Network Traffic Characteristic

Hojun Lee
hlee02@purros.poly.edu

Outline

- Motivation
- What is self-similarity?
- Behavior of Ethernet traffic
- Behavior of WAN traffic
- Behavior of WWW traffic

Motivation for network traffic study

- **Understanding network traffic behavior is essential for all aspects of network design and operation**
  - **Component design**
  - **Protocol design**
  - **Provisioning**
  - **Management**
  - **Modeling and simulation**

Three main reference papers

- **W. Leland, M. Taqqu, W. Willinger, D. Wilson, *On the Self-Similar Nature of Ethernet Traffic, IEEE/ACM* TON, 1994.**
- **V. Paxson, S. Floyd, *Wide-Area Traffic: The Failure of Poisson Modeling, IEEE/ACM* TON, 1995.**
- **M. Crovella, A. Bestavros, *Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes, IEEE/ACM* TON, 1997.**

In the past …

- Traffic modeling in the world of telephony was the basis for initial network models
  - Assumed Poisson arrival process
  - Assumed Poisson call duration
  - Well established queuing literature based on these assumptions
  - Enabled very successful engineering of telephone networks

What is self-similarity in nature?

- No natural length of a bust, at every time scale, similar looking traffic bursts are evident (structure repeats at all scales)
- Aggregating streams of such traffic intensifies the self-similarity instead of smoothing it
- Aggregation causes more burstsness and requires larger buffers (just as Stochastic processes are invariant to time, self-similar processes are invariant to scale)

Definition of self-similarity

- Consider a zero-mean stationary time series $X = (X_t; t = 1,2,3,\ldots)$, we define the *m*-aggregated series $X^{(m)} = (X_k^{(m)}; k = 1,2,3,\ldots)$ by summing X over blocks of size m. We say X is ***H-self-similar*** if for all positive *m*, $X^{(m)}$ has the same distribution as X rescaled by m^H => $X_t \stackrel{d}{=} m^{-H} \sum_{i=(t-1)m+1}^{tm} X_i$
- If X is *H*-self-similar, it has the same autocorrelation function r(k) as the series $X^{(m)}$ for all *m*. This is actually *distributional* self-similarity. ➔ $r(k) = E[(X_t - \mu)(X_{t+k} - \mu)]/\sigma^2$
- X(t) is ***exactly second-order self-similar*** with Hurst parameter H ($1/2 < H < 1$) if
  $$\gamma(k) = \frac{\sigma^2}{2}((k+1)^{2H} - 2k^{2H} + (k-1)^{2H}) \text{ for all } k \geq 1$$
  X(t) is ***asymptotically second-order self-similar*** if
  $$\lim_{m \to \infty} \gamma^m(k) = \frac{\sigma^2}{2}((k+1)^{2H} - 2k^{2H} + (k-1)^{2H})$$

Long-range dependence vs. SS

- **Values at any instant are typically nonnegligibly positively correlated with values at all future instants**
- **Return the definition of second-order self-similarity and its autocovariance**
- **Let *autocorrelation function*,** $r(k) = \dfrac{\gamma(k)}{\sigma^2}$

For 0 < *H* < 1, $H \neq \frac{1}{2}$ **, it holds**

➔ $r(k) \sim H(2H-1)k^{2H-2}, \quad k \to \infty$

Particularly, for , ½ < H < 1

➔ $r(k) \sim ck^{-\beta}$ **where** $0 < \beta < 1$ **and *c* > 0**

From this, $\beta = 2 - 2H$ **and** $\sum_{k=-\infty}^{\infty} r(k) = \infty$ **(LRD)**

➔ **If** $\sum_{k=-\infty}^{\infty} r(k) < \infty$ **, SRD(Short Range Dependence)**

Long-range dependence vs. SS cont'd

- ***Self-similar*** **processes are the simplest way to model processes with** ***long-range dependence*** **– correlations that persist (do not degenerate) across large time scales**
- **Degree of self-similarity is expressed as the speed of decay of series autocorrelation function using the Hurst parameter**
  - **$H = 1 - \beta /2$**
  - **For SS series with LRD, $½ < H < 1$**
  - **Degree of SS and LRD increases as $H \rightarrow 1$**

Heavy-tailed distribution

- **Definition: A random variable Z has a heavy-tailed distribution if , $P(Z > x) \sim cx^{-\alpha}$, $x \rightarrow \infty$**

  where $0<\alpha<2$ **=** ***tail index*** **or** ***shape parameter***

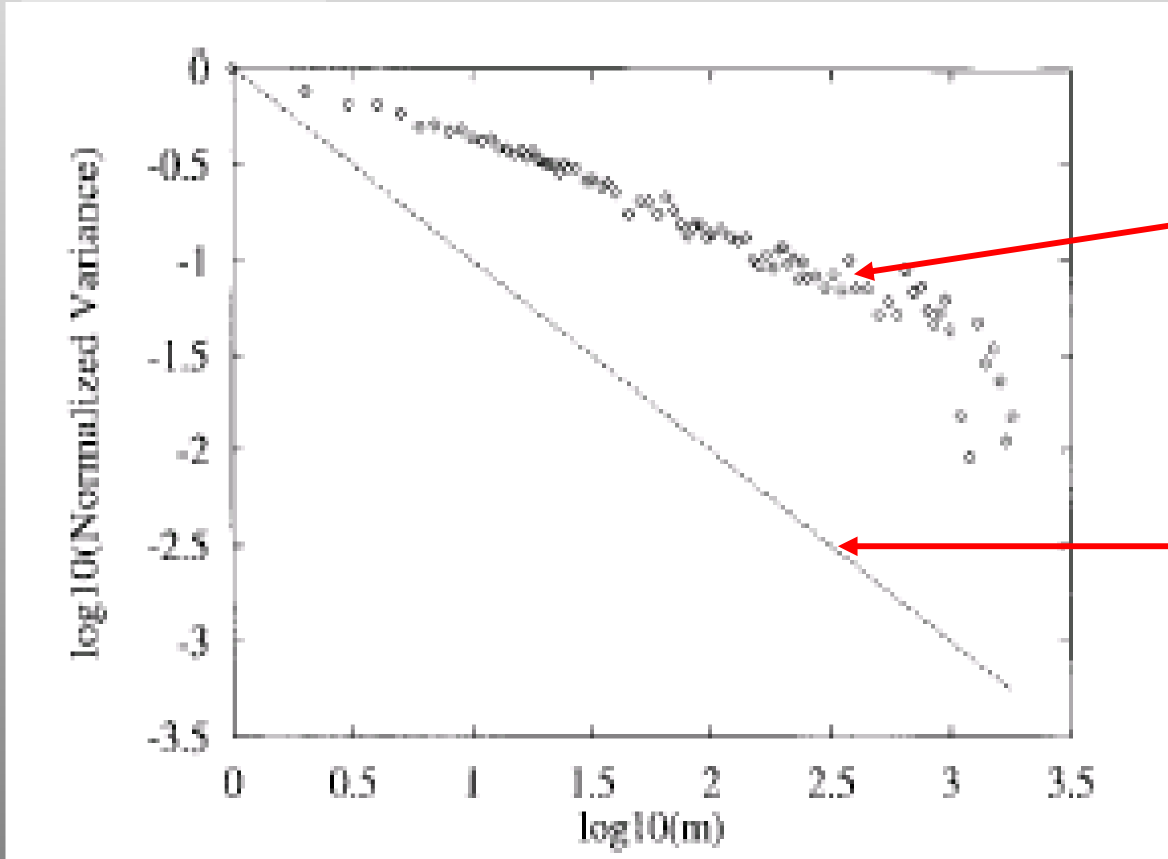
  c **= positive constant**
- **Tail of the distribution decays hyperbolically.**
- **Infinite variance for** $0<\alpha<2$
- **Unbounded mean for** $0 < \alpha \leq 1$
- **Frequently used heavy-tailed distribution is the Pareto distribution, whose distribution function is**

  $$P(Z \leq x) = 1 - \left(\frac{b}{x}\right)^{x} \quad , \quad b \leq x \quad where \;\; 0<\alpha<2$$
- ***Light-tailed distribution*****: exponential and Gaussian – which possess an exponentially decreasing tail.**

Graphical tests for self-similarity

- **Variance-time plots**
  - **Relies on slowly decaying variance of self-similar series**
  - **The variance of $X^{(m)}$ is plotted versus *m* on log-log plot**
  - **Slope ($-\beta$) greater than –1 is indicative of SS**
- **R/S plots**
  - **Relies on rescaled range (R/S)statistic growing like a power law with H as a function of number of points *n* plotted.**
  - **The plot of R/S versus *n* on log-log has slope which estimates H**
- **Periodogram plot**
  - **Relies on the slope of the power spectrum of the series as frequency approaches zero**
  - **The periodogram slope is a straight line with slope β – 1 close to the origin**

Graphical test examples – VT plot



slope = -0.48

then $\beta = 0.48$

Estimate H =

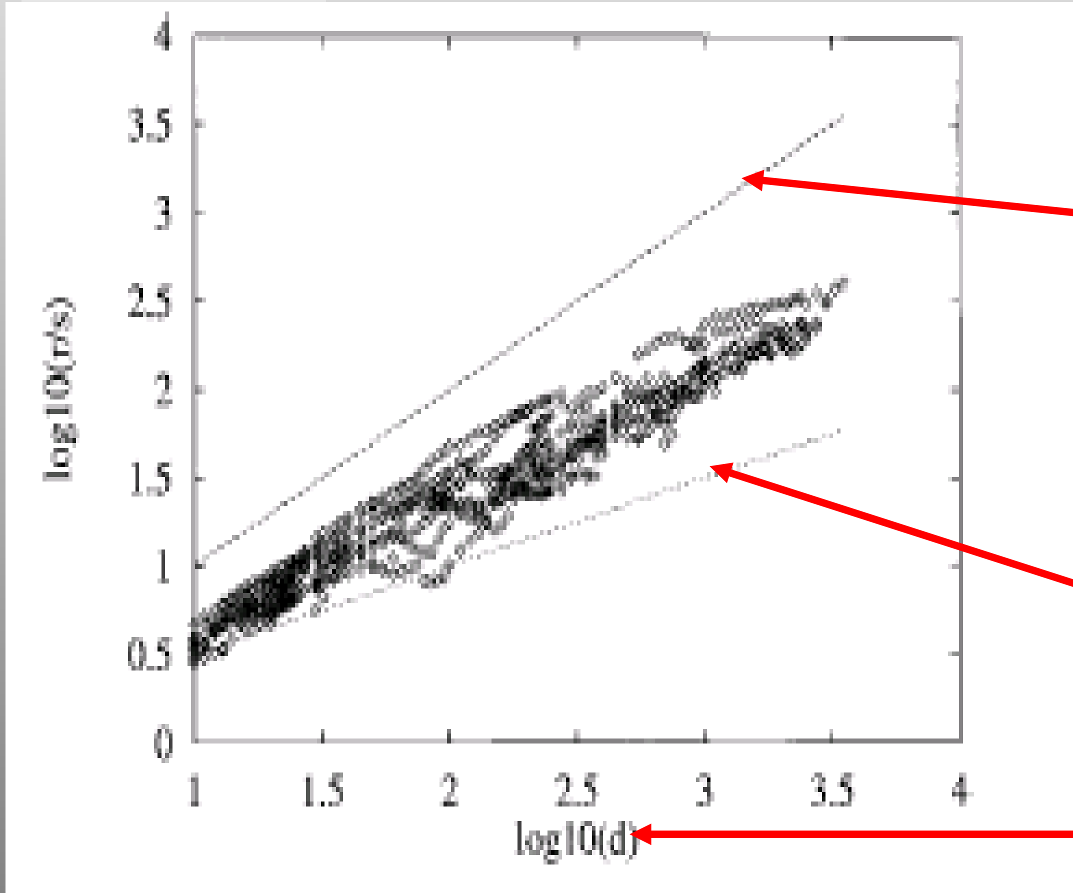
$1 - \beta/2 =$ **0.76**

$-\beta = -1$

-the variance of

$X^{(m)}$ is plotted against m on a log-log plot

Graphical test example – R/S plot



R = autocorrelation
S = variance

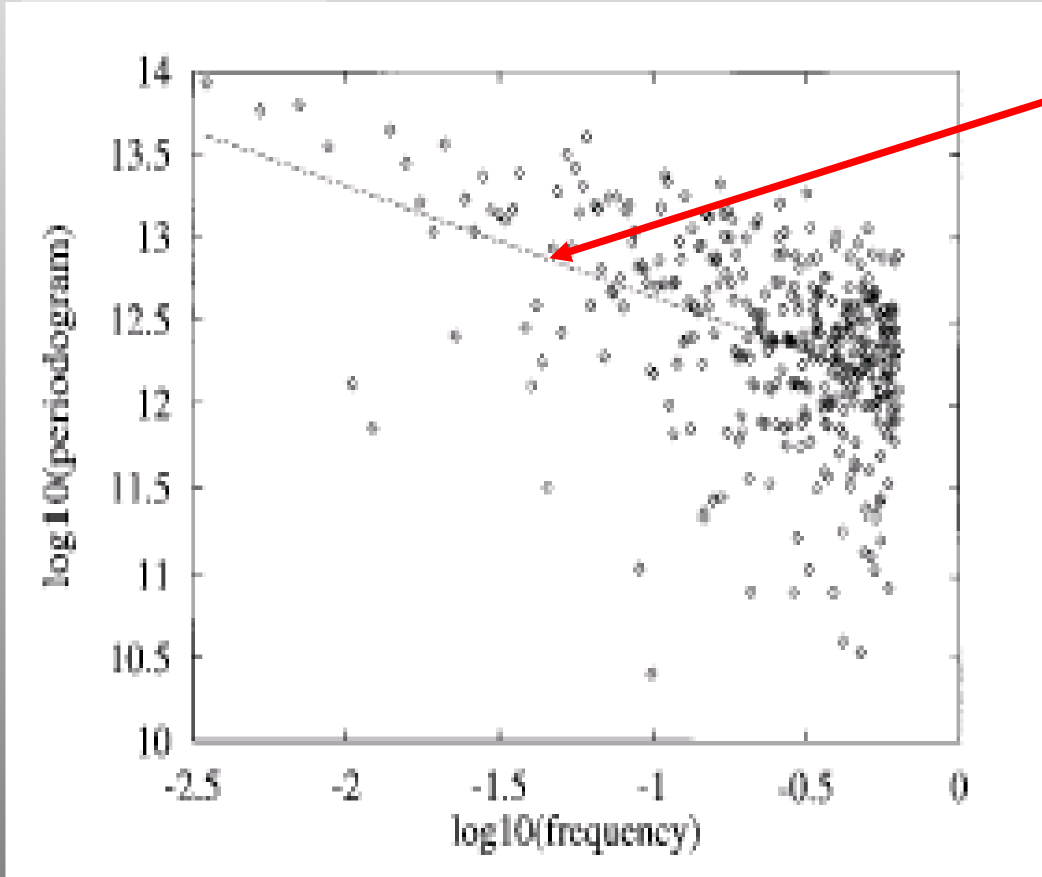
H = 1

Estimated H = **0.75**

H = 1/2

Degree of aggregation

Graphical test examples - Periodogram



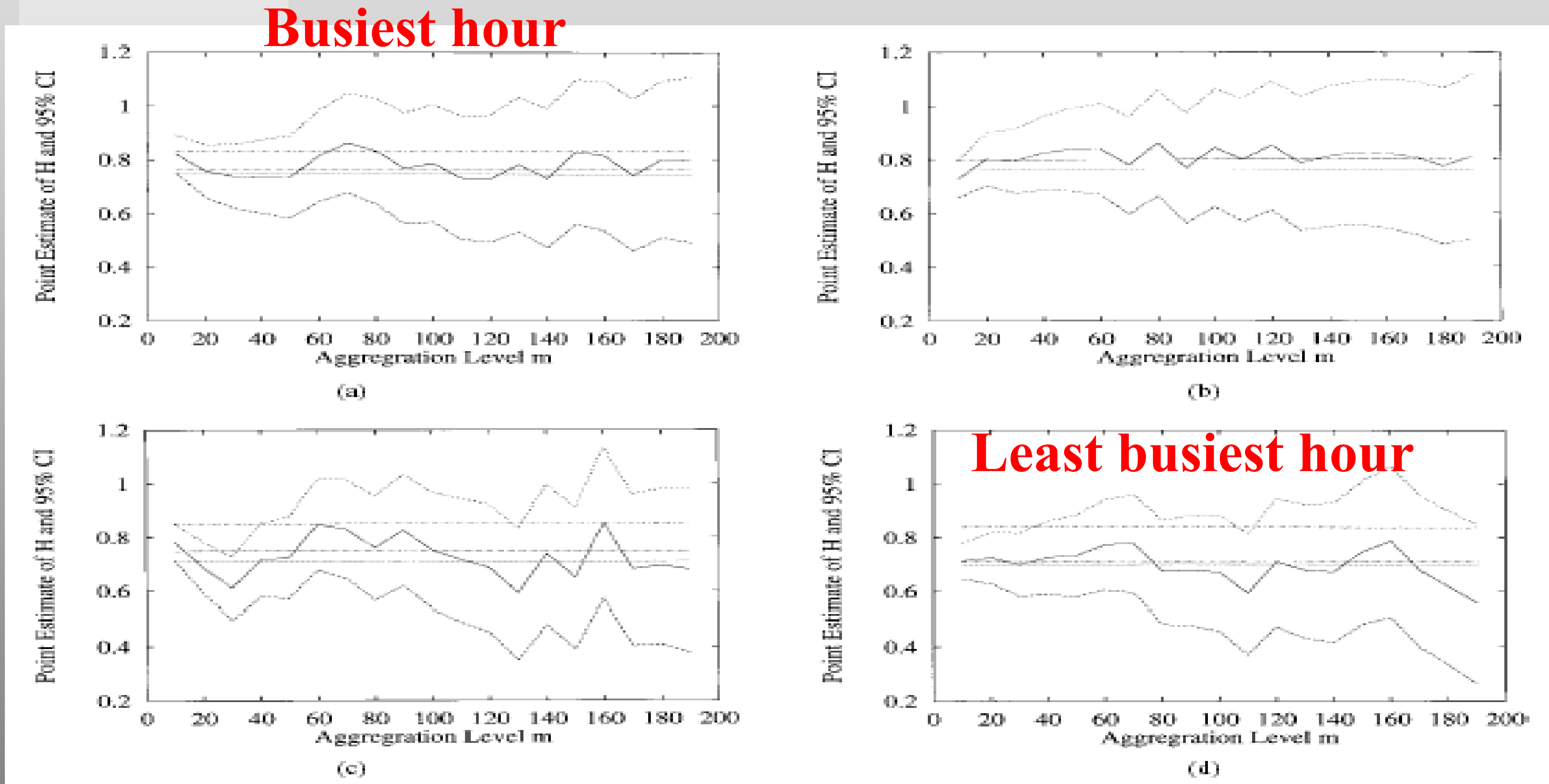
Slope = β-1 = 1-2H

In this case, slope = -0.66
then H = **0.83**

Non-graphical self-similarity test

- **Whittle's MLE Procedure**
  - **Provides confidence intervals for estimation of H (advantage)**
  - **Requires an underlying stochastic process for estimate (disadvantage)**
    - **Typical examples**
      - **FGN (Fractional Gaussian Noise) → exactly self-similar models**
      - **Fractional-ARIMA (Autoregressive Integrated Moving Average) → asymptotically self-similar models**
    - **FGN assumes no SRD(Short Range Dependence); however, F-ARIMA can assume a fixed degree of short-range dependence**

Non-graphical test example – Whittle estimator



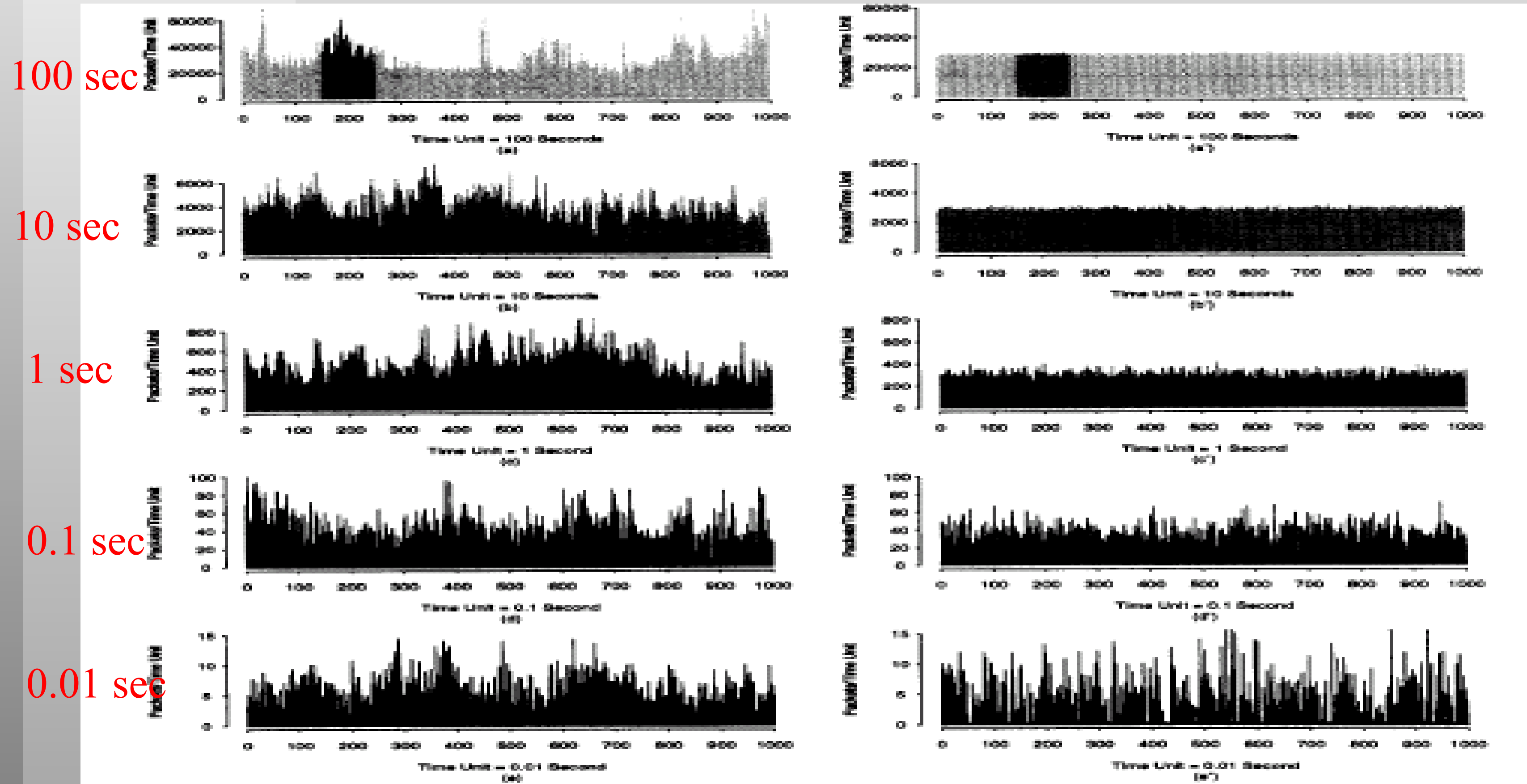
Analysis of Ethernet traffic

- **In 1989, Leland and Wilson begin taking high resolution traffic traces at Bellcore**
  - Ethernet traffic from a large research lab
  - 100 μ sec time stamps (update the version of monitor)
  - Packet length, status, 60 bytes of data
  - Mostly IP traffic (a little NFS)
  - Four data sets over three year period
  - Traces considered representative of normal use

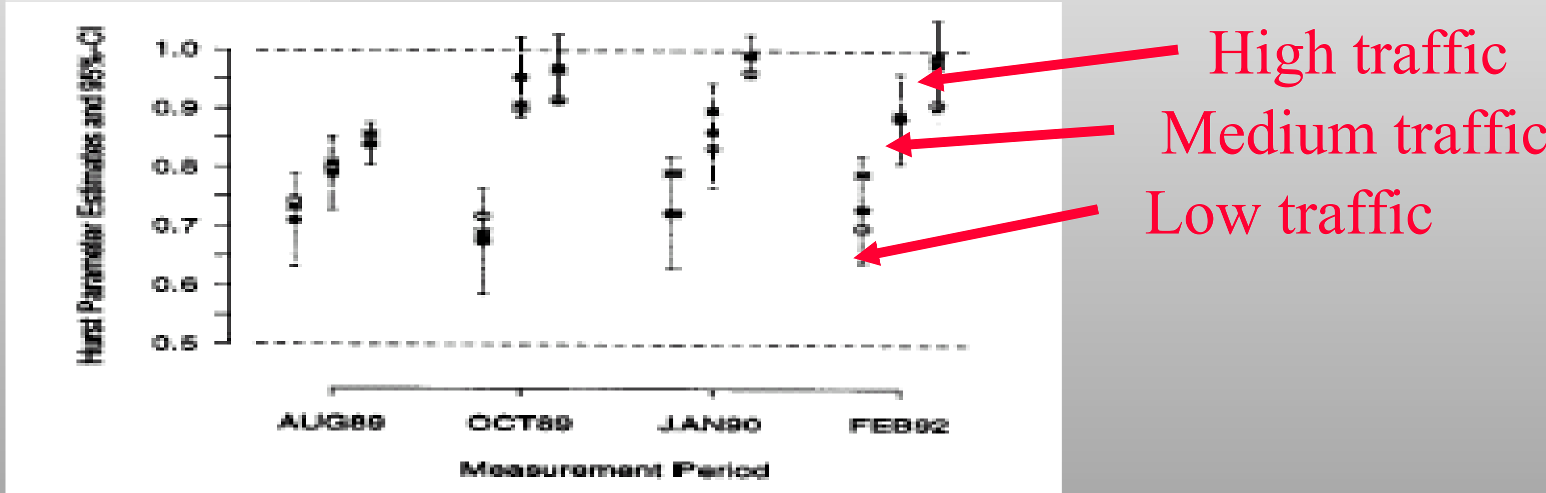
The packet count picture analysis

- A Poisson process
  - When observed on a fine time scale will appear bursty
  - When aggregated on a coarse time scale will flatten (smooth) to white noise
- A self-similar (fractal behavior) process
  - When aggregated over wide range of time scales will maintain its bursty characteristic

Pictorial proof of self-similarity (Ethernet Traffic)



Analysis of Ethernet traffic cont'd



- **Higher the load in the Ethernet traffic, higher the Hurst parameter**
- **Confidence Interval corresponding to H for the low traffic hours are typically wider than the normal and high traffic hours**
  - **Reason: Ethernet traffic during low traffic periods is asymptotically self-similar rather than exactly self-similar**

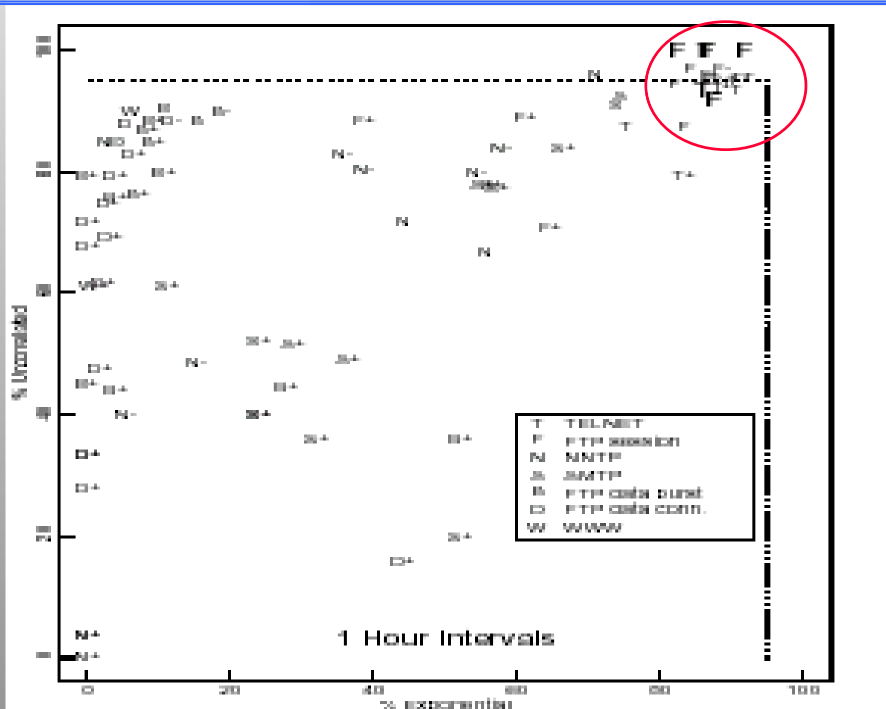
Major results of reference [1]

- **Analysis of traffic logs from perspective of packets/time unit found H to be between 0.8 and 0.95**
  - **Aggregation over many orders of magnitude**
  - **Initial looks at external traffic pointed to similar behavior**
- **First use of VERY large measurements in network research**
- **Very high degree of statistical rigor brought to bare on the problem**
- **Blew away prior notions of network traffic behavior**
  - **Ethernet packet traffic is self-similar**
- **Led to ON/OFF model of network traffic [WTSW97]**

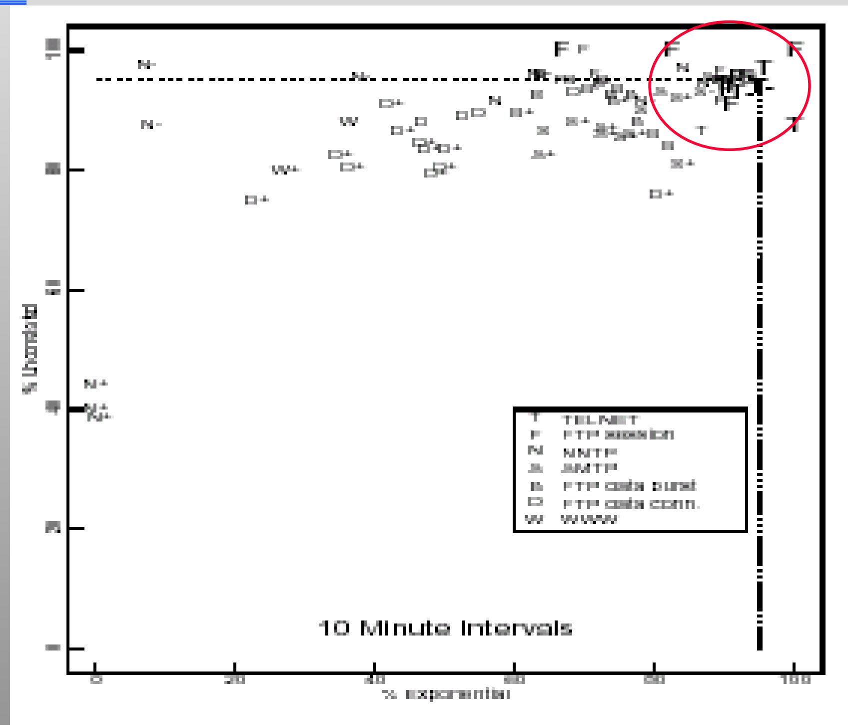
What about wide area traffic?

- Paxson and Floyd evaluated 24 wide-area traces
  - Traces included both Bellcore traces and five other sites taken between ’89 and ‘95
  - Focus was on both packet and session behavior
    - TELNET and FTP were applications considered
  - Millions of packets and sessions analyzed

Result of testing for Poisson arrivals



1 Hour Interval



10 Minute Interval

– **TELNET (T) and FTP connection(F) interarrivals are well modeled by a Poisson process**

TCP connection interarrivals

- The behavior analyzed was TCP connection start times
  - A simple statistical test was developed to assess accuracy of Poisson assumption
    - Exponential distribution of interarrivals
    - Independence of interarrivals
  - TELNET and FTP connection interarrivals are well modeled by a Poisson process
    - Evaluation over 1 hour and 10 minute periods
  - Other applications (NNTP, SMTP, WWW, FTP DATA) are not well modeled by Poisson

TELNET packet interarrivals

- The interarrival times of TELNET originator's packets (a user typing at a keyboard) was analyzed.
  - Process was shown to be heavy-tailed
    - $P[X > x] \sim x^{-\alpha}$ as x → inf. and $0 < \alpha < 2$
    - Simplest heavy-tailed distribution is the Pareto which is hyperbolic over its entire range
      - $p(x) = \alpha k^{\alpha} x^{-\alpha-1}$, α,k > 0, x >=k
      - If α =< 2, the distribution has infinite variance
      - If α =< 1, the distribution has infinite mean
      - It's all about the tail!
  - Variance-Time plots indicate self-similarity

TELNET session size (packets)

- **Size of TELNET session measured by number of originator packets transferred**
  - **Log-normal distribution was good model for session size in packets**
  - **Log-extreme has been used to model session size in bytes in prior work**
- **Putting this together with model for arrival processes results in a well fitting model for TELNET traffic**

FTPDATA analysis

- FTPDATA refers to data transferred after FTP session start
  - Packet arrivals within a connection are not treated
  - Spacing between DATA connections is shown to be heavy tailed
    - Bimodal (due to mget) and can be approximated by log-normal distribution
  - Bytes transferred
    - Very heavy tailed characteristic
    - Most bytes transferred are contained in a few transfers

Self-similarity of WAN traffic

- Variance-time plots for packet arrivals for all applications indicate WAN traffic is consistent with self-similarity
  - The authors were not able to develop a single Hurst parameter to characterize WAN traffic

The M/G/Inf. Model for generating self-similar traffic

- M/G/inf. Queue model considers customers that arrive at an infinite-server queue according to a Poisson process with rate ρ.
- In the count process $\{X_t\}_{t=0,1,2,\ldots}$ produced by M/G/Inf. Queue model, X_t gives the number of customers in the system at time t.
  - Reference: D. Cox and V. Isham, Point Processes, Chapman and Hall, 1980.

    shows: autocorrelation function *r(k)* for the count process is

$$r(k) = \operatorname{cov}\{X(t), X(t+k)\} = \rho \int_k^{\infty} (1 - F(x)dx$$

- If the service time has Pareto distribution with location parameter a and shape parameter β, for 1< β <2, then *r(k)* is the following:

$$r(k) = \rho \int_k^{\infty} \left(\frac{a}{x}\right)^{\beta} dx = \frac{\rho a^{\beta}}{\beta - 1} k^{(1-\beta)}$$

- Result: For Pareto service times and an arbitrary arrival rate ρ , the count process of the M/G/Inf. Model is ***asymptotically self-similar*** but **not exactly self-similar**.

major results of reference [2]

- **Verify that TCP *session* arrivals are well modeled by a Poisson process**
- **Showed that a number of WAN characteristics were well modeled by *heavy tailed* distributions**
- **Establish that *packet* arrival process for two typical applications (TELNET, FTP) as well as aggregate traffic is *self-similar***
- **Provide further statistical methods for generating self-similar traffic**

What about WWW traffic?

- Crovella and Bestavros analyze WWW logs collected at clients over a 1.5 month period
  - First WWW client study
  - Instrumented MOSAIC
    - ~600 students
    - ~130K files transferred
    - ~2.7GB data transferred

Self-similar aspects of Web traffic

- One difficulty in the analysis was finding stationary, busy periods
  - A number of candidate hours were found
- All four tests for self-similarity were employed
  - $0.7 < H < 0.8$

Explaining self-similarity

- Consider a set of processes which are either ON (transferring the data at *constant rate*) or OFF
  - The distribution of ON and OFF times are heavy tailed (α_1, α_2)
  - The aggregation of these processes leads to a self-similar process

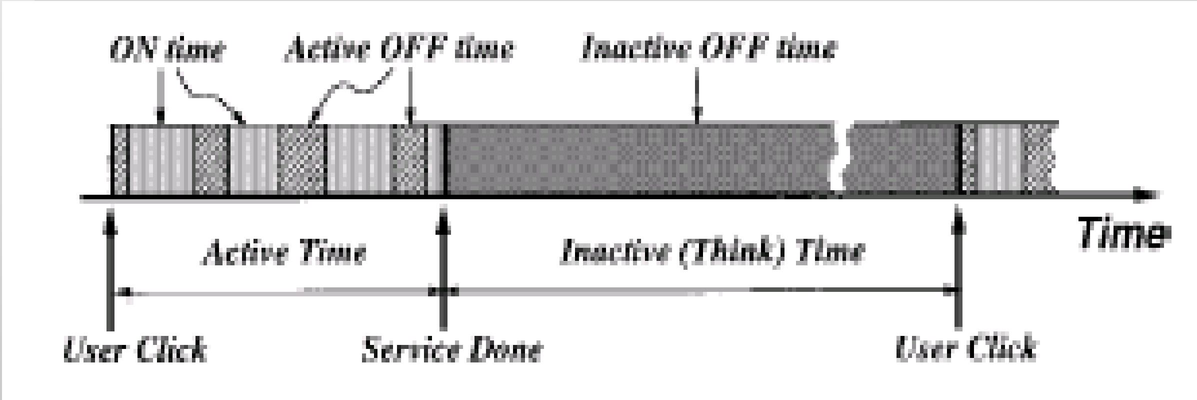
    ➔ $H = (3 - \min(\alpha_1, \alpha_2))/2$ [WTSW97]
- On-time: transmission duration of individual web-files
- Off-time: interval between transmission

Reference: **[WTSW97]➔** W. Willinger, M. S. Taqqu, R. Sherman, and D.V. Wilson, "self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM* Trans. Networking, vol 5.pp. 71-86 Feb, 1997.

Heavy tailed ON times and file sizes

- **Analysis of client logs showed that ON times were, in fact, heavy tailed**
  - $\alpha \sim 1.2$
- **This lead to the analysis of underlying file sizes**
  - $\alpha \sim 1.1$
  - **Similar to FTP data traffic where 0.9 <=** $\alpha <= 1.1$
- **Files available from UNIX file systems are typically heavy tailed**

Heavy tailed OFF times



- Analysis of OFF times showed that they are also heavy tailed; heavy-tailed nature of OFF time is a result of user think time (Inactive OFF) rather than machine-induced (Active OFF) delays
  - $\alpha \sim 1.5$
- Distinction between Active and Inactive(user think) OFF times
- ON times are more likely to be cause of self-similarity

Major results of reference [3]

- Established that WWW traffic was self-similar
- Modeled a number of different WWW characteristics (focus on the tail)
- Provide an explanation for self-similarity of WWW traffic based on underlying file size distribution

Where are we now?

- There is no mechanistic model for Internet traffic
  - Topology?
  - Routing?
- People want to blame the protocols for observed behavior
- Many people (vendors) chose to ignore self-similarity
- Lots of opportunity!!
- Current Research

1. G. Mansfield, T.K. Roy and N. Shiratori, "Self-similar and Fractal Nature of Internet Traffic Data", Infocom 2001.
2. B. Zwart, S. Bors, and M. Mandjes, "Exact queueing asymptotics for multiple heavy-tailed on-off flows," Infocom 2001.
3. C. Kotopoulos, N. Likhanov, and R.R. Maxumdar,"Asymptotic analysis of GPS systems fed by heterogeneous long-tailed sources," Infocom 2001.