# CHAPTER 2

# SELF-SIMILARITY AND LONG-RANGE DEPENDENCE

The traffic behavior in a network has serious implications for the design, control, and analysis of the network. By analyzing data collected, it was demonstrated that the Ethernet Local Area Network traffic is statistically self similar. [LTWW94] Asynchronous Transfer mode, high speed, cell relay, networks, mostly are used as backbones for the interconnection of enterprise networks composed of several LAN's, thus leading to self similar behavior of the traffic in ATM Networks. [PF95]

Self-similarity is the property we associate with fractals - the object appears the same regardless of the scale at which it is viewed. It 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 consisting of bursty subperiods separated by less bursty subperiods.

A phenomenon that is self similar looks the same or behaves the same when viewed at different degrees of "magnification " or different scales on a dimension. This dimension may be space (length, width) or time.

The commonly assumed models for network traffic, (*e.g.,* the Poisson distribution, etc.) did not fit the recorded traces, since these models were not able to capture the fractal behavior of the traffic. [TG97]

Were traffic to follow a Poisson or Markovian arrival process, it would have a characteristic burst length which would tend to be smoothed by averaging over a long enough time scale whereas measurements of real traffic indicate that significant traffic variance (burstiness) is present on a wide range of time scales.

The effect of self-similarity in network traffic is shown in [LTWW95], which compares a self-similar series, with a compound Poisson series with the same distributional characteristics. The paper shows that Poisson models for network traffic become essentially uniform when aggregated by a factor of 1,000; while actual network traffic shows no such decrease in variability over the same range of aggregation.

### 4.1 Definition: [S98]

The degree of self-similarity, defined via the Hurst parameter, typically depends on the utilization level of the network and can be used to measure the "burstiness " of the traffic. As H increases the degree of self similarity is increasing. $0.5 < H < 1$ .For H=0.5 and H>1, there is almost no self similarity. H is the measure of length of a long-range dependence of a stochastic process.

Continuous Time definition :

Suppose we define the Hurst parameter as H, then a stochastic process X(t) is statistically self similar with the parameter H, if for any real a>0, the process $a^{-H} X(at)$ has the same statistical properties as X(t) .This is expressed as three conditions .

i.e. Mean $E(X(t)) = E(X(at))/a^{H}$ ...(2.1)

Variance $Var(X(t)) = Var(X(at))/a^{2H}$ ...(2.2)

and Auto Correlation $r(t,s) = R(at,as)/a^{2H}$ ...(2.3)

Discrete Time definition :

A self-similar time series has the property that when aggregated (leading to a shorter time series in which each point is the sum of multiple original points) the new series has the same autocorrelation function as the original.

That is, given a stationary timeseries , $X = (X_t;\ t=0,1,2...)$ we define the m-aggregated series

$$X^{(m)} = (X_k^{(m)};\ k= 1,2,3\ldots)$$

by summing the original series X over nonoverlapping blocks of size m. Then if X is self-similar, it has the same autocorrelation function

$$r(k) = E[(X_t - \mu)(X_{t+k} - \mu)] \text{ as the series } X_{(m)} \text{ for all } m .$$

This means that the series is distributionally self-similar: the distribution of the aggregated series is the same (except for changes in scale) as that of the original. [B94, LTWW95]

If in particular we assume the autocorrelation function of an exactly second order self similar process to be $r(k) \sim k^{-\beta} L_1(k)$, as $k \rightarrow \infty$ $0<\beta<1$ ...(2.4)

and L is slowly varying at infinity , i.e , $\lim_{t\to\infty} L_1(tx)/L_1(t) = 1$ for all $x>0$ [LTWW94]

The process X is said to be exactly self similar with the parameter $\beta$ $(0<\beta<1)$ if for all m

$Var(X^{(m)}) = Var(X) / m^{\beta}$ Variance

$r_{X^{(m)}}(k) = r_X(k)$ Autocorrelation

the Hurst parameter, can be depicted as $H = 1-\beta/2$ .

For a stationary, ergodic process $\beta$=1 and the variance of the time average decays to zero at the rate of 1/m For a self similar process the variance of the time average decays more slowly.

The process X is said to be asymptotically self similar, if for all k large enough,

$$\mathrm{Var}\,(X^{(m)}) = \mathrm{Var}\,(X) / m^{\beta} \qquad \text{Variance}$$

$$r_{X^{(m)}}(k) = r_X(k) \quad \text{as } m \rightarrow \infty \qquad \text{Autocorrelation}$$

Thus the autocorrelation of the aggregated process has the same form as the original process.

## 4.2 Properties:

<u>Long Range Dependence:</u>

The existence of long range dependence was seen from the heuristic plots such as Variance Time plot and Autocorrelation plot. If a self-similar process has observable bursts on all time scales, it is said to exhibit long-range dependence; values at any instant are typically correlated with value at all future instants. [LTWW94]

A stochastic process satisfying relation (2.4) is said to exhibit long range dependence. Processes with long range dependence are characterized by an autocorrelation function that decays hyperbolically (as compared to the exponential decay exhibited by traditional traffic models). Hyperbolic decay is much slower than exponential decay, and since $\beta$<1 , the sum of the autocorrelation values of such series approaches infinity. The implication of this nonsummable autocorrelation is that, if we consider n samples from the series, then the variance does not decrease as a function of 'n' but by a value of $n^{-\beta}$.

This importance of this long range dependence was seen in studies conducted, which showed that the packet loss and delay behavior was very different in simulations using real traffic data rather that traditional network models. [CW91]

Long Range Dependence also has a serious impact on cell loss in Wide Area ATM Networks. From studies conducted [VMG9] the following results were obtained:

Let $\{X_{i,}\ i = 1,2\ldots\}$ be a covariance stationary process (eg: the sequence of cell interarrival times) with mean $\mu$, and variance $\sigma^2$ and the autocorrelation function

$$\rho\ (i - j) = E[(Xi - \mu)(Xj - \mu)] / \sigma^2$$

The long-range dependant process can be defined as [VMG9]:

$\{X_i\}$ is called long range dependent , if there exists a real number $\alpha \in (0,1)$ and a constant $c_1 > 0$ such that

$$\lim_{k \to \infty} \rho(k) / [c_1 k^{-\alpha}] = 1$$

If we use the parameter $H = 1 - \alpha / 2$ instead of $\alpha$, long range dependence occurs for $0.5 < H < 1$. The higher the value of H is, the stronger the long range dependence.

The above is an asymptotic definition, it determines the only the rate of convergence as the lag tends to infinity, but it does not specify the correlations for any fixed finite lags. This definition cannot be applied in practice.

Queuing properties, when LRD traffic is used as input: In case of LRD input, the tail of the complementary queue length distribution decays slower than exponentially. [VMG]

Impact of LRD on buffer dimensioning: LRD characteristic plays a significant role in queuing performance as well as in Buffer Dimensioning of packet networks [ENW96, LTG95, N95]. The effect of LRD traffic in ATM Network was studied in the paper [LS].

In this the input process consisted of infinite number of independent and identical ON / OFF mini sources where the ON and OFF follow the respective general distributions. This input was able to express long range dependent property and yet mathematically tractable [PM97]. Although other long range dependent traffic models (such as Fractional Brownian Motion [LTWW94, N95] and fractional autoregressive integrated moving average process [LTWW94] do exist, they are too computationally intensive to be tractable in general.

## Spectral Density [S98]:

The power spectrum of such a series is also hyperbolic. It obeys a power law near the origin:

$$S(\omega) \sim 1 / |\omega|^{\gamma} \qquad \text{as } \omega \rightarrow 0 , \qquad 0 < \gamma < 1$$

As it is known the power spectral density for a discrete time stochastic process is defined as follows:

$$S(\omega) = \sum_{k=\infty}^{-\infty} r(k)\, e^{-j2k\omega}; \qquad S(0) = \sum_{k=\infty}^{-\infty} r(k)$$

It can be shown that $\gamma = 1- \beta = 2H - 1$

In contrast short range dependent processes are characterised by a spectral density that remains finite as

$\omega \rightarrow 0$.

Heavy Tailed Distributions :

It is possible to define self-similar stochastic processes with distributions that are heavy tailed. Heavy tailed distributions can be used to characterize the probability densities that describe traffic processes such as packet interarrival times and burst lengths. The distribution of the random variable X is said to be heavy tailed if

$$1-F(x) = P[X \geq= x] \sim x^{-\alpha}, \text{ as } x \rightarrow \infty \quad , 0 < \alpha < 2 .$$

That is, regardless of the behavior of the distribution for small values of the random variable, if the asymptotic shape of the distribution is hyperbolic, it is heavy-tailed. A random variable with heavy tailed distribution exhibits an infinite variance.

**4.3 Modeling:**

If we consider modeling the self-similar traffic, this could be done by an asymptotically second order, heavy tailed distribution. Consider a large number of concurrent processes that are each either ON or OFF. At any point in time, the value of the time series is the number of processes in the ON state. If the distribution of ON and OFF times for each process is heavy-tailed, then the time series will be self-similar. Such a model could correspond to a network of workstations, each of which is either silent or transferring data at a constant rate.

According to the various properties of self similar traffic, many models were proposed to represent the self similar traffic.

Based on the continuous time definition, the Fractional Brownian Motion process was considered. An FBM process $B_H$ (t) can be defined as follows:

$$B_H(t) = X\, t^{H} \qquad (t > 0;\ 0.5 \leq H < 1)$$

where X is a normally distributed random variable with mean 0 and variance 1 and H is the parameter of the process . Thus $B_H$ (t) is a normally distributed random variable with zero mean. For H=0.5 it reduces to an ordinary Brownian motion. The probability density of a Brownian motion process has the form:

$$F_{BH}(x, t) = (1/ (2 \Pi t^{2H})^{0.5})\, e^{(-x^2/ 2t^{2H})}$$

To prove the FBM process is self-similar we have to prove three conditions.
Consider $B_H$ (a t), where $B_H$ (t) is the FBM process already defined.
Then $B_H (a t) = X (at)^H$ .
We easily see that $E [B_H (at)] = 0$, satisfying the first condition (2.1).
Also it can be proven that $Var (B_H (t)) = Var (B_H (at)) / a^{2H}$, satisfying the second condition (2.2)
The third condition listed (2.3) is also satisfied: $r_{BH} (at, as) = a^{2H} r_{BH} (t, s)$.
Thus the process is self-similar, but for H > 0.5, in this process, for large values of t, the strength of the correlation increases between past and future increments. This phenomenon is called persistence and is in conflict with what is normally assumed about stochastic phenomenon. Thus we turn to other models.

Heavy tailed distributions can be used to characterize self-similar traffic. The simplest heavy tailed distribution is the Pareto distribution with the parameters k and $\alpha$
($k, \alpha > 0$) , with density and distribution functions as

$$f(x) = 0 \qquad F(x) = 0 \qquad (x \leq k)$$

$$f(x) = (\alpha / k)\ (k / x)^{\alpha+1} \qquad F(x) = 1 - (k / x)^{\alpha} \qquad (x > k,\ \alpha > 0)$$

and a mean value $E[X] = \alpha\ k / \alpha - 1 \qquad (\alpha > 1)$

k specifies the minimum value that the random variable can take . The parameter $\alpha$ determines the mean and variance of the random variable: If $\alpha \leq 2$, then the distribution has infinite variance, and if $\alpha \leq 1$, it has infinite mean and variance.

### 4.4 Examples of occurrence of self-similarity:

Ethernet:

The theory of self-similarity of the traffic started with the analysis of the data collected on the Ethernet [LTWW94]. This report shattered the illusion that straightforward queuing analysis using the Poisson traffic assumption is adequate to model all network traffic. The paper reports the results of a detailed high-resolution (time accuracy of 20 microseconds) collection of Ethernet traffic measurements conducted between 1989 and 1992. The data consisted of 4 sets of traffic measurements, each representing between 20 and 40 consecutive hours of Ethernet traffic and consisting of total of well over 100 million packets. The data was collected from various Ethernet LAN's at Bellcore. From the plots comparing the actual measurements and the synthetic Poisson model we see that, for the Poisson model, as the data was aggregated, the traffic pattern smoothed out. Whereas this was not so for the actual measurements. It remained bursty at all scales with no natural length for the burst.

World Wide Web:

[CROV96] reports on a study of Web traffic that involved over half a million requests for Web documents. This study showed that the traffic pattern generated by the browsers was self-similar. The analysts modeled each browser as an ON / OFF source and found that the data fit very well to a Pareto distribution with ranging from 1.16 to 1.5.

Signalling System Number 7 Traffic:

The study reported in [MRW94] looked at the control signaling traffic generated on digital Telecommunication networks. The control signal protocol is SS7, used on ISDN and other digital networks .The studies showed that the Poisson model for the traffic was inadequate and self similar traffic models provided a better fit. Using the control signal data available the actual calls could be studied. The call duration was best characterized by a heavy tailed distribution.

Performance Implications of Self Similarity:

From the study conducted in the paper [LTWW94], an important discovery was that, the higher the load on the Ethernet, the higher the estimated Hurst parameter H or, higher the degree of self-similarity. This result became vital because at high load performance issues were more vital. Another important result yielded from this study was that the cell losses in the real traffic were more than the expected amount in the Poisson Model. This resulted in the redesign of switches and the increase of the buffer size.

Estimation of Self Similar Traffic:

A number of approaches have been adopted to determine if a time series of actual data is self-similar, and if so to estimate the self-similarity parameter H.

Variance Time Plot:

We know from previous assumptions, the aggregated time series is $X^{(m)}$ and its variance is

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

Where the self similarity parameter is $H = 1- \beta b /2$

The above can be rewritten as: $\log\,[Var\,(X^{(m)})] \sim \log\,[Var\,(X)] - \beta \log\,(m)$

Because $\log\,[Var\,(X)]$ is a constant independent of m, if we plot $Var\,(X^{(m)})$ versus m on a log - log graph, the result should be a straight line with the slope of $-\beta$.

Slope values between -1 and 0 indicate self-similarity.

The other methods are the R/S plot and periodogram periodogram plot. The fourth, more accurate method being the Whittles Estimator.

Conclusion:

Understanding the nature of traffic in high speed, high bandwidth communication systems such as B-ISDN (and therefore ATM Networks), is essential for engineering, operations and performance evaluation of these networks. The findings that

a) The Ethernet LAN Traffic and Traffic in ATM Networks is statistically self similar , and that,
b) The degree of self similarity can be measured in terms of Hurst Parameter "H" which is typically a function of the overall utilization of the network ,

affected many aspects of network design and network component designs.

Therefore the implications of self-similar nature of packet traffic for engineering, operations and performance evaluation of high-speed networks were ample:

(1) source models for individual Ethernet users are expected to show extreme variability in terms of interarrival of packets (i.e the infinite variance syndrome),
(2) commonly used measures for burstiness such as index of dispersion (for counts), the peak to mean ratio, or the coefficient of variation (for interarrival times) are no longer meaningful for self similar traffic but can be replaced by the Hurst parameter , and,
(3) the nature of congestion produced by the self similar network traffic models differs drastically from those predicted by standard formal models .

Self-similar processes could be modeled as either exactly or asymptotically second order self-similar processes. Models were given ATM cell traffic and sufficient conditions for exact and asymptotic self-similarity were derived. Mostly it was found that self-similar processes could be modeled by heavy tailed distributions, a very common one being the Pareto distribution.

Properties of the self similar traffic were observed for example, if number of sources increased the self similar nature increased, the long range dependence of the traffic increased as the Hurst parameter increased, and the buffer sizes used for normal Poisson models could not be used for self similar traffic etc. This particular property of long range dependence of self-similar traffic could affect cell loss ratio and Buffer dimensioning.

Even now research continues on the self-similar nature of Network traffic and the effect of it on various parameters like, cell loss, buffer size, peak rate enforcement etc.

References:

[LTWW94] : W.E. Leland , M.S. Taqqu , D.V. Wilson and Walter Willinger , *" On the Self-Similar Nature of Ethernet Traffic* (extended version ) , " IEEE / ACM Trans. Networking , vol. 2 , no. 1, pp. 1-15 ,1994

[PF95] : Vern Paxon and Sally Floyd , " *Wide Area Traffic : The failure of Poisson Modeling* , " IEEE / ACM Trans. Networking , Vol. 3 , pp. 226-244 , 1995 .

[TG97] : Boris Tsybakov and Nicolas D. Georganas , "*On Self Similar Traffic in ATM Queues :Definitions , Overflow Probability Bound, and Cell Delay Distribution* , " IEEE / ACM Trans. Networking , Vol. 5 , No. 3 , pp. 397-409 , 1997

[B94] : Jan Beran. " *Statistics for Long-Memory Processes* ". Monographs on Statistics and Applied Probability. Chapman and Hall, New York, NY, 1994.

[LTWW95] : Walter Willinger, Murad S. Taqqu, Will E. Leland, and Daniel V. Wilson. " *Self-similarity in high-speed packet traffic: Analysis and modeling of Ethernet traffic* measurements ". Statistical Science, 10(1): 67--85, 1995.

[LW91] : W. E. Leland and D. V. Wilson. " *High time-resolution measurement and analysis of LAN traffic: Implications for LAN interconnection* ". In Proceeedings of IEEE Infocomm '91, pages 1360--1366, Bal Harbour, FL, 1991.

[CB95] : Mark E. Crovella and Azer Bestavros "*Explaining World Wide Web Traffic Self-Similarity*". Technical Report TR-95-015.Revised.October 12, 1995

[S98] : William Stallings , " *Self Similar Traffic* " . High Speed Networks . Prentice-Hall International Inc. 1998

[VMG] : Attila Vidacs, Sandor Molnar and Geza Gordos ," *The impact of Long Range Dependence on cell loss in an ATM Wide Area Network* "

[ENW96] : A. Erramilli , O.Narayan, and W.Willinger , " *Experimental Queuing Analysis with long Range Dependent Packet Traffic*, " IEEE / ACM Trans. On Networking , Vol. 4 , No.2 , pp. 209-223 , 1996

[LTG95] : N.Likhanov, B.Tysbakov, and N.Georganas , " *Analysis of an ATM Buffer with Self-Similar ("Fractal") input Traffic* " , Proceedings of IEEE INFOCOM '95 , 1995

[N95] : Ilkka Norros , " *On the use of Fractional Brownian Motion in the Theory of Connectionless Networks* " , IEEE Journal on Selected Areas in Communications , Vol. 13 , No.6 , 1995

[LS] : George.C.Lin and Tatsuya Suda , "*On the impact of Long-Range-Dependent Traffic in Dimensioning ATM Network Buffer* " , [PM97] : M.Parulekar and A.Makowski,

" *M / G / ¥ Input Processes : A Versatile class of Models for Network Traffic* " , Proceedings of IEEE INFOCOM '97 , 1997.

[MRW94] : Duffy. D. McIntosh, A.Rosenstein .M and W.Willinger " *Statistical Analysis of CCSN / SS7 Traffic Data from Working CCS Subnetworks* ", IEEE Journal on Selected Areas in Telecommunication, April 1994.

[CROV96] : Mark .E.Crovella and A.Bestavros , "*Self similarity in World Wide Web traffic : Evidence and possible causes " ,* Proceedings, ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems , May 1996.