A Hierarchical Model for Teletraffic [1]

Vishal Misra and Wei-Bo Gong
Department of Electrical and Computer Engineering
University of Massachusetts, Amherst MA 01003
{vmisra,gong}@ecs.umass.edu

Abstract

Self-similarity over certain time scale range has been repeatedly observed in high speed network traffic. We propose a model motivated by the physical process of teletraffic generation to decipher such a phenomena. Our model is hierarchical in nature. We examine some properties of our model from a signal theoretic point of view, and explain why it will exhibit multi-scale or self-similar behavior. Simulations of synthetic traffic based on our model are presented and it's multi-scale properties are compared with that of the well-known Bellcore traffic traces. We present some preliminary conclusions on the actual nature of the traffic and the validity of our model.

1 Introduction

Teletraffic modeling is at a very interesting stage these days, with the seminal findings of Leland et. al. [1] changing the way we look at it. The key factor responsible for this is the observation of "scale-invariance" or "long range dependence"(LRD) in the traffic patterns [1],[2]. Implications of such traffic are manifold and are predicted to affect significantly network behavior like packet delay, buffer occupancy, cell loss rate etc. The modeling for this phenomena has given rise to two schools, one is the "self-similar" kind of models while the others are based on Markovian models. The first kind are asymptotically scale invariant and involve "heavy tailed" distributions of some sort in the model. The Markovian models, on the other hand, are (approximately) scale invariant only for a finite range of timescales. Examples of "self-similar" models found in [3], [4], [5] and in references therein. Markovian models have been proposed in [6] and [7].

Unlike other models, we attempt to directly model the physical process of teletraffic generation. Our model is hierarchical in nature, the hierarchy consisting of independent on-off processes. We show that the hierarchy naturally induces the self-similar behavior and also show that the distributions of the periods of the individual on-off processes are not critical in exhibiting the self-similar phenomena. We also view and analyze the model from a linear system and signal theoretic point of view, which yields fresh insights into the fundamental nature of traffic. Our model reflects the phenomena of "self-similarity" over a finite timescale range.

2 Motivation and the model

2.1 Self-Similarity

Before we begin our modeling, we'll briefly outline the concept of self-similarity. The concept of long range dependence and self-similarity are intimately related, and they are characterized by a slowly (polynomial) decaying autocorrelation function. If $x(t)$ is a wide sense stationary self-similar process, then the autocorrelation takes the form

$$R_X(t) \sim c_R t^{-(2-2H)}, t \to \infty \tag{1}$$

and equivalently, the power spectral density is of the form

$$S_X(\omega) \sim c_S |\omega|^{1-2H}, \omega \to 0 \tag{2}$$

Where H is the *Hurst* parameter of the long range dependent or self-similar process. H takes values between 0.5 and 1. The self-similar phenomena has also been referred to as "scaling-phenomena" or "scale-invariant behavior".

2.2 The life of a packet

Let's consider the transmission of a packet between a sender-receiver pair on a network, ethernet for instance. The packet is transmitted when a number of things simultaneously happen. Firstly, a session has to be in progress between a sender-receiver pair. Next, within the session the particular application intermittently requests/supplies data. The sender then starts sending the data according to a protocol like TCP. The transmission rate of the data is decided by the flow and congestion control mechanism. Finally, packets are allowed on an ethernet only one at a time, so each packet has to wait it's turn if the ethernet is busy, by the random back-off mechanism. As is evident, a number of different conditions have to be *true*, pretty much independent

[1] This work was supported in part by the National Science Foundation under Grant CCR-9506764, Grant EEC-9527422, Grant NCR-9523807 and Grant ANI-9809332, by the U.S. Army Research Office under Contract DAAH04-95-0148 and AAG55-98-1-0042 , and by Air Force Rome Laboratory under Contract F30602-97-C-0125.

of each other, for a packet to appear on the ethernet. Not only that, the time scales at which the events are occurring are disparate. Clearly, one should expect multiple timescale behavior from network traffic. It seems almost natural to define the packet transmission process as an on-off process which is a product of independent on-off processes operating at different time scales. We now define our model.

2.3 Hierarchical On-Off Process

Definition 1.1 An n-level Hierarchical on-off process (HOP) $Y(t)$ is defined by

$$Y(t) = \Pi_{i=1}^{n} X_i(t) \tag{3}$$

where each $X_i(t)$ is an independent on-off process.

We have made no assumption on the nature of the on-off processes asides from independence till now. Let's assume the component on-off processes to be Markovian and examine it's spectral properties.

Definition 1.2 A Markovian Hierarchical On-Off Process (MHOP) is a HOP where the component processes are Markovian.

We model our traffic as an aggregation of MHOPs. We also assume in our model that the timescales of different On-Off processes in an MHOP are disparate. We note that an MHOP can be described by a Markov process with augmented state. For example, let $n = 3$. The output of the process is 1 only when it's in the state 111. It could then also be equivalently thought of as a Hidden Markov Model, where the observation process has 7 Markov states mapped into one state ("off") and we directly observe the remaining eighth state ("on").

In the next section we explain how or why such a process would exhibit certain degree of self-similar behavior.

3 Spectral Properties of the MHOP

The autocorrelation function of a Markovian on-off process is given by

$$R_x(\tau) = p_{on}(1 - p_{on})e^{-(\lambda+\mu)|\tau|} + p_{on}^2 \tag{4}$$

Where λ is the transition rate from off to on and μ the rate in the reverse direction. They are related to p_{on} via the relation $p_{on} = \lambda/(\mu + \lambda)$. For the i'th process, let's denote $\lambda_i + \mu_i$ by ν_i, p_{on}^2 by k_{i1} and $p_{on}(1 - p_{on})$ by k_{i2}. The correlation of a product of two independent processes would then be just the product of the individual correlations, i.e.,

$$\begin{aligned} R_{x_1 x_2} &= R_{x_1} \cdot R_{x_2} \\ &= (k_{11} + k_{12}e^{-\nu_1|\tau|})(k_{21} + k_{22}e^{-\nu_2|\tau|}) \\ &= k_{11}k_{21} + k_{12}k_{21}e^{-\nu_1|\tau|} + k_{11}k_{22}e^{-\nu_2|\tau|} \\ &\quad + k_{12}k_{22}e^{-(\nu_1+\nu_2)|\tau|} \end{aligned}$$

The Fourier transform of the autocorrelation function gives the power spectral density, and thus we have

$$\begin{aligned} S_{x_1 x_2}(f) &= k_{11}k_{21}\delta(f) + \frac{2k_{12}k_{21}\nu_1}{(2\pi f)^2 + \nu_1^2} + \\ &\quad \frac{2k_{11}k_{22}\nu_2}{(2\pi f)^2 + \nu_2^2} \\ &\quad + \frac{2k_{12}k_{22}(\nu_1 + \nu_2)}{(2\pi f)^2 + (\nu_1 + \nu_2)^2} \end{aligned}$$

Let's look at the last two terms in the above expression. They can be rewritten as

$$(k_{22})(k_{11}\nu_2 + k_{12}(\nu_1 + \nu_2)) \cdot \frac{(2\pi f)^2 + (\nu_1 + \nu_2)\nu_2\rho}{((2\pi f)^2 + \nu_2^2)((2\pi f)^2 + (\nu_2 + \nu_1)^2)}$$

where

$$\rho = \frac{k_{11}(\nu_2 + \nu_1) + k_{12}\nu_2}{k_{12}(\nu_2 + \nu_1) + k_{11}\nu_2}$$

Viewed as a linear system, the above corresponds to a system having two poles at ν_2 and $\nu_1 + \nu_2$ and a zero at $\sqrt{\nu_2(\nu_1 + \nu_2)\rho}$ which lies between the two poles. In our model, we have assumed that the time scales of operations of the component processes are well separated. Thus $\nu_2 >> \nu_1$, which implies that both the poles and the zero are bunched very close together. This causes what is known as *a pole-zero cancellation* and we can approximate [1] the system with one having a single pole and no zero. Making that approximation and placing the resultant pole at ν_2, we get the power spectral density of the process as

$$S_{x_1 x_2} \approx k_{31}\delta(f) + \frac{2k_{32}\nu_1}{(2\pi f)^2 + \nu_1^2} + \frac{2k_{33}\nu_2}{(2\pi f)^2 + \nu_2^2} \tag{5}$$

where we have absorbed all the multiplicative constants into new constants k_{31}, k_{32} and k_{33}. The last two terms can be rewritten as

$$2(k_{32}\nu_1 + k_{33}\nu_2)\frac{(2\pi f)^2 + \nu_1\nu_2\rho}{((2\pi f)^2 + \nu_1^2)((2\pi f)^2 + \nu_2^2))} \tag{6}$$

where

$$\rho = \frac{k_{32}\nu_2 + k_{33}\nu_1}{k_{33}\nu_2 + k_{32}\nu_1}$$

Again viewing it as a linear system, this corresponds to a system which has two poles at ν_1 and ν_2 and a zero at $\sqrt{\nu_1\nu_2\rho}$ which lies between the two poles. Unlike the previous time, this time the poles and zero are well separated and there is no cancellation. However, the presence of the zero between the two poles leads to an interesting phenomena, which is the reason why MHOPs would exhibit self-similar behavior over a certain frequency range. The Bode

[1] Note that we are making the approximation to only explain the phenomena, the plots shown subsequently plot the exact spectrum.

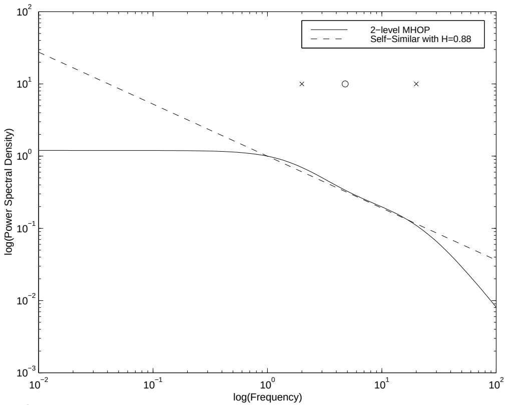

Figure 1: Power Spectral Density for 2-level MHOP and Self-Similar Process

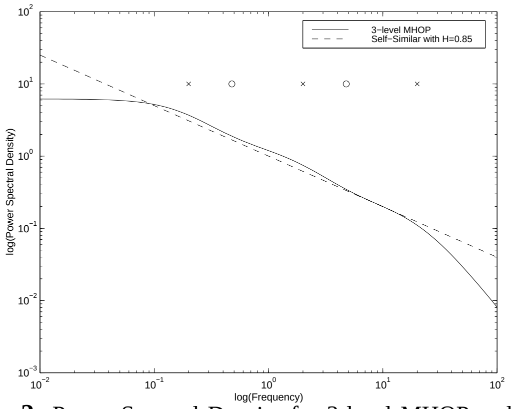

Figure 2: Power Spectral Density for 3-level MHOP and Self-Similar Process

(double-log) plot for a pole in a linear system corresponds to a flat line until the corner frequency (the pole) and then a straight line with a slope of -2 beyond that. The Bode plot of a zero is a flat line till the corner frequency and then a straight line with slope +2 beyond that. The effect of having a zero between two poles is that the rate of decay of the power spectral density slows down in the frequency region bounded on either side by the poles. If the zero is in exactly the middle of the two poles on the Bode plot (i.e. is the geometric mean of the two) then the decay of the power spectral density is like that of a $1/f$ process, which is self-similar. If the zero shifts to the left or right, then the decay is of the form $1/f^{\gamma}$, $\gamma < 1$ and $\gamma > 1$ respectively. This is shown in Figure 1. The two poles and the zero are marked out on the plot. We chose $\lambda = \mu$ for both levels and $\nu_2 = 10\nu_1$. Between the frequencies bounded by the two poles, the decay of the power spectral density slows down, and goes down with a slope less than 1. A reference power spectral density of a self-similar process corresponding to a Hurst parameter of 0.88 is plotted alongside. Clearly, the 2-level MHOP gives a close approximation to the self-similar process in the frequency range bounded by the poles. Thus, if we have an *n*-level MHOP, it would correspond to a cascade of such pole-zero systems and would thus tend to approximate the self-similar spectrum over the range of ν's of all the component processes. To illustrate that, we plot a similar figure (Figure 2) for a 3-level process, again marking out the poles and zeros and plotting a reference self-similar process. Again, the approximation to the self-similar process over the range of ν's (poles) is striking.

The range of self-similar processes that they approximate (via the Hurst parameter), corresponds to the range observed in network traffic [1], [8]. By changing the value of the *on* probability *p* in our component processes (thereby changing the λ/μ ratio) we can move the zero around between the poles and have considerable freedom in the range of self-similar slopes that we can approximate. Since our traffic model is an aggregation of independent MHOPs, the spectrum of the traffic would be a summation of the individual spectra. For homogeneous traffic, this corresponds to a simple scaling.

An important factor here is that this exhibition of the self-similar spectrum is not sensitive to the exponential distribution assumption. In the next 2 figures, we show the spectra for a HOP with composed of Erlang-2 distributed on-off sources and Hyper-exponentially distributed on-off sources. As can be seen, in the relevant frequencies (timescales) the processes still appears to be self-similar. This is important since the Markovian assumption is unrealistic for the TCP level on-off process and it is not memoryless. However, at higher timescales (low frequencies), the spectrum of the TCP process flattens out, appearing similar to a Markovian source with a high arrival rate. The disparity in the mean on-off periods of the various levels of the hierarchy makes the TCP spectra look similar to a Markovian one in the relevant frequencies. Under some general conditions, the spectrum of a phase-type distributed on-off process has a flat spectrum as $f \to 0$ and a $1/f^2$ like spectrum as $f \to \infty$. Around the mean of the on-off process, the transition (in the Bode plot) of the slope of the spectrum from 0 to -2 (20 dB/sec) takes place. For a large class of distributions, we can approximate the spectrum as being piecewise linear with the corner frequency governed by the mean of the process (per the asymptotes method familiar in linear system analysis). Under those conditions, the preceding analysis would hold and a hierarchy of such processes would exhibit "self-similarity" over a finite range of timescales.

Now that we have seen the capability of HOPs to approximate the *spectrum* of self-similar processes over a range, the next natural question is that do these processes exhibit the same scaling behavior? We ran several simulations and present the results in a following section.

4 Simulations and Analyses

To compare the behavior of MHOP with the Bellcore data set, we generated samples which were equal in length and with the same mean value as the various traces analyzed in

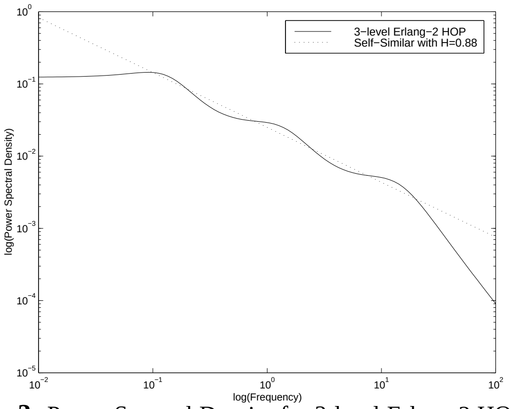

Figure 3: Power Spectral Density for 3-level Erlang-2 HOP and Self-Similar Process

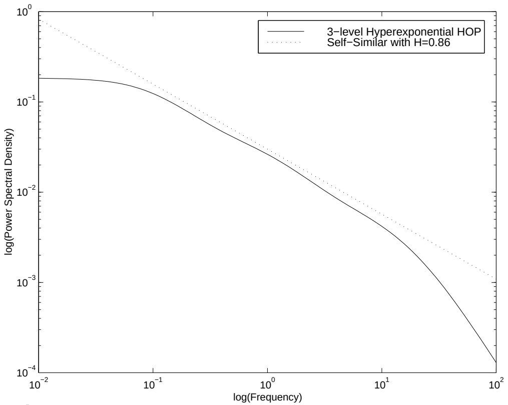

Figure 4: Power Spectral Density for 3-level Hyper-exponential HOP and Self-Similar Process

[1], [8]. For the plots shown, we used $\nu_2 = 10\nu_1$ and $\nu_3 = 30\nu_2$. For all the component processes, we had $\lambda/\mu = 0.8$. We used an aggregation of 16 independent such processes. We'll look at three plots , proposed previously, which indicate "self-similarity" at certain time scales. The first one is the "visual-proof". This is similar to the plot shown in [1]. We selected, completely at random, sections of the simulated traffic and the August Bellcore dataset and plotted them side by side at the same resolution level. Similar to the graphs plotted in [1], the plots at the highest resolution level have the same random noise term added to both to avoid the visually jarring quantization effect. This is shown in Figure 5. Visually at least, the simulated MHOP process exhibits a similar bursty behavior as the Bellcore dataset over all the levels plotted. Note that the length of the publicly available Bellcore August dataset limits the number of points we can plot at the coarsest resolution level.

The next plot (Figure 6) is the log(Variance) vs log(aggregation level) plots, again similar to those in [1]. Again, the behavior is strikingly similar, with the variances for both processes decaying at the same rate, slower than 1, with increasing aggregation level *m*.

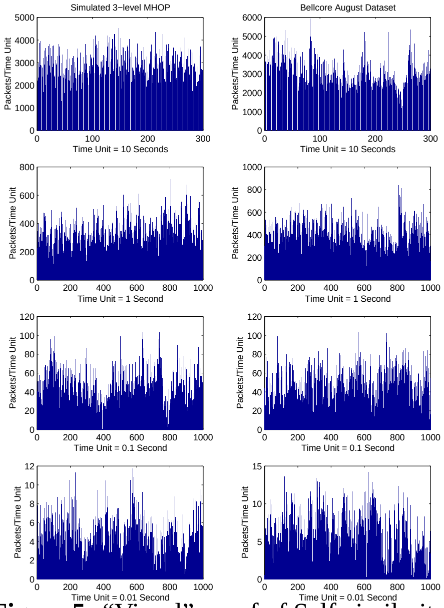

Figure 5: "Visual" proof of Self-similarity

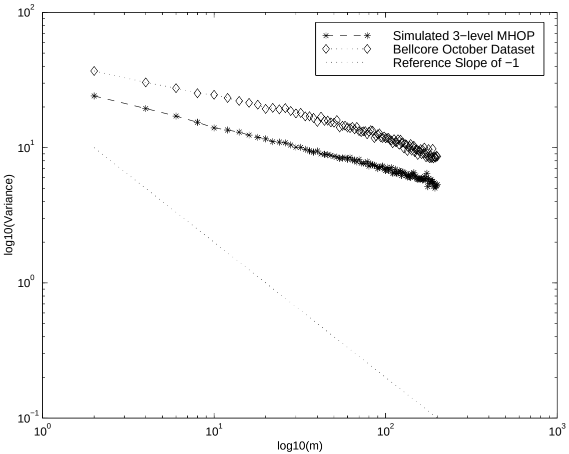

Figure 6: The variance vs aggregation level plot for the two processes. One of the plots has been shifted down to give a clearer picture

The final comparison is done via wavelet analysis. Wavelet analysis has become an important tool in studying self-similarity [8]. The details of wavelet analysis can be found in the reference, but to get a quick handle on the plots we have used, think of them as mirror images of magnitude Bode plots. The plots have scale on the x-axis, which corresponds to log-frequency on the Bode plot, and have variance of wavelet coefficients on the y-axis, which are an estimate of the power spectral density at the particular scale (frequency). The timescale increases from left to right, which corresponds to a decrease in the frequency and hence the plots appeared laterally inverted when compared to Bode plots. Linear regions in the plot correspond to power law behavior of the power spectral density, with the slope giving the estimate of the exponent of the frequency.

For a self-similar process, the plot of the variance of the wavelet coefficients vs the scale should turn out to be

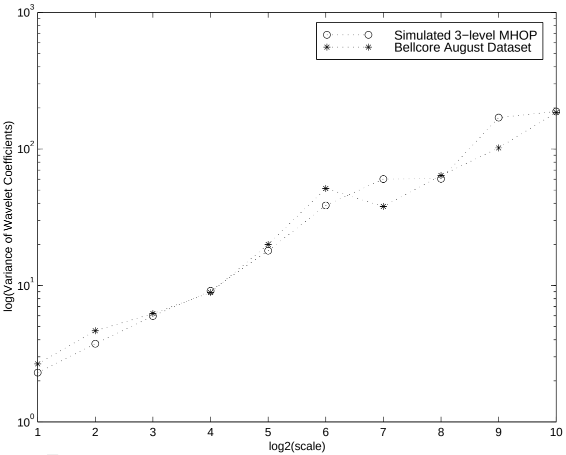

Figure 7: The variance scale plot for Bellcore Dataset and Simulated Process

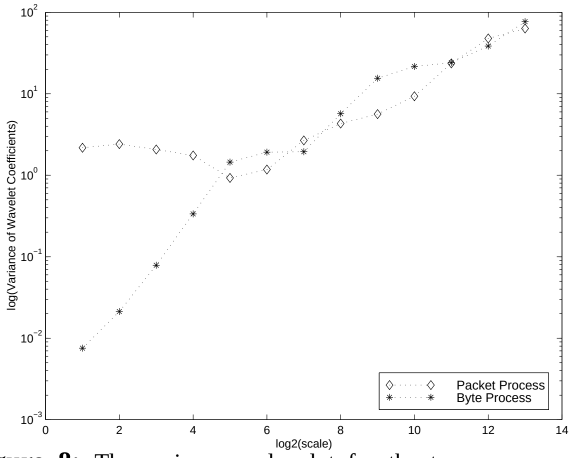

Figure 8: The variance scale plot for the two processes at high resolution (frequencies) for the Bellcore August Dataset. The time scales for two plots have been adjusted a little to show how the Byte process and Packet process differ in their departure from the scaling region.

a straight line, with the slope giving an estimate of the Hurst parameter. We analyzed the two signals using the Daubechies-4 wavelet. Simultaneous plot for the simulated MHOP and Bellcore August Dataset shows a strong resemblance in the properties of the two processes, yielding a nearly identical estimate of the Hurst parameter. This is seen in Figure 7.

5 Discussion

5.1 Extrapolation: What happens beyond the scale invariant behavior?

As has been noted by various researchers [2], [8], the scale invariant behavior occurs only over a finite range of scales. We have seen that behavior in our simulated model which matches that of actual traces. To go beyond what we have graphically shown in the previous section, we need to look at resolutions which are very *fine* (high frequency region) and resolutions which are very *coarse*. In the game of asymptotics, the dominant term wins. To look at low frequencies or coarse resolution, the term to be considered is the contribution from the highest level on-off process, whereas for high frequencies it is the lowest level on-off process. The coarse resolution analysis is limited by datalength. As far as the high resolution analysis goes, we have to first clearly define which process we are looking at. If we simply look at packet counts, then the process operating at the lowest level of our hierarchy would be a series of impulses, indicating the arrival time of a packet. Modeling the arrivals at that level as Poisson, the high frequency power spectrum of the process would be a flat line like white noise [2]. Again thinking of the variance-scale plot in the wavelet analyses as the mirror image of the power spectral density, we would expect the plot to have a flat region before the scaling behavior due to higher level processes starts. If instead of simple packet counts, we look at byte counts, in effect looking at the time periods where a packet is being transmitted or not, then our lowest level hierarchy again reduces to a simple on-off process. A Markovian model would then predict a roll off with a slope of -2 for frequencies much higher than the ν of the process. The mirror image analogy again indicates a region before the scaling starts with a higher slope than the one found in the scale-invariant region for the variance-scale plot of the wavelet analysis of the given trace. Both of the predictions based on our model are shown in Figure 8 and are also confirmed by the findings of Abry et al. [8].

Extrapolating at the other end of the spectrum, the very low frequencies, requires datasets much longer than what we have analyzed so far. Based on our model, we'd expect the spectrum to flatten out at frequencies much smaller than the ν of the highest level on-off process. On the variance-scale plot, this would correspond to a region of slope zero beyond the scale invariant region.

5.2 Conclusion

We have proposed a new model for network traffic. The model is motivated by physical arguments regarding the behavior of network traffic. Next we have shown the properties of our model which closely match those of "self-similar" process over a given time range. The time range depends on the choice of parameters of the model. Simulation runs of the model also show a close match to the observed dataset as far as "self-similarity" goes. Thus, our model provides a simulation tool which gives realistic traffic without paying the full price of "self-similar simulation". In our opinion, observation of scaling phenomena over a

[2] a roll off for very high frequencies can be expected as ideal, infinite bandwidth white noise is not found in the physical world due to *inertia*. In the Ethernet world, this inertia comes due to the fact that there is a minimum inter-arrival time which is the sum of the shortest packet-length and minimum required silence between frames.

certain range of time scales cannot and should not be extrapolated to all time scales, and can be well explained and approximated by non self-similar, physically realizable and realistic models as we have shown in the paper. Thus, if the real traffic is scale-invariant only over a finite range of time scales, then we could be looking at the wrong asymptotics if we assume true self-similarity or heavy tailed behavior. This could have an effect in evaluating the performances of networks as well as call admission and congestion control strategies. In addition, a source centric model is convenient for call admission decisions. We emphasize that unlike some other Markov models that also exhibit scale-invariance over a time scale range, our model is based on the physical structure of the traffic in the network. This feature is of fundamental importance because it enables us to explain various phenomena observed for real teletraffic. The higher levels of our hierarchy are independent of the underlying network, and depend on the nature of the application and user behavior. Thus, a change in transmission control strategy for instance won't significantly affect the "self-similar" behavior of a source on large timescales. The nature of applications or user behavior is not likely to change much with change in the transport protocol. Thus we can simply replace the lowest level of the hierarchy with an appropriate process and not have to change the complete model should the transport mechanism change. The need, in our opinion, is to study application and user patterns carefully to arrive at a reasonable mechanism to predict and regulate traffic. There is plenty of data available for researchers to look at structural models. The quest for parsimony, for instance characterizing the traffic just by the Hurst parameter, could be misleading and ignoring a lot of useful information available. For example, we collected WAN traffic data in the departmental LAN for seven consecutive days. In Figure 9 we show the histogram plots (on a double-log scale) of inter-arrival times for packets for 3 representative sources in the LAN (plots for other sources are very similar). The plot shows some interesting characteristics. For all the sources, there is a sort of "flat" region of inter-arrival times. The region starts from the time corresponding to approximately the minimum possible inter-arrival time (governed by shortest packet length and minimum silence period between frames) and goes on till about the averaged round trip time observed for connections of individual sources. Beyond that there is a sharp decline in the histogram plot. Thus, there are clear *zones* of operation of the traffic process, reinforcing the need to model and analyze them separately. We believe the first flat region corresponds to the TCP level on-off process, whereas the region beyond that corresponds to the inter-arrivals in the application and session level on-off processes. Structural information like that can be easily extracted from available data and reasonable models built for both simulations as well as analysis. Looking at sources (and traffic) in a hierarchical way also enables us to make some predictions about the traffic. For instance, if we employ some control strategy which has slow time dynamics, it might affect and shift the pole for the relevant on-off process. It might end up making the traffic even more long range dependent.

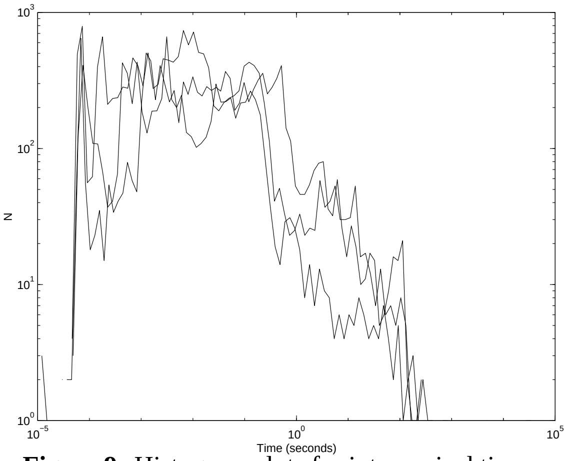

Figure 9: Histogram plots for inter-arrival times

References

[1] 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, pp. 1–15, February 1994.

[2] V. Paxson and S. Floyd, "Wide-area traffic: The failure of poisson modelling," *Proceedings of SIGCOMM*, 1994.

[3] 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.," *Proceedings of ACM SIGCOMM*, pp. 100–113, 1995.

[4] P. Pruthi, *An Application of Chaotic Maps to Packet Traffic Modeling*. PhD thesis, Royal Institute of Technology, Department of Teleinformatics, Kista, Sweden, 1995.

[5] K. Park, G. Kim, and M. Crovella, "On the relationship between file sizes, transport protocols, and self-similar network traffic," *Proceedings of the Fourth International Conference on Network Protocols*, 1996.

[6] S. Robert and J.-Y. L. Boudec, "A markov modulated process for self-similar traffic," in *Internationales Begegnungs und Forschungszentrum fuer Informatki, Schloss Dagsthul, Saarbrücken, Germany*, pp. 1–14, September 1995.

[7] A. T. Andersen and B. F. Nielsen, "An application of superpositions of two-state markovian sources to the modelling of self-similar behaviour," *Proceedings of the IEEE INFOCOM*, pp. 196–204, 1997.

[8] P. Abry and D. Veitch, "Wavelet analysis of long range dependent traffic," *IEEE Transactions in Information Theory*, vol. 44, pp. 2–15, January 1998.