ANALYSIS OF ETHERNET TRAFFIC

Yalemzewd Negash, G. Devarajan
Department of Electrical Engineering
Addis Ababa University

ABSTRACT

The purpose of this work is to show the self-similarity nature and long-range dependence of Ethernet network traffic. Different mathematical and graphical techniques are used to show this behavior. The result indeed shows the long-range dependence or the presence of long memory in Ethernet data traffic. A graphical proof of the self-similarity nature of the traffic is shown. Also Fractional Auto-Regressive Integrated Moving Average (FARIMA) model is developed to capture the long as well as the short memory properties of the collected Ethernet traffic data. The model is found to be in good agreement with the periodogram calculated from the data. The model could be used in different network application like congestion control in high-bandwidth networks, bandwidth allocation and the like. All the results in this work are supported by a rigorous statistical analysis of the collected data coupled with a discussion of the underlying mathematical and statistical properties of long memory processes.

INTRODUCTION

A common assumption in modeling computer networks is that packet arrivals occur as a Poisson process. However, data communication traffic levels fluctuate over time, and delays through congestion can occur even on lightly utilized links. These fluctuations can occur over very short periods of time giving rise to the concept of a burst of traffic. Bursts of traffic can be of intensity more than five times the average utilization so that if a user is trying to send data and it coincides with a burst the user will experience delays. Traffic that exhibits these wild fluctuations is known as "bursty" traffic [5].

Works done a Bellcore Research Laboratories has shown that network traffic is much more closely modeled by self-similar processes. They suggest that the Poisson process is inadequate as a model of the packet arrival process and that a fractal process is necessary to model the observed results.

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 towards this goal, it is important to know the traffic behavior of some of the expected major contributors to future high-speed network traffic.

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 sub-periods separated by less bursty sub-periods.

Leland and Wilson present a preliminary statistical analysis of the fractal nature of a high-quality data collected from Bellcore Morristown Research and Engineering Center and comment in detail on the presence of "burstiness" across an extremely wide range of time scales. This self-similar or apparently fractal-like behavior of aggregate Ethernet LAN traffic is very different from conventional telephone traffic.

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.

Ethernet is the most widely used local area network (LAN) technology [2]. The original and most popular version of Ethernet supports a data transmission rate of 10 Mb/s. Newer versions of Ethernet. called "Fast Ethernet" and "Gigabit Ethernet" support data rates of 100 Mb/s and 1 Gb/s (1000 Mb/s). An Ethernet LAN may use coaxial cable, special grades of twisted pair wiring, or fiber optic cable. "Bus" and "Star" wiring configurations are supported. Ethernet devices compete for access to the network using a protocol

called Carrier Sense Multiple Access with Collision Detection (CSMA/CD).

The main objectives of this work [1] are:

1. To establish in a statistically rigorous manner the self-similarity characteristic or, to use a more popular notion, the fractal nature of Ethernet traffic.
2. To illustrate some of the differences between self-similar models and the standard models for packet traffic considered in the literature.
3. To develop a model which is capable of capturing the long-range dependence as well as the short memory properties of the Ethernet data traffic.

Accordingly this work is divided into four sections described as follows:

In section two an introduction to Ethernet media access protocol is presented. At the end data collection method for this work is also presented.

Section three will look into the details of the statistics of stationary processes with long memory. Here the relevant theory is presented. Also a detailed discussion of the Fractional Auto Regressive Integrated Moving Average (fractional ARIMA) model is presented, which is of special interest to long memory.

Section four contains a discussion of the different testing methods for long memory. The R/S analysis, the Variance Plot method and the MLE are presented.

The final section contains the summary and conclusion part of this work. Here the results of the work are highlighted.

ETHERNET

In this section the different kinds of Ethernet media access protocol, the Carrier Sense Multiple Access with Collision detection (CSMA/CD) and the access mode for full-duplex Ethernet are discussed. CSMA/CD is what differentiates Ethernet from other LAN technologies. After CSMA/CD, the other mode of operation for Ethernet, which bypasses CSMA/CD, is the Full-duplex Ethernet, which allows a station to receive and transmit simultaneously. It is discussed after CSMA/CD Finally some statistical behaviors, on the collected data are analysed.

Ethernet Media Access Control

This section describes the two media access control protocols defined for Ethernet: "half-duplex", and "full-duplex".

Half-Duplex Ethernet (CSMA/CD Access Protocol)

Half-Duplex Ethernet is the traditional form of Ethernet that uses the CSMA/CD protocol[2]. With CSMA/CD two or more stations share a common transmission medium. To transmit a frame, a station must wait for an idle period on the medium when no other station is transmitting. It then transmits the frame by broadcasting it over the medium such that it is "heard" by all the other stations on the network. If another device tries to send data at the same time, a "**collision**" is said to occur. The transmitting station then intentionally transmits a "**jam sequence**" to ensure all stations are notified the frame transmission failed due to a collision. The station then remains silent for a random period of time before attempting to transmit again. This process is repeated until the frame is eventually transmitted successfully.

Full Duplex Ethernet

The release of the IEEE 802.3x standard defined a second mode of operation for Ethernet, called "full-duplex", that bypasses the CSMA/CD protocol. Full-duplex mode allows two stations to simultaneously exchange data over a point to point link that provides independent transmit and receive paths. Since each station can simultaneously transmit and receive data, the aggregate throughput of the link is effectively doubled. For example, A 10 Mb/s station operating in full-duplex mode provides a maximum bandwidth of 20 Mb/s.

Full-duplex operation is restricted to links meeting the following criteria:

- The physical medium must be capable of supporting simultaneous transmission and reception without interference. Media

specifications which meet this requirement are: 10-Base-T, 10Base-FL, 100Base-TX, 100Base-FX, 100Base-T2, 1000Base-CX, 1000Base-SX, 1000Base-LS, and 1000Base-T. The following media specification cannot support full duplex: 10Base5, 10Base2, 10Base-FP, 10Base-FB, and 100Base-T4.

- Full-duplex operation is restricted to point to point links connecting exactly two stations. Since there is no contention for a shared medium, collisions cannot occur. Frames may be transmitted providing the required separation of the minimum inter-frame gap.
- Both stations on the link must be capable of, and be configured for full-duplex operation.

Full-duplex operation offers several major advantages:

- Throughput is doubled by permitting simultaneous transmit and receive.
- The efficiency of the link is improved by eliminating the potential for collisions.
- Segment lengths are no longer limited by the timing requirements of half-duplex Ethernet that ensure collisions are propagated to all stations within the required 512 bit times. For example, 100Base-FX is limited to 412 meter segment length in half-duplex mode, but may support segment lengths as long as 2 km in full-duplex mode.

The Data from an Ethernet LAN

In this section a close look at the basic statistical properties of the collected data for further investigation is done. Figs. 1 and 2 show us the plot of the data in two time units.

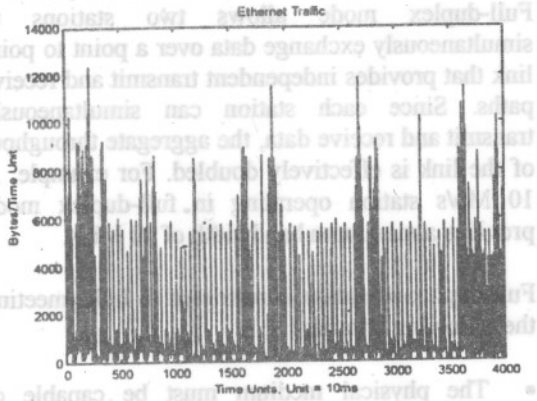

Figure 1 Ethernet Traffic (Bytes per Time Unit) for Time Unit = 10ms

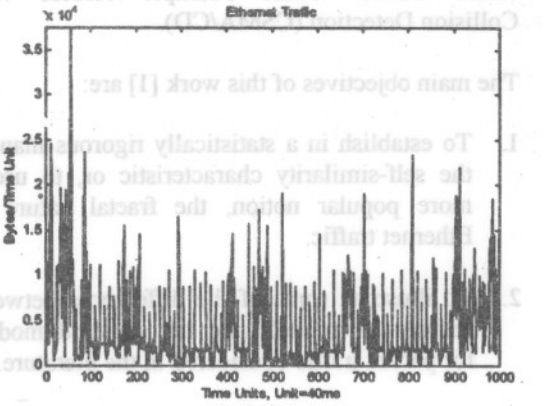

Figure 2 Ethernet Traffic (Bytes per Time Unit) for Time Unit = 40ms

Both figures show bytes count in two different time scales, 10ms, 40ms. The second data, given in Fig. 2 is obtained by increasing the time resolution by 4.

Intuitively the two plots look very similar to one another (in a distribution sense) and are distinctively different from white noise i.e., an independently, identically distributed (iid) sequence of random variables. We also see that there exists a "burst" like traffic in all time scales. From the plots we can see also the absence of a natural length of a "burst": at every time scale bursts consist of bursty sub-periods separated by less bursty sub-periods.

The autocorrelation function of the above data sets are given in Figs. 3 and 4

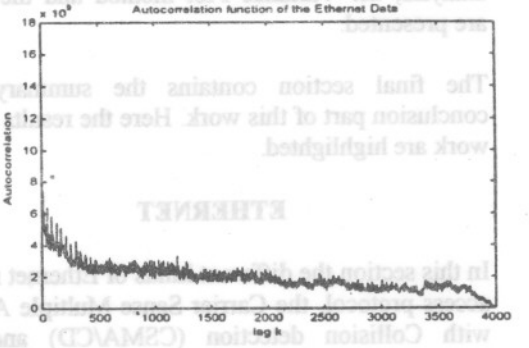

Figure 3 Autocorrelation function of the Ethernet Data (10 ms data)

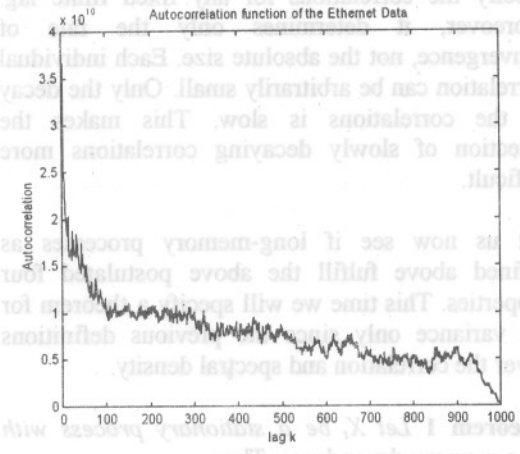

Figure 4 Autocorrelation function of the Ethernet Data (40ms data)

From these two figures the inference drawn are:

- The autocorrelation function look similar, hence we can say that the data sets look similar in the distribution sense.
- The autocorrelation function does not decay exponentially, a slow decay is observed. Hence the data set has Long Memory, or is Long Range Dependence (LRD).
- The sequence is not **iid**.

The long-range dependence nature, scale invariant "bursty" nature of the Ethernet traffic is drastically different from both conventional telephone and packet traffic

The above observation lead to the conclusion that Ethernet traffic is statistically self-similar, that none of the commonly used traffic models is able to capture this fractal behavior. So looking deep into the statistics of self-similar processes is necessary. The next section will focus on the theory of the statistics for self-similar and Long-memory processes.

STATIONARY PROCESSES WITH LONG MEMORY

In this section the statistics of stationary processes with long-memory is discussed. As we have seen in section 2, the statistical data from an Ethernet LAN shows long-range dependence. So in order to model the packet arrivals on an Ethernet LAN, the model should posses long-memory property. This section will give us some insight to the properties of long memory processes and serve as a gateway to the methods of estimation and modeling of long-memory processes. A very detailed analysis of these processes can be found on Jan Beran's "Statistics of Long-Memory Processes," [4]

Let $X_t=(X_t:t=0,1,2,\ldots)$ be a covariance stationary (sometimes called wide-sense stationary) stochastic process; that is, a process with constant mean $\mu = E[X_t]$, finite variance $\sigma^2=E[(X_t - \mu)^2]$. Let's define some of the parameters used in the statistics of X_t:

The sample mean is given by

$$\bar{X} = n^{-1}\sum_{i=1}^{n} X_i \qquad (1)$$

The autocovariance between X_i and X_j is given by

$$\gamma(i,j) = E[(X_i - \mu)(X_j - \mu)] \qquad (2)$$

and the autocorrelation between X_i and X_j is given by

$$\rho(i,j) = \frac{\gamma(i,j)}{\sigma^2} \qquad (3)$$

A spectral density f defined by

$$f(\lambda) = \frac{\sigma^2}{2\pi}\sum_{k=-\infty}^{\infty}\rho(k)e^{jk\lambda} \qquad (4)$$

Processes with long range dependence (or with long memory) have the following common features:

- The variance of the sample mean seems to decay to zero at a slower rate than n^{-1}. In a good approximation, the rate is proportional to $n^{-\alpha}$ for some $0<\alpha<1$.
- The sample correlation decays to zero at a rate that is in good approximation proportional to $k^{-\alpha}$ for some $0<\alpha<1$.
- Near the origin, the logarithm of the periodogram I(λ) plotted against the logarithm of the frequency appears to be randomly scattered around a straight line with negative slope.

We can reformulate the above properties as mathematical conditions on the stationary process:

- The variance of sample mean $var(\bar{X})$ is asymptotically equal to a constant c_{var} times $n^{-\alpha}$ for some $0<\alpha<1$.

- The correlation $\rho(k)$ is asymptotically equal to a constant c_ρ times $k^{-\alpha}$ for some $0<\alpha<1$.
- The spectral density $f(\lambda)$ has a pole at zero that is equal to a constant c_f times $\lambda^{-\beta}$ for some $0<\beta<1$.

A slight generalization of these conditions may be obtained by replacing the proportionality constants c_{var}, c_ρ, c_f by so-called slowly varying functions, i.e., functions such that for any $t \in R$, $L(tx)/L(x) \to 1$ as $x \to \infty$ or as $x \to 0$, respectively. Through out this work, the above conditions are used. Thus, the following definition is used for a stationary process with long memory or long-range dependence:

Definition 1 *Let X_t be a stationary process for which the following holds: There exists a real number $\alpha \in (0,1)$ and a constant c_ρ such that*

$$\lim_{k\to\infty} \rho(k)/[c_\rho k^{-\alpha}] = 1 \qquad (5)$$

Then X_t is called a stationary process with long memory or long-range dependence or strong dependence, or a stationary process with slowly decaying or long-range correlations.

For reasons, discussed later, the parameter $H = 1 - \alpha / 2$ will also be used instead of α. In terms of this parameter, long memory occurs for $\frac{1}{2} < H < 1$. Knowing the covariance (or correlations and variances) is equivalent to knowing the spectral density f. Therefore, long-range dependence can also be defined by imposing a condition on the spectral density.

Definition 2 *Let X_t be a stationary process for which the following holds: There exists a real number $\beta \in (0, 1)$ and a constant c_f such that*

$$\lim_{\lambda\to 0} f(\lambda)/[c_f |\lambda^{-\beta}|] = 1 \qquad (6)$$

Then X_t is called a stationary process with long memory or long-range dependence or strong dependence.

It is important to note that the definition of long-range dependence by 1 [or 2] is an asymptotic definition. It only tells us something about the ultimate behavior of the correlations as the lag tends to infinity. In this generality, it does not specify the correlations for any fixed finite lag. Moreover, it determines only the rate of convergence, not the absolute size. Each individual correlation can be arbitrarily small. Only the decay of the correlations is slow. This makes the detection of slowly decaying correlations more difficult.

Let us now see if long-memory processes as defined above fulfill the above postulated four properties. This time we will specify a theorem for the variance only since the previous definitions cover the correlation and spectral density.

Theorem 1 *Let X_t be a stationary process with long-memory dependence. Then*

$$\lim_{n\to\infty} \mathrm{var}(\sum_{i=1}^{n} X_i)/[c_\gamma n^{2H}] = \frac{1}{H(2H-1)} \qquad (7)$$

Self-similar processes

Self-similar processes were introduced by Kolmogorov in 1941 in a theoretical context. Statisticians do not seem to have been aware of the existence or statistical relevance of such processes, until Mandelbrot and his co-workers introduced them into statistics.

The basic idea of self-similarity is much older. Mandelbrot refer, in his famous book "The Fractal Geometry of Nature", for example to Leonardo da Vinci's drawings of turbulent flows that exhibit coexistent "eddies" of all sizes and thus self-similarity. A geometric shape is called self-similar in a deterministic way if the same geometric structures are observed [6]. In the context of stochastic processes, self-similarity is defined in terms of the distribution of the process.

Definition 3 *Let Y_t be a stochastic process with continuous time parameter t. Y_t is called self-similar with self-similarity parameter H, if for any positive stretching factor c, the rescaled process with time scale ct, $c^{-H}Y_{ct}$ is equal in distribution to the original process Y_t.*

This means that, for any sequence of time points $t_1, \ldots, t_k$, and any positive constant c, $c^{-H}(Y_{ct_1}, \ldots, Y_{ct_k})$ has the same distribution as $(Y_{t_1}, \ldots Y_{t_k})$. Thus, typical sample paths of a self-similar process look qualitatively the same,

irrespective of the distance from which we look at them. In contrast to deterministic self-similarity, it does not mean that the same picture repeats itself exactly as we go closer. It is rather the general impression that remains the same.

Stationary increments of self-similar processes

Let Y_t is a self-similar process with self-similarity parameter H. The property

$$Y_t =_d t^H Y_1 \quad (\text{for } t > 0), \qquad (8)$$

Where $=_d$ is equality in distribution, implies the following limiting behavior of Y_t as t tends to infinity

1 If $H<0$, then $Y_t \to_d 0$ (where $\to_d$ is convergence in distribution).
2 If $H = 0$, then $Y_t =_d Y_1$.
3 If $H>0$ and $Y_t \neq 0$, then $|Y_t| \to_d \infty$

Analogously, for t converging to zero, we have

1 If $H<0$ and $Y_t \neq 0$, then $|Y_t| \to_d \infty$
2 If H=0, then $Y_t =_d Y_1$
3 If H>0, then $Y_t \to_d 0$

If we exclude the trivial case $Y_t \equiv 0$, then these properties imply that Y_t is not stationary unless $H = 0$. The exception $H = 0$ is not interesting, as it implies that for all $t>0$, Y_t is equal to Y_1 with probability 1. For the purpose of modeling data that look stationary, we need only to consider self-similar processes with stationary increments. The range of H can be restricted to $H>0$. The reason is that if the increments of a self-similar process are stationary, then the process is mathematically pathological for negative values of H. More specifically, for $H < 0$, Y_t is not a measurable process. The only exception is the trivial case where $Y_t = Y_1 = 0$ with probability 1. So in the following, we consider positive values of H only, in particular, $Y_0 = 0$ with probability 1.

The form of the covariance function $\gamma_y(t,s) = \text{cov}(Y_t, Y_s)$ of a self-similar process Y_t with stationary increments follows from these two properties. To simplify notation, assume $E(Y_t) = 0$. Let $s<t$ and denote by $\sigma^2 = E[(Y_t - Y_{t-1})^2] = E(Y_1^2)$ the variance of the increment process $X_t = Y_t - Y_{t-1}$. Then

$$E[(Y_t - Y_s)^2] = E[(Y_{t-s} - Y_0)^2] = \sigma^2 (t-s)^{2H}$$

On the other hand,

$$E[(Y_t - Y_s)^2] = E[Y_t^2] + E[Y_s^2] - 2E[Y_t Y_s]$$
$$= \sigma^2 t^{2H} + \sigma^2 s^{2H} - 2\gamma_y(t,s)$$

Hence,

$$\gamma_y(t,s) = \frac{1}{2}\sigma^2 [t^{2H} - (t-s)^{2H} + s^{2H}]. \quad (9)$$

Similarly, the covariances of the increment sequence $X_I = Y_I - Y_{I-1}$ $(I=1,2,3, \ldots)$ are obtained. The covariance between X_I *and* X_{I+k} $(k>0)$ is equal to

$$\gamma(k) = \text{cov}(X_i, X_{i+k}) = \text{cov}(X_1, X_{k+1})$$
$$= \frac{1}{2} E[(\sum_{j=1}^{k+1} X_j)^2 + (\sum_{j=2}^{k} X_j)^2 - (\sum_{j=1}^{k} X_j)^2$$
$$- (\sum_{j=2}^{k+1} X_j)^2]$$
$$= \frac{1}{2}\{E[(Y_{k+1} - Y_0)^2] + E[(Y_{k-1} - Y_0)^2] -$$
$$E[(Y_k - Y_0)^2] - E[(Y_k - Y_0)^2]\}$$

Using self-similarity, we obtain the formula

$$\gamma_y(k) = \frac{1}{2}\sigma^2 [(k+1)^{2H} - 2(k)^{2H} + (k-1)^{2H}]. \quad (10)$$

for $k \geq 0$ and $\gamma(k) = \gamma(-k)$ for $k<0$. The correlations are given by

$$\rho(k) = \frac{1}{2}[(k+1)^{2H} - 2(k)^{2H} + (k-1)^{2H}]. \quad (11)$$

for $k \geq 0$ and $\rho(k) = \rho(-k)$ for $k<0$.

The asymptotic behavior of $\rho(k)$ follows from Taylor expansion: First note that

$$\rho(k) = \frac{1}{2} k^{2H} g(k^{-1})$$

where $g(x) = (1+x)^{2H} - 2 + (1-x)^{2H}$

If $0<H<1$ and $H\neq 1/2$, then the first non-zero term in the Taylor expansion of $g(x)$, expanded at the origin, is equal to $2H(2H-1)x^2$. Therefore, as k tends to infinity, $\rho(k)$ is equivalent to $H(2H-1)k^{2H-2}$, i.e.,

$$\rho(k)/[H(2H-1)k^{2H-2}] \to 1 \text{ as } k \to \infty$$

For $1/2 < H < 1$, this means that the correlations decay to zero so slowly that

$$\sum_{k=-\infty}^{\infty} \rho(k) = \infty \qquad (12)$$

The process $X_i (i = 1,2,...)$ has long memory. For $H=1/2$, all correlations at non-zero lags are zero, i.e., the observations X_i are uncorrelated. For $0<H<1/2$, the correlations are summable. In fact a more specific equation holds, namely,

$$\sum_{k=-\infty}^{\infty} \rho(k) = 0 \qquad (13)$$

We conclude this section by nothing an appealing property of stationary increments of self-similar processes: The sample mean can be written as

$$\bar{X} = n^{-1}\sum_{i=1}^{n} X_i = n^{-1}(Y_n - Y_0) =_d n^{-1}n^H(Y_1 - Y_0).$$

Therefore, instead of the asymptotic Eq. (7), we obtain for each sample size the exact equality

$$\text{var}(\bar{X}) = \sigma^2 n^{2H-2}.$$

For H=1/2, this is the classic result $\text{var}(\bar{X}) = \sigma^2 n^{-1}$. Moreover, if X_i is a Gaussian process with mean μ and variance σ^2, then $n^{1-H}(\bar{X} - \mu)/\sigma$ is a standard normal random variable. This can be used to calculate tests and confidence intervals for μ.

In the next section our attention will focused on a model which can be used to model long-memory processes.

Fractional ARIMA Models

ARIMA (Auto-Regressive Integrated Moving Average) models were introduced by Box and Jenkins. Because of their simplicity and flexibility, it became very popular in applied time series analysis. The theory of statistical inference for these processes is well developed. Fractional ARIMA models are a natural extension of the classic ARIMA models.

Recal the definition of ARMA and ARIMA processes. To simplify notations, $\mu = E(X_t) = 0$ is assumed. Let B be the backshift operator defined by $BX_t = X_{t-1}, B^2 X_t = X_{t-2}, \ldots$. In particular, differences can be expressed in terms of the backshift operator as $X_t - X_{t-1} = (1-B)X_t \ldots$ Let p and q be integers. Defining the polynomials

$$\phi(x) = 1 - \sum_{j=1}^{p} \phi_j x^j$$

and

$$\varphi(x) = 1 + \sum_{j=1}^{q} \varphi_j x^j$$

Assume that all solutions of $\phi(x) = 0$ and $\varphi(x) = 0$ are outside the unit circle. Furthermore, let $\varepsilon_t (t = 1,2,...)$ be iid normal variables with zero expectation and variance σ_ε^2. An ARMA(p,q) model is defined to be the stationary solution of

$$\phi(B)X_t = \varphi(B)\varepsilon_t \qquad (14)$$

If instead Eq. (14) holds for the *dth* difference $(1-B)^d X_t$, then X_t is called an ARIMA(p,d,q) process. The corresponding equation is

$$\phi(B)(1-B)^d X_t = \varphi(B)\varepsilon_t. \qquad (15)$$

Note that an ARMA(p,q) process is also an ARIMA(p,0,q) process. If d is larger that or equal to 1, then the original series X_t is not stationary. To obtain a stationary process, X_t must be differenced d times.

Equation (15) can be extended to non-integer values of d in the following way:

Definition 4 *Let X_t be a stationary process such that*

$$\phi(B)(1-B)^d X_t = \varphi(B)\varepsilon_t$$

for some $-1/2 < d < 1/2$. Then X_t is called a fractional ARIMA(p,d,q) process.

The range that is interesting in the context of long-memory processes is $0 \le d < \frac{1}{2}$. The upper bound $d<½$ is needed, because for $d \ge ½$, the process is not stationary, at least not in the usual sense. In particular, the usual definition of the spectral density of X_t would lead to a non-integrable function. For the range $½ \le d \le 1$, one can still define "spectral density," by using a more general definition. Also note that, the case $d > ½$ can be reduced to the case $-½ < d \le ½$ by taking appropriate differences.

Eq. (11) can be interpreted in several ways. For instance, it can be written as

$$(1-B)^d X_t = \bar{X}_t, \qquad (16)$$

Where $\bar{X}_t$ is an ARMA process defined by

$$\bar{X}_t = \phi^{-1}(B)\varphi(B)\varepsilon_t, \qquad (17)$$

This means that, after passing X_t through the fractional difference operator (or finite linear filter) $(1-B)^d$, we obtain a ARMA process. On the other hand, we can write

$$X_t = \phi(B)^{-1}\varphi(B)X_t^*, \qquad (18)$$

Where X_t^* is a fractional ARIMA(0,d,0) process defined by

$$X_t^* = (1-B)^{-d}\varepsilon_t \qquad (19)$$

That is X_t is obtained by passing a fractional ARIMA(0,d,0) process through an ARMA filter. The parameter d determines the long-term behavior, whereas p, q, and the corresponding parameters in $\phi(B)$ and $\varphi(B)$ allow for more flexible modeling of short-range properties.

The spectral density of a fractional ARIMA process follows directly from Eq. (15). That is,

$$f_{ARMA}(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\left|\frac{\varphi(e^{j\lambda})}{\phi(e^{j\lambda})}\right|^2$$

the spectral density of the ARMA process $\bar{X}_t$. Recall that if X_t is obtained from a process Y_t with spectral density f_Y by applying the linear filter $\sum a(s)Y_{t-s}$, then the spectral density of X_t is equal to $|A(\lambda)|f_Y(\lambda)$, where $A(\lambda) = \sum a(s)e^{js\lambda}$.

From (3.11) we then obtain the spectral density of X_t:

$$f(\lambda) = \left|1-e^{j\lambda}\right|^{-2d} f_{ARMA}(\lambda) \qquad (20)$$

Note that $\left|1-e^{j\lambda}\right| = 2\sin\frac{\lambda}{2}$. Because $\lim_{\lambda\to 0}\lambda^{-1}(\sin\frac{\lambda}{2}) = 1$, the behavior of the spectral density at the origin is given by

$$f(\lambda) \approx \frac{\sigma_\varepsilon^2}{2\pi}\frac{|\varphi(1)|^2}{|\phi(1)|^2}|\lambda|^{-2d} = f_{ARMA}(0)|\lambda|^{-2d} \qquad (21)$$

Thus, for $d > 0$, the spectral density has a pole at zero. Comparing this with our notation in the previous sections we see that

$$d = H - \tfrac{1}{2}. \qquad (22)$$

For $d = 0$, X_t is an ordinary ARMA(p,q) process with bounded spectral density. Long-range dependence occurs for

$$0 < d < \tfrac{1}{2}. \qquad (23)$$

In order to transform X_t into a process with a bounded spectral density, on has to apply the linear filter $(1-B)^d$. For $-½ < d < 0$, $f(0)=0$ so that the sum of all correlations is zero, and this case is of less practical importance.

ESTIMATION OF LONG AND SHORT MEMORY

The phenomenon of long memory was observed in applications long before appropriate stochastic models were known [6]. Several heuristic methods to estimate the long-memory parameter H were suggested. Best known is the *R/S* statistics, which was first proposed by Hurst in a hydrological context. Other methods include the log-log correlogram, the log-log plot of $\mathrm{var}(\bar{X}_n)$ versus n, the semi-variogram, and least squares regression in the spectral domain. These methods are mainly useful as simple diagnostic tools. Short memory processes are discussed and used extensively in many statistical applications. The case where long

as well as short memory dependence exists is not a well developed parts of applied statistics.

The *R/S* Statistics

The famous hydrologist Hurst noticed some long memory characteristics when he was investigating the question of how to regularize the flow of Nile River. More specifically, his discovery can be described as follows: Suppose we want to calculate the capacity of a reservoir such that it is ideal for the time span between t and $t + k$. To simplify matters, assume that time is discrete and that there are no storage losses (caused by evaporation, leakage, etc.)[6]. By ideal capacity we mean that we want to achieve the following: that the outflow is uniform, that at time $t + k$ the reservoir is as full as time t, and that the reservoir never overflows. Let X_i denote the inflow at time i and $Y_j = \sum_{i=1}^{j} X_i$ be the cumulative inflow up to time j. Then the ideal capacity can be shown to be equal to

$$R(t,k) = \max_{0\le i\le k}[Y_{t+i} - Y_t - \tfrac{i}{k}(Y_{t+k} - Y_t)] - \min_{0\le i\le k}[Y_{t+i} - Y_t - \tfrac{i}{k}(Y_{t+k} - Y_t)] \quad (24)$$

$R(t,k)$ is called the adjusted range. In order to study the properties that are independent of the scale, $R(t,k)$ is standardized by

$$S(t,k) = \sqrt{k^{-1}\sum_{i=t+1}^{t+k}(X_i - \bar{X}_{t,k})^2} \quad (25)$$

where $\bar{X}_{t,k} = k^{-1}\sum_{i=t+1}^{t+k} X_i$. Note that $S^2(t,k)$ is equal to $(k-1)/k$ times the usual sample variance of $X_{t+1}, \ldots, X_{t+k}$. The ratio

$$\frac{R}{S} = \frac{R(t,k)}{S(t,k)} \quad (26)$$

is called the rescaled adjusted range or *R/S*-statistic. Hurst plotted the logarithm of *R/S* against several values of k. He observed that, for large values of k, *log R/S* was scattered around a straight line with slope that exceeded ½. In probabilistic terminology this means that for large k,

$$\log E[R/S] \approx a + H\log k, \text{ with } H > \tfrac{1}{2}. \quad (27)$$

This empirical finding was in contradiction to results for Markov processes, mixing processes and other stochastic processes that where usually considered at that time. For any stationary process with short-range dependence, *R/S* should behave asymptotically like a constant times $k^{\frac{1}{2}}$. Therefore, for large values of k, *log R/S* should be randomly scattered around a straight line with slope ½. Hurst's finding that for the Nile River data, and for many other hydrological, geophysical, and climatological records, *R/S* behaves like a constant times k^H for some $H > ½$, is known under that name *Hurst effect*.

Let $Q = Q(t,k) = R(t,k)/S(t,k)$ be the *R/S* statistic defined earlier. To estimate the long-memory parameter, the logarithm of Q is plotted against k. For each k, there are n-k replicates. The *R/S* method can be summarized as follows [4]:

1. Calculate Q for all possible (or for a sufficient number of different) values of t and k.
2. Plot *log Q* against *log k*.
3. Draw a straight line $y = a + b\log k$ that corresponds to the "ultimate" behavior of the data. The coefficients a and b can be estimated, for instance, by least squares or "by eye". Set $\hat{H} = \hat{b}$.

The following difficulties arise: How do we decide from which k on the "ultimate behavior" starts? How uncertain is the estimate of H? In particular, for finite samples, the distribution of Q is neither normal nor symmetric. This makes estimation by eye more difficult. Also, it raises the question of whether least squares regression is appropriate. The exact distribution of Q seems to be difficult to derive and depends on the actual distribution of the data generating process. The values of Q for different time points t and lags k are not independent from each other. The exact description of the dependence would be very complicated and possibly model-dependent. Finally, for large lags k, only very few values of Q can be calculated. Because of these problems, it seems difficult to define a fully "automatic" R/S methodology, and to derive results on statistical inference based on the method.

Variance Plot

As it was noticed in the previous section, one of the striking properties of long-memory processes is that the variance of the sample mean converges slower to zero than n^{-1}. From Theorem 3.1 we have

$$\text{var}(\bar{X}_n) \approx cn^{2H-2}, \quad (28)$$

where $c>0$. This suggests the following method for estimating H:

1. Let k be an integer. For different integers k in the range $2 \le k \le n/2$, and a sufficient number (say m_k) of sub-series of length k, calculate the sample means $\bar{X}_1(k),\ldots,\bar{X}_{m_k}(k)$ and the overall mean
$$\bar{X}(k) = m_k^{-1}\sum_{j=1}^{m_k}\bar{X}_j(k). \quad (29)$$
2. For each k, calculate the sample variance of the sample means $\bar{X}_j(k)$ $(j = 1,\ldots, m_k)$:
$$s^2(k) = (m_k - 1)^{-1}\sum_{k=1}^{m_k}(\bar{X}_j(k) - \bar{X}_j(k))^2$$
3. Plot $\log s^2(k)$ against $\log k$.

For large values of k, the points in this plot are expected to be scattered around a straight line with negative slope $2H-2$. In the case of short-range dependence or independence, the ultimate slope is 2(½) - 2 = -1. Thus, the slope is steeper (more negative) for short-memory processes. The problems in this method are in principle the same as for the *R/S* plot.

Least squares regression in the spectral domain

Least squares regression in the spectral domain exploits the simple form of the pole of the spectral density at the origin:

$$f(\lambda) \approx c_f|\lambda|^{1-2H} \quad (|\lambda| \to 0) \quad (30)$$

Equation (3.7) can be written as

$$\log f(\lambda) \approx \log c_f + (1-2H)\log|\lambda|. \quad (31)$$

Recall that, for fixed frequency $\lambda \neq 0$, the periodogram $I(\lambda)$ is an asymptotically unbiased estimate of f, i.e., we have

$$\lim_{n\to\infty} E[I(\lambda)] = f(\lambda). \quad (32)$$

Usually, $I(\lambda)$ is calculated the Fourier frequencies

$$\lambda_{k,n} = \frac{2\pi k}{n}, k = 1,\ldots, n^* \quad (33)$$

where n is the integer part of $(n-1)/2$. For short-memory processes, it is well know that for a finite number of frequencies $\lambda_1,\ldots,\lambda_k \in (0,\pi)$, the corresponding periodiogram ordinates $I(\lambda_1),\ldots,I(\lambda_k)$ are approximately independent exponential random variables with means $f(\lambda_1), \ldots, f(\lambda_k)$. For long-memory processes, this result can be stated as:

$$[I(\lambda_1),..I(\lambda_k)] \to_d [f(\lambda_1)\zeta_1,\ldots,f(\lambda_k)\zeta_k] \quad (34)$$

where $\zeta_1,\ldots,\zeta_k$ are independent standard exponential random variables. This together with (4.8), leads to the approximate equation

$$\log I(\lambda_{k,n} \approx \log c_f + (1-2H)\log\lambda_{k,n} + \log\zeta_k \quad (35)$$

where ζ_k are independent standard exponential random variables. Not that

$$E(\log\zeta_k) = -C = -0.577215\ \ldots,$$

where C is the Euler constant. Define

$$y_k = \log I(\lambda_{k,n}),$$
$$x_k = \log\lambda_{k,n},$$
$$\beta_0 = \log c_f - C, \beta_1 = 1-2H,$$

and the "error" term

$$e_k = \log\zeta_k + C$$

Then (3.12) can be written as

$$y_k = \beta_0 + \beta_1 x_k + e_k \quad (36)$$

This is a regression equation with independent identically distributed errors e_k with zero mean. The coefficients β_0 and β_1 may be estimated, for instance, by least squares regression. The estimate of H is then set equal to

$$\hat{H} = \frac{1 - \hat{\beta}_1}{2} \quad (37)$$

Several problems arise with this approximate method:

1. The notion of long memory is an asymptotic one. Often the spectral density might be proportion to λ^{1-2H} in a small neighborhood of zero only. By wrongly assuming that this proportionality is correct in the whole interval $[-\pi, \pi]$, the estimate of H can be highly biased.

2. If $\hat{\beta}_0$ and $\hat{\beta}_1$ are based on all Fourier frequencies, then this does not matter asymptotically. It might, however, have an influence on the finite sample properties of $\hat{H}$, or it matters if only a small number of the smallest frequencies are used.

3. The distribution of e_k is highly skewed. A least squares estimator will therefore be inefficient compared to an estimator that uses this property.

Problem 1 can be solved, for example by estimating the least squares line [7] from the periodogram ordinates at low frequencies only. Clearly, this can be done only at the cost of lower precision. Also, because only small frequencies are considered, problem 2 need to be taken more seriously. The third problem cannot be solved without abandoning the ordinary least squares method. For parametric models, efficient maximum likelihood type methods can be obtained by applying weighted least squares with appropriate weights.

An estimation of $\hat{\beta}_0$ and $\hat{\beta}_1$ based on all Fourier frequencies is of little practical importance, so we do not discuss it further here. The least squares method becomes attractive when the focus is on estimating the pole (i.e., H *and* c_f) only by considering a certain number of the smallest Fourier frequencies. The advantage of this method is that it is easier to derive the asymptotic distribution. In contrast to maximum likelihood estimation, almost no model assumption is necessary.

Maximum Likelihood Estimator

In this section a maximum likelihood estimator called Whittle estimator is discussed. A Fraction Auto-Regressive Integrated Moving Average (FARIMA) model to fit the periodiogram of the collected data is used. The FARIMA model will have the following parameters, (0,d,0).

Consider a series X_t of length n, which obeys the following rule:

$$\phi(B)(1-B)^d (X_t - \mu) = \varphi(B)\varepsilon_t \quad (38)$$

where ε_t are Gaussian white noise with variance σ_ε^2. The roots of the polynomials $\phi(z)$ and $\varphi(z)$ are assumed to lie outside the unit circle. Then the process X_t is stationary for $d < ½$ and invertible for $d > -1$. The spectrum

$$f(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\left|\frac{\varphi(e^{j\lambda})}{\phi(e^{j\lambda})}\right|^2 \left|1 - e^{j\lambda}\right|^{-2d} \quad (39)$$

is infinite at the origin for $d > 0$ and zero for $d < 0$. Long memory is associated with $d > 0$. If $d < 0$, the process is said to have intermediate memory.

The Whittle likelihood, WL

Let the Whittle likelihood be defined as [8]

$$\log L_W(\theta, \sigma_\varepsilon^2) = \quad (40)$$
$$-\sum_{j=1}^{m} \log f(\lambda_j \mid \theta, \sigma_\varepsilon^2) - \tfrac{1}{2}\sum_{j=1}^{m} \frac{I(\lambda_j)}{f(\lambda_j \mid \theta, \sigma_\varepsilon^2)}$$

Where L_W is the whittle version of the likelihood function. And $I(\lambda_j)$ denotes the periodogram at the j-*th* Fourier frequency, $\lambda_j (\lambda_j = 2\pi \frac{j}{n}, j = 1, \ldots, m$

$$I(\lambda_j) = \frac{1}{n}\left|\sum_{t=1}^{n} X_t e^{-jt\lambda}\right|^2 \quad (41)$$

m is the largest integer contained in (n-1)/2. It is discrete time version of the Whittle function. In the ARMA case it may be interpreted e.g. as the likelihood associated with the asymptotic distribution of the periodogram. On the other hand, if the term $\sum_{j=1}^{m} \log f(\lambda_j \mid \theta, \sigma_\varepsilon^2)$ is dropped the asymptotic properties remain the same, and it becomes the Yule-Walker estimator for $AR(p)$ process. The reduced form of L_W with respect to the error variance σ_ε^2 is

$$\log L_W^*(\theta) = m \log(2\pi) - m \log\left[\frac{1}{m}\sum_{j=1}^{m} \frac{I(\lambda_j)}{g(\lambda_j)}\right] - \sum_{j=1}^{m} \log g(\lambda_j) - m \tag{42}$$

$$with\ \sigma_\varepsilon^2 = \sigma_\varepsilon^{2*} = \frac{1}{m}\sum_{j=1}^{m} \frac{I(\lambda_j)}{g(\lambda_j)}$$

$$where\ f(\lambda) = \sigma_\varepsilon^2 g(\lambda)/2\pi$$

with $g(\lambda) = g(\lambda \mid \theta)$

This estimator is denoted by WL.

Minimizing the above equation amounts to minimizing the sum of the ratios $\frac{I(\lambda_j)}{f(\lambda_j \mid \theta, \sigma_\varepsilon^2)}$ with respect to θ. We follow the following to steps:

1. Minimize $Q(\theta) = \sum_{j=1}^{m} \frac{I(\lambda_j)}{f(\lambda_j \mid \theta, \sigma_\varepsilon^2)}$ with respect to θ.
2. Set $\sigma_\varepsilon^{2*} = \frac{4\pi}{n} Q(\theta)$

Constructing a FARIMA model

The applicability of FARIMA models is dependent on the ease with which they may be fitted to an observed time series. Selection and estimation for FARIMA model is more difficult than that of standard ARMA models [3,4]. When both long- and short-range correlation structures are present in the data, their behavior is hard to distinguish. Additionally, likelihood estimation techniques for FARIMA model present computational problems and often result in significant finite sample biases. However, methods for constructing ARMA models are now well established. FARIMA model is the extended version of ARMA model in nature, so it will be helpful to solve model-building problems of FARIMA by applying the effective methods for ARMA processes.

Transfer the FARIMA problem to the ARMA problem by splitting FARIMA model into its "fractional differenced" and "ARMA" part. For a FARIMA process X_t,

$$\phi(B)\Delta^d X_t = \varphi(B)\varepsilon_t \tag{43}$$

Then obtain

$$W_t = \Delta^d X_t \tag{44}$$

where $W_t = [\phi(B)/\varphi(B)]\varepsilon_t$

So, if the differencing parameter d can be obtained in advanced and the fractional differencing operator Δ^d can be implemented in practice, W_t, the fractionally differenced X_t, can be evaluated as an ARMA model.

The following is the procedure used to fit a FARIMA model to the data.

Step 1. Estimating fractional differencing parameter d

Hurst parameter H can be estimated by several methods such as Variance-time analysis, R/S analysis and periodogram-based analysis. The strength of long-range dependence measured by d is the same as that by H upon $H = d + 0.5$. Then we can estimate parameter d from this relationship.

Step 2. Fractional differencing on X_t

The calculation of W_t from X_t involves a finite approximation to the infinite sum in the definition of fractional differencing operator. The exact fractional differencing operator for the X_t series with mean value μ is

$$\Delta^d (X_t - \mu) = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k (X_t - \mu) = \sum_{j=0}^{\infty} \varpi_j (X_{t-j} - \mu) \tag{45}$$

where

$$\varpi_j = \frac{(-1)^j \Gamma(1+d)}{\Gamma(1+j)\Gamma(1+d-j)}$$

This involves the unobserved quantities $X_0, X_{-1}, \ldots$. We have assumed causality here.

After using the operator Δ^d with estimated parameter d on X_t, we can obtain W_t.

Step 3. Model identification and parameter estimation for the FARIMA model.

Now W_t can be analyzed as an ARMA(p,q) model using conventional method. We select ARMA(1,1). Once the order p and q are selected, we can estimate all the parameters of the desired FARIMA model by Durbin recursion method[3].

SUMMARY AND CONCLUSION

In this research work, Ethernet traffic from ECA's External network, which is connected using a shared media (hub), is analyzed. Graphical as well as mathematical methods were used to show its behavior. The results from each of the anaysis methods are presented.

SUMMARY

In this section the results are summarized in detail and the observation are represented graphically.

R/S analysis

The result obtained after analyzing the collected data using the Rescaled Range analysis shows the self-similar behavior of the Ethernet traffic. The value of H (Hurst parameter) obtained was ***0.8087***.

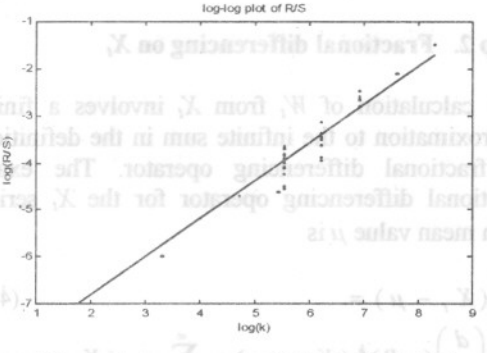

Figure 5 log-log plot of the R/S statistics

Figure 5 depicts the log-log plot of R/S. It shows an asymptotic slope that is distinctly different from 0.5 and is easily estimating using least squares to be 0.8087.

Variance Plot

The result obtained using the Variance Plot method proves that the data has long memory or it is long-range dependence. The estimated Hurst parameter equals ***0.7811***.

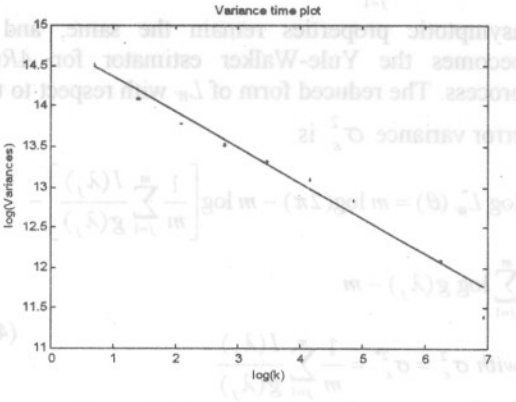

Figure 6 Variance-time plot

The variance-time curve, given in Fig. 6, shows an asymptotic slope that is clearly different from –1 and easily estimated to be –0.4378 (using least squares), resulting in a practically identical estimate of the Hurst parameter H of about 0.7811.

Least Squares Regression in the Spectral Domain

An important practical problem that remains unsolved with this method is how to choose the lower limit (l) and the upper limit (m) for finite samples. Depending on the choice l and m, results can differ considerably. Increasing m reduces the variance of H but increases the bias. On the other hand, reducing m increases the variance but reduces the bias. The estimated H for four combinations of l and m is shown in the Table 1.

Table 1: Results from the Least squares regression in the spectral domain

l	m	H
1	360	0.8202
1	400	0.7851
10	360	0.7947
10	400	0.7504

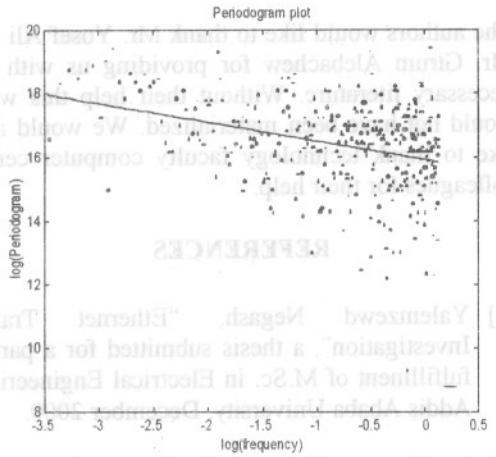

Figure 7 Estimation of H by least squares regression, based on the Fourier frequencies $\lambda_l, \ldots, \lambda_{l+m}$.

Note that the value of H increases the more one concentrates on low frequencies only. This might be an indication for strong long-range dependence.

Looking at the periodogram plot, Fig. 7, corresponding to the time series, it may be observed that although there are some pronounced peaks in the high frequency domain of the periodogram, the low-frequency part is characteristic for a power-low 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 of -0.64 which results in a Hurst parameter estimate of about 0.8202.

Whittle Estimation

In this analysis a fractional *ARIMA(0,d,0)* model is fitted to the collected Ethernet data. The estimate of *H* was equal to 0.80. The plot of the periodogram and the fitted spectral density, shown in Fig. 8, show a good agreement between data and both fitted models. The fractional ARIMA(0,d,0) process has only long memory property. So it might not have a good approximation for the short memory part which is observed at large frequencies.

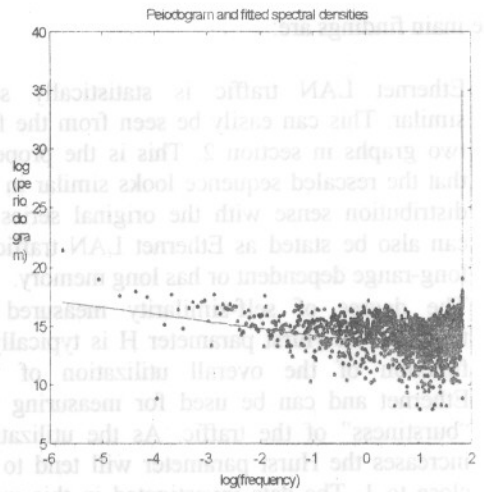

Figure 8 Periodogram and fitted Fractional ARIMA (0,0.297,0) spectral densities.

Fitting FARIMA(p,d,q) Model.

The result for the fitted FARIMA(3,0.297,3) model is plotted along with the periodogram in Fig. 9.

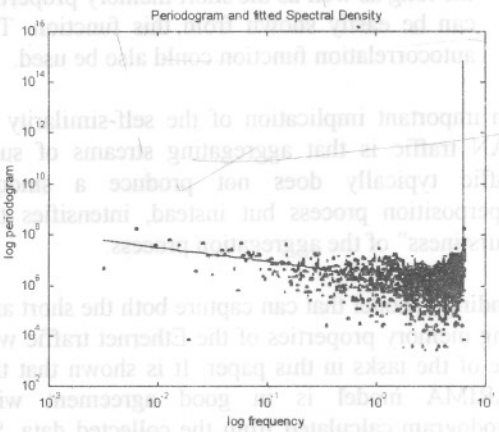

Figure 9 Periodogram and fitted Fractional ARIMA (3,0.297,3) spectral densities.

A fractional ARIMA(3,0.297,3) process is a process with 3 poles and 3 zeros and fractionally differenced with a differencing parameter 0.297. It may be observed that the model has fitted both the long-and short-ranges with the periodogram appropriately. So we can say that FARIMA process is capable of simultaneous modeling long-range and short-range behavior of network traffic.

CONCLUSION

The main findings are:

- Ethernet LAN traffic is statistically self-similar. This can easily be seen from the first two graphs in section 2. This is the property that the rescaled sequence looks similar in the distribution sense with the original series. It can also be stated as Ethernet LAN traffic is long-range dependent or has long memory.
- 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. As the utilization increases the Hurst parameter will tend to be close to 1. The data investigated in this work has a Hurst parameter of around 0.8.
- Traffic characteristics using a FARIMA model has been investigated. This model has the capability of capturing the long as well as short memory properties of the traffic pattern. And as it is shown the spectral density of the FARIMA model considered in this work fits nicely with the periodogram calculated from the data. Periodogram fitting is used because the long as well as the short memory properties can be easily shown from this function. The autocorrelation function could also be used.

An important implication of the self-similarity of LAN traffic is that aggregating streams of such traffic typically does not produce a smooth superposition process but instead, intensifies the "burstiness" of the aggregation process.

Finding a model that can capture both the short and long memory properties of the Ethernet traffic was one of the tasks in this paper. It is shown that the FARIMA model is in good agreement with periodogram calculated from the collected data. So we can use a FARIMA model, for example, in traffic prediction problems.

ACKNOWLEGDEMENT

The authors would like to thank Mr. Yosef Ali and Mr. Girum Alebachew for providing us with the necessary literature. Without their help this work would not have been materialized. We would also like to thank technology faculty computer center colleagues for their help.

REFERENCES

[1] Yalemzewd Negash, "Ethernet Traffic Investigation", a thesis submitted for a partial fulfillment of M.Sc. in Electrical Engineering, Addis Ababa University, December 2000.

[2] TechFest – "Ethernet Technical Summary", 1999.

[3] Monson H. Hayes, "Statistical Digital Signal Processing and Modelling," John Wiley & Sons, INC., 1996.

[4] J. Beran, "Statistical Methods for Long-Memory Processes", Monographs on Statistics and Applied Probability 61, CHAPMAN & HALL/CRC, 1994.

[5] W.Leland, et al., "On the self-similarity nature of Ethernet traffic," IEEE/ACM Transactions on Networking, 1993.

[6] Jens Feder , "Fractuals," Plenum Press, 1988.

[7] Sanford Weisberg, "Applied Linear Regression, second edition, " John Wiley & Sons, 1985.

[8] I. V. Basawa, B.L.S. Prakasa Rao, "Statistical Inference for Stochastic Processes," Acadamic Press, London, 1980.