesis of self-similar traffic, that is, by merging (multiplexing) numerous traffic sources, each generating Pareto-distributed inter-arrival times. Simulation experiments are conducted to demonstrate the relative effects of (various types of) bursty traffic on a single queue, in Section 8. The experimental set up for the study of multi-hop networks is a 30 node static network. These experiments and results are reported in Section 9. Two experiments are conducted with the AR model-based bursty type of traffic. The first is a statistically multiplexed routing algorithm with constant routing parameters (nonadaptive routing). The second experiment uses an adaptive algorithm (recently developed by the authors in [12]) that continuously reacts to the ill effects of bursty traffic. The large extent of improvement due to the adaptive approach is clearly brought out. Results are also compared with those of similar experiments fielding different types of traffic, published in [12]. Major conclusions on bursty traffic, its effects on data networks,

and how to control them are pointed out in Section 10.

2 Survey of Bursty Traffic Models

In classical teletraffic theory, telephone traffic has been modeled with Poisson call arrivals and exponential service times. Since Poisson processes have elegant analytical properties, they have been employed widely in queuing analysis of telecommunication systems. However, Gusella [10], Leland *et al.* [15] and Willinger *et al.* [28] have shown that data traffic in data networks has properties that are vastly different from voice traffic in telecommunication networks. They perform statistical analysis on observed real life Ethernet traffic and determine that packet arrivals in data networks are extremely bursty. The burstiness spans several time scales, with a high variance in burst sizes and in the time interval between bursts. Furthermore, the sequence of inter-arrival times of data packets shows strong autocorrelations. The autocorrelation function decays slowly giving rise to what is known as *long range dependence* (LRD). These traffic characteristics cannot be modeled adequately using Poisson processes [23]. Therefore, researchers have suggested that network traffic be modeled using processes known as self-similar processes that capture the properties of burstiness and LRD. Further justification of this suggestion is provided by Crovella and Bestavros [4] and Garrett and Willinger [9]. [4] reports that data traffic on the World Wide Web (WWW) is self-similar and provides explanations as to the possible cause. The studies therein show that WWW document size distributions, caching algorithms of the web browser and user preferences in following the displayed links have a significant impact on the nature of the generated data traffic. The superimposition of many file transfers in a local area network, with the above dynamics influencing each such transfer, contributes to traffic self-similarity. [9] provides evidence that a Variable Bit Rate (VBR) video sample exhibits considerable burstiness and LRD.

An overview of traffic modeling techniques, for use in discrete-event network simulations, is provided by Frost and Melamed [8]. They subjectively discuss the properties of different traditional and newer, bursty traffic models. Self-similar traffic models, in particular, are discussed in greater detail by Popescu in [25]. He discusses the mathematical properties of self-similar processes, their relation to fractal processes and techniques for the generation of synthetic self-similar traffic traces. The various methods of estimation of the *Hurst* parameter H , which is a measure of burstiness of the traffic, are also discussed. According to [25], bursty traffic models can be categorized as either single source or aggregate traffic models, depending on the application that they model. Aggregate traffic models result in LRD.

2.1 Single source traffic models

A single source traffic model is useful in modeling traffic generated by a specific (single) application in a client-server environment. The packet inter-arrival times are modeled using

a heavy-tailed distribution such as the *Pareto* distribution, or the *log-normal* distribution. Over limited time scales of observation, the tails of both the Pareto as well as the log-normal distribution appear similar. The packet sizes (or file transfer sizes) are modeled with either the Pareto or the log-normal distribution. The model chosen depends on the application. For instance, based on their studies, Paxson and Floyd [23] report that the number of packets in TCP connections can be modeled well with a log-normal distribution whereas the Pareto distribution is better suited for modeling inter-arrival times for packets within a TCP session. The Pareto distribution is characterized by the density function

$$f_X(x) = \begin{cases} \left(\frac{\gamma}{\beta}\right) \left(\frac{\beta}{x}\right)^{\gamma+1} & \text{if } x \geq \beta \\ 0 & \text{if } x < \beta \end{cases} \quad (1)$$

for $\gamma > 0$ and $\beta > 0$. If we select $1 < \gamma \leq 2$, then the distribution has a finite mean and an infinite variance. The Hurst parameter H , which is a measure of burstiness, is related to γ by the equation $H = \frac{3-\gamma}{2}$. The value of H lies in $[0.5, 1)$, for infinite variance.

If Y is a Gaussian random variable, $Z = e^Y$ is called a log-normal random variable with the density function

$$f_Z(z) = \frac{1}{\sigma z \sqrt{2\pi}} e^{-\frac{[\ln(z)-\mu]^2}{2\sigma^2}} \quad (2)$$

where σ and μ are the standard deviation and the mean, respectively, of Y . The tail of the log-normal distribution is skewed to the right, and the extent of the skew increases with increase in σ or μ .

2.2 Traffic models with LRD

Aggregate traffic models are used in modeling aggregated or merged traffic from several sources or applications (running on either the same or different nodes in a network). Since aggregate traffic models exhibit LRD, we survey literature on traffic models with LRD properties and some suggested techniques for the synthesis of such traffic.

The book by Beran [1] defines a long memory or a long range dependent process as a stationary sequence of random variables X_i whose autocovariance sequence

$$\rho(k) = E[X_i X_{i+k}] - (E[X_i])^2 \quad (3)$$

varies hyperbolically, satisfying

$$\lim_{k \rightarrow \infty} \frac{\rho(k)}{ck^{-\alpha}} = 1 \quad (4)$$

where $\alpha \in (0, 1)$ and c is a positive constant. Beran [1] deals with limit theorems in time and frequency domain concerning such processes, and heuristic, time domain, and frequency domain estimation of long memory.

The book edited by Park and Willinger [22] is a collection of chapter contributions by several authors on the topics of measurement, modeling, performance analysis, and control

of self-similar network traffic. In the overview chapter of the book, a stochastic process $Y(t)$ is defined to be self-similar with a Hurst parameter $0 < H < 1$, if $Y(t)$ and $a^{-H}Y(at)$ have the same finite dimensional distributions. The second order self-similarity of a stochastic process is defined in terms of the structure of autocorrelation coefficients of a sequence of stationary increments of the stochastic process, over successive non-overlapping time intervals. The definition specifies the autocorrelation coefficients $R(k)$ of the increments to be of the form

$$R(k) = \frac{\sigma^2}{2}[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}] \quad (5)$$

for $k \geq 1$. For $\frac{1}{2} < H < 1$,

$$r(k) = \frac{R(k)}{\sigma^2} \sim H(2H-1)k^{2H-2}, \quad k \rightarrow \infty. \quad (6)$$

That is, $r(k)$ asymptotically behaves hyperbolically. For $k \geq 1$, $r(k)$ is bounded from below by the function $s(x) = r(x+1) < r(x)$. Since $\int_{x=1}^{\infty} s(x)dx = \infty$,

$$\sum_{k=1}^{\infty} r(k) = \infty. \quad (7)$$

When $r(k)$ behaves hyperbolically and is not infinite summable, the stationary sequence of increments is said to be long range dependent. A Gaussian self-similar process with stationary increments is defined as Fractional Brownian Motion (FBM) in [22]. Section 4.2 of this paper discusses FBM further.

The relation between heavy tailed distributions and self-similarity is brought out in [22] as follows. Consider a data packet source that is in the *ON* state for the duration of transmission of all the bits in the packet and in the *OFF* state between the end of transmission of one packet and the start of the next. The *ON* and *OFF* time periods are modeled to be continuous non-negative random variables. All the *ON* and *OFF* times are mutually independent. All the *ON* times are identically distributed; so are all the *OFF* times. At least one of the *ON* and *OFF* times follows a heavy-tailed probability distribution; that is, $P[T > x]$ varies as $cx^{-\alpha}$ as $x \rightarrow \infty$ and where $1 < \alpha < 2$. Let N number of packet sources be merged at a fictitious point so that many packets can be simultaneously present at that point. Let $Y(t)$ be the total amount of traffic (the accumulation of all the bits in all the packets) received at the merging point between the time 0 and t . Let ρ be the ratio of the expected *ON* time to the expectation of the sum of *ON* and *OFF* times. Then [22] states that $Y(Tt)$ behaves statistically as

$$\rho NTt + CN^{\frac{1}{2}}T^H B_H(t) \quad (8)$$

for large N and large T . In the above, $B_H(t)$ is the fractional Brownian motion with parameter $H = \frac{3-\alpha}{2}$ and $C > 0$ is a quantity depending only on the distributions of *ON* and *OFF* times. It appears that the authors of the book use the standardized version of FBM for $B_H(t)$ in the above. That is, the variance of $B_H(t)$ is assumed to be 1 at $t = 1$,

though their original definition of FBM (given two paragraphs above, in this paper) does not specify the constant of proportionality.

Following the overview chapter, the first part of the book [22] discusses the wavelet based approach to modeling and estimation with a multifractal view of self-similar traffic. One of the conclusions is that special care is required in simulations to avoid erroneous conclusions, attributable to finite sample effects when dealing with heavy tailed properties of these stochastic processes.

The second part of [22] is on the queuing behavior of finite and infinite buffer systems when fed with LRD data input. Exact results on queues are very hard to obtain due to the non-Markovian nature of the system. The general results are that the queue length densities of unlimited buffer systems decay slower than exponentially. One of the consequences of this result is that if we provision large buffers to reduce packet losses, the higher probability of longer queues results in very large delays. Most of the results on queue-based results are asymptotic, such as upper and lower bounds for the tail of the queue length distributions as the queue length approaches infinity. Even in finite buffer studies, results are on bounds for buffer overflow probabilities as the buffer capacities tend to infinity.

The third part of the book [22] deals with traffic control issues. In the open loop control approach, one suggestion is to reserve enough bandwidth and no buffer to avoid the effects of long queues and delays. For feedback control, exploiting the correlation structure at multiple time scales is suggested as an alternative, to reduce congestion. Tuan and Park (Chapter 18 in [22]) conduct extensive analysis of various traffic traces with long range dependence. They show that the traffic levels are strongly correlated in the one to five second time intervals, by discretizing the observed traffic intensity into eight levels. Through additional experiments, they conclude that the one to five second time interval region is suited for updates in congestion control schemes.

Leland *et al.* [15] suggest that a large number of *ON-OFF* renewal-rewards heavy-tailed processes may be superimposed (merged) to generate bursty traffic. The rewards are restricted to values of either 0 or 1. Taquu *et al.* [27] prove that the aggregation of such *ON-OFF* sources results in asymptotically self-similar traffic. As mentioned earlier in this subsection, the limiting process in [22] corresponding to this result is given by Equation (8). Furthermore, the *ON* and *OFF* periods need not have the same distribution. Norros [20] proposes the use of Fractional Brownian Motion (FBM) for generating bursty traffic. However, true FBM traces are extremely difficult to generate. Therefore, an algorithm that approximates FBM, called the Random Midpoint Displacement (RMD) algorithm [14], is used. RMD generates an FBM trace in the time interval $[0, T]$ as follows. The interval $[0, T]$ is subdivided recursively. At the midpoint of each interval, a zero mean Gaussian displacement is provided. The variance of the displacement can be scaled to yield an approximate FBM trace. Rambaldi and Pinazza [26] propose the generation of approximate FBM traces as follows. In generating a sample point in an FBM trace, they do not consider the correlational contributions of all previous sample points individually. Instead, they

“block” together the previous sample points in the path and consider only the combined correlational contribution of those points. Another variation in the use of FBM is based on its wavelet transform as in Flandrin [7]. This method involves the computation of the wavelet coefficients corresponding to the FBM wavelet transform. Next, these computed wavelet coefficients are utilized along with the inverse wavelet transform of FBM to yield FBM-like paths.

Bursty traffic can also be generated based on the Auto-Regressive Integrated Moving-Average (ARIMA) model by Box and Jenkins [2]. A variation of ARIMA called the Fractional ARIMA (FARIMA) model (Beran, [1]) can be used to model both short range as well as long range dependence. FARIMA is asymptotically self-similar [3], but tends to have high computational complexity. A relatively fast (but approximate) method of generating self-similar traffic traces is by means of the Fast Fourier Transform (FFT) proposed in [24]. It involves approximation of the driver process of FBM, called the Fractional Gaussian Noise (FGN) process [19]. The power spectrum of an FGN process is approximated by a sequence of complex numbers. The approximate FGN sequence is obtained by computing the inverse Discrete Time Fourier Transform (DTFT) of this approximate power spectrum. A very different approach is taken by Erramilli and Singh in [5]. They show how broadband traffic can be modeled with low order chaotic systems (using deterministic chaotic maps).

In the schemes surveyed above, the generation of accurate traces depends on utilizing more terms in computation and hence requires massive computing power. Consequently, a procedure with relatively low overheads and high accuracy in generating a bursty traffic trace is of considerable interest. We develop such a procedure in Section 5, for the case of limited scale self-similarity. The next two sections develop the properties of bursty traffic and self-similarity.

3 Distinction Between Smooth and Bursty Traffic

Let λ be the arrival rate of traffic (in bytes per unit time). Then, $\lambda\tau$ is the average amount of traffic (in bytes) received in the time interval τ . Let $X_i(\tau)$ be the fluctuation around the mean $\lambda\tau$. Therefore, $X_i(\tau) + \lambda\tau$ represents the amount of traffic received in the i th interval of successive, equal, and non-overlapping time intervals, each time interval being τ long. We concentrate our attention on $X_i(\tau)$, the variation around the mean, and refer to this variation itself as the traffic. Therefore, in the following discussion, unless otherwise specified, random processes are zero mean.

$\frac{X_i(\tau)}{\tau}$ is the traffic arrival rate averaged over the i th such interval. Consider the cumulative traffic received over k such successive time intervals. We now form non-overlapping time intervals, with each interval made of k original intervals. Let $Y_i(k\tau)$ be the traffic received in the i th of such aggregated, larger intervals. Then, $\frac{Y_i(k\tau)}{k\tau}$ is the time average of the traffic rate received in the i th aggregated interval. The dimensions of both $\frac{X_i(\tau)}{\tau}$ and $\frac{Y_i(k\tau)}{k\tau}$ are “bytes per second.”

Case 1: Smooth Traffic

Let $X_i(\tau)$ be an independent, identically distributed (*iid*) sequence of random variables indexed by i . Var denotes the variance of the argument random variable. Let

$$Var[X_i(\tau)] = \sigma^2(\tau). \quad (9)$$

$$Y_i(k\tau) = \sum_{j=l}^{k+l-1} X_j(\tau) \text{ for some } l. \quad (10)$$

From Probability theory [21],

$$Var\left[\frac{X_i(\tau)}{\tau}\right] = \frac{\sigma^2(\tau)}{\tau^2} \quad (11)$$

and

$$Var[Y_i(k\tau)] = k\sigma^2(\tau). \quad (12)$$

Then,

$$\begin{aligned} Var\left[\frac{Y_i(k\tau)}{k\tau}\right] &= \frac{k\sigma^2(\tau)}{k^2\tau^2} \\ &= \frac{\sigma^2(\tau)}{k\tau^2}. \end{aligned} \quad (13)$$

Therefore,

$$Var\left[\frac{Y_i(k\tau)}{k\tau}\right] = \frac{Var\left[\frac{X_i(\tau)}{\tau}\right]}{k}. \quad (14)$$

In the above equation, both $\frac{X_i(\tau)}{\tau}$ and $\frac{Y_i(k\tau)}{k\tau}$ are random variables corresponding to the traffic received per unit time. However, $Y_i(k\tau)$ is the traffic averaged over an interval k times the size of the interval used to average $X_i(\tau)$. Since the amounts of traffic received over successive intervals are assumed to be independent in this case, the fluctuations over the different (smaller) intervals tend to cancel out to quite some extent when considering the traffic aggregated over many such intervals. Hence we observe the $\frac{1}{k}$ type of drop in the variance of $\frac{Y_i(k\tau)}{k\tau}$, in comparison with the variance of $\frac{X_i(\tau)}{\tau}$. This phenomenon is observed widely in practice. That is, as the scale of averaging increases, the averaged (or blurred) observations tend to exhibit smaller variance (smaller by a factor proportional to the linear scale of blurring), and appear smooth. However, data traffic over the Internet appears to defy this principle up to long intervals of time-averaging. Thus, for data traffic, the variance of the corresponding $\frac{Y_i(k\tau)}{k\tau}$ does not decrease linearly with k , but decreases much more slowly. We deal with such a case of bursty traffic in the following discussion.

Case 2: Bursty Traffic

If the components of $\{X_i(\tau)\}$ are statistically correlated (as opposed to being statistically independent) in a particular way, the cancellation of fluctuations can get decelerated as

follows. Consider the autocorrelation sequence formed from the sequence $X_i(\tau)$. We denote this by $R_X(k, \tau)$. That is,

$$R_X(k, \tau) = E[X_i(\tau)X_{i+k}(\tau)]. \quad (15)$$

The stationarity of $\{X_i(\tau)\}$ ensures that $E[X_i(\tau)X_{i+k}(\tau)]$ is not a function of i . In addition,

$$R_X(k, \tau) = R_X(-k, \tau). \quad (16)$$

Note that if $\{X_i(\tau)\}$ is an independent sequence,

$$R_X(k, \tau) = \sigma_X^2(\tau) \quad \text{and} \quad (17)$$

$$R_X(k, \tau) = 0, \quad \forall k \neq 0. \quad (18)$$

In some cases of dependent sequences, $R_X(k, \tau)$ may be non-zero up to $k = m$, for a small integer m . In some other cases, $R_X(k, \tau)$ may decay as an exponential function of k . In these two cases, if we average $X_i(\tau)$ over many intervals, the variance diminishes rapidly as the number of intervals used for aggregation increases. On the other hand, for some profiles of $R_X(k, \tau)$, the variance of the data averaged over many intervals can decrease very slowly as demonstrated below. The variance of $\frac{Y_i(k\tau)}{k\tau}$ is evaluated as follows.

$$\begin{aligned} \text{Var}\left[\frac{Y_i(k\tau)}{k\tau}\right] &= \frac{1}{k^2\tau^2} \text{Var}[Y_i(k\tau)] \\ &= \frac{1}{k^2\tau^2} E\left\{[Y_i(k\tau)]^2\right\} \\ &= \frac{1}{k^2\tau^2} \sum_{j=1}^k \sum_{l=1}^k E[X_j(\tau)X_l(\tau)] \\ &= \frac{1}{k^2\tau^2} \left\{ \sum_{j=1}^k \sum_{l=1}^k R_X(j-l, \tau) \right\} \\ &= \frac{1}{k^2\tau^2} [kR_X(0, \tau) + 2 \sum_{l=1}^{k-1} (k-l)R_X(l, \tau)] \\ &= \frac{\sigma^2(\tau)}{k\tau^2} + \frac{2}{k^2\tau^2} \sum_{l=1}^{k-1} (k-l)R_X(l, \tau). \end{aligned} \quad (19)$$

If the second term (the summation) on the right side of Equation (19) is zero, then

$$\text{Var}\left[\frac{Y_i(k\tau)}{k\tau}\right] = \frac{\sigma^2(\tau)}{k\tau^2} \quad (20)$$

corresponding to the *iid* sequence $X_i(\tau)$. If $R_X(k, \tau)$ decays rapidly (exponentially), then the second term on the right side of Equation (19) is also bounded by an exponentially decaying function of k . This set of conditions corresponds to smooth traffic. However, if $R_X(k, \tau)$ are positive and large for $k > 0$, from Equation (19), we observe that the variance

of $\frac{Y_i(k\tau)}{k\tau}$ will be much larger than in the case of smooth traffic. The argument cannot be extended to make $\frac{R_X(k,\tau)}{\sigma^2(\tau)} = 1$, $k > 0$, since this would render $X_i(\tau)$ to be identical and fully correlated for all values of k , with no variations at all. If we want to extend this property for large values of k , then $\frac{R_X(k,\tau)}{\sigma^2(\tau)}$ should decay slowly, for example, as $k^{-\alpha}$ with $0 < \alpha < 1$.

Similarly, we can study the effects and required properties of variations and correlations for an arbitrarily sized window of averaging. That is, instead of examining the statistical properties of the sequence $Y_i(k\tau)$ over a time window of size $k\tau$ with an integer k , we can examine the statistical properties of $Y_i(\delta\tau)$, with a positive real δ . In an abstract model, we can let $\tau \rightarrow 0$, resulting in traffic that arrives as a continuous function of time. This is in contrast to the earlier described model in which we studied increments in traffic received over successive, equal time intervals.

4 Self-similar Processes

4.1 Definition of self-similarity

The above discussion suggests that if $R_X(k, \tau)$ are *positive* and *large* for $k > 0$, then $Var[\frac{Y_i(k\tau)}{k\tau}] \gg \frac{1}{k} Var[\frac{X_i(\tau)}{\tau}]$. If this property is realized, the averaged traffic that is received, will still exhibit considerable variations (burstiness) over a large range of time scales and will not appear smoother as the time scale increases. These ideas can be quantified using the concepts of self-similar processes studied by Mandelbrot in [17]. Mandelbrot's definition of self-similarity is as follows.

Let $Z(t)$ be a stochastic process and consider any t_0 and δ , ($\delta > 0$). If the processes $Z(t_0 + \tau) - Z(t_0)$ and $\delta^{-H}[Z(t_0 + \delta\tau) - Z(t_0)]$ have the same finite joint distributions, we say that the stochastic process $Z(t)$ is self-similar with the parameter H . Mandelbrot's definition is valid for $H \geq 0$. The concept of self-similarity is used by Mandelbrot [18] to describe naturally occurring fractals. Self-similar processes exhibit structural similarities in their statistics over a large number of different time scales. That is, even if the time scale of measurement is changed, the statistics of the process remain *similar*, (not identical). In this section, we enunciate some of the properties and consequences of the above definition in our special case of data traffic. As mentioned earlier, it is convenient to restrict $Z(t)$ to be a zero mean process. Any required mean value can be added at convenient points of mathematical (and correspondingly physical) transformations. For zero mean processes, the autocorrelation and the autocovariance functions are the same. In addition, we consider stationary increments.

4.2 Fractional Brownian motion

An example family of self-similar stochastic processes is the Fractional Brownian Motion (FBM), defined as follows. $Y(t)$ is an FBM process with parameter $0.5 \leq H < 1$, if the

increment $Y(t+t_0) - Y(t_0)$ is a stationary Gaussian process with zero mean and variance t^{2H} . Using the notation $\mathcal{N}(\eta, \sigma^2)$ for a Gaussian random variable with a mean η and variance σ^2 , we can express the rate of FBM averaged over an interval δ as

$$\frac{Y(t+\delta) - Y(t)}{\delta} = \delta^{H-1} \mathcal{N}(0, 1). \quad (21)$$

Therefore,

$$\begin{aligned} \frac{d}{dt} Y(t) &= \lim_{\delta \rightarrow 0} \frac{Y(t+\delta) - Y(t)}{\delta} \\ &= \lim_{\delta \rightarrow 0} \delta^{H-1} \mathcal{N}(0, 1). \end{aligned} \quad (22)$$

Since $0.5 \leq H < 1$, $\delta^{H-1} \rightarrow \infty$ as $\delta \rightarrow 0$.

4.3 Properties of self-similar processes

In [19], Mandelbrot and Van Ness show the following properties in addition to the above.

1. FBM has continuous sample paths with probability 1, but is almost surely not differentiable.
2. Any non-constant mean square continuous self-similar Gaussian process must be FBM.
3. Non-Gaussian self-similar processes must necessarily have increments with infinite variance.

These properties imply that the most useful self-similar process is FBM.

4.4 Discrete time Fractional Gaussian Noise and its properties

FBM is not mean square differentiable. However, its increments over any time interval $\delta > 0$ have finite variances. We can construct a time series (or a sequence of Gaussian random variables) corresponding to FBM increments over successive δ . The increments $V_i(\delta) = a[Y(i\delta) - Y((i-1)\delta)]$ of the FBM process $aY(t)$ have variance $\sigma^2(\delta) = a^2\delta^{2H}$. The sequence $V_i(\delta)$ constitutes the discrete Fractional Gaussian Noise (FGN) process. Consider the autocorrelation sequence of $V_i(\delta)$.

$$R_V(k, \delta) = a^2 E \left[\left(\frac{Y((k+1)\delta) - Y(k\delta)}{\delta} \right) \left(\frac{Y(\delta) - Y(0)}{\delta} \right) \right]. \quad (23)$$

We make use of the algebraic identity

$$(a-b)(c-d) = \frac{1}{2} \left[(a-d)^2 - (a-c)^2 + (b-c)^2 - (b-d)^2 \right]$$

to rewrite Equation (23) as follows. The symbol a in the above algebraic identity is a dummy variable, not to be confused with the scale factor a in Equation (23).

$$\begin{aligned}
R_V(k, \delta) &= a^2 E \left[\frac{1}{2\delta^2} \left((Y((k+1)\delta) - Y(0))^2 - (Y((k+1)\delta) - Y(\delta))^2 \right. \right. \\
&\quad \left. \left. + (Y(k\delta) - Y(\delta))^2 - (Y(k\delta) - Y(0))^2 \right) \right] \\
&= \frac{a^2}{2\delta^2} [Var[Y((k+1)\delta) - Y(0)] - Var[Y((k+1)\delta) - Y(\delta)] \\
&\quad + Var[Y(k\delta) - Y(\delta)] + Var[Y(k\delta) - Y(0)]] \\
&= \frac{a^2}{2\delta^2} [|\delta(k+1)|^{2H} - |\delta k|^{2H} + |\delta(k-1)|^{2H} - |\delta k|^{2H}] \\
&= \frac{a^2}{2\delta^{2-2H}} [|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}]. \tag{24}
\end{aligned}$$

Therefore, the normalized autocorrelation sequence of $V_i(\delta)$ is

$$\begin{aligned}
r_V(k, \delta) &= \frac{R_V(k, \delta)}{Var[V_i(\delta)]} \\
&= \frac{\frac{a^2}{2\delta^{2-2H}} [|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}]}{a^2 \delta^{2H}} \\
&= \frac{1}{2\delta^2} [|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}] \\
&\sim \left[\frac{H(2H-1)}{\delta^2} \right] k^{2H-2}, \quad k \rightarrow \infty. \tag{25}
\end{aligned}$$

The final limiting form of $r_V(k, \delta)$ in Equation (25) can be verified by using L'Hospital's rule twice to evaluate the limit of $\frac{\delta^2 r_V(k, \delta)}{H(2H-1)} k^{2H-2}$ as $\frac{1}{k} \rightarrow 0$. As discussed after Equation (6), $r_V(k, \delta)$ in Equation (25) is also not infinite summable over $k = 1, 2, \dots$

4.5 Problems in generating pure FBM

The most useful self-similar process is FBM, as noted at the end of Section 4.3. Most publications in literature on FBM-based self-similar traffic model the cumulatively arriving bursty traffic as a sample path of the continuous time FBM process. Continuous time FGN is the derivative of FBM and forms its driver process. Continuous time FGN has unbounded instantaneous variance. The continuous time FBM process can be discretized over non-overlapping windows. The increments form an FGN sequence with finite variance. The cumulative addition of successive random variables in such an FGN sequence is a discretized FBM sequence. Many publications report the realization of bursty traffic through the approximation of continuous time FGN. In order to satisfy self-similarity, discretized FGN must follow the autocorrelation function specified in Equation (24) in the previous subsection. As pointed out after Equation (22), the instantaneous variance of continuous time FGN is infinity and therefore, continuous time FGN is not suitable for generating self-similar traffic.

The autocorrelations given in Equation (24) are non-summable over the infinite range of k , even for $\delta > 0$. With the use this property, it is easy to show that the required autocorrelation profile is not realizable with a linear shift invariant, causal and stable discrete time system excited by a white Gaussian input sequence (see Kulkarni [11] for details). Therefore, generating discrete time FGN is also very difficult.

5 An Auto-Regressive (AR) Model and Procedure to Generate Bursty Traffic

This section develops an N^{th} order AR model to generate limited-scale bursty traffic. The traffic generated by the AR model has the autocorrelation profile given in Equation (24) over a limited scale. The results of experiments conducted on a single queue with traffic (inter-arrival times) generated by the AR model are also reported.

5.1 Self-similarity over limited time scale

Mandelbrot [17] emphasizes that the properties of self-similarity were never meant to hold rigidly when the time lag is extremely small (i.e. $\delta \rightarrow 0$) or extremely large (i.e. $k\tau \rightarrow \infty$), in reality. Therefore, in our proposed model in Section 5, we consider increments over only non-overlapping time intervals of non-zero δ . This restricts our attention to discrete time sequences. Furthermore, instead of allowing the time difference $k\tau$ to stretch to infinity, we limit it to some $k\tau_{Max}$. Therefore, in our model, the autocorrelation statistics of the process satisfy the properties of self-similarity over time intervals ranging from δ through $k\tau_{Max}$. Thereafter, the profile of the autocorrelations can fall off in *some* exponential manner. We emphasize that $k\tau_{Max}$ can be as large as desired, depending on how long the LRD persists in the real life application being modeled.

Let $V(i)$ be the driver sequence generating self-similar increments. That is, $V(i)$ is a discretized FGN process with increments measured at $k\delta$ time instants. It is shown in Equation (24) that the autocorrelation sequence of $V_i(\delta)$ is

$$R_V(k) = \frac{a^2}{2\delta^{2-2H}}(|k+1|^{2H} + |k-1|^{2H} - 2|k|^{2H}) \quad \text{for } k = 1, 2, \dots, N. \quad (26)$$

The corresponding normalized autocorrelation sequence (normalized with respect to $r_V(0)$) is given by

$$r_V(k) = \frac{1}{2\delta^2}(|k+1|^{2H} + |k-1|^{2H} - 2|k|^{2H}) \quad \text{for } k = 1, 2, \dots, N. \quad (27)$$

Equation (27) defines $r_V(k)$ up to *some* specified large N . For a given value of δ , beyond $k = N$, the value of $r_V(k)$ in Equation (27) falls off in some exponential manner. The smallest value of δ is the one for which the wide-sense self-similarity (profile of autocorrelation sequence) just starts to fail. Therefore, $N\delta$ is the largest significant lag over which long range dependence (LRD) persists.

5.2 Linear prediction (N^{th} order auto-regressive) model

Let

$$Y(i) = \sum_{j=1}^N Y(i-j)a_j + a_0 X_i \quad (28)$$

where X_i is a white sequence. The linear prediction coefficients $\{a_j\}$, are efficiently computed using a recursive procedure due to Durbin (see Makhoul [16]). The coefficients are functions solely of H and the autocorrelation function used for their computation is given by Equation (24). The computation of these coefficients is very stable even for $H = 0.9$ and $N = 10,000$. All the linear prediction coefficients are positive. X_i is the input provided to the system that generates $Y(i)$. The X_i s can have any zero mean and finite variance distribution.

Intuitively too, a discrete time self-similar model is more appropriate for data traffic than a continuous time model. For instance, $Y(i)$ could either be the rate at which the i^{th} packet arrives, or it could be the overall data rate in the i^{th} discrete time interval. $Y(i)$ can also model the sequence of inter-arrival times. This concludes the study of bursty traffic models and their properties. The rest of the paper deals with the performance of queues and data networks fielding different types of bursty traffic.

5.3 Experiments on a single queue with arrivals generated by the AR model

We perform experiments on a single queue using the AR model to generate the inter-arrival times as follows. Let $Z(i)$ denote the inter-arrival time sequence. Then $Z(i)$ is modeled as

$$Z(i) = \beta + \alpha Y(i), \quad (29)$$

where $Y(i)$ is a zero mean, unity variance AR sequence of Equation (28), $Z(i)$ has mean β and variance α^2 . Thus, the variance of the inter-arrival times can be controlled by choosing an appropriate value for α . However, since the inter-arrival times are constrained to be non-negative, β is chosen such that $\beta > -\alpha Y_{\min}$, where Y_{\min} is the lowest value in the sequence $Y(i)$. Two different distributions are used for service times; the exponential distribution and the Pareto distribution.

Figure 1 shows the mean queue length for different values of the normalized load ρ , when the variance of $Z(i)$ is 1. For $H = 0.6$, the queue saturates at 98-99% of the load when the service time distribution is exponential, which is roughly the same behavior as that exhibited by an $M/M/1$ queue. For the same value of H (that is, $H = 0.6$), Pareto service times result in the queue saturating at lower loads (90-95%). For $H = 0.9$, the saturation load for exponential and Pareto service times is 60% and 85% respectively. Therefore the higher value of H results in performance figures that are worse than those observed for a lower value of H .

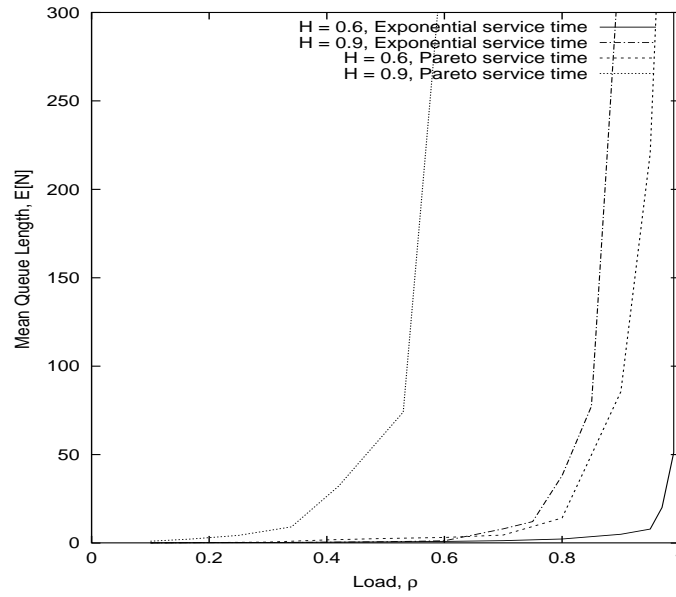


Figure 1: Mean queue length $E[n]$ as a function of the normalized load ρ , (AR traffic)

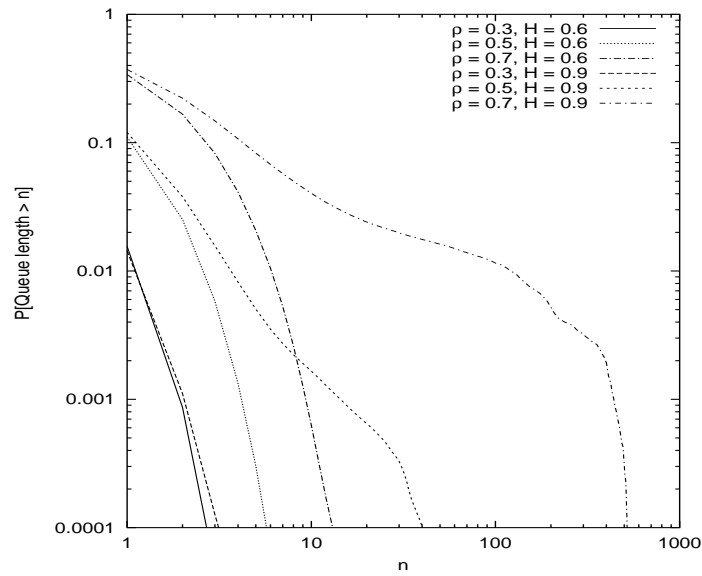


Figure 2: $P[N > n]$ Vs n , (AR inter-arrival times, Exponential service times)

Figures 2 and 3 show the corresponding complementary queue length distributions for Poisson and Pareto distributed service times, when the inter-arrival times are generated with the AR model. In both figures, for $H = 0.6$, the tail of the distribution drops off sharply at a relatively low value of queue length, for a given ρ . On the other hand, for $H = 0.9$, for all the tested values of ρ , the tail of the complementary queue length distribution is much longer. Figure 3 is similar to Figure 2, except that the tail of the complementary queue length distribution is much heavier in the former, indicating that the high variance in the service times has an adverse effect on the performance of the queue. In Figure 2, at $\rho = 0.7$, the divergence in the tails is very striking, with the tail corresponding to $H = 0.9$ being an order of magnitude larger than the one for $H = 0.6$. In general, for a given ρ , longer queue lengths are more common for higher values of H , regardless of the value of ρ , or the service time distribution.

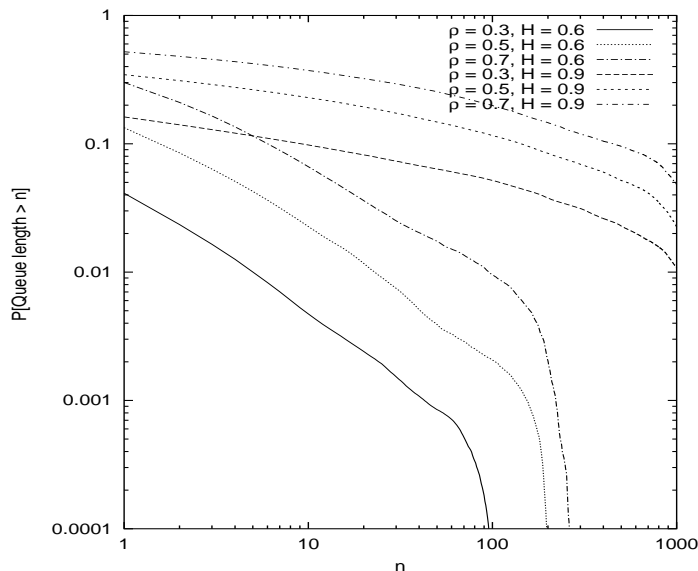


Figure 3: $P[N > n]$ Vs n , (AR inter-arrival times, Pareto service times)

6 Impact of Pareto Inter-Arrival Times on a Single Queue

The Pareto distribution exhibits a heavy tail and therefore, Pareto inter-arrival times can be used as another model of bursty traffic, as mentioned in Section 2.1. Networks usually have a fixed small upper limit on the allowable packet sizes (in bits) and exponentially distributed packet sizes are reasonable in many cases. Such traffic fed to a constant bit-rate processor results in $G/M/1$ queues. Consider a $G/M/1$ queuing system with a service rate μ . Let α with $(0 < \alpha < 1)$ be the proportion of arrivals that finds the server busy at the arrival time instant. Then, from the analysis of $G/M/1$ queues in Wolff [29], α is determined

by

$$\alpha = \mathcal{L}_t\{\mu(1 - \alpha)\} \quad (30)$$

where, $\mathcal{L}_t(s)$ represents the Laplace Transform with parameter s . That is,

$$\alpha = \int_0^\infty e^{-\mu(1-\alpha)t} f_T(t) dt. \quad (31)$$

where $f_T(t)$ is the probability density function of the inter-arrival times. Let ρ be the normalized load experienced by the queue. Then, $\alpha \in (0, 1)$ is the unique solution to Equation (31) if $\rho < 1$ (see [29]).

The density function of a Pareto random variable is in Equation (1). Consider the case of $1 < \gamma \leq 2$ and $\beta > 0$ resulting in a finite mean and an infinite variance. Note that the Hurst parameter $H = \frac{3-\gamma}{2}$. Therefore, if the inter-arrival time T is a Pareto random variable, then Equation (31) can be rewritten as

$$\alpha = \int_\beta^\infty e^{-\mu t + \mu t \alpha} \left(\frac{\gamma}{\beta}\right) \left(\frac{\beta^{\gamma+1}}{t^{\gamma+1}}\right) dt. \quad (32)$$

Kulkarni [11] carries out the required integration by parts three times and concludes that

$$\begin{aligned} \alpha &= \gamma [\mu(1 - \alpha)]^\gamma \\ &\left\{ \frac{1}{\gamma e^{\mu(1-\alpha)} [\mu(1 - \alpha)]^\gamma} - \frac{1}{\gamma(\gamma - 1) e^{\mu(1-\alpha)} [\mu(1 - \alpha)]^{\gamma-1}} \right. \\ &\quad + \frac{1}{\gamma(\gamma - 1)(\gamma - 2) e^{\mu(1-\alpha)} [\mu(1 - \alpha)]^{\gamma-2}} \\ &\quad \left. - \frac{1}{\gamma(\gamma - 1)(\gamma - 2)} [\Gamma(-\gamma + 3) - \Gamma^*(-\gamma + 3, \mu(1 - \alpha))] \right\} \quad (33) \end{aligned}$$

where $\Gamma^*(-\gamma + 3, \mu(1 - \alpha)) = \Gamma(-\gamma + 3)P(-\gamma + 3, \mu(1 - \alpha))$. Here, $P(-\gamma + 3, \mu(1 - \alpha))$ is the incomplete gamma function.

Equation (33) can be solved numerically (in Matlab¹) to yield the profile of α curves for different values of the Hurst parameter H and the normalized load ρ . The curves are plotted in Figure 4. For low values of ρ , we observe that $\alpha < \rho$, upto $H = 0.9$. In other words, for low values of ρ , the load actually observed (α) in the queue by incoming packets is lower than the long term load average (ρ). For $H = 0.5$ and $H = 0.6$, $\alpha < \rho$ over most of the normal operating range of the queue (i.e. $\rho < 0.75$). However, for high values of H , that is, for $H = 0.9$ and $H = 0.99$, $\alpha < \rho$ only for very low values of ρ . As ρ increases, the actual load observed by the incoming traffic is considerably higher than ρ itself. This implies that a $G/M/1$ queue with Pareto traffic arrivals saturates at relatively low values of ρ , for high values of H . Indeed, Figure 5 illustrates this property.

¹Matlab is a registered trademark of *The Mathworks*.

The plots in Figure 5 are obtained as follows. From Wolff [29], the number of packets in the system at arrival time instants has a modified Geometric distribution with parameter α . Also, the expected number (time average) of packets in the system is given by the following function of α and ρ :

$$E[N] = \frac{\rho}{1 - \alpha}. \quad (34)$$

The values of α corresponding to the values of ρ (over closely spaced intervals, $0 < \rho < 1$) are computed with Equation (33) for $H = 0.5, 0.6, 0.7, 0.8, 0.9$ and 0.99 . Then, for each such computed α , the corresponding expected number of customers is computed with Equation (34) and plotted. The resulting plots show that for low values of H (that is, for $H = 0.5, 0.6$ and 0.7), the performance curves are similar to those observed in an $M/M/1$ queuing system. For $H = 0.9$, saturation occurs at barely 60-70% of the load. In general, as H increases, the corresponding load threshold at which saturation occurs, diminishes rapidly.

According to Erramilli *et al.* [6], the value of H observed in real life traffic is approximately 0.9. Therefore, in real life, we expect that network queues will saturate at approximately 60-70% of the load. An important conclusion of this section is that even though the $G/M/1$ queue fed by traffic with Pareto interarrivals exhibits Geometric buffer occupancies, the mean occupancy is a very steep function of the load, for high values of the Hurst parameter.

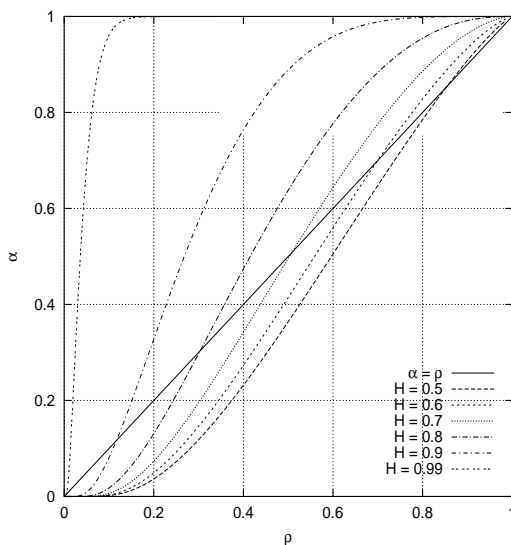


Figure 4: α Vs ρ , single-source Pareto traffic

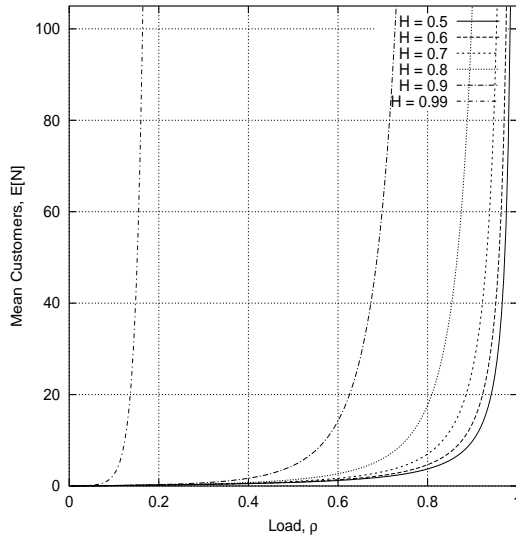


Figure 5: $E[N]$ Vs ρ , G/M/1 queue with single-source Pareto traffic

7 Synthetic Self-similar Traffic Using Merged Pareto Sources

The self-similar traffic model used in this section is based on a Pareto random variable in a different way from that in the above section [6], [28]. The data at a station is considered to be the result of multiplexing numerous data sub-sources. Each such data sub-source has intervals of *ON* and *OFF* time periods. A packet of data (contiguous sequence of bits) flows into the multiplexer during an *ON* period. The *ON* as well the *OFF* time periods are modeled as *iid* Pareto random variables. In numerous different observations of real life data traffic, the underlying Hurst parameter of the Pareto distribution has been very close to 0.9 (see [6], [28]). Therefore, X , each of the *ON* and the *OFF* time periods has the following probability density function:

$$\begin{aligned} f_X(x) &= 1.2x^{-2.2}, \quad x > 1 \\ &= 0, \quad x < 1. \end{aligned} \tag{35}$$

The numerical constants 1.2 and 2.2 are the consequences of the choice of 0.9 for the Hurst parameter, so as to approximate Ethernet traffic characteristics reported in [6] and [28]. The mean of X , $E[X] = 6$ and X has an unbounded variance. Such multiplexed Pareto traffic tends to FBM as pointed out in Section 2.2.

Once a packet is stored and transmitted, the *ON* time of a packet at the output of the transmitter is proportional to the *ON* time of the input packet and the transmitter speed. Therefore, equivalently, a sequence of packets in a data sub-source is a sequence

of tuples of arrival time instants and packet lengths. Each inter-arrival time is the sum of two *iid* Pareto random variables. Each packet length corresponds to one of the above two random numbers. Thus, if X_1 is the *ON* time and X_2 is the *OFF* time, the inter-arrival time is X_1+X_2 and the packet length is X_1 . Inter-arrival times can be scaled with a factor a (without changing the Hurst parameter) to generate data packets of any desired rate. In our simulation experiments, we multiplexed 100 such data sub-sources to yield one data source. The resulting data input rate at a node is $\frac{100}{12a}$ packets per unit time. The packet size can also be similarly scaled without changing the Hurst parameter. Therefore, if a packet of size X_1 data units is stored (X_1 is a random variable as above) and transmitted with a speed of μ data units per unit time, the expected packet *ON* time at the output of the transmitter will be $\frac{6}{\mu}$ time units. A data unit can be, for example, one kilobits.

Intuitively, in the above model, the time for the next arrival depends on the most recent arrival from *each* subsource. Since the elapsed time in each subsource is heavy-tailed, the dependence in inter-arrival times is also heavy-tailed. Therefore, the unbounded variance of each Pareto subsource gives rise to LRD of the merged scheme.

8 Response of Single Queues with Different Types of Bursty Traffic

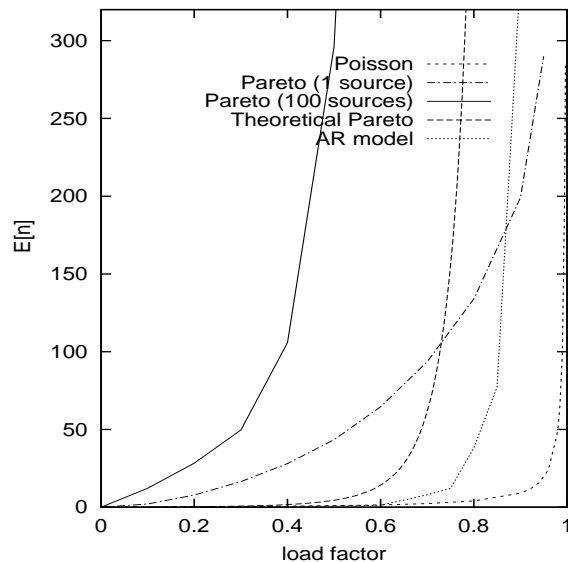


Figure 6: Buffer occupancy Vs load for a single queue

We simulate the behavior of a single queue using the following four inter-arrival time distributions: Poisson, single-source Pareto, 100-source multiplexed Pareto and the AR model-based traffic. In addition, the theoretical load curve for the case wherein inter-

arrival times are Pareto distributed, is also provided for comparison. The service times are exponentially distributed for Poisson, Pareto and AR model based arrivals. When the inter-arrival times are generated using multiplexed Pareto sources, the service times have Pareto distribution. The value of the Hurst parameter selected for the Pareto, multiplexed Pareto and AR model-based traffic is 0.9. The resulting performance curves (refer to Figure 6) clearly illustrate that whereas Poisson traffic saturates the queue at around 98-99% of the offered load, the 100-source multiplexed Pareto traffic saturates the queue at barely 50 - 60% of the load. Queue saturation for the single-source Pareto distribution and the AR model occurs between these two extremes. The traffic generated by the AR model saturates the queue at approximately 80% of the load. The curve obtained by simulation of the queue with single-source Pareto traffic (with $H = 0.9$) varies somewhat from the theoretical curve shown in Figure 5. This variation can be attributed to the following reason. The Pareto random variable has finite mean and infinite variance (for $1 < \gamma \leq 2$). However, the infinite variance cannot be realized practically when generating Pareto random numbers using a computer. In various simulation runs, we found that the steepness of the self-similar curve varied significantly so that saturation occurred at anywhere between 40 - 90% of the load, depending on the value of the selected Hurst parameter. Therefore, in real life data networks, due to the burstiness of network traffic, the actual observed performance of queues is significantly worse than the results that are predicted using Poisson models.

9 Multi-Hop Network Operation with AR Model Based Bursty Traffic

Research work and the results reported here are part of a larger effort that also developed adaptive performance optimization techniques in static and dynamic networks [11]. In a general dynamic network, wireless nodes physically move around and/or can be inactive over some time periods. Such networks are known as ad hoc networks. Under such dynamic conditions, route discovery and maintenance are required, in addition to performance optimization. A common approach to solve all these problems and a specific solution algorithm have been developed for this purpose. It is called the “Statistically Multiplexed Adaptive Routing Technique” (SMART). A very important conclusion from this study is that under bursty data traffic conditions, even static networks benefit enormously from the use of the adaptive performance optimization approach. In this section, we report simulation experiments and results on static network operation fielding the AR model-based bursty traffic. Simulation work based on smooth traffic and other forms of bursty traffic are reported by the authors in [12]. We also discuss below the present results and the corresponding results for other traffic models appearing in [12]. Results on dynamic networks are reported in Kulkarni and Dattatreya [12], [13].

SMART gathers estimates of delays between a source node and a destination node through different possible links originating from the source node. The source node uses

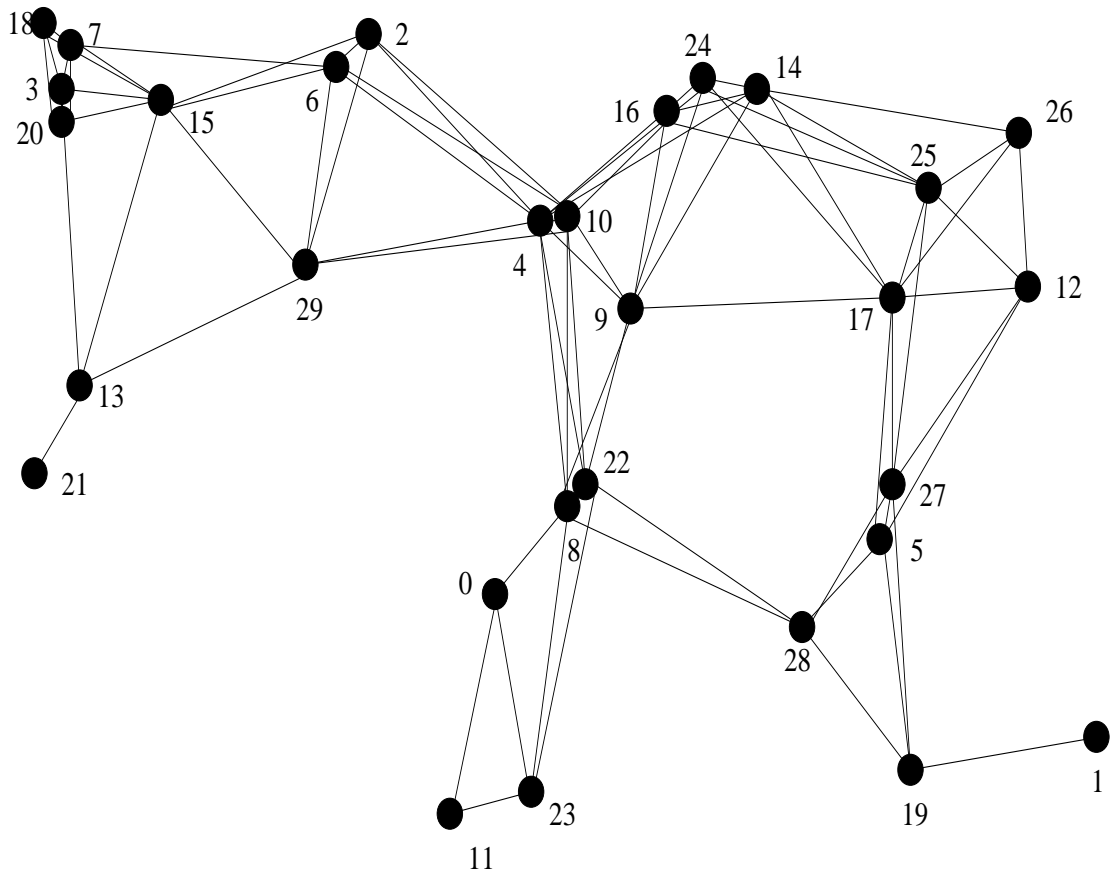


Figure 7: Experimental network)

these delay estimates to calculate routing probabilities. These routing probabilities are used to randomly select the link through which the source node sends a packet to one of its neighbors for forwarding to the destination. Every intermediate node along the path uses a similar policy for forwarding packets. Thus, routing decisions are purely local. This is called statistical multiplexing and it distributes traffic between a source and a destination over many different paths. The delay estimates are frequently obtained on-line, as the network operates. For example, in the network of Figure 7, node 28 has only 5 possible choices for routing to the eventual destination of node 24. Each intermediate link also uses the same local approach to forward traffic to the destination. Therefore, in Figure 7, two packets routed from the node 28 through the same node 5 for the eventual destination of node 24 can take different routes, for example, through nodes 17 and 12, respectively, beyond node 5, to reach the final destination. This is known as statistical multiplexing. Thus, though decisions are local and limited to the outgoing links available at a node, numerous combinations of the overall paths are available for communication between a pair of source and destination nodes. At every node, SMART maintains and updates routing probabilities for forwarding decisions through different links, for every destination. Many of these probabilities may be zero so that from a node A to a destination B , not all outgoing links from A are used. For example, to reach node 24 in Figure 7 from node 28, the probability of the node 28 routing through node 19 may turn out to be zero. In the adaptive routing technique, these probabilities are frequently updated as the network operates. Bursty traffic has a tendency to saturate node buffers in an erratic way and the adaptive approach that continuously reacts to such effects tremendously improves the overall performance.

The static network used for simulation experiments is the one in Figure 7. The service rate of processing and forwarding at every node is 20 kilobits per seconds. The packet size is Pareto distributed with a Hurst parameter of 0.9 and an average length of 64 bytes. SMART updates the routing probabilities at intervals of 30 seconds, at each node. Node operations are not synchronized in any way. Each node in the network has a buffer size of 150 packets. Data packets are dropped if they encounter full buffers. In addition, due to statistical multiplexing, a packet can reach an intermediate node C after visiting all the neighbors of C thus having no unvisited neighbor to move away from C . In such a case, the packet is considered to be cornered and dropped. In order to assess the effects of cornering, no attempt is made to recover a cornered packet. In practice, such attempts are possible and they will reduce the packet drop rate. Communication is by means of a datagram service with no acknowledgments, since the objective is to study the network layer performance. Each experiment is conducted for five hours of network operation.

The traffic load at a node includes the data traffic generated by itself, the traffic it receives from other nodes because the node under consideration is the final destination, and the traffic it receives from other nodes for forwarding to different nodes. The measured traffic at a node over a time interval can vary considerably from the mean node traffic (averaged over all nodes) depending on that node's interconnection to the rest of the network. Additional variations in the short term mean are imposed by the traffic pattern itself. The

network-wide traffic is defined as the mean traffic processed by a node, averaged over all the nodes in the network and averaged over time. The network-wide load is the ratio of the network-wide traffic to the processing capacity of a node; note that all the nodes have identical processing capacities in this study. The network-wide load for all experiments in this study is 50%.

The data packet inter-arrival times are generated using the AR model in the manner described in Section 5.3. We tested the performance of SMART in both the adaptive and the nonadaptive modes of operation. In the adaptive operation, the routing probabilities are updated every 30 seconds (at each node) as described earlier. In the non-adaptive operation, the routing probabilities are updated as earlier in the first 300 seconds of the network operation. Thereafter, the routing probabilities are kept fixed so that routing does not respond to network status.

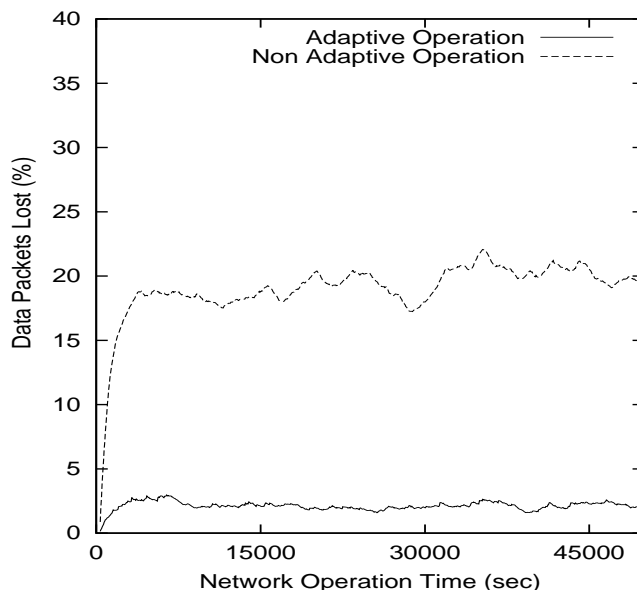


Figure 8: Percentage of data packet loss Vs time, load = 50%, (AR traffic on network)

The results of the experiments are illustrated in Figures 8 and 9. Figure 8 shows (the moving average of) the percentage of data packet loss against the time line of network operation. After the initial transient, the packet losses fluctuate between 17 - 22 % for the nonadaptive approach and around only 2 - 3% for the adaptive operation. In [12] we report results on a similar experiment with Poisson traffic and merged-Pareto based self-similar traffic. Performance figures from the experiments reported here and from comparable experiments in [12] are listed in Table I. The packet losses for Poisson traffic and nonadaptive operation is around 10% and quite steady. The packet losses for Poisson traffic and adaptive operation is around only 3%. The packet losses for the merged-Pareto based self-similar traffic and nonadaptive operation fluctuates wildly over 10 - 35%. The adaptive operation

for the same traffic could bring it down to the range of 2 - 10%. There are two clear observations. The adaptive operation is very effective in reducing the packet losses, for all forms of data traffic. The effect of burstiness of the AR model-based traffic is between those of the smooth Poisson traffic and the merged-Pareto based self-similar traffic. This is intended in the sense that the AR model-based traffic is designed to possess limited scale burstiness.

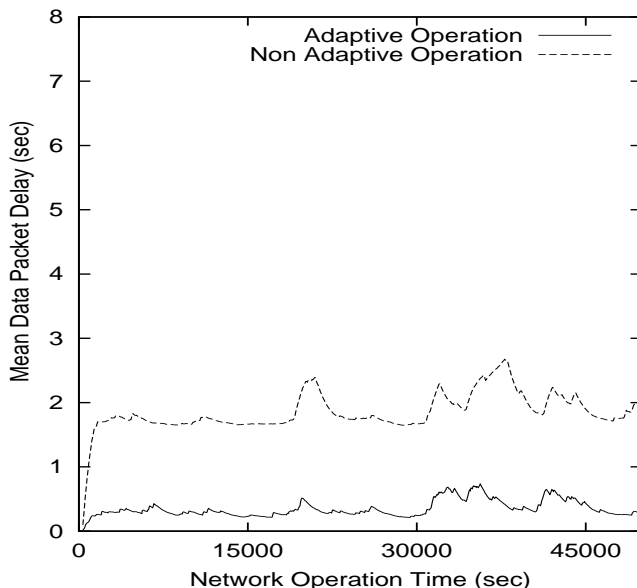


Figure 9: Mean data packet delay Vs time, load = 50%, (AR traffic on network)

Figure 9 shows the (moving average of) mean packet delay as the network operates, again for the two cases of nonadaptive and adaptive network operation fielding the AR model-based bursty traffic. The mean delay for the nonadaptive operation fluctuates between 1.75 - 2.75 seconds. For the same traffic, the adaptive operation results in the mean delay fluctuating between 0.2 - 0.7 second.

The following is a comparison of these results with the corresponding ones for smooth and merged Pareto-based self-similar traffic reported in [12]. For Poisson traffic and non-adaptive operation, the mean delay is very close to 0.064 second. For Poisson traffic and adaptive operation, the mean delay is around 0.058 second. The mean delay for merged Pareto-based self-similar traffic and nonadaptive operation fluctuates wildly between 0.3 - 6.5 seconds. The mean delay for the same traffic and adaptive operation is in the range 0.3 - 1.0 second. Again, similar comparative observations as in the study of packet losses hold for the three traffic types and the two types of network operation (adaptive and nonadaptive). In addition, the mean delay for the smooth Poisson traffic is considerably smaller than for other traffic types, for both adaptive and nonadaptive network operation.

Table I
Comparison of performances across types of traffic and operation

Traffic type	Packet loss		Mean delay	
	Nonadaptive	Adaptive	Nonadaptive	Adaptive
Poisson	$\approx 10\%$	$\approx 3\%$	≈ 0.064 sec.	≈ 0.058 sec.
AR model	17 - 22%	2 - 3%	1.75 - 2.75 sec.	0.2 - 0.7 sec.
Self-similar	10 - 35%	2 - 10%	0.3 - 6.5 sec.	0.3 - 1.0 sec.

10 Conclusion

A systematic development of the properties of traffic processes shows that very high variance and/or LRD in the traffic received over a given time interval results in burstiness. The hyperbolic decay of LRD is a characteristic of self-similarity. The most useful self-similar process is FBM. Its driver process, the continuous time FGN process has unbounded instantaneous variance. Furthermore, FBM is a continuous process over the continuous time domain. This may not be suitable in modeling a sequence of ordered pairs of inter-arrival times and packet sizes. The autocorrelation sequence of discrete time FGN is not summable. Therefore, even discrete time FGN cannot be realized as the output of a linear shift invariant causal and stable system. These properties and the fact that in reality, we may not need LRD persisting over unbounded time intervals have led to the development of our high order AR model for bursty traffic. Such a model incorporates limited scale burstiness.

The analysis of queues with a dependent sequence of inter-arrival times is very difficult. The simplest case of burstiness using Pareto inter-arrival times with unbounded variance is analyzed. Results show the somewhat strange behavior of the equivalent load as the real normalized load varies from 0 through 1. When H takes high values (bursty traffic), the equivalent load is much higher than the real normalized load, resulting in degradation of queuing performance. Experiments conducted on a single queue with different types of bursty traffic (inter-arrival times) show that in *all* cases, the buffer occupancies are higher than in the Poisson traffic case, at all values of the real normalized load. Furthermore, the high variance of the Pareto-distributed service times contributes to further performance degradation. Experiments conducted with two different types of bursty (AR model-based

and multiplexed Pareto) traffic on a static network indicate that network performance too, is significantly worse for the bursty traffic type than that for Poisson traffic. The extent of degradation due to the traffic generated by the AR model is between those due to Poisson and self-similar traffic. In all cases of traffic types, the network performance improves significantly when the network operation is switched from the nonadaptive type to the adaptive type. The improvement is much more pronounced in the cases of bursty traffic.

References

- [1] J. Beran, *Statistics for long-memory processes*. NY: Chapman and Hall, 1994.
- [2] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. San Francisco, CA: Holden Day Inc., 1970.
- [3] D. R. Cox, "Long-range dependence: A review," In H. A. David and H. T. David, eds., *Statistics: An Appraisal, Proceedings of the 50th Anniversary Conference*, pp. 55-74, The Iowa State University Press, IA, 1984.
- [4] M. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: Evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835-846, Dec. 1997.
- [5] A. Erramilli and R. P. Singh, "Application of deterministic chaotic maps to model packet traffic in broadband networks," *Proceedings of the 7th International Teletraffic Congress (ITC) Specialists Seminar*, pp. 8.1.1-8.1.3, Morristown, NJ, Oct. 1990.
- [6] A. Erramilli, O. Narayan and W. Willinger, "Experimental queueing analysis with long-range dependent packet traffic," *IEEE/ACM Transactions on Networking*, vol 4, pp. 209-223, Apr. 1996.
- [7] P. Flandrin, "Wavelet analysis and synthesis of fractional Brownian Motion," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 910-917, Mar. 1992.
- [8] V. S. Frost and B. Melamed, "Traffic modeling for telecommunications networks," *IEEE Communications Magazine*, vol. 32, no. 3, pp. 70-81, Mar. 1994.
- [9] M. W. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," *ACM Computer Communication Review*, vol. 24, no. 4, pp. 269-280, Oct. 1994.
- [10] R. Gusella, "A measurement study of diskless workstation traffic on Ethernet," *IEEE Transactions on Communications*, vol. 38, no. 9, pp. 1557-1568, Sep. 1990.
- [11] S. S. Kulkarni, "Adaptive load-balancing over multiple routes in mobile ad hoc networks," Ph. D. Dissertation, Department of Computer Science, University of Texas at Dallas, Richardson, TX, December 2002.

- [12] S. S. Kulkarni and G. R. Dattatreya, "SMART: Statistically multiplexed adaptive routing technique for ad hoc networks," *Wireless Networks: The journal of Mobile Communication, Computation, and Information*, vol. 10, pp. 89 - 101, Mar. 2004.
- [13] S. S. Kulkarni and G. R. Dattatreya, "Adaptive control of heterogeneous ad hoc networks," To appear in the journal *Wireless Communications and Mobile Computing*,
- [14] W. Lau, A. Erramilli, J. Wang and W. Willinger, "Self-similar traffic generation: The random midpoint displacement algorithm and its properties," *Proceedings of ICC '95*, pp. 466-472, 1995.
- [15] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1-15, Feb. 1994.
- [16] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561-580, Apr. 1975.
- [17] B. B. Mandelbrot, "Self-similar error clusters in communication systems and the concept of conditional stationarity," *IEEE Transactions on Communication Technology*, vol. 13, pp. 71-90, Mar. 1965.
- [18] B. B. Mandelbrot, *The fractal geometry of nature*. NY: W. H. Freeman, 1983.
- [19] B. B. Mandelbrot and J. W. Van Ness, "Fractional Brownian motion, fractional noises and applications," *SIAM Review*, vol. 10, no. 4, pp. 422-437, Oct. 1968.
- [20] I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks," *IEEE Journal on Selected Areas of Communications*, vol. 13, no. 6, pp. 953-962, Aug. 1995.
- [21] A. Papoulis, *Probability, random variables, and stochastic processes*. NY: McGraw-Hill, 1984.
- [22] K. Park and W. Willinger (eds.), *Self-similar network traffic and performance evaluation*, NY: John Wiley and Sons, 2000.
- [23] V. Paxson and S. Floyd, "Wide area traffic: The failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226-244, Jun. 1995.
- [24] V. Paxson, "Fast, approximate synthesis of fractional Gaussian noise for generating self-similar network traffic," *ACM Computer Communication Review*, vol. 27, no. 5, pp. 5-18, Oct. 1997.
- [25] A. Popescu, "Traffic self-similarity," (invited) tutorial, *IEEE International Conference on Telecommunications (ICT)*, Bucharest, Romania, Jun. 2001.

- [26] S. Rambaldi and O. Pinazza, “An accurate fractional Brownian motion generator,” *Physica A*, vol. 208, no. 1, pp. 21-30, Jul. 1994.
- [27] M. S. Taqqu, W. Willinger and R. Sherman, “Proof of a fundamental result in self-similar traffic modeling,” *ACM Computer Communication Review*, vol. 27, pp. 5-23, Apr. 1997.
- [28] 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,” *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, Feb. 1997.
- [29] R. W. Wolff, *Stochastic modeling and the theory of queues*, ch. 8, pp. 381-404, Englewood Cliffs, NJ: Prentice-Hall, 1989.