Proceedings of the 1997 Winter Simulation Conference
ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson

FAST AND PHYSICALLY-BASED GENERATION OF SELF-SIMILAR NETWORK TRAFFIC WITH APPLICATIONS TO ATM PERFORMANCE EVALUATION

Ashok Erramilli
Qmetrix, Inc.
18 Marlow Road
E. Brunswick, NJ 08816, U.S.A.

Parag Pruthi
Qmetrix, Inc.
18 Marlow Road
E. Brunswick, NJ 08816, U.S.A.

Walter Willinger
AT&T Labs–Research
Florham Park, NJ 07932, U.S.A.

ABSTRACT

Self-similarity concepts relate statistical properties of processes observed at different time scales through judicious scaling of time and space. They have recently been shown to be ideally suited to account for the surprising scaling properties that measured network traffic (e.g., number of packets/bytes per time unit) exhibits over a wide range of time scales, from milliseconds to seconds to minutes and beyond. The observed self-similar property in measurements from working packet networks is in sharp contrast to commonly made assumptions about the bursty nature of network traffic and challenges many of the traditional approaches to traffic and performance modeling. In this paper, we illustrate how the self-similar finding gives rise to new mathematical results that (i) clear the way for physically-based approaches to network traffic modeling, (ii) can be combined with high-performance computing capabilities to yield new and fast (i.e., linear in the number of observations) methods for generating self-similar traces, and (iii) provide new insights into the potential performance implications that self-similar traffic can have on the design of network equipment and on the perceived quality-of-service experienced by some of the dominant applications and services. In particular, studying the cell loss dynamics (rather than the traditional long-term cell loss rate) observed at an ATM switch that is fed by self-similar traffic, we discuss the impact of network traffic self-similarity on broadband services such as VBR video and on popular network protocols such as TCP/IP.

1 INTRODUCTION

Recent empirical studies of high-resolution traffic measurements from a variety of different communications networks (e.g., see Willinger, Taqqu, Erramilli 1996 and references therein) have provided ample evidence that actual network traffic is *fractal* in nature in that it exhibits statistical features over many timescales. In particular, these studies have demonstrated that measured traffic rates (i.e., number of packets or cells or bytes per time unit) in LAN/MAN/WAN environments, where data transfer rates typically vary between $1.5 - 155$ Mbps, exhibit surprising scaling properties over a wide range of time scales; that is, actual network traffic looks statistically the same in the small (i.e., at small time scales, on the order of milliseconds or seconds) and in the large (i.e., at time scales on the order of seconds and beyond), and no natural length of a "burst" is discernible: at every time scale ranging from milliseconds to seconds to minutes and beyond, bursts have the same qualitative appearance and cause the resulting traffic to exhibit fractal-like characteristics.

The observed self-similarity properties in measurements from working packet networks is in sharp contrast to commonly made modeling choices in today's traffic theory and practice (where the focus remains on reproducing the bursty behavior of network traffic time scale by time scale) and challenges traditional approaches to traffic and performance modeling. At the same time, it provides new insights into the dynamic nature of actual network traffic, gives rise to novel modeling approaches that take into account the specific features of the underlying networking structure and hence allows for plausible physical explanations of observed traffic characteristics in the networking context. For example, not only can the observed self-similar nature of Ethernet LAN traffic at the aggregate level (i.e., aggregated over all active hosts on the network; see Leland, Taqqu, Willinger, Wilson 1994) be effectively and parsimoniously described and captured by self-similar stochastic processes, but it can in fact be reduced to a simple *ON/OFF* or *busy/idle* behavior at the microscopic level (i.e., for the traffic generated by the individual hosts), with the distinctive feature that the du-

rations of the *ON*- and/or *OFF*-periods themselves vary over a wide range of time scales (see Willinger, Taqqu, Sherman, Wilson 1997). In the WAN context, the observed self-similar characteristic at the aggregate level (see Paxson, Floyd 1995) can be directly related to high variability phenomena at the microscopic level, where in this case, microscopic refers to the level of the individual connections/applications that generate the overall traffic (e.g., WWW, FTP, TELNET; for details, see Willinger, Paxson, Taqqu 1997).

In this paper, we first review in Section 2 one of the main mathematical results in self-similar traffic modeling that clears the way for physically-based approaches and provides simple physical explanations for the presence of self-similar traffic patterns in modern high-speed network traffic, which are, in addition, fully consistent with traffic measurements at the different layers in the networking hierarchy. In Section 3, we illustrate how this new mathematical result can be combined with modern high-performance computing capabilities to yield novel and highly efficient algorithms for synthetically generating self-similar traffic traces. We outline the advantages of this approach, and describe three specific implementations. In Section 4, we discuss the application of these methods to evaluate ATM performance, and to gain insights into the nature of loss processes with self-similar traffic.

2- SELF-SIMILARITY

2.1- Definitions

For a covariance-stationary process $X = (X_i : i \geq 1)$, consider the aggregated processes $X^{(m)}$ with level of aggregation $m \geq 1$, defined by $X^{(m)}(k) = m^{-1}(X_{(k-1)m+1} + \cdots + X_{km}); k \geq 1$. Self-similarity concepts relate statistical properties of X to those of $X^{(m)}$ through judicious scaling of time and space. Following Cox (1984), we call X *exactly self-similar with self-similarity parameter* H if

$$X \stackrel{\mathrm{d}}{=} m^{1-H}X^{(m)}, \quad m \geq 1, \quad 0 < H < 1, \qquad (1)$$

where the equality in (1) means that X and $m^{1-H}X^{(m)}$ have the same finite-dimensional distributions. X is said to be *asymptotically self-similar* if (1) holds as $m \to \infty$. Similarly, we call a covariance-stationary process X *exactly second-order self-similar* or *asymptotically second-order self-similar* (with self-similarity parameter H) if $m^{1-H}X^{(m)}$ has the same variance and autocorrelation function as X, for all m, or as $m \to \infty$. Fractional Gaussian noise with $1/2 < H < 1$ is the standard example of an exactly self-similar (Gaussian) process with self-similarity parameter H.

Mathematically, autocorrelations that decay hyperbolically (i.e., X exhibits *long-range dependence*), a spectral density that exhibits the $1/f^{\alpha}$-phenomenon around the origin, and variances of the arithmetic mean that decrease more slowly than the reciprocal of the sample size are different manifestations of the property that the underlying process X is statistically self-similar with self-similarity parameter $1/2 < H < 1$. All four notions concern statistical properties of X on all time scales and relate them through proper scaling behavior. Intuitively, the most striking feature of (exactly or asymptotically) self-similar or (exactly or asymptotically) second-order self-similar processes with $1/2 < H < 1$ is that their aggregated processes $X^{(m)}$ possess a non-degenerate autocorrelation function $r^{(m)}$ as $m \to \infty$. This behavior is in stark contrast to the conventional short-range dependent processes, all of which have the property that their aggregated processes $X^{(m)}$ tend to second-order pure noise as $m \to \infty$; that is, they satisfy $r^{(m)}(k) \to 0, k > 0$. For further details, see Leland, Taqqu, Willinger, Wilson (1994).

2.2- Physical Explanations

The observed self-similarity properties in measurements from working packet networks is in sharp contrast to commonly made modeling choices in today's traffic theory and practice (where the focus remains on reproducing the bursty behavior of network traffic time scale by time scale) and challenges traditional approaches to traffic and performance modeling. At the same time, it provides new insights into the dynamic nature of actual network traffic, gives rise to novel modeling approaches that take into account the specific features of the underlying networking structure and hence allows for plausible physical explanations of observed traffic characteristics in the networking context. For example, the observed self-similar nature of Ethernet LAN traffic at the aggregate level (i.e., aggregated over all active hosts on the network; see Leland, Taqqu, Willinger, Wilson 1994) can be reduced to a simple *ON/OFF* or *busy/idle* behavior at the microscopic level (i.e., for the traffic generated by the individual hosts), with the distinctive feature that the durations of the *ON*- and/or *OFF*-periods themselves vary over a wide range of time scales.

Mathematically, variability over a wide range of time scales for single random variables such as the durations of successive *ON*- or *OFF*-periods of an active network host can be efficiently modeled using *heavy-*

tailed distributions with infinite variance. To this end, a random variable X with distribution function F is called *heavy-tailed (with index $\alpha > 0$)* if

$$1 - F(x) = P[X > x] \approx cx^{-\alpha}, \text{ as } x \to \infty, \qquad (2)$$

where c is a finite positive constant that does not depend on x. Such distributions are also called *hyperbolic* or *power-law distributions*, and include, among others, the well-known class of *Pareto distributions*. Note that the case $1 < \alpha < 2$ is of special interest and concerns heavy-tailed distributions with finite mean but infinite variance. Intuitively, infinite variance distributions allow random variables to take exceptionally large values with non-negligible probabilities and hence allow for compact descriptions of high variability phenomena that dominate traffic-related measurements at all layers in the networking hierarchy.

To explain self-similarity at the macroscopic level via high variability at the microscopic level, consider M i.i.d. sources, each with its own reward process $W^{(m)} = (W^{(m)}(t), t \geq 0)$, where $W^{(m)}(t) = 1$ or 0, depending on whether source m is in an *ON*-period or *OFF*-period, where the distributions of the *ON/OFF*-periods satisfy the heavy-tailed property (2) with $\alpha = \alpha_{ON}$ and $\alpha = \alpha_{OFF}$, respectively. Let

$$W_M^*(Tt) = \int_0^{Tt} (\sum_{m=1}^{M} W^{(m)}(u))du \qquad (3)$$

denote the aggregated (over all sources) cumulative packet counts in the interval $[0, Tt]$. Willinger et al.- (Willinger, Taqqu, Sherman, Wilson 1997; Taqqu, Willinger, Sherman 1997) determined the statistical behavior of the stochastic process $W_M^* = (W_M^*(Tt), t \geq 0)$ for large M and T and proved the following fundamental theorem in self-similar traffic modeling.

Theorem.- *As $M \to \infty$ and then $T \to \infty$, the aggregate cumulative packet count process $W_M^* = (W_M^*(Tt), t \geq 0)$ satisfies*

$$\frac{\left(W_M^*(Tt) - TM\frac{\mu_1}{\mu_1+\mu_2}t\right)}{T^H L^{1/2}(T) M^{1/2}} \to \sigma_{\lim} B_H(t), \qquad (4)$$

where $H = (3 - \min(\alpha_{ON}, \alpha_{OFF}))/2$, $\sigma_{\lim}$ is a constant, and where the convergence is in the sense of the finite-dimensional distributions.

Heuristically, the Theorem states that the mean level given by $TM(\mu_1/(\mu_1 + \mu_2))t$ provides the main contribution for large M and T. Fluctuations from that level are given by the fractional Brownian motion $\sigma_{\lim} B_H(t)$ scaled by a lower order factor $T^H L(T)^{1/2} M^{1/2}$.

In the WAN context, the observed self-similar characteristic at the aggregate level (see Paxson, Floyd 1995) can also be related to high variability phenomena at the microscopic level. However, in this case, microscopic refers to the level of the individual connections/applications that generate the overall traffic (e.g., WWW, FTP, TELNET; for details, see Willinger, Paxson, Taqqu 1997).

3- SYNTHETIC TRAFFIC GENERATION

There are over a dozen techniques to generate self-similar traffic - exact generation using Cholesky decomposition, the Random Midpoint Displacement (RMD) method, FFT based generation, the Hoskings method, to name a few. The various techniques have features that make them well suited for certain applications (e.g., accuracy in the exact method) but unsuitable for others (e.g., the exact method is only practical for short traces). As we demonstrate in the next section, the class of physically-based generation methods is well-suited for network simulations, and is an obvious choice for ATM performance evaluation.

3.1- Physically-based generation

There are several distinct techniques in this class of generation methods, but the common feature is to mimic the physical basis of self-similarity in network traffic by superposing the output of a large number of individual ON/OFF sources. There are several features of this class of generation methods that makes them well-suited for network simulations: i) they are efficient - generating N samples requires $O(N)$ complexity, while several of the other techniques are $O(N^2)$ (e.g., Hoskings method) or worse; ii) they enable "on-the-fly" generation of traffic, in contrast to other methods (notably, the RMD and FFT methods) which require the generation of the entire trace before the simualtion can proceed; iii) by varying M (the number of sources) and T (the aggregation period), they offer the complete range of speed vs accuracy trade-offs; iv) depending on the specific context, network simulations require generation at the level of individual sources, or of limited aggregates, or of a large number of sources; by definition, this class of methods incorporates all these scenarios as special cases v) some methods (e.g., RMD, FFT) occasionally result in negative "arrivals" - while this is a limitation of all Brownian motion traffic models, the physically-based generation methods avoid this problem entirely vi) these methods are well suited for parallel simulation for e.g., each ON/OFF source can be implemented on a separate processor.

We will discuss three specific methods that are based on the principle of physically-based generation. The first method is based on simulating an $M/G/\infty$ queueing system, with Poisson arrivals, an infinite number of servers (i.e., a pure delay system), and a service time distribution with infinite variance. It is well-known that the queue length distribution of such a system is Poisson, and that the queue length process is LRD (Cox 1984); thus a discrete sampling of the continuous time $M/G/\infty$ queue will yield an approximation to a self-similar process. Physically, this system is an abstract representation of the arrival of TCP connections to a network (which is observed to be Poisson), and the empirically observed heavy-tailed densities that characterize session durations. Details of the traffic patterns within a session (such as TCP dynamics) are abstracted out, and traffic is effectively assumed to arrive at a constant rate within the session. If the service time distribution is taken to be Pareto (Equation (2) for $x \geq \beta$; 0 otherwise), the average value of the queue length process is:

$$\frac{\lambda\beta\alpha}{\alpha - 1} \tag{5}$$

and its autocorrelation function is:

$$r(k) = r_0 k^{-(2-2H)} \tag{6}$$

Speed vs accuracy trade-offs are controlled by the relative rate of sampling / aggregating the $M/G/\infty$ queue vs the Poisson rate of arrivals and the average service rate. If the arrival rate relative to the sampling interval is high, in effect many events are being simulated to generate a sample point.

A second method models the ON/OFF sources explicitly with heavy-tailed sojourn times. Samples y_i are drawn from a uniform density and transformed to follow a Pareto density as follows:

$$\tau_i = c/{y_i}^{1/\alpha} \tag{7}$$

For simplicity, the ON/OFF times can be assumed to follow Pareto distributions with the same exponent between 1 and 2. In the ON state, packets are once again assumed to arrive at a constant rate. Physically, each of the sources can be interpreted as representing a permanent virtual circuit or PVC, and the ON/OFF periods represent successive idle and busy periods on the PVC. As before, aggregation first in sources reduces the effects of higher-order statistics (in the limit leading to a Gaussian), and then aggregating in time attenuates non-scaling high-frequencies (in the limit leading to an exactly self-similar process).

The third method is a deterministic implementation of the above method, using the notion of *deterministic chaotic maps*. In this approach, a single source is modeled by a chaotic map which describes the evolution of a bounded state variable. A packet is generated every time the state variable is in a specified sub-interval. The map and its parameters can be carefully chosen to model the full range of ON/OFF behavior (from periodic, to exponential to Pareto, for example). For the example, the following so-called "fixed point double intermittency" map generates sojourn times in the ON/OFF states that are heavy-tailed (Pruthi 1995):

$$f_1(x) = \frac{x}{(1 - c_1 x^{m-1})^{1/m-1}}, \quad x \leq d \;\; OFF - state; \tag{8}$$

$$f_2(x) = 1 - \frac{1 - x}{(1 - c_2(1 - x)^{m-1})^{1/m-1}}, \quad x > d \;\; ON - state. \tag{9}$$

Aggregate flows can be derived from this map in a number of ways. The most direct approach is to superpose the outputs of a large number of such maps (or alternately, one can use invariant densities which characterize the limiting distribution of the state variable also to describe aggregation). Simulation time can be sharply reduced by calculating sojourn times directly, rather than calculating the evolution of the map at every step. The above fixed point double intermittency map exhibits interesting scaling behavior, and one can directly express the map in terms of an evolution of successive ON/OFF periods.

Nominally, all three methods provide batch generation of traffic arrivals, and one can distribute the arrivals within the generation interval in a number of ways: deterministic spacing, or uniform distribution (which is equivalent to a process that is locally Poisson), or by using RMD (which is fundamentally an interpolation scheme) to extend the scaling behavior to within the generation interval. One motivation for using self-similar traffic models is of course that such fine timescale structure is relatively inconsequential in determining performance (see Erramilli, Narayan and Willinger 1996).

4- ATM PERFORMANCE EVALUATION

In this section, we will illustrate the use of physically based generation methods in ATM performance evaluation. The complexity of ATM traffic controls, the nature of ATM traffic (mixtures of a number of service categories and priority levels) as well as the need

to estimate stringent performance objectives (e.g., cell loss rates of the order of 10^{-7} or less) makes ATM performance evaluation particularly challenging. While recent analytical results e.g, Narayan (1997), on the queueing behavior induced by self-similar traffic are promising, in-depth evaluation of ATM peformance still requires detailed simulations. Analytical results are useful in validating simulations, and as discussed below, in extrapolating simulation results.

4.1 ATM Simulations Using Physically-Based Generation

Traditionally, ATM simulations have relied on hierarchical models of traffic behavior, e.g., cell level, burst level, connection level and so on. Rather than modeling "one timescale at a time" explicitly, self-similar traffic models describe the fluctuations in traffic intensity in a unified, parsimonious manner. Physically-based generation is particularly well suited for ATM performance evaluation. The efficient and "on the fly" nature of the generation scheme allows the simulation of long traces needed to estimate small probabilities adequately. A second aspect of traffic characterization is the need to model spatial dependencies in traffic, beyond temporal characteristics such as correlations. In output buffered ATM switches, for example, cell losses do not occur until multiple input ports generate traffic to the same output port for an interval of time. Persistence in spatial patterns is therefore important to model. Physically based methods implicitly capture this dependence. For example, one can represent a single VC as carrying traffic from one or more heavy-tailed ON/OFF sources, with one or more VCs carrying traffic over a VP. All the traffic in an ON period for a given source would then be bound to the same output port. In contrast, generating traffic in bulk and uniformly distributing it over a number of output ports will lead to an underestimation of cell losses.

Self-similarity can also be exploited to reduce simulation time. Nominally, the FBM model is described by three parameters - a mean rate m, the Hurst parameter H, and a "peakedness" parameter a that quantifies the magnitude of fluctuations. In addition, an ATM queue, at the highest abstraction level, can be described by a link capacity C and a finite buffer B. Even this very abstract representation of ATM traffic contention therefore requires exploration of cell loss rate as a function of 5 parameters, which can be prohibitive even with fast traffic generation methods. Scaling relationships can be used to extrapolate performance from one combination of parameters to other combinations. Two sets of parameters (m_1, a_1, H, C_1, B_1) and (m_2, a_2, H, C_2, B_2) will result in the same buffer-level exceedance probability provided:

$$\frac{C_1 - m_1}{(a_1 m_1)^{1/2H}} {B_1}^{(1-H)/H} = \frac{C_2 - m_2}{(a_2 m_2)^{1/2H}} {B_2}^{(1-H)/H} \tag{10}$$

Thus if the dependence of CLR on the mean rate (for example) was determined for one combination of (a, H, C, B) through exhaustive simulations, the CLR vs mean rate curve for other values of these parameters can be directly estimated from the above scaling relation. Note that the scaling relation is fairly general in that it holds for all values of buffer sizes and for any self-similar process. The limitation is that the scaling relation does not apply when the Hurst parameter changes.

4.2 On the Dynamics of Cell Losses in an ATM Switch

Cell losses are a primary source of performance degradation in ATM networks, and it is invariably characterized by an average rate in performance evaluation. One basic lesson from the study of fractal processes is that average rates are not very meaningful in describing bursty processes, and cell losses are notoriously bursty. We illustrate this point in this section using simulations based on chaotic maps.

Cell losses can have a dramatic impact on application level performance, beyond the actual cells lost, because of the way applications react to losses. The performance of many video coders/decoders is tightly coupled not only to average cell losses but also the manner in which losses occur (i.e., if losses occur in bursts or are completely random). Similarly, flow control algorithms like TCP employ techniques to reduce congestion and various versions of TCP employ different techniques for handling and recovering from packet losses; and the performance of all these schemes depends upon the nature of the losses. Therefore, a more complete characterization of the cell loss process is essential for evaluating the performance of data transmission (as the performance of flow control algorithms depends upon the nature of the losses) and real-time services (as audio/video coders are optimized for dealing with certain prescribed loss characteristics).

We simulate a simple Permanent Virtual Circuit on a ATM switch where the output link has a finite FIFO buffer leading to losses under heavy loads. The input into the switch is assumed to be self-similar and is modeled by an aggregation of chaotic maps as

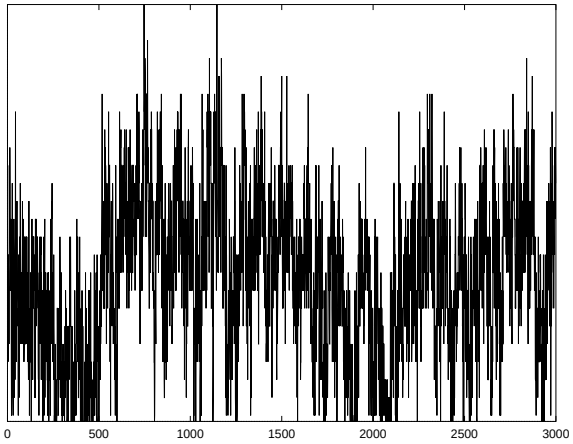

Figure 1: A Synthetically Generated Traffic Process Showing Number of Cells per Unit Time on a Permanent Virtual Circuit Connection. Note the Highly Bursty Nature of the Arrivals as well as the Existence of Significant Power in the Low-Frequency Component Indicating Long-Range Correlations.

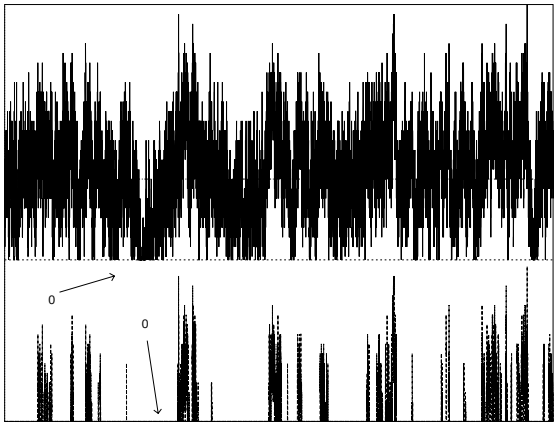

Figure 2: The Packet Arrival Process (top panel) and Resulting Cell Loss Process (bottom) Indicating the Highly Bursty Nature of Cell Losses. Cell Losses Can Be Either Zero or Very Large Due to the Highly Variable Nature of the Input Traffic. For Illustration We Have Shown the Loss Process for $H = 0.85$ at 80% Utilization but Similar Results Are Otained Irrespective of H (in the range $0.5 < H < 1$) and Utilization.

described in Pruthi (1995). The parameters for the input traffic are taken from various emperical studies and in particular the Hurst parameter H is > 0.5.

Figure 1 shows a sample traffic trace into the ATM switch, specifically, a time series of the number of cells per unit time (which is some multiple of a cell time slot). Note the highly bursty nature of the incident traffic, and in particular note the presence of low-frequency components which indicate the presence of $1/f$-noise characteristic of self-similar processes. This low-frequency component gives rise to cell losses that are highly clustered and bursty.

In Figure 2 we simultaneously show both the input trace to the ATM switch (top) and the resulting loss process (bottom). Note that the zero of the input process is marked by a line at the center of the figure (pointed to by a arrow) and the zero of the cell loss process is marked by the line at the bottom of the figure (also pointed to by another arrow). Also note that both figures are to scale which illustrates that the higly variable nature of the input can lead to a large number of losses in a short period of time (although the average number can be very small). For the illustration shown in this figure the mean number of cells lost was 3.97 whereas the peak cell loss was as high as 100. This illustration clearly shows the highly bursty and clustered nature of cell losses with self-similar input traffic and the need for new metrics other than averages to characterize such processes. Similar results are obtained for other choices of parameters.

Overflow processes in traditional telephony are described using mean and peakedeness parameters. However, overflow processes resulting from self-similar traffic flows have not been studied before and analytical results are not available at this time. Our empirical results provide some intuition on the nature of the cell loss process. One description of the loss process is possible in terms of the well-known packet train process. To distinguish between inter cell loss times within a burst, and interval between bursts, we define nonoverlapping intervals of size T, and declare an interval to be lossless if there are no losses within that interval. A lossy period is defined as a complement of a loss-less period. For a fixed T (within a reasonable range, the particular value does not matter) we count the number of consecutive lossless intervals between loss intervals. Based upon these counts we generate a cumulative distribution function which characterizes the inter-loss time. The simulations indicate that the inter-loss time process is indeed heavy-tailed (in the weak sense that they are heavier than the exponential), as shown by the empirically determined distribution functions in Figure 3. For the empirical data set, the Weibullian distribution (lines) shows a good fit

$$P(L > x) = \alpha e^{-\gamma * x^{\beta}} \tag{11}$$

to the two plots (points) with the parameter $\beta < 1$. We remark that these are typical results for vari-

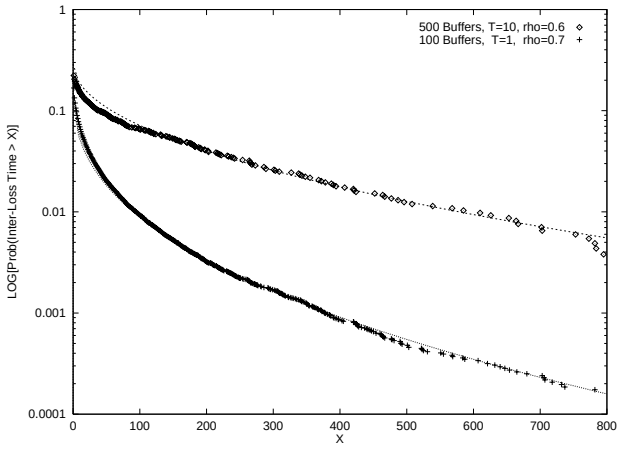

Figure 3: The CDF for the Inter-Loss Intervals; i.e., the Probability of the Number of Lossless Intervals Between Losses $> X$.- Note the Semi-Log Scale. Shown Are the CDFs for Two Cases (points): (a) With 500 Buffers and T=10 and 60% Utilization (top), and (b) With 100 Buffers and T=1 and 70% Utilization (bottom).- Also Shown Are Weibullian Fits to the Two Curves (lines).

ous parameter values and possess a Weibullian density. However, given the difficulties of distinguishing between Weibullian densities and heavier power law densities over a limited range of data samples, more extensive analysis is required to establish the nature of the heavy-tailed distribution. However, the heavy-tailed nature of loss events clearly shows the need for evaluating current metrics (such as means and variances) for QoS and flow control and question their applicability in quantifying actual losses and their usefulness as measures of performance.

4.3- The Effects of Traffic Shaping

It is expected that sessions seeking guaranteed levels of performance from ATM networks will be required to offer traffic conforming to a traffic contract, which is enforced using policing mechanisms. Shaping is required to ensure that the offered traffic is indeed conforming. An issue in ATM performance evaluation is therefore the impacts of shaping and policing on traffic characteristics, and on user perceived performance. Physically-based generation models provide direct insights on these issues. Consider a situation in which each VC can be represented by an individual heavy-tailed ON/OFF source, and the impacts of peak rate shaping on such a source. The net effect of peak rate shaping is to lower the rate at which traffic is generated in the ON state while prolonging the duration of the ON state. Applying the results outlined in Section 2, it is seen that the exponent of the sojourn times, and hence the Hurst parameter, is unchanged, whereas the magnitude of fluctuations (indicated by $\sigma_{\lim}$) is reduced. Note that application level throughputs are proportionately decreased, and the net effect of shaping can be to move the throughput bottleneck from network resources to the shaper itself. Such an analysis can also be used to evaluate the impacts of shaping when VC traffic resembles an aggregate of ON/OFF sources, and for sustained cell rate shaping etc.

5 CONCLUSIONS

The ubiquity of fractal traffic features spanning many timescales in network traffic traces, as well as their relevance in determining network performance, motivates the search for fast synthetic generation methods. A particularly attractive class of methods for network performance evaluation is based on the physical basis of the observed self-similarity in traffic, which arises from aggregating a large number of sources with infinite variance busy and idle periods. This class of generation methods has numerous advantages in network performance simulations, including linear time in speed, suitability for parallel simulations, non-negative sample values, and versatility. The methods have been applied in ATM performance evaluation to evaluate cell loss rates, where scaling relationships can be used to reduce simulation effort. Finally, we used simulations to examine the nature of cell losses, and in particular illustrated the bursty nature of cell losses which cannot adequately be described by long-term average rates. Future work in this area could include optimized methods for real-time generation at OC-12 rates and above (for use in load boxes), techniques to exploit self-similarity in performance simulations further (for example, using "coarse graining" to reduce simulation time), parallel simulation methods, and engineering applications (e.g., evaluating routing vs switching trade-offs).

ACKNOWLEDGMENT

A. Erramilli and W. Willinger were partially supported by the NSF-grant NCR-9628067 at the University of California at Santa Cruz.

REFERENCES

Cox, D. R. 1984. Long-range dependence: A review. In: *Statistics: An Appraisal*, ed. H. A. David and

H. T. David, 55–74. Iowa State University Press, Ames, Iowa.

Erramilli, A., O. Narayan, and W. Willinger. 1996. Experimental queueing analysis with long-range dependent packet traffic. *IEEE ACM Transactions on Networking* 4:209–223.

Erramilli, A., R. P. Singh, and P. Pruthi. 1995. An application of deterministic chaotic maps to model packet traffic. *Queueing Systems* 20:171–206.

Lau, W.-C., A. Erramilli, J. L. Wang, and W. Willinger. 1995. Self-similar traffic generation: The random midpoint displacement algorithm and its properties. In *Proceedings of the ICC '95*, 466–472, Seattle, WA. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Leland, W. E., M. S. Taqqu, W. Willinger, and D. V. Wilson. 1994. On the self-similar nature of Ethernet traffic (Extended Version). *IEEE/ACM Transactions on Networking* 2(1):1–15.

Narayan, O. 1997. Exact asymptotic queue length distribution for fractional Brownian traffic. Submitted to *Advances in Performance Analysis*, ed. V. Ramaswami.

Paxson, V. 1995. Fast approximation of self-similar network traffic. Preprint.

Paxson, V., and S. Floyd. 1995. Wide-area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking* 3:226–244.

Pruthi, P. 1995. An application of chaotic maps to packet traffic modeling. Ph.D. thesis, Department of Teleinformatics, Royal Institute of Technology, Stockholm, Sweden.

Willinger, W., M. S. Taqqu, and A. Erramilli. 1996. A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks. In *Stochastic Networks: Theory and Applications*, ed. F. P. Kelly, S. Zachary, and I. Ziedins, 339–366. Oxford, UK: Clarendon Press.

Willinger, W., V. Paxson and M. S. Taqqu. 1997. Self-similarity and heavy tails: Structural modeling of network traffic. In: *A Practical Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions*, ed. R. Adler, R. Feldman and M. S. Taqqu, Birkhauser Verlag, Boston (to appear).

Willinger, W., M. S. Taqqu, R. Sherman, and D. V. Wilson. 1997. Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Transactions on Networking* 5(1):71–86.

AUTHOR BIOGRAPHIES

PARAG PRUTHI is currently with the Network Management and Research department of Qmetrix International Inc, East Brunswick, NJ. From 1987-1996 he was with Bellcore where he worked on many aspects of communications. He left Bellcore from 1993-1995 and received his Ph.D. in telecommunications from the the Royal Institute of Technology, Stockholm, Sweden in 1995. His research intrests include analysis of high-speed communication network traffic and modeling traffic flows in broadband/wireless networks. He is a member of IEEE.

ASHOK ERRAMILLI is currently the President of Qmetrix International Inc., after a number of years at Bellcore, where he managed a research group engaged in traffic analysis and network design. His research interests include fractal traffic modeling, traffic management methods for broadband and cellular networks, simulation methods, and traffic engineering economics. He is a member of IEEE.

WALTER WILLINGER is with the Information Sciences Research Center at AT&T Labs-Research, Florham Park, NJ. From 1986-1996, he was with Bellcore Applied Research, Morristown, NJ. His research interests include mathematical modeling of traffic dynamics in modern high-speed communications networks, analysis of large data sets, mathematical finance, and nonstandard analysis. He is a member of SIAM, IMS, IEEE, INFORMS and the Bernoulli Society.