On the Self-Similar Nature of Ethernet Traffic

Will E. Leland† — *wel@bellcore.com*

Murad S. Taqqu§ — *murad@bu-ma.bu.edu*

Walter Willinger† — *walter@bellcore.com*

Daniel V. Wilson† — *dvw@bellcore.com*

†Bellcore
445 South Street
Morristown, NJ 07960-6438

§Department of Mathematics
Boston University
Boston, MA 02215

Abstract

We demonstrate that Ethernet local area network (LAN) traffic is statistically *self-similar*, that none of the commonly used traffic models is able to capture this fractal behavior, and that such behavior has serious implications for the design, control, and analysis of high-speed, cell-based networks. Intuitively, the critical characteristic of this self-similar traffic is that there is no natural length of a "burst": at every time scale ranging from a few milliseconds to minutes and hours, similar-looking traffic bursts are evident; we find that aggregating streams of such traffic typically intensifies the self-similarity ("burstiness") instead of smoothing it.

Our conclusions are supported by a rigorous statistical analysis of hundreds of millions of high quality Ethernet traffic measurements collected between 1989 and 1992, coupled with a discussion of the underlying mathematical and statistical properties of self-similarity and their relationship with actual network behavior. We also consider some implications for congestion control in high-bandwidth networks and present traffic models based on self-similar stochastic processes that are simple, accurate, and realistic for aggregate traffic.

1. INTRODUCTION

The main objectives of this paper are (i) to establish in a statistically rigorous manner the *self-similarity* characteristic or, to use a more popular notion, the *fractal* nature of the high time-resolution Ethernet traffic measurements of Leland and Wilson (1991), (ii) to illustrate some of the most striking differences between self-similar models and the standard models for packet traffic currently considered in the literature, and (iii) to demonstrate some of the serious implications of self-similar network traffic for the design, control, and performance analysis of high-speed, cell-based communications systems.

Intuitively, self-similar phenomena display structural similarities across all (or at least a very wide range of) time scales. In the case of Ethernet LAN traffic, self-similarity is manifested in the absence of a natural length of a "burst"; at every time scale ranging from a few milliseconds to minutes and hours, bursts consist of bursty subperiods separated by less bursty subperiods. We also show that the degree of self-similarity (defined via the *Hurst parameter*) typically depends on the utilization level of the Ethernet and can be used to measure "burstiness" of LAN traffic. The term "self-similar" was formally defined by Mandelbrot. For applications and references on the theory of self-similar processes, see Mandelbrot (1983) and the extensive bibliography by Taqqu (1985). For an early application of the self-similarity concept to communications systems, see the seminal paper by Mandelbrot (1965).

In this paper, we use very high quality, high time-resolution LAN traffic data collected between August 1989 and February 1992 on several Ethernet LANs at the Bellcore Morristown Research and Engineering Center (MRE). Leland and Wilson (1991) present a preliminary statistical analysis of this unique high-quality data and comment in detail on the presence of "burstiness" across an extremely wide range of time scales: traffic "spikes" ride on longer-term "ripples", that in turn ride on still longer term "swells", etc. This self-similar or apparently fractal-like behavior of aggregate Ethernet LAN traffic is very different both from conventional telephone traffic and from currently considered formal models for packet traffic (e.g., pure Poisson or Poisson-related models such as Poisson-batch or Markov-Modulated Poisson processes (Heffes and Lucantoni (1986)), packet-train models (Jain and Routhier (1986)), and fluid flow models (Anick et al. (1982)), etc.). These differences require a new look at modeling the traffic and performance of broadband networks. For example, our analysis of the Ethernet data shows that the generally accepted argument for the "Poisson-like" nature of aggregate traffic, namely, that aggregate traffic becomes smoother (less bursty) as the number of traffic sources increases, has very little to do with reality. In fact, the burstiness (degree of self-similarity) of LAN traffic typically intensifies as the number of active traffic sources increases, contrary to commonly held views.

Because of the growing market for LAN interconnection services, LAN traffic is rapidly becoming one of the major potential traffic contributors for high speed networks of the future such as B-ISDN. Another expected major contributor is variable-bit-rate (VBR) video service. Since VBR video traffic has recently been shown to display the same self-similarity property as LAN traffic (see Beran et al. (1992)), self-similar models provide simple, accurate, and realistic descriptions of traffic scenarios to be encountered during high-bandwidth network deployment.

In light of this new understanding of the nature of broadband traffic, we also address some of the serious implications of self-similar traffic for issues related to the design, control, and performance analysis of high-speed, cell-based networks. As one specific example, we consider the area of congestion management and show that the nature of congestion produced by self-similar traffic differs drastically from that predicted by traffic models currently considered in the literature and is far more complex than has been typically assumed in the past. As a result, proposed congestion control schemes that work well assuming conventional traffic models typically perform less than satisfactorily in a self-similar traffic environment. Finally, we mention two approaches for modeling self-similar network traffic.

The paper is organized as follows. In Section 2, we first briefly describe the available Ethernet traffic measurements and comment on the changes of the Ethernet environment during the measurement period from August 1989 to February 1992. In

Section 3, we give the mathematical definition of self-similarity, identify classes of stochastic models which are capable of accurately describing the self-similar behavior of the traffic measurements at hand, and present statistical techniques for dealing with self-similar data. Section 4 describes our statistical analysis of the Ethernet data, with emphasis on testing for self-similarity. We illustrate our statistical methods with a variety of different sets of Ethernet traffic data, taken at different times during the measurement period, with quite different user populations and gross traffic rates. We typically deal with time series with hundreds of thousands of observations and are, therefore, in the unique situation to rely on statistical results known to hold asymptotically in the number of observations. Finally, in Section 5 we discuss the significance of self-similarity for traffic engineering, and for operation, design, and control of B-ISDN environments. Among the implications discussed are (i) infinite variance source models for individual Ethernet users, (ii) inadequacies of commonly used notions of "burstiness", and (iii) better understanding of the nature of congestion for broadband network traffic. We conclude with a brief discussion of two different approaches for modeling self-similar network traffic.

2. TRAFFIC MEASUREMENTS

The monitoring system used to collect the data for the present study is custom built, records all packets seen on the Ethernet under study with accurate timestamps, and will do so for very long runs without interruption. The monitoring system is designed so that traffic analysts need make no *a priori* decisions as to what they are searching for when the data is taken other than how much of each packet is to be saved. The monitor was custom-built in 1987/88 and has been in use to the present day with one upgrade. The original version is described at length, including extensive testing of its capacity and accuracy, in Leland and Wilson (1991). There is only one major difference between the two versions: the updated version records timestamps accurate to 20 μs versus the original version's accuracy of 100 μs.

2.1 BELLCORE'S NETWORK ENVIRONMENT

The MRE environment is probably typical of a research or software development environment where workstations are the primary machines on people's desks. Table 1 gives a summary description of the traffic data analyzed later in the paper. We consider 4 sets of traffic measurements, each one representing between 20 and 40 consecutive hours of Ethernet traffic and each one consisting of tens of millions of Ethernet packets. The data were collected on different intracompany LAN networks at different periods in time over the course of approximately 4 years (August '89, October '89, January '90, and February '92). The traffic was mostly from services that used the Internet Protocol (IP) suite for such capabilities as remote login or electronic mail, and the Network File System™ (NFS) protocol for file service from servers to workstations. While it is not our intent to provide here a detailed description of the particular MRE network segments under study, some words about the network environment from which each data set was taken are appropriate.

The first two data sets were collected from a typical workgroup or laboratory network which was isolated from the rest of the Bellcore network by a router. At the time of collection of the first (August '89) data set, the laboratory consisted of about 140 people, most of whom had diskless Sun-3™-class workstations on their desks. The network in the laboratory consisted of two cable segments separated by a bridge, implying that not all traffic within the laboratory could be seen by the monitor. The hosts on this network consisted of workstations, their file servers, and a pair of DEC 8650™-class minicomputers. Only a small number of hosts were reduced instruction set (RISC) based. However, by the time the second data set (October '89) was collected, a massive upgrade of the Sun-3 class machines to RISC-based SPARCstation 1™ and DEC 3100-class machines had taken place on this network, along with a small increase in the number of hosts (from about 120 who spoke up during the first collection to about 140 in the second set). This extensive upgrade is the reason for the large increase in traffic volume seen between the first two data sets. Note, for example, that the "busy hour" from this October '89 data set is indeed busy: 30.7% utilization as compared to 15.1% during the August '89 busy hour; similar increases can also be observed for the low and normal hours. Less than 5% of the total traffic observed in these two data sets was from conversations with hosts in the rest of Bellcore or the outside world; during the busy periods this figure was more typically 1-2% of the total traffic observed.

The third data set, taken in January 1990 (row 3 in Table 1), came from an Ethernet cable that linked the two wings of the MRE facility that were occupied by a second laboratory. At the time the data set was collected, this second laboratory comprised about 160 people engaged in work similar to the first laboratory. This particular Ethernet segment was unique in that it was also the segment serving Bellcore's link to the outside Internet world. Thus the traffic on this cable was from several sources: (i) two very active file servers (Sun 4/490™'s) directly connected to the segment; (ii) traffic (file service and remote login) between the two wings of this laboratory, (iii) traffic between the laboratory and the rest of Bellcore, and (iv) traffic between Bellcore as a whole and the larger Internet world. This Ethernet segment was specifically monitored to capture this last type of traffic, which we term *external* traffic. We studied both the aggregate and external traffic from this and the last data set, but as we shall see in Section 4 the external traffic was no different from internal traffic as far as our analysis is concerned. This segment was separated from both the Bellcore internet and the two wings of the laboratory by bridges, and from the outside world by a vendor-controlled router programmed to pass anything with a Bellcore address as source or destination. In contrast to the two earlier data sets, over 1200 hosts spoke up during the 40 hour monitoring period on this segment.

The last data set, taken in February 1992 (row 4 in Table 1) was taken from the Ethernet "backbone" in the MRE facility. Because of rising concern about network security, Bellcore's connection to the Internet world was moved to a Bellcore-controlled "firewall" security router directly connected to the building backbone. Many major workgroups and laboratories also inserted routers between their networks and the backbone; previously they had been either directly connected or bridged. The traffic on this backbone cable therefore consisted of (i) traffic between workgroups and laboratories within the MRE facility, including traffic between the two laboratories previously discussed, (ii) traffic from some individual hosts still directly connected to the backbone, (iii) traffic from MRE to other Bellcore locations via a mesh of bridged interlocation links, and finally (iv) *external* traffic from Bellcore to the outside Internet world. Because there is very little workstation to fileserver traffic on this cable, the overall load on this cable is the lowest of any of the three data collection points considered. The most radical difference between this data set and the others is that the traffic is primarily router to router rather than host to host. About 600 hosts spoke up during this measurement period (down from about 1200 active hosts during the January '90 measurement period), and the five most active hosts were routers. The updated version of the monitor was used to collect

Traces of Ethernet Traffic Measurements					
Measurement Period		data set	total number of bytes	total number of packets	Ethernet utilization
AUGUST 1989	total (27.45 hours)		11,448,753,134	27,901,984	9.3%
Start of trace: Aug. 29, 11:25am	low hour (6:25am-7:25am)	AUG89.LB	224,315,439		5.0%
		AUG89.LP		652,909	
End of trace: Aug. 30, 3:10pm	normal hour (2:25pm-3:25pm)	AUG89.MB	380,889,404		8.5%
		AUG89.MP		968,631	
	busy hour (4:25pm-5:25pm)	AUG89.HB	677,715,381		15.1%
		AUG89.HP		1,404,444	
OCTOBER 1989	total (20.86 hours)		14,774,694,236	27,915,376	15.7%
Start of trace: Oct. 5, 11:00am	low hour (2:00am-3:00am)	OCT89.LB	468,355,006		10.4%
		OCT89.LP		978,911	
End of trace: Oct. 6, 7:51am	normal hour (5:00pm-6:00pm)	OCT89.MB	827,287,174		18.4%
		OCT89.MP		1,359,656	
	busy hour (11:00am-12:00am)	OCT89.HB	1,382,483,551		30.7%
		OCT89.HP		2,141,245	
JANUARY 1990	total (40.16 hours)		7,122,417,589	27,954,961	3.9%
Start of trace: Jan. 10, 6:07am	low hour (Jan. 11, 8:32pm-9:32pm)	JAN90.LB	87,299,639		1.9%
		JAN90.LP		310,038	
End of trace: Jan. 11, 10:17pm	normal hour (Jan. 10, 9:32am-10:32am)	JAN90.MB	182,636,845		4.1%
		JAN90.MP		643,451	
	busy hour (Jan. 11, 10:32am-11:32am)	JAN90.HB	711,529,370		15.8%
		JAN90.HP		1,391,718	
FEBRUARY 1992	total (47.91 hours)		6,585,355,731	27,674,814	3.1%
Start of trace: Feb. 18, 5:22am	low hour (Feb. 20, 1:21am-2:21am)	FEB92.LB	56,811,435		1.3%
		FEB92.LP		231,823	
End of trace: Feb. 20, 5:16am	normal hour (Feb. 18, 8:21pm-9:21pm)	FEB92.MB	154,626,159		3.4%
		FEB92.MP		524,458	
	busy hour (Feb. 18, 11:21am-12:21am)	FEB92.HB	225,066,741		5.0%
		FEB92.HP		947,662	

Table 1. Qualitative description of the sets of Ethernet traffic measurements used in the analysis in Section 4.

this data set. Overall, the 4 data sets cover a wide range of network utilizations and host populations over a 4 year period.

3. SELF-SIMILAR STOCHASTIC PROCESSES

The presentation below of the mathematical and statistical properties of self-similar processes closely follows Cox (1984) and Beran et al. (1992).

3.1 SELF-SIMILARITY BY PICTURE

For 27 consecutive hours of monitored Ethernet traffic from the August 1989 measurements (first row in Table 1), Figure 1 depicts a sequence of simple plots of the packet counts (i.e., number of packets per time unit) for 5 different choices of time units. Starting with a time unit of 100 seconds (a), each subsequent plot is obtained from the previous one by increasing the time resolution by a factor of 10 and by concentrating on a randomly chosen subinterval (as indicated by the darker shade). Recall that the time unit corresponding to the finest time scale is 10 milliseconds (e). Observe that all plots look intuitively very "similar" to one another (in a distributional sense) and are distinctively different from white noise (i.e., an independent and identically distributed sequence of random variables). Notice also the scaling property (y-axis) and the absence of a natural length of a "burst": at every time scale ranging from milliseconds to minutes and hours, bursts consist of bursty subperiods separated by less bursty subperiods. This scale-invariant or "self-similar" feature of Ethernet traffic is drastically different from both conventional telephone traffic and from stochastic models for packet traffic currently considered in the literature. The latter typically produce plots of packet counts which are indistinguishable from white noise after aggregating over a few hundred milliseconds. This pictorial "proof" of the self-similar nature of Ethernet packet traffic suggests that Ethernet traffic on one time scale is statistically identical (at least with respect to its second-order statistical properties) to Ethernet traffic on a different time scale and, thus, motivates the use of self-similar stochastic processes for traffic modeling purposes.

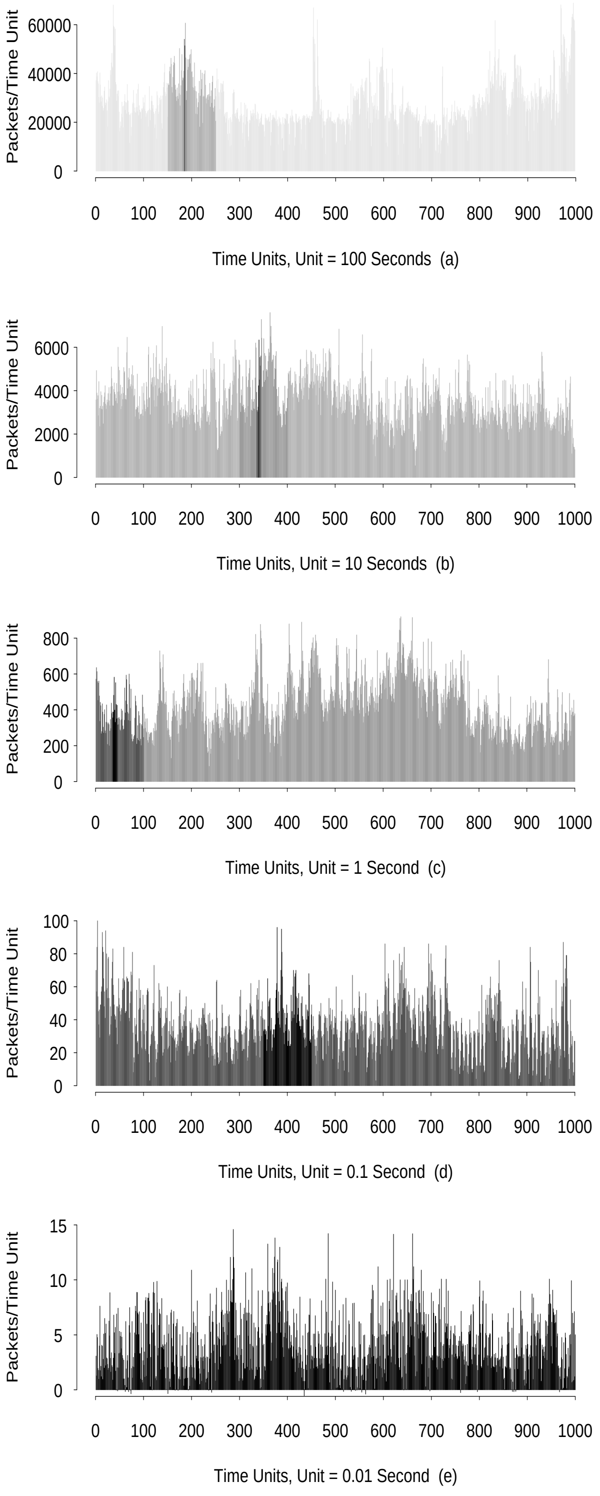


Figure 1 (a)—(e). Pictorial "proof" of self-similarity: Ethernet traffic (packets per time unit for the August '89 trace) on 5 different time scales. (Different gray levels are used to identify the same segments of traffic on the different time scales.)

3.2 THE MATHEMATICS OF SELF-SIMILARITY

The notion of self-similarity is not merely an intuitive description, but a precise concept captured by the following rigorous mathematical definition. Let $X = (X_t : t = 0, 1, 2, ...)$ be a *covariance stationary* (sometimes called *wide-sense stationary*) stochastic process; that is, a process with constant mean $\mu = E[X_t]$, finite variance $\sigma^2 = E[(X_t - \mu)^2]$, and an autocorrelation function $r(k) = E[(X_t - \mu)(X_{t+k} - \mu)] / E[(X_t - \mu)^2]$ $(k = 0, 1, 2, ...)$ that depends only on k. In particular, we assume that X has an autocorrelation function of the form

$$r(k) \sim a_1 k^{-\beta}, \quad as\ k \to \infty, \tag{3.2.1}$$

where $0 < \beta < 1$ (here and below, $a_1, a_2, \cdots$ denote finite positive constants). For each $m = 1, 2, 3, \cdots$, let $X^{(m)} = (X_k^{(m)} : k = 1, 2, 3, ...)$ denote a new time series obtained by averaging the original series X over non-overlapping blocks of size m. That is, for each $m = 1, 2, 3, \cdots$, $X^{(m)}$ is given by $X_k^{(m)} = 1/m\ (X_{km-m+1} + \cdots + X_{km})$, $(k \geq 1)$. Note that for each m, the aggregated time series $X^{(m)}$ defines a covariance stationary process; let $r^{(m)}$ denote the corresponding autocorrelation function.

The process X is called *exactly (second-order) self-similar* with self-similarity parameter $H = 1 - \beta/2$ if the corresponding aggregated processes $X^{(m)}$ have the same correlation structure as X, i.e., $r^{(m)}(k) = r(k)$, for all $m = 1, 2, \cdots$ $(k = 1, 2, 3, \cdots)$. In other words, X is exactly self-similar if the aggregated processes $X^{(m)}$ are indistinguishable from X—at least with respect to their second order statistical properties. An example of an exactly self-similar process with self-similarity parameter H is *fractional Gaussian noise* (FGN) with parameter $1/2 < H < 1$, introduced by Mandelbrot and Van Ness (1968).

A covariance stationary process X is called *asymptotically (second-order) self-similar* with self-similarity parameter $H = 1 - \beta/2$ if $r^{(m)}(k)$ agrees asymptotically (i.e., for large m and large k) with the correlation structure $r(k)$ of X given by (3.2.1). The *fractional autoregressive integrated moving-average processes* (fractional ARIMA(p,d,q) processes) with $0 < d < 1/2$ are examples of asymptotically second-order self-similar processes with self-similarity parameter $d + 1/2$. (For more details, see Granger and Joyeux (1980), and Hosking (1981).)

Intuitively, the most striking feature of (exactly or asymptotically) self-similar processes is that their aggregated processes $X^{(m)}$ possess a nondegenerate correlation structure as $m \to \infty$. This behavior is precisely the intuition illustrated with the sequence of plots in Figure 1; if the original time series X represents the number of Ethernet packets per 10 milliseconds (plot (e)), then plots (a) to (d) depict segments of the aggregated time series $X^{(10000)}$, $X^{(1000)}$, $X^{(100)}$, and $X^{(10)}$, respectively. All of the plots look "similar", suggesting a nearly identical autocorrelation function for all of the aggregated processes.

Mathematically, self-similarity manifests itself in a number of equivalent ways (see Cox (1984)): (i) the variance of the sample mean decreases more slowly than the reciprocal of the sample size (*slowly decaying variances*), i.e., $var(X^{(m)}) \sim a_2 m^{-\beta}$, as $m \to \infty$, with $0 < \beta < 1$; (ii) the autocorrelations decay hyperbolically rather than exponentially fast, implying a non-summable autocorrelation function $\sum_k r(k) = \infty$ (*long-range dependence*), i.e., $r(k)$ satisfies relation (3.2.1); and (iii) the spectral density $f(\cdot)$ obeys a power-law behavior near the origin (*1/f-noise*), i.e., $f(\lambda) \sim a_3 \lambda^{-\gamma}$, *as* $\lambda \to 0$, with $0 < \gamma < 1$ and $\gamma = 1 - \beta$.

The existence of a nondegenerate correlation structure for the aggregated processes $X^{(m)}$ as $m \to \infty$ is in stark contrast to typical packet traffic models currently considered in the literature, all of which have the property that their aggregated processes $X^{(m)}$ tend to second order pure noise, i.e., $r^{(m)}(k) \to 0$, as $m \to \infty$ ($k = 1, 2, \cdots$). Equivalently (see Cox(1984)), they can be characterized by (i) a variance of the sample mean that decreases like the reciprocal of the sample mean, (ii) an autocorrelation function that decreases exponentially fast, implying a summable autocorrelation function $\sum_k r(k) < \infty$ (*short-range dependence*), or (iii) a spectral density that is bounded at the origin.

Historically, the importance of self-similar processes lies in the fact that they provide an elegant explanation and interpretation of an empirical law that is commonly referred to as *Hurst's law* or the *Hurst effect*. Briefly, for a given set of observations $(X_k: k = 1, 2, ..., n)$ with sample mean $\bar{X}(n)$ and sample variance $S^2(n)$, the *rescaled adjusted range* or the *R/S statistic* is given by $R(n)/S(n) = 1/S(n)\,[\max(0, W_1, W_2, ..., W_n) - \min(0, W_1, W_2, ..., W_n)]$, with $W_k = (X_1 + X_2 + \cdots + X_k) - k\bar{X}(n)$, $k = 1, 2, ..., n$. Hurst (1955) found that many naturally occurring time series appear to be well represented by the relation $E[R(n)/S(n)] \sim a_4 n^H$, as $n \to \infty$, with *Hurst parameter* H "typically" about 0.73. On the other hand, if the observations X_k come from a short-range dependent model, then Mandelbrot and Van Ness (1968) showed that $E[R(n)/S(n)] \sim a_5 n^{0.5}$, *as* $n \to \infty$. This discrepancy is generally referred to as the *Hurst effect* or *Hurst phenomenon*.

Finally, for an attempt to explain self-similarity in terms of representing some underlying physical process, we refer to a construction originally introduced by Mandelbrot (1969) (see also Taqqu and Levy (1986)) of self-similar processes based on aggregating many simple renewal reward processes exhibiting inter-renewal times with infinite variances (i.e., "heavy-tails"). That is, the distribution of the inter-renewal times U satisfies $P[U \geq u] \sim a_6 u^{-\alpha}$, *as* $u \to \infty$, $1 < \alpha < 2$ (e.g., stable (Pareto) distributions with parameter $1 < \alpha < 2$). Producing self-similarity by aggregating more and more i.i.d. copies of elementary renewal reward processes relies crucially on this "heavy-tail property" of the inter-renewal times and provides an intuitive explanation for the occurrence of self-similarity in high-speed network traffic (see Section 5).

3.3 THE STATISTICS OF SELF-SIMILARITY

Since slowly decaying variances, long-range dependence, and a spectral density obeying a power-law behavior are different manifestations of one and the same property of the underlying covariance stationary process X, namely that X is (asymptotically or exactly) self-similar, we can approach the problem of testing for and estimating the degree of self-similarity from three different angles: (1) time-domain analysis based on the R/S-statistic, (2) analysis of the variances of the aggregated processes $X^{(m)}$, and (3) periodogram-based analysis in the frequency-domain. This subsection provides a brief discussion of the graphical R/S analysis and briefly mentions methods (2) and (3). We will illustrate the use of these methods in our analysis of the Ethernet data in Section 4 below.

The objective of the R/S analysis of an empirical record is to infer the degree of self-similarity H (Hurst parameter) in relation (3.1.6) for the self-similar process that presumably generated the record under consideration. In practice, R/S analysis is based on a heuristic graphical approach (originally described in detail in Mandelbrot and Wallis (1969)) that tries to exploit as fully as possible the information in a given record. The following graphical method has been used extensively in the past. Given a sample of N observations $(X_k : k = 1, 2, 3, \cdots, N)$, one subdivides the whole sample into K non-overlapping blocks and computes the rescaled adjusted range $R(t_i, n)/S(t_i, n)$ for each of the new "starting points" $t_1 = 1$, $t_2 = N/K + 1$, $t_3 = 2N/K + 1$, $\cdots$ which satisfy $(t_i - 1) + n \leq N$. Here, the R/S-statistic $R(t_i, n)/S(t_i, n)$ is defined as above with W_k replaced by $W_{t_i+k} - W_{t_i}$ and $S^2(t_i, n)$ is the sample variance of $X_{t_i+1}, X_{t_i+2}, \cdots, X_{t_i+n}$. Thus, for a given value ("lag") of n, one obtains many samples of R/S, as many as K for small n and as few as one when n is close to the total sample size N. Next, one takes logarithmically spaced values of n, starting with $n \approx 10$. Plotting $log(R(t_i, n)/S(t_i, n))$ versus $log(n)$ results in the *rescaled adjusted range plot* (also called the *pox diagram of R/S*). When the parameter H is well defined, a typical rescaled adjusted range plot starts with a transient zone representing the nature of short-range dependence in the sample, but eventually settles down and fluctuates in a straight "street" of a certain slope. Graphical R/S analysis is used to determine whether such asymptotic behavior appears supported by the data. In the affirmative, an estimate $\hat{H}$ of the self-similarity parameter H is given by the street's asymptotic slope (typically obtained by a simple least squares fit) which can take any value between 1/2 and 1. For practical purposes, the most useful and attractive feature of the R/S analysis is its relative robustness against changes of the marginal distribution. This feature allows for practically separate investigations of the self-similarity property of a given empirical record and of its distributional characteristics.

We have observed that for self-similar processes, the variances of the aggregated processes $X^{(m)}$ ($m = 1, 2, 3, \cdots$) decrease linearly (for large m) in log-log plots against m with slopes arbitrarily flatter than -1. The so-called *variance-time plots* are obtained by plotting $\log(\mathrm{var}(X^{(m)}))$ against $\log(m)$ ("time") and by fitting a simple least squares line through the resulting points in the plane, ignoring the small values for m. Values of the estimate $\hat{\beta}$ of the asymptotic slope between -1 and 0 suggest self-similarity, and an estimate for the degree of self-similarity is given by $\hat{H} = 1 - \hat{\beta}/2$.

The absence of any limit law results for the statistics corresponding to the R/S analysis or the variance-time plot makes them inadequate for a more refined data analysis (e.g., requiring confidence intervals for the degree of self-similarity H, model selection criteria, and goodness of fit tests). In contrast, a more refined data analysis is possible for maximum likelihood type estimates (MLE) and related methods based on the periodogram and its distributional properties. In particular, for Gaussian processes, Whittle's approximate MLE has been studied extensively and has been shown to have desirable statistical properties (Fox and Taqqu (1986), and Dahlhaus (1989)). Combined, Whittle's approximate MLE approach and the aggregation method discussed earlier give rise to the following operational procedure for obtaining confidence intervals for the self-similarity parameter H. For a given time series, consider the corresponding aggregated processes $X^{(m)}$ with $m = 100, 200, 300, \cdots$, where the largest m-value is chosen such that the sample size of the corresponding series $X^{(m)}$ is not less than about 100. For each of the aggregated series, estimate the self-similarity parameter H via the Whittle estimate. This procedure results in estimates $\hat{H}^{(m)}$ of H and corresponding, say, 95%-confidence intervals of the form $\hat{H}^{(m)} \pm 1.96\hat{\sigma}_H$, where $\hat{\sigma}_H^2$ is given by a known central limit theorem result (see Dahlhaus (1989)). Finally, we plot the estimates $\hat{H}^{(m)}$ of H together with their 95%-confidence intervals versus m. Such plots will typically vary widely for small aggregation levels, but will stabilize after a while and fluctuate around a constant value, our final estimate of the self-similarity parameter H.

4. THE SELF-SIMILAR NATURE OF ETHERNET TRAFFIC

In this section, we establish in a statistically rigorous manner (using the graphical and statistical tools described in the previous section) the self-similar nature of (internal) Ethernet traffic and of some of its major components (e.g., external traffic and external TCP traffic). For each of the 4 measurement periods described in Table 1, we identified what are considered "typical" low-, medium-, and high-activity hours. With the resulting data sets, we are able to investigate features of the observed traffic (e.g., self-similarity) that persist across the network as well as across time, irrespective of the utilization level of the Ethernet.

4.1 ETHERNET TRAFFIC OVER A 1-DAY PERIOD

We first consider the August ’89 snapshot of Ethernet traffic (row 1 in Table 1) and analyze the 3 subsets AUG89.LB, AUG89.MB, and AUG89.HB. Each sequence contains 360000 observations, and each observation represents the number of bytes sent over the Ethernet every 10 milliseconds.

Figure 2 depicts the pox plot of R/S (a), the variance-time curve (b), and the periodogram plot (c) corresponding to the sequence AUG89.MB. The pox plot of R/S (Figure 2 (a)) show an asymptotic slope that is distinctly different from 0.5 (lower dotted line) and 1.0 (upper dotted line) and is easily estimated (using the points) to be about 0.79. The variance-time curve (Figure 2 (b)), which has been normalized by the corresponding sample variance, shows an asymptotic slope that is clearly different from -1 (dotted line) and is easily estimated to be about -0.40, resulting in a practically identical estimate of the Hurst parameter H of about 0.80. Finally, looking at the periodogram plot corresponding to the time series AUG89.MB, we observe that although there are some pronounced peaks in the high-frequency domain of the periodogram, the low-frequency part is characteristic for a power-law behavior of the spectral density around zero. In fact, by fitting a simple least-squares line using only the lowest 10% of all frequencies, we obtain a slope estimate $\hat{\gamma} \approx 0.64$ which results in a Hurst parameter estimate of about 0.82. Thus, together the 3 graphical methods suggest that the sequence AUG89.MB is self-similar with self-similarity parameter $H \approx 0.80$. Moreover, Figure 2 (d) indicates that the normal hour Ethernet traffic of the August ’89 data is (exactly) self-similar rather than asymptotically self-similar (see Section 3.2). Figure 2 (d) shows the estimates of the Hurst parameter H for selected aggregated time series derived from the sequence AUG89.MB, as a function of the aggregation level m. For aggregation levels $m = 1,\ 5,\ 10,\ 50,\ 100,\ 500,\ 1000$, we plot the Hurst parameter estimate $\hat{H}^{(m)}$ (based on the pox plots of R/S ("$*$"), the variance-time curves ("○"), and the periodogram plots ("□")) for the aggregated time series $X^{(m)}$ against the logarithm of the aggregation level m. Notice that the estimates are extremely stable and practically constant over the depicted range of aggregation levels $1 \le m \le 1000$. Thus, in terms of their second-order statistical properties, the aggregated series $X^{(m)} (m \ge 1)$ can be considered to be identical and produce, therefore, realizations that have similar overall structure and look very much alike.

In addition to the sequence AUG89.MB, we also analyzed the sequences AUG89.LB and AUG89.HB. While both series behave very much like the sequence AUG89.MB, the resulting Hurst parameter estimates $\hat{H}$ differ slightly; $\hat{H} \approx 0.75$ for AUG89.LB, and $\hat{H} \approx 0.85$ for AUG89.HB. This difference suggests that although Ethernet traffic over approximately a 24-hour period is self-similar, the degree of self-similarity depends on the utilization level of the Ethernet and increases as the utilization increases. Finally, we also analyzed the sequences AUG89.LP, AUG89.MP, and AUG89.HP, i.e., the time series representing the total number of Ethernet packets rather than the total number of bytes in every 10 millisecond interval. Not surprisingly, the packet traffic is also self-similar, but with slightly larger H-values than the corresponding bytes/time unit data. For a more detailed analysis of Ethernet traffic at the packet level, see Section 4.2 below.

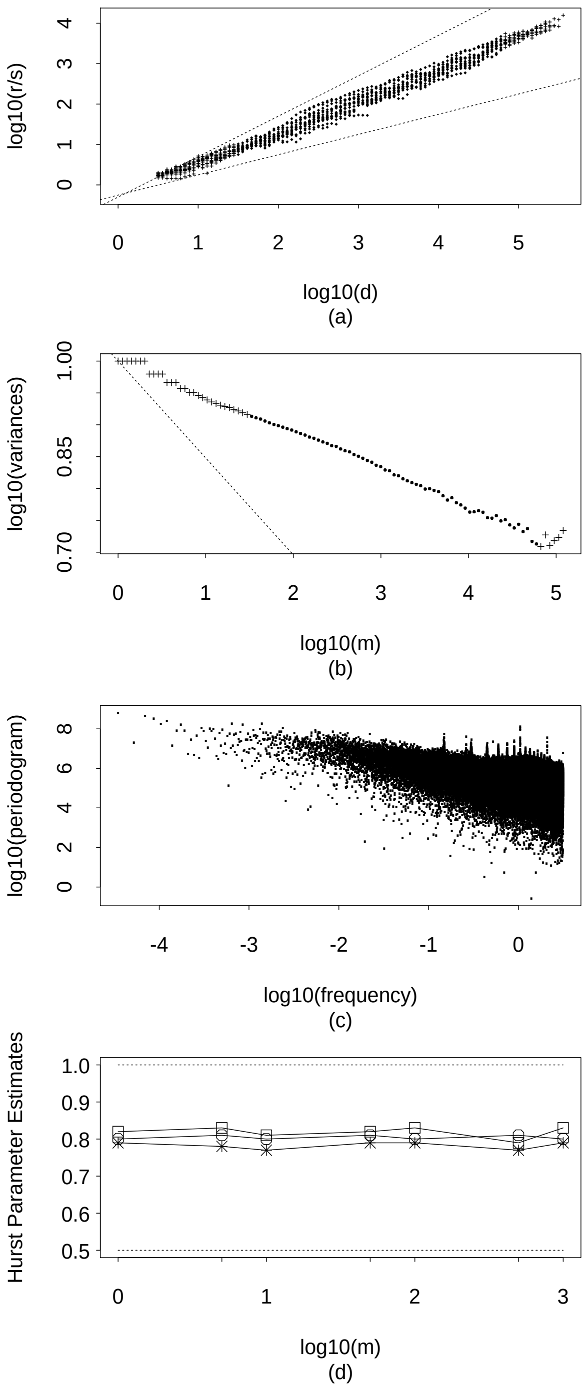


Figure 2 (a)—(d). Graphical methods for checking for self-similarity of the sequence AUG89.MB. ((a) pox plot of R/S, (b) variance-time plot, (c) periodogram plot, and (d) Hurst parameter estimates as a function of the aggregation level m ($*$ is method (a), ○ is method (b), and □ is method (c).)

4.2 ETHERNET TRAFFIC OVER A 4-YEAR PERIOD

As discussed in Section 2, Ethernet LANs are generally known to change significantly during the course of a few years. By analyzing additional data sets (see Table 1), similar to the August ’89 ones but taken at different points in time and at different physical locations within the network, we show below that Ethernet traffic is self-similar, irrespective of when and where the data were collected in the Bellcore Morristown network during the 4-year period August ’89—February ’92.

In contrast to Section 4.1, our analysis below results in point estimates of the self-similarity parameter H together with their respective 95%-confidence intervals. As discussed in Section 3.3, such a refined analysis resulting in confidence intervals for H is possible if maximum likelihood type estimates (MLE) or related estimates based on the periodogram are used instead of the mostly heuristic graphical estimation methods illustrated in the previous section. Plots (a)—(d) of Figure 3 show the result of the MLE-based estimation method when combined with the method of aggregation. For each of the 4 sets of traffic measurements described in Table 1, we use the time series representing the packet counts during normal traffic conditions (i.e., AUG89.MP in (a), OCT89.MP in (b), JAN90.MP in (c), and FEB92.MP in (d)), and consider the corresponding aggregated time series $X^{(m)}$ with $m = 100, 200, 300, ..., 1900, 2000$ (representing the packet counts per 1, 2, ..., 20 seconds, respectively). We plot the Hurst parameter estimates $\hat{H}^{(m)}$ of H obtained from the aggregated series $X^{(m)}$, together with their 95%-confidence intervals, against the aggregation level m. Figure 3 shows that for the packet counts during normal traffic loads (irrespective of the measurement period), the values of $\hat{H}^{(m)}$ are quite stable and fluctuate only slightly in the 0.85 to 0.95 range throughout the aggregation levels considered. The same holds for the 95%-confidence interval bands, indicating strong statistical evidence for self-similarity of these 4 time series with degrees of self-similarity ranging from about 0.85 to about 0.95. The relatively stable behavior of the Hurst parameter estimates $\hat{H}^{(m)}$ for the different aggregation levels m also confirms our earlier finding that Ethernet traffic during normal traffic hours can be considered to be exactly self-similar rather than asymptotically self-similar. Plots (a)—(d) of Figure 3 suggest that this property holds irrespective of when and where the Ethernet was monitored. For exactly self-similar time series, determining a single point estimate for H and the corresponding 95%-confidence interval is straightforward and can be done by visual inspection of plots such as the ones in Figure 3 (see below). Notice that in each of the four plots in Figure 3, we added two lines corresponding to the Hurst parameter estimates obtained from the pox diagrams of R/S and the variance-time plots, respectively. Typically, these lines fall well within the 95%-confidence interval bands which shows that for these long time series considered here, graphical estimation methods based on R/S or variance-time plots can be expected to be very accurate.

In addition to the 4 normal hour packet data time series, we also applied the combined MLE/aggregation method to the other traffic data sets described in Table 1. Figure 4 (a) depicts all Hurst parameter estimates (together with the 95%-confidence interval corresponding to the choice of m discussed earlier) for each of the 12 packet data time series, while Figure 4 (b) summarizes the same information for the time series representing the number of bytes per 10 milliseconds during a typical low, normal, and busy hour for each of the four measurement periods. We also included in these summary plots the Hurst parameter estimates obtained via the R/S analysis ("$*$") and variance-time plots ("○") in order to indicate the

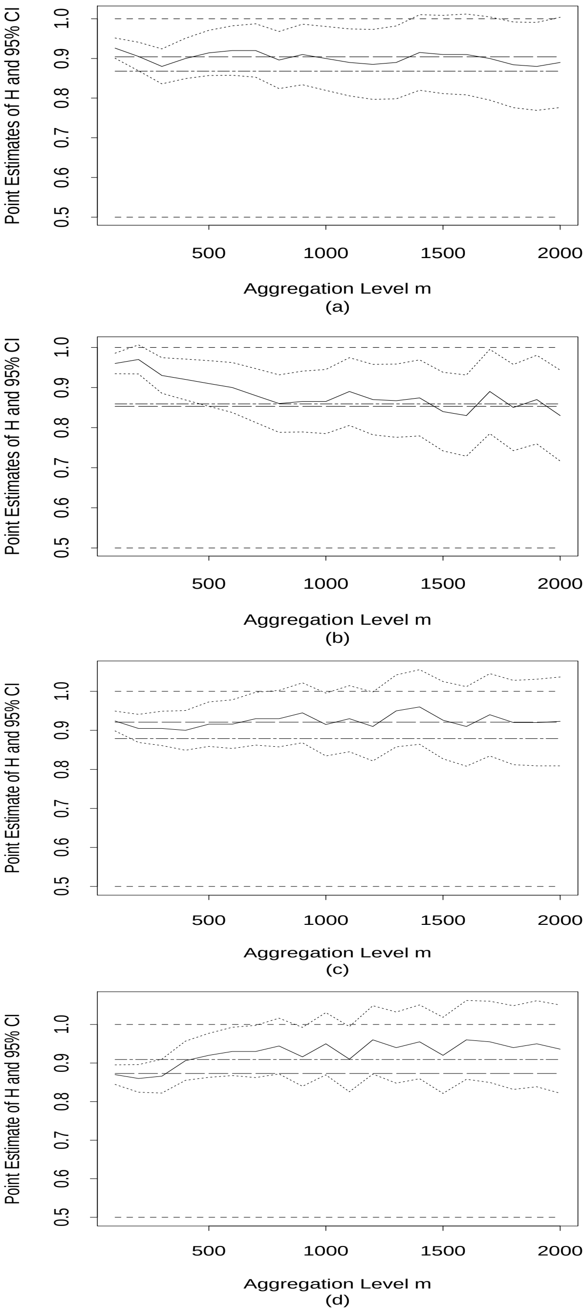


Figure 3 (a)—(d). Periodogram-based MLE estimate $\hat{H}^{(m)}$ of H (solid line) and 95%-confidence intervals (dotted lines), as a function of the aggregation level m, for sequences AUG89.MP (a), OCT89.MP (b), JAN90.MP (c), and FEB92.MP (d). For example, plot (a) shows that $m \approx 300$ is an appropriate aggregation level for sequence AUG89.MP, yielding a point estimate $\hat{H} = \hat{H}^{(300)} = 0.90$ and a 95%-confidence interval [0.85,0.95]. For comparison, we also added to each plot the estimate of H based on the variance-time plot (–·–·–) and the R/S-based estimate of H (– – –).

accuracy of these "graphical" estimators when compared to the statistically more rigorous Whittle estimator ("•"). Concentrating first on the packet data, i.e., Figure 4 (a), we see that despite the transition from mostly host-to-host workgroup traffic during the August ’89 and October ’89 measurement

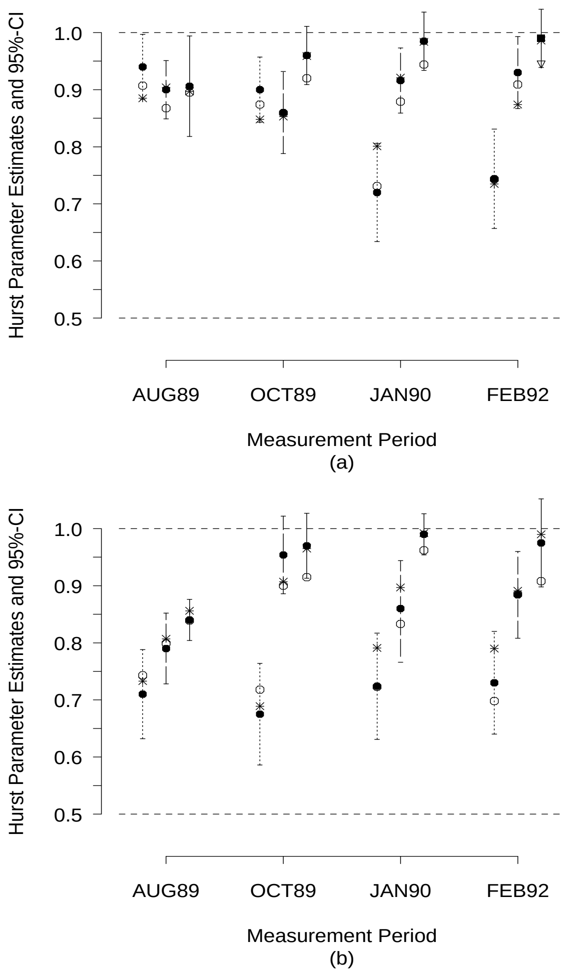


Figure 4 (a)—(b). Summary plot of estimates of the Hurst parameter *H* for all data sets (representing the number of packets (a) and the number of bytes (b)) in Table 1. (● denotes the periodogram-based MLE estimate and corresponding 95%-confidence interval, ○ the point estimate of *H* based on the asymptotic slope of the variance-time plot, and $*$ the point estimate of *H* based on the asymptotic slope of the pox plot of R/S.)

periods, to a mixture of host-to-host and router-to-router traffic during the January ’90 measurement period, to the predominantly router-to-router traffic of the February ’92 data set, the Hurst parameter corresponding to the typical normal and busy hours, respectively, are comparable, with slightly higher *H*-values for the busy hours than for the normal traffic hours. This latter observation might be surprising in light of conventional traffic modeling where it is commonly assumed that as the number of sources (Ethernet users) increases, the resulting aggregate traffic becomes smoother. In contrast to this generally accepted argument for the "Poisson-like" nature of aggregate traffic, our analysis of the Ethernet data shows that, in fact, the aggregate traffic tends to become less smooth (or, more bursty) as the number of active sources increases (see also our discussion in Section 5). In fact, while there were about 120 hosts that spoke up during the August ’89 or October ’89 busy hour, we heard from about 1200 hosts during the January ’90 high traffic hour; the comparable number of active hosts during the February ’92 busy hour was around 600. The major difference between the early (pre-1990) measurements and the the later ones (post-1990, i.e., January ’90 and February ’92) can be seen during the low traffic hours. Intuitively, low period router-to-router traffic consists mostly of machine-generated packets which tend to form a much smoother arrival process than low period host-to-host traffic, which is typically produced by a smaller than average number of actual Ethernet users, i.e., researchers working late hours.

Next, turning our attention to Figure 4 (b), i.e., the Hurst parameter estimates for the bit rates, we observe that as in the case of the packet data, the degree of self-similarity *H* increases as we move from low to normal to high traffic hours. Moreover, while there is practically no difference between the two post-1990 data sets, the two pre-1990 sets clearly differ from one another but follow a similar pattern as the post-1990 ones. The difference between the August ’89 and October ’89 measurements can be explained by the transition from diskless to "dataless" workstations (workstations with the operating system on a local disk but all user files on a remote fileserver) that occurred during the latter part of 1989 (see Section 2). Except during the low hours, the increased computing power of many of the Ethernet hosts causes the Hurst parameter to increase and gives rise to a bit rate that nearly matches the self-similar feature of the corresponding packet process. Also note that the 95%-confidence intervals corresponding to the Hurst parameter estimates for the low traffic hours are typically wider than those corresponding to the estimates of *H* for the normal and high traffic hours. This widening indicates that Ethernet traffic during low traffic periods is asymptotically self-similar rather than exactly self-similar.

The Ethernet traffic analyzed in Section 4.2 is called *internal* traffic and consists of all packets on a LAN. In addition to this internal traffic, we also analyzed *external* or *remote* Ethernet traffic, and *external TCP* traffic, the portion of external traffic using the Transmission Control Protocol (TCP) and IP. Repeating the same laborious statistical analysis for these important components of internal Ethernet traffic, we find that in terms of its self-similar nature, external traffic and external TCP traffic do not differ from the internal traffic studied earlier, and that our findings for the internal traffic apply directly.

5. SIGNIFICANCE OF SELF-SIMILARITY FOR TRAFFIC ENGINEERING

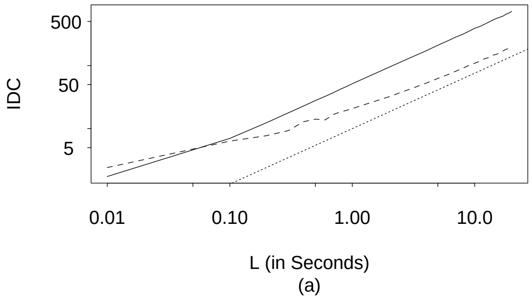
Our measured data show dramatically different statistical properties than those predicted by the stochastic models currently considered in the literature. Almost all these models are characterized by an exponentially decaying autocorrelation function. As a result, they give rise to a Hurst parameter estimate of $\hat{H} = .50$, producing variance-time curves, R/S plots, and frequency domain behavior strongly disagreeing with the self-similar behavior of actual traffic (see Section 4). In terms of the aggregation procedure described above, the theoretical models have the property that typically, after aggregating over non-overlapping blocks of about 10-100 observations, the aggregated series become indistinguishable from second-order pure noise. The fact that one can distinguish clearly—with respect to second-order statistical properties—between the existing models for Ethernet traffic and our measured data is surprising and clearly questions some of the modeling assumptions that have been made in the past. Potential traffic engineering implications of this distinction are currently under intense scrutiny. In this section, we emphasize three direct implications of the self-similar nature of packet traffic for traffic engineering purposes: modeling individual Ethernet sources, inadequacy of conventional notions of "burstiness", and effects on congestion management for packet networks. We conclude with some guidelines toward modeling self-similar traffic and suggest some open problems.

5.1 ON THE NATURE OF TRAFFIC GENERATED BY AN INDIVIDUAL ETHERNET USER

In Section 4, we showed that irrespective of when and where the Ethernet measurements were collected, the traffic is self-similar, with different degrees of self-similarity depending on the load on the network. We did so without first studying and modeling the behavior of individual Ethernet users (sources). Although historically, accurate source modeling has been considered an absolute necessity for successful modeling of aggregate traffic, we show here that in the case of self-similar packet traffic, knowledge of fundamental characteristics of the aggregate traffic can provide new insight into the nature of traffic generated by an individual user. To this end, we recall Mandelbrot's construction of self-similar processes (see Mandelbrot (1969)) by aggregating many simple renewal reward processes, which provides a physical explanation for the visually obvious (see Figure 1) and statistically significant (see Figure 4) self-similarity property of Ethernet LAN traffic in terms of the behavior of individual Ethernet users. In fact, the renewal rewards for one such process represent the amount of traffic (in bytes or packets) generated by a single user during successive time intervals whose lengths obey the "heavy-tail" property discussed in Section 3.2. If bytes are the preferred units, the renewal reward process source model resembles the popular class of fluid models (see Anick et al. (1982)). On the other hand, if we think of packets as the underlying unit of information, the renewal reward process is basically a packet train model in the sense of Jain and Routhier (1986). For simplicity, one can restrict the "rewards" to the values 0 and 1, where 0 means that the corresponding source is inactive and 1 means that it is active (and generating traffic at a fixed rate). Note that the crucial property that distinguishes this model from fluid models and packet train models is that the lengths of the inactive/active periods are *heavy-tailed* in the sense of Section 3.2. Intuitively, this property means that there is no characteristic length of a busy period or packet train: individual inactive/active periods can be arbitrarily long with significant probability (for evidence in support of this "heavy-tail behavior" in related traffic studies, see the recent work by Meier-Hellstern et al. (1991)). Mandelbrot (1969) and Taqqu and Levy (1986) showed that aggregating the traffic of many such source models produces a self-similar superposition process with self-similarity parameter $H = (3 - \alpha)/2$, where α is the parameter that characterizes the "thickness" of the tail of the distribution.

5.2 SELF-SIMILARITY AND SOME COMMONLY USED NOTIONS OF BURSTINESS

Section 4 has shown that data sets with higher self-similarity H satisfy intuitive notions of higher "burstiness". Similarly, the greater the variability (smaller α) of the inactive/active periods in our individual source model, the higher the H and the burstier the aggregate traffic. The fact that the self-similarity parameter H captures the intuitive notion of burstiness in a mathematically rigorous manner can be contrasted with the behavior of many commonly used measures of "burstiness", including the *index of dispersion*, the *peak-to-mean ratio*, and the *coefficient of variation*. The index of dispersion (for counts) has recently attracted considerable attention (e.g., Heffes and Lucantoni (1986)) as a measure for capturing the variability of traffic over different time scales. For a given time interval of length L, the index of dispersion for counts (IDC) is given by the variance of the number of arrivals during the interval of length L divided by the expected value of that same quantity. the IDC increases monotonically throughout a time span that covers nearly 5 orders of magnitude (Figure 5), in stark contrast to conventional



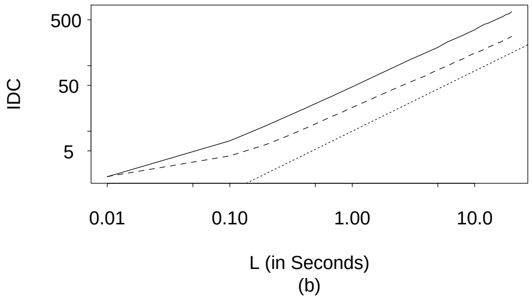


Figure 5 (a)—(b). Index of dispersion for counts (IDC) as a function of the length L of the time interval over which IDC is calculated, for the January '90 busy hour (a) and the February '92 busy hour (b). The solid lines are the IDC curves for the sequences JAN90.HP and FEB92.HP (internal traffic) and the dashed lines depict the IDC curves for the corresponding hours of external traffic. The dotted lines are the IDC curves predicted by a self-similar model fitted to the time series JAN90.HP and FEB92.HP, respectively. Note that the asymptotic slopes of the solid lines agree with the slopes of the dotted lines.

traffic models where the IDC is either constant or converges to a fixed value rapidly. Simple self-similar traffic models with parameter H are easily shown to produce a monotonically increasing IDC proportional to L^{2H-1}, exactly matching the straight-line appearance of the log-log plot of the actual data. For self-similar traffic, both the peak-to-mean ratio and the coefficient of variation are unsatisfactory measures: essentially any peak-to-mean ratio is possible, depending on the length of the interval over which the peak is determined, and essentially any ratio of the standard deviation of interarrival times to the expected value is possible, depending on the sample size.

5.3 CONGESTION MANAGEMENT IN THE PRESENCE OF SELF-SIMILAR TRAFFIC

In order to illustrate the effect of self-similar traffic on basic architectural issues concerning high-speed, high-bandwidth communications systems of the future, we revisit some aspects of congestion management first explored using simulation by Fowler and Leland (1991) and Leland and Wilson (1991). Because of the statistical groundwork established in Section 4, their conclusions about the nature of congestion and the task of congestion management for B-ISDN provide convincing evidence for the significance of self-similar network traffic for engineering future integrated high-speed networks. In light of the same self-similar behavior of VBR video traffic (see Beran et al. (1992)), their conclusions are likely to also apply to more heterogeneous B-ISDN traffic.

Leland and Wilson (1991) consider the access control scheme proposed for Switched Multimegabit Data Service (SMDS) on public B-ISDN. SMDS is a connectionless data service where packets arriving from a LAN are buffered at the interface and delivered to the cell-based network at some maximum rate, subject to traffic shaping intended to reduce the burstiness of the submitted LAN traffic. Simple conventional models based on the observed external LAN traffic suggest that proposed SMDS quality of service requirements can readily be met. However, packet loss and delay behaviors differ radically between trace-driven simulations based on the actual traffic measurements and those based on these formal traffic models. In particular, overall packet loss decreases very slowly with increasing buffer capacity, in sharp contrast to Poisson-based models where losses decrease exponentially fast with increasing buffer size. Moreover, packet delay always increases with buffer capacity, again in contrast to the formal models where delay does not exceed a fixed limit regardless of buffer size. This distinctive loss and delay behavior can be precisely explained in terms of the frequency domain manifestation of self-similarity (see Section 3.2). Because both low and high frequencies are significant, heavy losses and long delays occur during long time-frame bursts (due to the presence of low frequencies) and can, therefore, not be dealt with effectively by larger buffers. This observation is also backed by a recent analytic study by Norros (1992) who studies a model for connectionless traffic based on fractional Brownian motion. The mathematical properties of self-similarity also explain the results of Fowler and Leland (1991), who also observed the ineffectiveness of buffering to manage congestion and went on to observe that when congestion occurs, losses are severely concentrated and are far greater than the background loss rate. While many formal standard network traffic models provably show that congestion control "works" (e.g., large buffers provide protection against congestion and average loss rates are a sensible quality of service measure) self-similar traffic models reveal a far more challenging picture for broadband congestion management.

5.4 PARSIMONIOUS MODELS FOR SELF-SIMILAR TRAFFIC

Self-similarity is often explained as being equivalent to the existence of a multilevel hierarchy of underlying mechanisms. For packet traffic, it is practically impossible to demonstrate why such mechanisms should result, for example, in an asymptotic power law for the autocorrelations of the form (3.2.1). Even if their physical reality could be established, the resulting models for packet traffic are likely to have a large number of parameters. Similarly, conventional modeling approaches that stress the importance of source modeling produce highly overparameterized models for aggregate traffic. Two alternative approaches are far more *parsimonious* for self-similar traffic, yielding models with a small number of parameters where every parameter can be given a physically meaningful interpretation.

As we have noted in Section 4, self-similar stochastic models fit Ethernet traffic very well using very few parameters. For example, FGN is characterized by just 3 parameters (mean μ, variance σ^2, and H), each with a natural physical interpretation. Parameter estimation techniques are known for FGN and fractional ARIMA models, but often turn out to be computationally too intensive for large data sets. However, Section 4 illustrates how to estimate the parameter H for large data sets, and methods to adapt the existing parameter estimation techniques to long time series are currently being studied. Another promising approach to modeling packet traffic uses deterministic *chaotic maps* (Erramilli and Singh (1990)). Generating a packet stream using this approach is appealingly easy; however, the problem of deriving an appropriate nonlinear chaotic map based on a set of actual traffic measurements currently requires considerable experimenting. Developing more rigorous statistical estimation methods for dynamical systems has recently attracted considerable attention in the statistics literature (e.g. Berliner (1992) and references therein). Both approaches offer a simple description of the complex packet traffic generation process, and each yield a single parameter that describes the fractal-like nature of traffic (H and the fractal dimension, respectively). While traditional performance modeling favors the use of stochastic input models, studying arrival streams to queues that are generated by non-linear chaotic maps may well provide new insight into the performance of queueing systems where the arrival processes exhibit fractal-like properties.

6. CONCLUSIONS

Understanding the nature of traffic in high-speed, high-bandwidth communications systems such as B-ISDN is essential for engineering, operations, and performance evaluation of these networks. In a first step toward this goal, it is important to know the traffic behavior of some of the expected major contributors to future high-speed network traffic. In this paper, we analyze LAN traffic offered to a high-speed public network supporting LAN interconnection, an important and rapidly growing B-ISDN service. The main findings of our statistical analysis of a few hundred million high quality, high time-resolution Ethernet LAN traffic measurements are that (i) Ethernet LAN traffic is statistically self-similar, irrespective of when during the 4-year data collection period 1989-1992 the data were collected and where they were collected in the network, (ii) the degree of self-similarity measured in terms of the Hurst parameter H is typically a function of the overall utilization of the Ethernet and can be used for measuring the "burstiness" of the traffic (namely, the burstier the traffic the higher H), and (iii) major components of Ethernet LAN traffic such as external LAN traffic or external TCP traffic share the same self-similar characteristics as the overall LAN traffic.

An important implication of the self-similarity of LAN traffic is that aggregating streams of such traffic typically does not produce a smooth ("Poisson-like") superposition process but instead, intensifies the burstiness (i.e., the degree of self-similarity) of the aggregation process. Thus, self-similarity is both ubiquitous in our data and unavoidable in future, more highly aggregated, traffic. However, none of the currently common formal models for LAN traffic is able to capture the self-similar nature of real traffic. We briefly mention two novel methods for modeling self-similar LAN traffic, based on stochastic self-similar processes and deterministic nonlinear chaotic maps, that provide accurate and parsimonious models.

Implications of the self-similar nature of packet traffic for engineering, operations, and performance evaluation of high-speed networks are ample: (i) source models for individual Ethernet users are expected to show extreme variability in terms of interarrival times of packets (the infinite variance syndrome), (ii) the Hurst parameter provides a more satisfactory measure of "burstiness" for self-similar traffic than such commonly used measures as the index of dispersion, the peak-to-mean-ratio, or the coefficient of variation (which become ill-defined for self-similar traffic), and (iii) the nature of congestion produced by self-similar network traffic models differs drastically from that predicted by standard formal models and displays a far more complicated picture than has been typically assumed in the past. Finally, in light of the same self-similar behavior recently observed in VBR video traffic—another major contributor to future high-speed network traffic—the more complicated nature of congestion due to the self-similar traffic behavior can be expected to persist even when we move toward a more heterogeneous B-ISDN environment. Thus, we believe based on our measured traffic data that the success or failure of, for example, a proposed congestion control scheme for B-ISDN will depend on how well it performs under a self-similar rather than under one of the standard formal traffic scenarios.

ACKNOWLEDGEMENTS

This work could not have been done without the help of J. Beran and R. Sherman who provided the S-functions that made the statistical analysis of an abundance of data possible. We also acknowledge many helpful discussions with A. Erramilli about his dynamical systems approach to packet traffic modeling. M. S. Taqqu was supported, at Boston University, by ONR grant N00014-90-J-1287.

REFERENCES

1. D. Anick, D. Mitra, M.M. Sondhi, "Stochastic Theory of a Data-Handling System with Multiple Sources", *Bell System Technical Journal 61*, 1871-1894, 1982.

2. J. Beran, "Statistical Methods for Data with Long-Range Dependence", *Statistical Science 7,* No. 4, 1992.

3. J. Beran, R. Sherman, M. S. Taqqu, W. Willinger, "Variable-Bit-Rate Video Traffic and Long-Range Dependence", accepted for publication in *IEEE Trans. on Communication,* subject to revisions, 1992.

4. L. M. Berliner, "Statistics, Probability and Chaos", *Statistical Science 7,* 69-90, 1992.

5. D. R. Cox, "Long-Range Dependence: A Review", in: *Statistics: An Appraisal*, H. A. David and H. T. David (Eds.), The Iowa State University Press, Ames, Iowa, 55-74, 1984.

6. R. Dahlhaus, "Efficient Parameter Estimation for Self-Similar Processes", *Ann. Statist. 17,* 1749-1766, 1989.

7. A. Erramilli, R. P. Singh, "Application of Deterministic Chaotic Maps to Model Packet Traffic in Broadband Networks", *Proc. 7th ITC Specialists Seminar*, Morristown, NJ, 8.1.1-8.1.3, 1990.

8. H. J. Fowler, W. E. Leland, "Local Area Network Traffic Characteristics, with Implications for Broadband Network Congestion Management", *IEEE Journal on Selected Areas in Communications 9*, 1139-1149, 1991.

9. R. Fox, M. S. Taqqu, "Large-Sample Properties of Parameter Estimates for Strongly Dependent Stationary Gaussian Time Series", *Ann. Statist. 14*, 517-532, 1986.

10. C. W. J. Granger, R. Joyeux, "An Introduction to Long-Memory Time Series Models and Fractional Differencing", *J. Time Series Anal. 1,* 15-29, 1980.

11. H. Heffes, D. M. Lucantoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance", *IEEE Journal on Selected Areas in Communications 4*, 856-868, 1986.

12. J. R. M. Hosking, "Fractional Differencing", *Biometrika 68*, 165-176, 1981.

13. H. E. Hurst, "Methods of Using Long-Term Storage in Reservoirs", *Proc. of the Institution of Civil Engineers*, Part I, 519-577, 1955.

14. R. Jain, S. A. Routhier, "Packet Trains: Measurements and a New Model for Computer Network Traffic", *IEEE Journal on Selected Areas in Communications 4,* 986-995, 1986.

15. W. E. Leland, D. V. Wilson, "High Time-Resolution Measurement and Analysis of LAN Traffic: Implications for LAN interconnection", *Proceedings of the IEEE INFOCOM'91*, Bal Harbour, FL, 1360-1366, 1991.

16. B. B. Mandelbrot, "Self-Similar Error Clusters in Communication Systems and the Concept of Conditional Stationarity", *IEEE Trans. Communications Technology COM-13*, 71-90, 1965.

17. B. B. Mandelbrot, "Long-Run Linearity, Locally Gaussian Processes, H-Spectra and Infinite Variances", *Intern. Econom. Rev.* 10, 82-113, 1969.

18. B. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, New York, 1983.

19. B. B. Mandelbrot, J. W. Van Ness, "Fractional Brownian Motions, Fractional Noises and Applications", *SIAM Review 10*, 422-437, 1968.

20. B. B. Mandelbrot, J. R. Wallis, "Some Long-Run Properties of Geophysical Records", *Water Resources Research 5*, 321-340, 1969.

21. K. Meier-Hellstern, P. E. Wirth. Y-L Yan, D. A. Hoeflin, "Traffic Models for ISDN Data Users: Office Automation Application", in: *Teletraffic and Datatraffic in a Period of Change* (Proc. 13th ITC, Copenhagen, 1991), A. Jensen, V. B. Iversen (Eds.), North Holland, 167-172, 1991.

22. I. Norros, "Studies on a Model for Connectionless Traffic, Based on Fractional Brownian Motion", *COST24TD(92)041*, 1992.

23. M. S. Taqqu, "A Bibliographical Guide to Self-Similar Processes and Long-Range Dependence", in: *Dependence in Probability and Statistics*, E. Eberlein and M. S. Taqqu (Eds.), Birkhauser, Basel, 137-165, 1985.

24. M. S. Taqqu, J. B. Levy, "Using Renewal Processes to Generate Long-Range Dependence and High Variability", in: *Dependence in Probability and Statistics*, E. Eberlein and M. S. Taqqu (Eds.), Progress in Prob. and Stat. Vol. 11, Birkhauser, Boston, 73-89, 1986.