Exploring regularities and self-similarity in Internet traffic

FRANCESCO PALMIERI and UGO FIORE
Centro Servizi Didattico Scientifico
Università degli studi di Napoli "Federico II"
Complesso Universitario di Monte S. Angelo, Via Cinthia 5
NAPOLI - ITALY

Abstract: - The characterization and quantitative description of Internet traffic is becoming increasingly important in light of the rapid growth being observed in the size and usage of the network. Modeling of modern teletraffic and large telecommunications networks with fractal stochastic processes has been investigated by analyzing traffic measurements collected on the "Federico II" university WAN link to the National Research Network backbone, and using them to test the existence of fractal behavior and specific properties. A fractal character with approximate self-similarity and statistical long-range dependence has been observed in the Internet access traffic measurements. This approach leads to some interesting insights about whether and how regularities in Internet user behavior can be noticed and exploited. The derived characterization could be seen as a first step to a generalized parametric Internet traffic model.

Key-Words: - Internet traffic modeling, self-similarity, long-range dependency

1 Introduction

Internet engineering and management depend on an understanding of the characteristics of network traffic patterns. Increased traffic volumes has resulted in congested routes, often leading to delays on the order of hundreds of seconds. As a result, much work is being done to understand the end-to-end performance related issues for the Internet to identify appropriate traffic descriptors and properties and design traffic engineering rules. Statistical models are needed that can generate traffic that mimics closely the observed behavior on live Internet wires. Models can be used on their own or combined with network simulators for a wide variety of tasks such as traffic forecasting and future network capacity planning. The challenge of model development is immense.

Fractal geometry, frequently portrayed as the opposite to mathematically elegant and tractable Euclid geometry, has revealed the whole new horizon on the way our minds observe and interpret the various shapes of nature we see everyday. This is attributed to its simplicity to explain the extreme variability of natural shapes which would be otherwise difficult, if not impossible, to be described by its Euclidean counterpart. Fractal analysis is now, the most promising technique for network traffic characterization. The high variability and burstiness characteristics of the Internet traffic make the fractal techniques the most attractive modeling tool to describe such a complex phenomenon.

In this paper we show the results of the statistical analysis on the traffic samples recorded for a year on the IP over ATM link that connects the "Federico II" university WAN link to the National Research Network backbone (GARR). Our results demonstrate that the WAN traffic exhibits fractal scaling, indicating that the traffic could be well approximated by a weakly stationary, self-similar process.

2 Internet traffic characterization

Traffic in the Internet results from the uncoordinated actions of a very large population of users. There are also a far greater number of protocols and applications, each with its own traffic characteristics, and new applications can arise at any time. There are a great variety of network connectivity types, architectures and equipment, and, accordingly, different kind of traffic flows. There are no standard network topologies around which all design efforts can be based, and the topologies that exist are subject to

constant change. Perhaps the most serious and most surprising characteristic that the network traffic exhibits is burstiness.

2.1 The fractal nature of Internet traffic

It has been suggested that Internet traffic is far too complicated to be modeled using the techniques developed for the telephone network or for computer systems [1]. Traditionally, networks have been described by generalized Markovian processes. Markov and Poisson models, typically used in queuing analysis, are characterized by limited memory of the past in which they assume variations only in limited time scales. They do not allow for long-range dependence; rather, they explicitly show only short-range dependence. Furthermore, in a Markovian model, smoothing of bursty data is possible. In fact traffic bursts who have a characteristic length and averaging over a long period of time results in a smooth data stream.

Recently it has been demonstrated [2] that real packet traces do not obey to the above models in most cases and instead, the notion of self-similarity, that is the typical fractal behavior in terms of resemblance or correspondence between the parts of an object obtained by scaling and the object as a whole, can be used to explain the extreme burstiness of the Internet traffic with a surprisingly small number of parameters. Burstiness is a qualitative concept, but it can be described analytically as self-similarity over multiple time scales. In fact burstiness is the result of phenomena at several levels or protocol layers, which interact to produce it. Roughly speaking, the distribution of file sizes (an application layer statistic) affects transmission time statistics, which, in turn, shows up as self-similarity at the transport and network layers, and gives rise to long-range dependence, finally leading to burstiness seen at the link layer.

From a statistical point of view, network traffic exhibits other fractal characteristics in its second-order statistics such as slowly decaying variance and long-range dependency over a wide range of time and frequency scales. This means that the variance of the sample mean decreases much more slowly than the reciprocal of the sample size, or in other terms that the distribution of the traffic process decay more slowly than exponential (i.e. a Poisson distribution), and autocorrelation exhibit an hyperbolic ("long range") rather than exponential ("short range") decay.
Such properties observed over a wide range of time scales suggested that fractals are the most appropriate mathematical tool to describe certain aspects of network behavior.

2.2 The need for traffic modeling

It is important to be able describe this traffic succinctly in a statistical manner which is useful for network engineering. The traffic process can be described in terms of the characteristics and properties of a number of objects, including packets, bursts, flows, etc. depending on the time scale of relevant statistical variations. The preferred choice for modelling purposes depends on the object to which traffic controls are applied. Conversely, in designing traffic controls it is necessary to bear in mind the facility of characterizing the implied traffic object. The most relevant and interesting characteristics to be taken into account in the formulation of a fractal model of Internet traffic are:

- Long-range dependence (LRD)
- Self-similarity
- Infinite variance
- Burstiness

3 Essential concepts

To clarify our terminology, we briefly summarize the definition and deeper significance of some of the basic concepts we introduced in the above sections that will be extensively used in the analysis of our real word traffic samples.

3.1 Self-similarity

The notion of fractal behavior, in the sense of an object that it has similarity on all scales is translated into the stochastic analysis by the definition of Self-Similar processes with Stationary Increments. Let $X = \{X_k : k> 0\}$ a stochastic stationary process in the discrete-time domain, representing the amount of data transmitted in consecutive short time periods, and let $X^{(m)} = \{X_k^{(m)} : k\geq 1\}$ its aggregate form obtained by averaging the X_k over adjacent, non

overlapping blocks of size m:

$$X_k^{(m)} = m^{-1} \sum_{n=(k-1)m+1}^{km} X_n$$

The stochastic process X satisfies exactly the *self-similarity* property, if X_k and $m^{1-H} X_k^{(m)}$ have identical finite-dimensional distributions for all $m \geq 1$ that is:

$$X_k \overset{d}{=} m^{1-H} X_k^{(m)}$$

Where $\overset{d}{=}$ means the equality of the finite dimensional distributions. The parameter H is referred to as *Hurst parameter* or *Hurst exponent* and represents the degree of self-similarity in the observed sample. When the value of the Hurst parameter is between 0.5 and 1 the sample is said to be self similar (values of H closer to 1 indicate a high degree of self-similarity).

Furthermore, second order *self-similarity* is satisfied when X and $m^{1-H} X^{(m)}$ have the same variance and autocorrelation. Second order self-similarity manifests itself in several equivalent ways, one of them is that the spectral density of the process decays as f^{1-2H} at the origin as $f \rightarrow 0$.

3.2 Long-range dependency

This refers to the degree of dependence of samples taken at one time on those of an earlier time. It is gauged quantitatively by the *autocorrelation function*. Autocorrelation of a stationary process measures the degree of correlation of nearby and far-off events, i.e., the ability to predict (in a statistical sense) process values removed in time from any given time t. For short-range dependent traffic, which is non-bursty, the autocorrelation function falls off quickly with time, usually exponentially. For long range dependent traffic, it falls off much more slowly, usually obeying some type of power law. In detail, let $X = \{X_k : k > 0\}$ a stationary process in the discrete-time domain with mean $\mu \equiv E[X_k]$, variance $\sigma^2 \equiv E[(X_k - \mu)^2]$ and normalized autocorrelation function:

$$r(k; t) \equiv E[(X_n - \mu)(X_{n+k} - \mu)]/\sigma^2$$

The process X is said to be a long-range dependent (LRD) process if for fixed t its autocorrelation function $r(k; t)$ is non-summable [8], i.e.

$$\sum_{k=0}^{\infty} r(k;t) = \infty$$

Because the behavior of the tail of $r(k; t)$ completely determines its summability the details of how $r(k; t)$ decays with k engender much interest. This originates another widely employed definition of long-range dependency [9] which employs also the Hurst parameter:

$$r(k; t) \sim f(t) k^{-(2-2H)}, \text{ as } k \rightarrow \infty$$

where $f(t)$ is a positive function independent of k which emphasize the dependence of the autocorrelation function on the time scale t in addition to the lag k. The *Hurst* parameter, that in this case must be in the range $0.5 < H < 1$, implying self similarity, completely characterizes the above relation.

The same concept may be more evident in the frequency domain when examining the Fourier transform of the autocorrelation:

$$S(\omega;t) = 2\pi^{-1} \sum_{k=-\infty}^{\infty} e^{-jk\omega} r(k;t)$$

The value of the coefficients $r(k;t)$ falls off either exponentially or slower, with increasing k. If the rate is slower than exponential, the traffic exhibits long-range dependence.

3.3 Slowly decaying variance

Considering the aggregate form of the above stationary process in the discrete-time domain, defined by $X^{(m)} = \{X_k^{(m)} : k \geq 1\}$, the process exhibits slowly decaying variance when the variance of the sample mean decreases much more slowly than the reciprocal of the sample size, that can be represented as:

$$Var[X^{(m)}] \sim m^{-\beta}$$

for $0 < \beta < 1$ and sufficiently large m. This can be easily detected by plotting $Var[X^{(m)}]$ against m on a log-log plot which in the literature [9]is referred to as the variance-time plot. If this plot forms a straight line with an absolute slope less than unity over a wide range of m. then we say

the process X exhibits the slowly decaying variance property.

4 Analyzing real Internet traffic

The traffic measurements analyzed in this work are obtained at the junction where the "Federico II" university (Naples, Italy) connects to the GARR network (Italian academic and research network) through a 34Mbps ATM CBR link. The data, collected via SNMP from the border router, represent the aggregate traffic flowing from the university local area networks towards the national backbone. Average and maximum input and output rate have been collected during a year, on a daily basis (in 5 min. samples) and a progressive statistical consolidation by averaging the samples on respectively 30 minutes, 2 hours and 1 day basis has been performed to obtain the aggregate weekly, monthly and yearly trends.

Daily and weekly traffic observation reveal intensity levels (in bits/sec) averaged over periods of 5 to15 minutes, which are relatively predictable from day to day. Systematic intensity variations occur within the day reflecting user activity. It is possible to detect a busy period during which the traffic intensity is roughly constant. Similar considerations can be done by scaling on the weekly and monthly statistics.

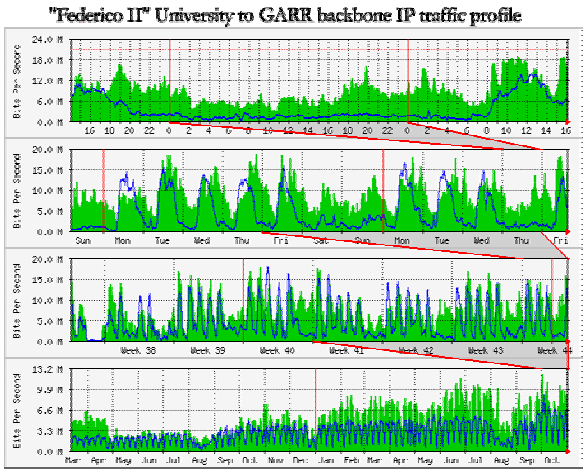

Figure 1: "Federico II" Internet traffic sampling

The above graphics (figure 1) clearly show at a glance the fractal behavior of Internet traffic. Self-similarity can be readily observed in this figure, which shows the similar appearance of the various traffic patterns, regardless of the time scale.

4.1 Self-similarity analysis

To gauge the self-similarity of packet traffic from empirical data, several measures rooted in fractal dimensionality have been devised. These measures make use of the scaling factor discussed earlier. The principal methods that will be used in our analysis are:

- Variance/Time Plot [1]
- Whittle Estimator [3]
- Rescaled Range (R/S) [5]
- Periodogram [6]

4.1.1 Hurst estimation

The *Hurst* parameter was estimated by means of the variance-time method and the *Whittle*'s estimator [3]. Almost all samples gave values of the *Hurst* parameter higher than 0.8, that is the telltale sign of *fractal* behavior.

AGGREGATION LEVEL	HURST PARAMETER
5 min	*0,9767*
30 min	*0,9584*
2 hour	*0,9814*
1 day	*0,7705*

Table 1: Hurst calculated on the 4 aggregated sets

4.1.2 Periodograms

Further confirmation of the above statements comes from the periodogram analysis, where, at different time scales specific cycles can be detected. The log-scaled graphics below, referring to the "Average Input" samples, exhibit a clear x^{-b} shape, which implies self-similarity. Furthermore, a closer look at the periodograms shows the presence – not surprisingly – of 7-days, and 24-hours cycles at the larger time scales. However, at finer grain some minor cycles can be noticed that we were unable to explain, for instance a 3-hour cycle and a 20-minutes cycle.

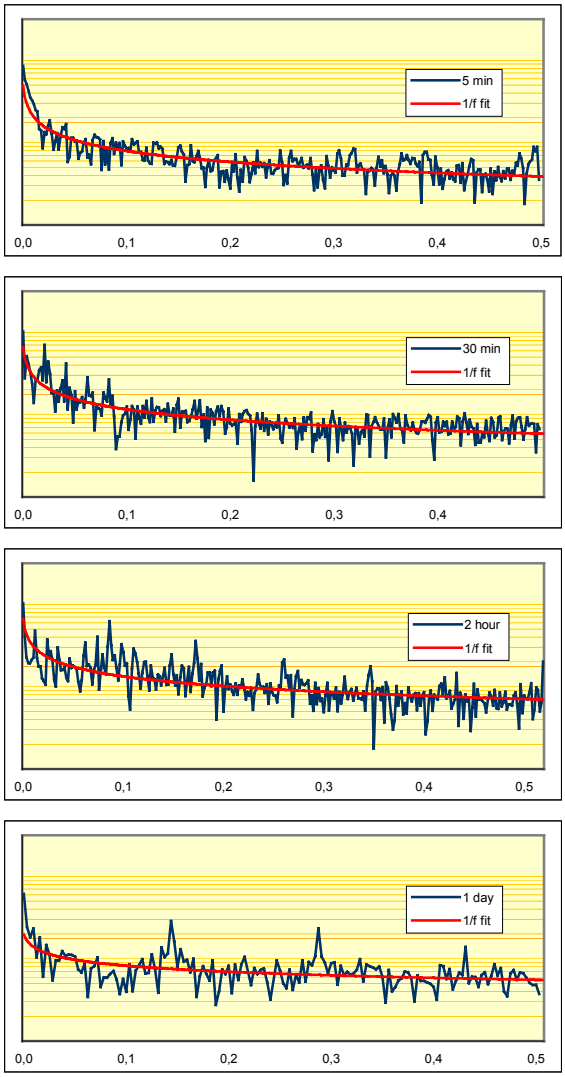

Figure 2: Periodograms at four time scales

Even clearer evidence results from Figure 3, where the major peaks in the periodograms are plotted. As one can see, the peak distribution scales regularly with aggregation level, with the partial exception of the 5-minutes time scale, where the effects of many different frequencies tend to appear. This result, combined with the previous strong evidence in time cycles can become a very significant matter in traffic trend forecasting.

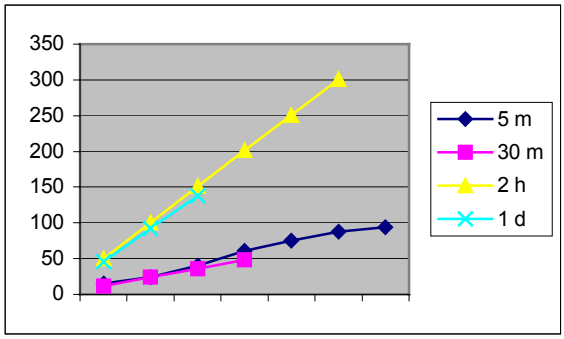

Figure 3: Periodogram peaks at four time scales

The log-log R/S graphs are pretty linear, as shown in figure 4 below. We can thus rely on the inherent self-similarity of traffic to understand its trends and oscillations.

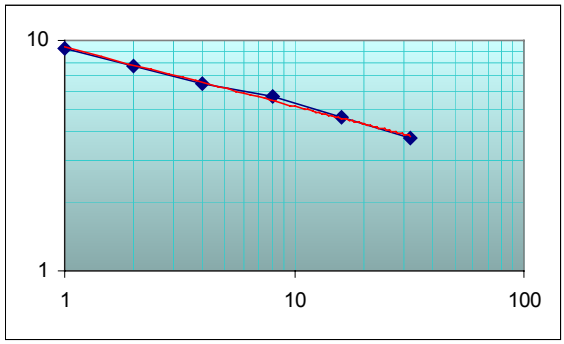

Figure 4: Log-log R/S graph

4.2 Long-range dependence

The Fourier transform of the autocorrelations, shown in figure 5, exhibits a rate far from slower than exponential. This implies the clear evidence of long-range dependency.

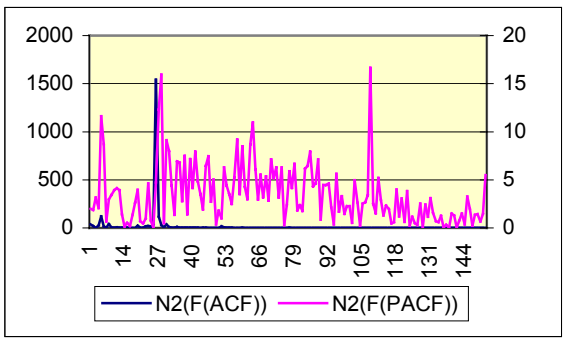

Figure 5: Autocorrelations in the frequency domain

5 Conclusions

We analyzed a wide collection of samples of

aggregated WAN traffic collected on the high-speed link that connects the Federico II University to the Internet. Fractal behavior and Long-range dependence has been detected respectively through the estimation of the Hurst parameter and the study of R/S and Fourier transform of autocorrelation diagrams. These results comfort us in our purpose of applying a fractal model for predicting traffic peaks. This would greatly improve the capacity planning activities and let the network organization evolve in accordance with the traffic trends.

References:

[1] W. Willinger, and V. Paxson, "Where Mathematics meets the Internet", *Notices of the American Mathematical Society*, vol. 45, no. 8, Aug. 1998, pp. 961–70.

[2] V. Paxson and S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling", *IEEE/ACM Transactions on Networking*, Vol. 3 No. 3, June 1995, pp. 226-244.

[3] V. Paxson, "Fast, Approximate Synthesis of Fractional Gaussian Noise for Generating Self-Similar Network Traffic," *Computer Communications Review*, Vol. 27 N. 5, pp. 5-18, October 1997.

[4] K. Park, G. Kim, and M. Crovella, "On the Effect of Traffic Self-Similarity on Net-work Performance," *Proceedings 1997 SPIE International Conference On Perfor-mance and Control of Network Systems*, 1997.

[5] K. Park, G.T. Kim, and M. Crovella, "The Relationship Between File Sizes, Transport Protocols, and Self-Similar Network Traffic," *Boston University, Technical Report* BU-CS-96-016, 1996.

[6] A. Erramilli, O. Narayan, and W. Willinger, "Experimental Queueing Analysis with Long-Range Dependent Traffic," *IEEE Transactions Networking* 4(2), pp. 209-222, 1996.

[7] M. S. Taqqu, "Self-similar processes", *S. Kotz and N. Johnson Editors, Encyclopedia of Statistical Sciences*, volume 8. Wiley, New York, 1987.

[8] D.R. Cox, "Long-range dependence. A review.", *H.A. David and H.T. Davis Editors, Statistics, An Appraisal*, The Iowa State University Press, Ames, Iowa, 1984, pp. 55-74.

[9] W. Leland *et al.*, "On the Self-Similar Nature of Ethernet Traffic", *IEEE Transactions on Networking*, vol 2, no. 1, Feb 1994, pp. 1–15.