Characterizing and Modeling Network Traffic Variability

Sarat Pothuri, David W. Petr, Sohel Khan†
Information and Telecommunication Technology Center
Electrical Engineering and Computer Science Department, University of Kansas, Lawrence, KS 66045
†Research, Architecture and Design Group, Sprint, Overland Park, KS 66212

***Abstract*— This paper advocates using the recently introduced Index of Variability (IDV) and a new measure, Peak Rate Variability (PRV), to characterize the variability (burstiness) in real communications network traffic over the entire range of time scales. Further, we suggest the general hyperexponential interarrival distribution as a model suitable for network traffic and evaluate the ability of the third-order hyperexponential model to capture IDV, PRV, and queuing characteristics. Although the hyperexponential interarrival distribution holds promise for network traffic modeling, in part due to its analytical tractability, we conclude that hyperexponential models with order larger than 3 will be required to adequately model the burstiness of real network traffic.**

I. Introduction

Since the publication of [1], there has been intense interest in the self-similarity (fractal nature) and long-range-dependence (LRD) of communications network traffic. One mathematical definition is that a stationary process X_n is self-similar with Hurst parameter H if it satisfies the following scale-invariant behavior in the sense of finite-dimensional distributions [2]

$$X = m^{1-H} X^{(m)} \tag{1}$$

where $X^{(m)}$ is the aggregated process derived from X_n by averaging the X_n values in non-overlapping blocks of m instants, replacing each block by its sample mean. The constant H ($0 < H < 1$) is a measure of the degree of self-similarity of the process. If $0.5 < H < 1$, then the process is said to exhibit LRD; if $0 < H < 0.5$, it exhibits short-range dependence. For H=0.5, the process consists of uncorrelated samples.

One focus of past research has been on methods for accurately estimating the value of H from a given traffic trace. For example, if a process is asymptotically second-order self-similar, we will have

$$\sigma^2_{X^{(m)}} \sim c \cdot m^{-2(1-H)} \quad as\ m \to \infty \tag{2}$$

and the plot of $\log(\sigma^2_{X^{(m)}})$ versus $\log(m)$, known as the aggregated variance-time plot, becomes a straight line as $m \to \infty$. Hence one simple way to estimate H is by constructing the aggregated variance-time plot, checking if the plot becomes "approximately" a straight line as $m \to \infty$, and if so, estimating the slope of the line.

In section II, we discuss two relatively new characterizations of network traffic variability, both of which are functions of time scale: Index of Variability (IDV), which is a generalization of the Hurst parameter, and Peak Rate Variability (PRV). In section III, we demonstrate that the hyperexponential interarrival distribution can yield a variety of IDV functions. In section IV, we construct a 3^{rd} order hyperexponential model from traffic trace data and evaluate the similarity between the trace and the model in terms of IDV, PRV, and queuing delay. We present our conclusions in section V.

Research sponsored by Sprint Corp.

II. Traffic Variability as a Function of Time Scale

A. Index of Variability

Asymptotically second-order self-similar processes are appropriately characterized by the scalar parameter H and are sometimes known as mono-fractal processes. However, for many network traffic processes, the variance-time plot may not tend to a straight line. These are referred to as multi-fractal processes [3]. Hence, another metric is needed, and [4] has suggested a generalization of the Hurst parameter, known as the Index of Variability (IDV), that captures the degree of self-similarity over all time scales.

IDV is related to the Index of Dispersion for Counts that has been frequently used for describing the variability of network traffic over different time scales. IDV is a function of time scale (or aggregation level) τ. IDV may be defined as

$$\mathrm{IDV}(\tau) = 0.5 \frac{\mathrm{d}\log(\sigma^2_{X^{(\tau)}})}{\mathrm{d}\log(\tau)} + 1 \tag{3}$$

From (2) and (3), we can see that when a random process is asymptotically second-order self-similar, the IDV becomes a constant across all large time scales at a value that is equal to the Hurst parameter H of the process.

We now introduce a practical method for estimating IDV from a given traffic trace. For illustration, we use traces consisting of counts of Asynchronous Transfer Mode (ATM) cells (fixed-length packets) in 5 ms intervals for a period of 24 hours. Although the higher-level protocols and applications being carried by these cell streams is not known exactly, it is likely that these traces represent either IP over ATM or IP over Frame Relay over ATM.

Assuming that a given trace is a representative sample function from a stationary, ergodic random process, we can estimate the IDV of the process as follows. For each time scale τ, we

consider the cell counts in the non-overlapping τ-second intervals to be sample values of a random variable and then estimate the variance of the random variable. A plot of the log of these variances vs. $log(\tau)$ corresponds to the aggregated variance-time plot previously discussed. Values of τ should be chosen to be linearly spaced on a log scale, ranging from the smallest possible value (5 ms in our case) to the largest value that allows accurate variance estimation.

The solid line of Fig. 1 shows the resulting variance-time plot for a particular sample trace. The largest time scale considered is 1000 s so that variance estimates are based on no fewer than 86 sample values. Note that the curve does not tend to a straight line as τ increases. Note also that the significant amount of "noise" in the curve would make direct estimation of the derivative exceptionally noisy. For this reason, it is appropriate to fit a polyomial curve (8^{th} order in our case) to the variance-time plot, as shown by the dashed line in Fig. 1. This also allows direct computation of the derivative, and hence the IDV.

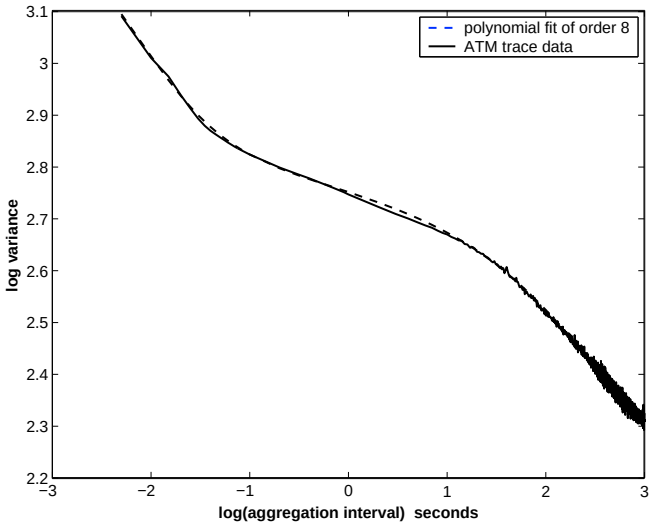

Fig. 1. Variance-Time Plot and Polynomial Fit

The resulting IDV for the trace is shown in Fig. 2, along with an IDV calculated for a trace of a Poisson arrival process (H=0.5), and another trace for fractional Gaussian noise (fGn), known to have a theoretical Hurst parameter of H=0.90. We can see that the IDV of the latter two traces tends to their respective H values (although the fGn IDV tends to fall off somewhat, perhaps due to approximate methods used for trace synthesis [5]), but the IDV of the ATM traffic trace is not constant over any significant range.

B. Peak Rate Variability

As discussed in [1], the scalar quantity *peak-to-mean ratio* is not a particularly good characterization of traffic burstiness because the value obtained depends critically on the time scale used for the calculation of the peak rate. Rather than dismissing

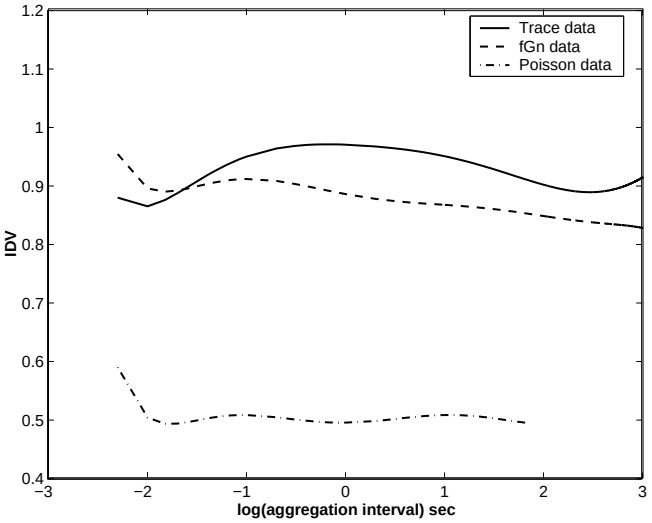

Fig. 2. Index of Variability Comparison

peak-to-mean ratio as a means of characterizing traffic variability, however, we propose here that it should be generalized to Peak Rate Variability (PRV), which is the peak rate of a process as a function of time scale (or aggregation level).

For each aggregation level τ, we calculate the average traffic rate in each non-overlapping τ-second interval, then choose the maximum of these values to be the peak rate at time scale τ. Plotting these as a function of τ yields the PRV curve.

PRV values can be expressed in units of bits or bytes per second if packet lengths are considered, otherwise in units of packets per second. For the constant-length ATM cells of our trace data, the two are related by a constant, so we choose b/s units. Fig. 3 shows PRV curves for the ATM traffic trace (solid line), a trace derived from a Poisson arrival process (dash-dot line), and a trace derived from a hyperexponential model (dashed line, to be discussed in section IV), each with the same mean rate. In all cases, we have assumed that each packet is an ATM cell.

Note that the Poisson PRV decreases smoothly toward its mean, as might be expected. However, the PRV for the ATM trace data reveals some very interesting burstiness characteristics. The peak rate of the ATM trace remains almost constant at nearly 10 Mb/s from a time scale of less than 100 ms (0.1 s) to more than 10 s, then drops rapidly for larger time scales. Even at a time scale of 1 hour (3600 s), the peak rate (largest of the 24 1-hour average rates) of approximately 4 Mb/s is significantly above the mean rate of 2.2 Mb/s (5183 cells/s). The PRV clearly reveals that this ATM traffic trace, which is typical of the ones we have observed, exhibits substantial burstiness across a very wide range of time scales.

The PRV curves reveal details about the peak rate behavior of traffic traces, but it would still be desirable to have a scalar measure of traffic burstiness. In our analysis of ATM traffic traces, we have found max-to-min peak ratio, defined as the peak rate on a 5-ms time scale (max) divided by the peak rate on a 1-hour time scale (min), to be a useful measure. For example, for hun-

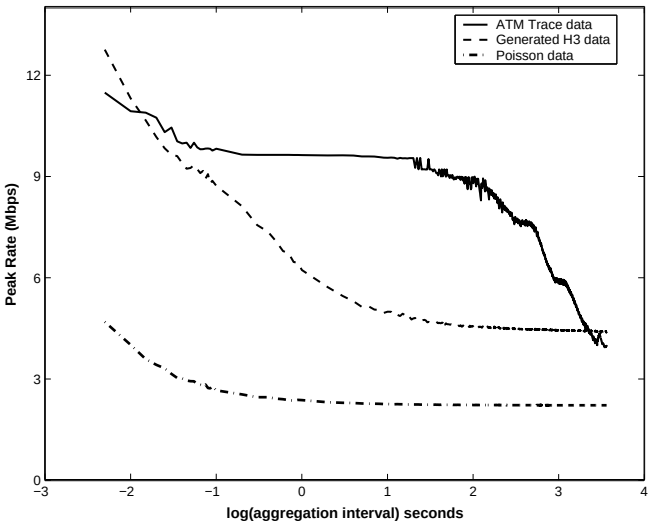

Fig. 3. Peak Rate Variation (PRV) Comparison

dreds of ATM Virtual Circuit Connection traffic traces, we have found a strong correlation between max-to-min peak ratio and the minimum (1-hour) peak rate. Linear regression yields the following relationship with a coefficient of determination of r^2 = 0.85.

$$log(ratio) = -0.66 \cdot log(1\_hour\_peak\_rate) + 1.3 \quad (4)$$

The consistency of this relationship, if it can be shown to hold across a broad class of connections, could allow one to estimate the peak rate on a small time scale from measurement of peak rate on a large time scale. This could be extremely useful for traffic management since small time-scale measurements are quite costly relative to large time-scale measurements.

III. Hyperexponential Model for Traffic Variability

Once the characteristics of network traffic have been determined, a very important next step is to find mathematical models that exhibit characteristics similar to those found in network traffic. This section discusses the potential utility of the hyperexponential distribution as a relatively simple and robust model for packet interarrival times. An n^{th} order hyperexponential probability density function (pdf) H_n for a random variable X (representing interarrival time in our case) is given by:

$$f_X(x) = w_1\lambda_1 e^{-\lambda_1 x} + w_2\lambda_2 e^{-\lambda_2 x} + ... + w_n\lambda_n e^{-\lambda_n x} \quad (5)$$

In this paper, we will will focus on the 3^{rd} order hyperexponential distribution H_3 that has six parameters (w_i and λ_i) and four degrees of freedom since the weights w_i must sum to unity and we wish to fix the mean interarrival time at a particular value $1/\lambda$ given by $1/\lambda = \sum w_i/\lambda_i$.

The IDV of the H_3 interarrival distribution can be calculated using the methods given in [4]. Here, we illustrate the variety of IDV curves that can be obtained from the H_3 interarrival distribution. We begin by considering a *balanced* H_3 distribution, which reduces the degrees of freedom to two by further requiring $w_i/\lambda_i = 1/3\lambda$. In Fig. 4, we allow only a single degree of freedom by requiring $1/\lambda$ = 1/5183 and fixing w_2 = 0.75. Fig. 4 shows that the balanced H_3 distribution can produce unimodal IDV curves with values very close to 1.0 over time scales spanning several orders of magnitude. Note that increasing the mean interarrival time $1/\lambda_1$ of the first term in the H_3 pdf to very large values (and correspondingly reducing the weight w_1 of this term to very small values) is required to extend the time scale region for which the IDV value is close to 1.0. Note that all of the IDV curves eventually approach the asymptotic Hurst parameter value of H=0.5 for any hyperexponential distribution. However, with large enough $1/\lambda_1$ values (small enough w_1 values), the balanced H_3 interarrival distribution can be made to be *practically indistinguishable* from a second-order self-similar process with Hurst parameter close to 1.0.

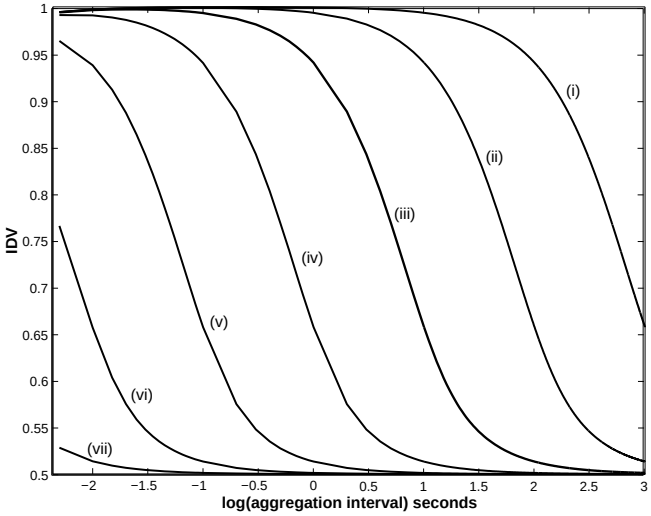

Fig. 4. IDVs of Balanced H_3 Interarrival Distribution with $w_2 = 0.75$; the w_1 values of the curves range from 1e-7 for curve (i) to 0.1 for curve (vii), incrementing by an order of magnitude for each curve.

However, as shown in Fig. 2, IDV curves of real traffic can exhibit more complicated structure than the simple unimodal form of Fig. 4. We now introduce a new form of the general H_n distribution that we call *doubly-balanced* because in addition to the conditions imposed by the balanced definition, we further require $w_i/w_{i+1} = k$ for some constant k and for $i = 1, 2...n-1$. A doubly-balanced H_3 distribution has only a single degree of freedom. Fig. 5 shows that doubly-balanced H_3 interarrival distributions can have bi-modal IDV curves, with the value of k controlling the location (in time scale) of the "valley" between the two "hills".

IV. Evaluation of H_3 Traffic Model

Having established that the H_3 interarrival distribution can exhibit a variety of IDV curve shapes, we proceed to more fully

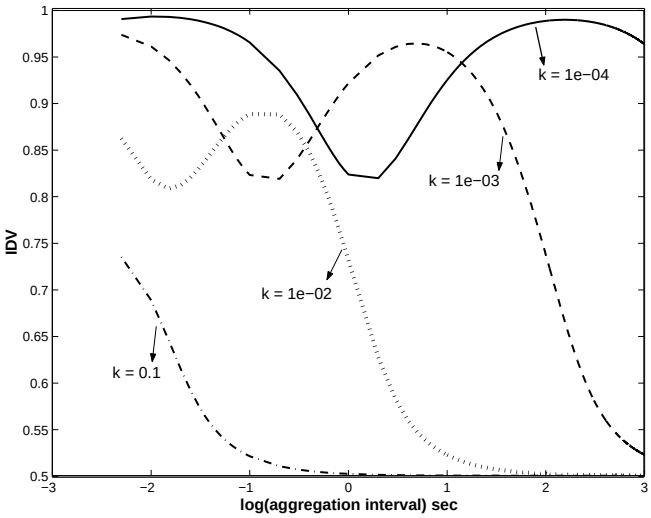

Fig. 5. IDVs of Doubly-Balanced H_3 Interarrival Distribution

evaluate its suitability as a traffic model for capturing the variability present in real network traffic. The evaluation consists of comparisons between an actual (ATM) traffic trace with a fairly challenging IDV curve and an H_3 model that attempts to match the IDV of the actual trace. The measured trace and the model are then compared using IDV, PRV, and simple queuing performance.

A. Index of Variability Evaluation

We begin by matching the IDV of the chosen ATM traffic trace as closely as possible with an H_3 model using the AMPL optimization tool [6]. We constrain the mean rate of the H_3 model to match the mean rate of the selected traffic trace (5183 arrivals/sec), leaving four degrees of freedom in the H_3 parameter space. At each sampled point of the traffic trace IDV curve (approximately 1000 points in all), we define the error value to be the magnitude of the difference between the measured trace IDV and the calculated H_3 IDV. The objective function to be minimized is then defined as the maximum of these error values.

Since iterative optimization tools can only find relative minima, and since our objective function contains many relative minima, it is necessary to run the optimization program a number of times with randomly (but reasonably) chosen initial points. This process produced a number of "solutions" with approximately the same objective function but with a variety of IDV shapes. We selected one from this set that seemed to match the shape of the trace IDV the best.

Fig. 6 shows the results. The solid curve is the IDV of the original measured traffic trace, and the dashed curve is the theoretical IDV of the chosen H_3 interarrival model. The parameters of the H_3 model are λ_1 = 82.821, λ_2 = 19242, λ_3 = 0.01037, w_1 = 0.003701, w_2 = 0.9962989, and $w_3 = 10^{-7}$. We can see that there is a reasonably good match between the original IDV and the model IDV across most time scales, with the largest difference coming at the largest time scales. We hypothesize that a better match could be obtained with higher-order hyperexponential models, but the analytic expressions for IDV of such models quickly become unwieldy.

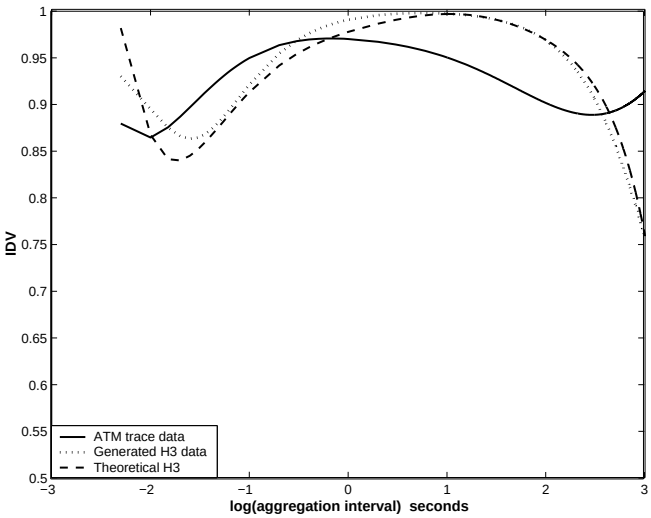

Fig. 6. IDV Comparison Between Measured Traffic Trace and H_3 Model

The dotted curve shows the IDV of a synthetic trace generated using the H_3 model. The good agreement between the theoretical H_3 IDV and the trace-measured H_3 IDV validates our trace-based IDV calculation methodology discussed in section II.

B. Peak Rate Variability Evaluation

We next compare the Peak Rate Variability (PRV) of the original trace and a trace from the selected H_3 model. For this comparison, we associate an ATM cell with every arrival from the H_3 trace and compare PRV in units of Mb/s (which is equivalent to a pkts/s PRV comparison in this case). From the previous Fig. 3 we see that the H_3 PRV (dashed line) falls off smoothly like the PRV of the Poisson arrival process (dash-dot line), but with considerably larger peak rates at all time scales, even though the mean rates are matched. Relative to the ATM trace PRV (solid line), the PRV match of the H_3 trace is quite good at both very small and very large time scales. However, the ATM PRV levels off at relatively small time scales before dropping rapidly at larger time scales, resulting in significantly larger peak rates than the H_3 trace at intermediate time scales.

C. Queuing Performance Evaluation

A major advantage of using the hyperexponential distribution to model network traffic is its relative ease of analysis. In addition to obtaining closed-form expressions for the IDV of a hyperexponential model, we can also obtain analytic performance predictions based on the hyperexponential model. In this section, we use G/M/1 queuing results (see, for example, [7])

to obtain mean and variance of packet delay for an H_3 arrival model with exponential service times, then compare these analytic results with simulation results for an H_3 traffic trace and a real traffic trace.

As derived in [7], the total delay (queuing plus service time) for a G/M/1 queue is exponentially distributed, hence its mean and standard deviation are identical. Finding the parameter of the exponential delay distribution requires solving a nonlinear equation involving the Laplace transform of the arrival distribution, which is readily obtained for the hyperexponential distribution.

In order to make comparisons with the analytic results, we associate exponentially distributed service times with the arrivals listed in the H_3 and real traffic trace files, even though the real traffic trace was gathered from an ATM (fixed packet size) link. Also, the trace files list number of arrivals in each 5 ms interval, so the simulation spaces each set of arrivals evenly throughout the associated 5 ms interval.

Fig. 7 shows the results for total delay (queuing plus service time) as a function of normalized load. Note that delay is on a log scale. First, we see good agreement between the analysis (circles) and simulation results (dashed line) for the H_3 model across the entire load range. We also see good agreement between the delays for the real trace and for the H_3 model at very low and very high loads. For intermediate loads (from about 0.3 to about 0.5), the real trace simulated delays are significantly larger than the H_3 model delays. This delay behavior is foreshadowed by the PRV curve of Fig. 3, as follows.

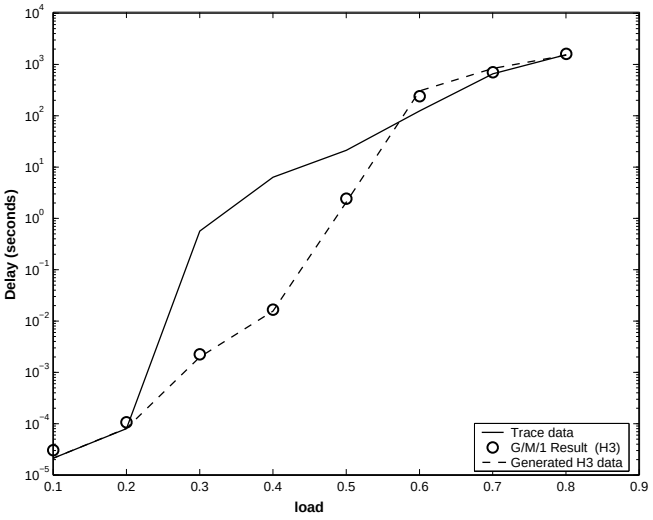

Fig. 7. Delay Comparison Between Measured Traffic Trace and H_3 Model

For a normalized load of 0.2, the service rate of the queue is approximately 11 Mb/s, which exceeds the peak rates of *both* the real trace and the H_3 trace for all but the very smallest time scales. Thus we would expect the mean delay to be quite small for both. Along the same lines, at a normalized load of 0.6, the service rate of the queue is approximately 3.7 Mb/s, which is smaller than the peak rates of both the real and the H_3 trace even for the largest time scale of one hour. The result is that the queue grows rapidly during these relatively long periods of time, producing huge average delays (hundreds of seconds) for both. However, for an intermediate load of 0.4, for example, the service rate is 5.5 Mb/s, larger than the H_3 peak rates at the larger time scales, but significantly smaller than the real trace peak rates for all but the very largest time scales. Thus we should not be surprised at the reasonably small delays (approximately 10 ms) for the H_3 model and the very large delays (approximately 10 s) for the real trace. Clearly, a better queuing match would be obtained with a model that has a PRV curve closer to the one for the real traffic. This underscores the utility of the PRV as a tool for characterizing traffic.

V. Conclusions

We have argued that single-parameter traffic characterizations, even those that can capture long-range dependence, are inadequate for characterizing the complexities of network traffic variability over the entire range of time scales. We advocate measures such as the index of variability (IDV) and the newly-introduced peak rate variability (PRV) that provide insight into traffic characteristics as a function of time scale.

The family of hyperexponential interarrival distributions, even though they all have asymptotic Hurst parameters of H = 0.5, are nonetheless promising network traffic models due to their relative analytic simplicity and the variety of IDV curves that they can exhibit. They can, in fact, be made to be practically indistinguishable from long-range dependent processes with H approaching 1.0.

However, our investigations of the H_3 model indicate that its four degrees of freedom limit its ability to adequately model traffic that is highly bursty over a broad range of time scales, such as that exhibited by ATM network traffic traces. Further research is required into methods of order selection and parameter selection for the hyperexponential distribution to produce desired traffic characteristics as measured by IDV, PRV or other metrics.

References

[1] W. E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, February 1994.

[2] K. Park and W. Willinger, *Self-Similar Network Traffic and Performance Evaluation*, 1^{st} ed., John Wiley & Sons, New York, 2000.

[3] V. J. Ribeiro, R. H. Riedi and R. G. Baraniuk, *Wavelets and Multifractals for Network Traffic Modeling and Inference*, ICASSP 2001.

[4] G. Y. Lazarou, "On the variability of Internet traffic," Ph. D. dissertation, University of Kansas, 2000.

[5] V. Paxson, "Fast Approximation of Self-Similar Network Traffic," Technical Report LBL-36750/UC-405, April 1995.

[6] R. Fourer, D. M. Gay, and B. W. Kernighan, *AMPL: A Modeling Language for Mathematical Programming*, Boyd and Fraser, 1993.

[7] R. Nelson, *Probability, Stochastic Processes, and Queuing Theory*, Springer-Verlag, 1995.