Characteristics Analysis of Web Traffic with Hurst Index

Gagandeep Kaur, Dr. Vikas Saxena, Prof. J. P. Gupta

Abstract— **Today's computer networks are highly dynamic to change especially Internet. Simulating Internet behavior is therefore a difficult and cumbersome task. The web remains the largest application class on wide-area links. Generation of natural web traffic behavior is an essential component for simulation models. Failure of Poisson and Markov models has led to development of new models. In the past decade self-similarity has been proved to be an important feature in the network traffic. Hurst index is used to test for self-similarity of a process. Recently, PackMime has been introduced to NS-2 for HTTP traffic generation. In this paper statistical analysis method has been used to estimate the Hurst index. Simulations have been carried out using PackMimeHTTP. The results obtained are tested for self-similarity by finding the H index from the slope of the least-squares fitting line of the variance-time plot.**

Index Terms- **HTTP traffic, Hurst Index, Self-Similarity**

I. INTRODUCTION

The self similarity or long range dependence (LRD) of network traffic was found by Lelan et al. [1] in 1993. Since then a lot of studying has been done mainly in queuing performance by [2]. Both [1] and [2] argue against use of Markov based models for traffic engineering techniques and stress on the need to introduce self-similarity or fraction traffic in the models. The degree of self-similarity is measured with Hurst parameter, *H*. According to the properties of the self-similar process, there are three ways to estimate the *H* value: R/S statistic, Variance Time (VT) plots and periodogram analysis [1].

Networking research uses simulation as a very important tool for testing of the proposed algorithms and methods. Experimentation involves use of real or simulated network running varied applications like file transfer, web browsing, or peer-to-peer file sharing. Modeling the synthetic traffic for the network therefore plays a very crucial role to check the behavior of the applications.

The web remains a dominant application on wide area networks. Study by Sprint [3] found that on 16 of 19 OC-48 links traffic is mainly consisting of web traffic. It is therefore essential to study the characteristics of web traffic and see how it moves in the network. Recently, PackMime has been introduced in NS-2 for synthetic generation of HTTP 1.0 traffic in the network [4]. This paper tries to test for the self-similarity in the network traffic generated with PackMime by measuring *H* index and plotting VT chart.

The rest of the paper is organized as follows: Section II presents preliminaries. Section III gives the proposed study Section IV shows simulation results of the Hurst parameter estimation and conclusion in Section V.

Manuscript received on June 25, 2010.

Gagandeep Kaur is PhD student at CS&IT Deptt, Jaypee Institute of Information Technology,Noida,India (email:gagandeep.kaur@jiit.ac.in)

Dr. Vikas Saxena is Asstt. Professor at CS&IT Deptt, Jaypee Institute of Information Technology,Noida,India (email:vikas.saxena@jiit.ac.in)

Prof. J P Gupta is Vice Chancellor at Jaypee Institute of Information Technology,Noida,India

II. PRELIMINARIES

The flavor of self-similarity is that the same pattern is repeated at different levels of aggregation by time sequences. Therefore, if an object is self-similar, its parts, when magnified, resemble-in a suitable sense-the shape of the whole. All self-similar phenomena have two properties: a) There is structure at the smallest of smallest scales. b) The structure repeats. Therefore, a self-similar process contains replicas of itself at different scales.

Stochastic process is opposite of deterministic process. It says that even if the initial condition is known, there are many possibilities the process might go to, but some paths may be more probable than the others. Therefore, we need to look out for stochastic nature of network traffic and self-similarity property of it. Network traffic does not possess exact resemblance of their parts with the whole at finer details. If we adopt a view that traffic series are sample paths of stochastic processes and relax the measure of resemblance of the rescaled time series, then it may be possible to expect similarity of mathematical object.

Second order statistics are statistical properties that capture burstiness or variability, and the autocorrelation function is a yardstick with respect to which scale-invariance can be defined.

A. Hurst Parameter:

Hurst parameter is named after H. E. Hurst, a hydrologist who spent a lifetime studying Nile and other rivers and the problems related to water storage [5].

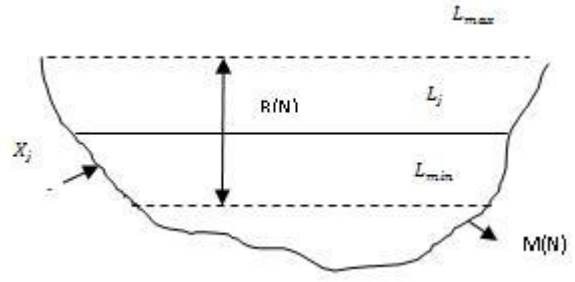

Figure 1: Hurst Analysis of Reservoir

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

In the Fig. [1] X_j means inflow during year j, M(N) means constant yearly outflow, L_j is reservoir level at the end of the year j. L_{max} and L_{min} represent the maximum and minimum levels. The range R(N)= $L_{max}(N)$ - L_{min}(N) = $\max_{1\le j\le N} X_j - \min_{1\le j\le N} X_j$. Hurst developed a normalized, dimensionless measure to characterize variability of the data:

$$R/S \sim (N/2)^H \quad H > 0.5$$

$$S = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(X_j - M(N))^2}$$

With a log-log plot, Hurst found that data points fell on a straight line and the slope of the line is *H*.

Recently in the study of Ethernet traffic by [1], Wide Area Network by [6], World Wide Web by [7], VBR video traffic by [8], the traffic was found to be self similar with a Hurst parameter *H* of 0.9. Farther experiments have also proved that if H>0.5 then the process is long range dependent.

B. Self-Similarity

Let $X = \{X(t), t \in Z\}$ be a wide-sense stationary process with constant mean μ, finite variance σ^2, and autocorrelation function $r(k)$ that depends only on $k, (k \in Z)$. Their definitions are given as follows:

$$\mu = E[X(t)], \quad \sigma^2 = E[(X(t) - \mu)]^2$$

$$r(k) = E[(X(t) - \mu)(X(t+k) - \mu)]/\sigma^2$$

Let $X^{(m)}$=$\{X^{(m)}(t), t \in Z_+\}$ denote the aggregate process of X at aggregation level $m(m \in Z_+)$. That is, for each m, $X^{(m)}$ is given by

$$X^{(m)}(t) = \frac{1}{m}\sum_{i=m(t-1)+1}^{mt} X(i), t \in Z_+$$

V^m and $r^{(m)}(k)$ denote the variance and autocorrelation function of $X^{(m)}$, respectively. The second-order stationary process can be categorized into the following:

a. A second-order stationary process *X* is called *exactly second-order self-similar* with $H = 1 - \beta/2$, if its autocorrelation function is

$$r(k) = g(k), k \in Z_+$$

where

$$g(k) \triangleq \frac{1}{2}[(k+1)^{2-\beta} - (2k)^{2-\beta} + (k-1)^{2-\beta}]$$

$$, 0 < \beta < 1, k \in Z_+$$

b. A second-order stationary process X is called *long range dependent* with $H = 1 - \beta/2, 0 < \beta < 1$, if its autocorrelation function satisfies

$$r(k) \sim ck^{-\beta}, k \to \infty$$

where c is a positive constant.

c. A second-order stationary process X is called *strong asymptotical second-order self-similar* with $H = 1 - \beta/2, 0 < \beta < 1$, if the variance of $X^{(m)}$ satisfies

$V^m \sim cm^{-\beta}$, m→ ∞

where c is a positive constant.

d. A second-order stationary process X is called *asymptotical second-order self-similar* with $H = 1 - \beta/2, 0 < \beta < 1$, if

$\lim_{m\to\infty} r^m(k) = g(k), k \in Z_+$

Based on the above equations we can say that process X is exactly second-order self-similar implies it is long range dependent which implies that it is strong asymptotical second-order self-similar. We can also say that a strong asymptotical second-order self-similar process further implies an asymptotical second-order self-similar process.

B.1 Characteristics of Self-Similar Process

A self-similar process follows following characteristics [1].

a. Long-range dependence: The autocorrelations decay hyperbolically rather than exponentially and $\sum_k r(k) = \infty$

b. *Hurst effect:* The *R/S* statistical analysis is characterized by a power law $E[\frac{R(n)}{S(n)}] \sim c_1 n^H$ as $n \to \infty$ with $0.5 < H < 1$.

c. *Slowly decaying variances:* The variances of the aggregate process are decaying more slowly than the reciprocal of the sample size, i.e $V^m \sim c_2 m^{2H-2}$ as m→ ∞ with $0.5 < H < 1$.

d. l/f--noise: The spectral density *f(.)* obeys a power law near the origin, i.e. $f(\lambda) \sim c_3 \lambda^{1-2H}$ as $\lambda \to 0$ with $0.5 < H < 1$, where $f(\lambda) = \sum_k r(k) e^{ik\lambda}$.

B.2 Aggregation of Self-Similar Processes

Let $X(t), t \in Z, i \in [1, N]$ be N uncorrelated stochastic process with Hurst parameter $H_i, i \in [1, N]$ for each process. Let $S_N(t) = \sum_{i=1}^{N} X_i(t), t \in Z$ denote the aggregate process of $X(t), t \in Z, i \in [1, N]$, and then we have three important characteristics of aggregated self-similar processes.

a. If $X(t), t \in Z, i \in [1, N]$ are exactly self-similar, $S_N(t)$ is exactly self-similar with Hurst parameter H if $H_1 = \cdots = H_N = H$, otherwise $S_N(t)$ is asymptotically self-similar with $H = \max(H_1 \ldots H_N)$.

b. If $X(t), t \in Z, i \in [1, N]$ are long range dependent, $S_N(t)$ is long range dependent with parameter $H = \max(H_1 \ldots H_N)$

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

c. If $X(t), t \in Z, i \in [1, N]$ are strong second-order self-similar, $S_N(t)$ is strong second-order self-similar with parameter $H = \max(H_1 \ldots H_N)$.

These theorems prove that the aggregation of multiple uncorrelated self-similar processes retains the self-similarity characteristic.

C. Hurst Parameter Estimation with V-T Plot

The relation between the variance of an aggregate process and the block size m is given by:

$$V^{(m)} \sim cm^{-\beta} 0 < \beta < 1 \text{ as } m \to \infty$$

By plotting the $\log(V^{(m)})$ versus $\log(m)$ and its least squares line, the slope of the fitting line is the estimate of Hurst parameter. Using the relationship $H = 1 - \beta/2$, H can be estimated. The initial block size and the number of groups must be large enough. Fig. [2] outlines the procedure for calculating H using variance-time (VT) plots.

1. Collect a sample of N observations .The size of N should be sufficiently large.
2. Divide the N observations into K non-overlapping blocks of size m, such that $X=(X_k(m), i = 1 \ldots K)$.
3. Calculate the independent mean of each block by
$$X_k^{(m)} = (X_{k1} + \cdots X_{k1+m-1})/m$$
4. Compute the mean for aggregated series
$$X^{(m)} = (X_k^{(m)}, k = 1 \ldots K)$$
5. Calculate the independent variance of $X^{(m)}$ by
$V_k^{(m)} = [X_k^{(m)} - \frac{1}{K}\sum_{k=1}^{K} X_k^{(m)}]^2$
Compute mean variance for the series by
$$V^{(m)} = [\frac{1}{K}(V_k^{(m)}), k = 1 \ldots K]$$
6. Repeat the steps 2,3,4,5 and 6 for different block sizes to get new values of $V^{(m)}$
7. Plot the logarithm of versus to get VT plot.
8. Using least squares fit plot the straight line.
9. Plot the values obtained in step 9 to get slope of line equal to 2-2H.

Figure 2: V-T plotting steps

III. PROPOSED STUDY

As stated earlier, the purpose of our study is to give characteristics analysis of aggregated HTTP traffic. Recently PackMime has been introduced to NS2 for generation of synthetic web traffic. In this paper we try to show self-similar nature of second order HTTP traffic by measuring the Hurst index. PackMime has been used for our work.

The network architecture considered is depicted in Fig. [3]. The network consists of two ns nodes "n0" and "n1" linked to each other by 10Mbps duplex link. Each node represents two object clouds namely client cloud and server cloud independent of each other. Each cloud handles multiple HTTP connections at a time. For each HTTP connection, PackMime-HTTP object creates a new web client and a new web server, sets up a TCP connection between the client and server, has the client sends an HTTP request, and sets a timer to expire when the next new connection should begin. The time between new connections is governed by the connection rate parameter "rate". New connections are started according to the connection arrival times without regard to the completion of previous requests, but a new request between the same client and server pair begins only after the previous request-response pair has been completed. Each web client controls the HTTP request sizes that are transferred and listens for the HTTP responses. Each web server listens for an HTTP request from its associated client and controls the response sizes that are transferred. The process is repeated until the requests are exhausted. Then the server sends a FIN.

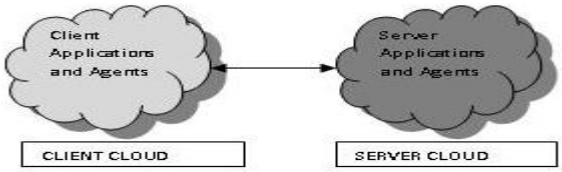

Figure 3: Network Architecture

We have carried out simulations for five minutes and collected data of approximately 5800 observations. Fig. [4] gives code snippet of the tcl script used in our test.

```
set CLIENT 0
set SERVER 1
remove-all-packet-headers;
add-packet-header IP TCP;
set ns [new Simulator];
.
// data storage allocation et al.//
.
set n(0) [$ns node]
set n(1) [$ns node]
$ns duplex-link $n(0) $n(1) 10Mb 0ms DropTail
# SETUP PACKMIME
set rate 15
set pm [new PackMimeHTTP]
$pm set-client $n(0);
$pm set-server $n(1);
$pm set-rate $rate;
$pm set-http-1.1;
.
\\ Set PackMime variables like flow arrival
rate, request size,response size,etc.\\
.
$ns at 0.0 "$pm start"
$ns at 300.0 "$pm stop"
$ns at 301.0 "exit 0"
$ns run
```

Figure 4: Sample Code

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

IV. RESULTS

Past decade has seen failure of Poisson and Markov models for network traffic modeling [2]. Self-similarity has come up as an important characteristic of network traffic [1],[6-8].The Hurst Index is used to measure degree of self-similarity in a process. We have collected aggregated HTTP traffic data of 5800 observations and analyzed it for the presence of self-similarity. The collected datum consists of client-server request size and response sizes. Fig. [5] shows the traffic trace for window size of 0.1 sec.

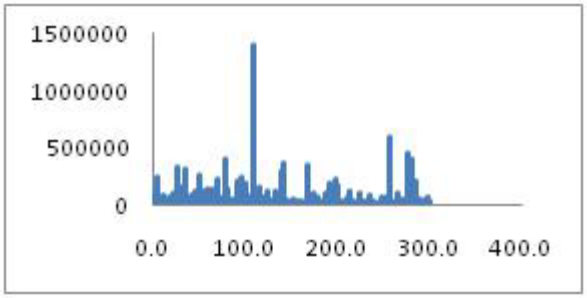

Figure 5: Traffic Traces (0.1s)

For measuring H index we had sample of N (N=5800(approx)) observations subdivided into K non-overlapping blocks with block size *m* respectively. Using the steps stated in Fig. [2], we calculated the individual means and then mean for the aggregated series. These values were used to compute sample individual variances and then aggregated variance was calculated. The corresponding $\log(V^{(m)})$ values are computed and listed in Table 1.

Table 1: log(V(m)) values versus block sizes

Block_size(m)	Block_num(K)	$\log(V^{(m)})$
15	389	6.219092
25	233	6.541934
50	116	6.821668
100	58	7.107834
125	46	7.248423
130	44	7.25801
150	38	7.317601
160	36	7.355102
200	29	7.454712
250	23	7.56699
500	11	7.897348
1000	5	8.348951

By using variance-time plots, we plotted the $\log(V^{(m)})$ versus $\log(m)$. Since the data values suffer from residual error therefore method of least squares fitting was used to trace the slope of the line in Fig [6].The slope of the fitting line is estimate of β. By using the equation $H = 1 - \beta/2$, H can be estimated, i.e. $\beta = 0.63$, H =0.68

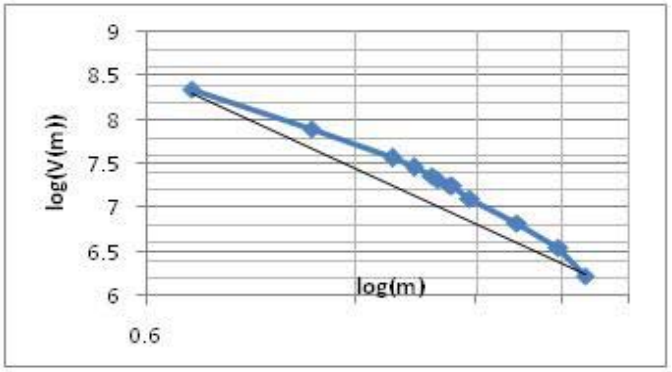

Figure 6: V-T plot and Least Squares Line Fit

V. CONCLUSIONS

The studies by [1], [6-8] have shown the network traffic to be self-similar with values of Hurst parameter *H* falling in the range of 0.5<H<1. As per section II self-similar process exhibits the characteristics of long range dependence, Hurst effect, slowly decaying variances, and the spectral density that follows the power law. Based on our findings presented in section IV and using the least squares fitting line for removing residual error we found value of H=.68. It is also visible from the values of Table 1 that the variance is slowly decaying instead of exponential decay. Therefore, we can conclude that the HTTP traffic follows the property of self-similarity at different scales.

Today's research in the areas of computer networks ranging from network traffic behaviors, network modeling, intrusion detection systems to quality of service for the end users largely depends on the study of traffic synthesizers. The better and efficient design of these models therefore asks for a look into changing trends in the traffic. In this paper we have shown the presence of self-similarity in the HTTP traffic. This result is of importance while applying wavelets on web traffic for designing intrusion detection systems.

REFERENCES

[1] W. E. Leland, M. *S.* Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethemet traffic (extended version)," *IEEEACM Trans. Networking, vol. 2, pp. 1-15, 1994.*

[2] K. Park and W. Willinger, " Self-Similar Network Traffic and Performance Evaluation", New York: Wiley Interscience, 2000.

[3] S. Bhattacharyya, C. Diot, R. Gass, E. Kress, S. Moon, A. Nucci, D. Papagiannaki, and T. Ye, "Packet Trace Analysis", Sprint Labs, Tech. Rep., 2003, ipmon.sprintlabs.com.

[4] J. Cao, W. S. Cleveland, Y. Gao, K. Jeffay, F. D. Smith, and M. C. Weigle, "Stochastic Models for Generating Synthetic HTTP Source Traffic", *Proc. IEEE INFOCOM*, pp. 1547-1558, 2004.

[5] H. Hurst, "Long-Term Storage of Reservoirs: An Experimental Study", *Trans. of the American Society of Civil Engineers*, pp. 770-799, 1951.

[6] Paxson and *S.* Floyd, "Wide Area Traffic: A Failure of Poison Modeling", *IEEEACM Trans. Networking*, pp. 226-244, 1995.

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[7] M. Crovella and A. Bestavros. “Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes”. *IEEEACM Trans. on Networking*, pp. 160-169, 1996.
[8] M. Garrett and W. Willinger, “Analysis, modeling and generation of self-similar VBR video traffic”. *Proc. ACM SIGCOMM*, pp. 269-280, 1994.

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)