QUEUE LENGTH-BUSY TIME DISTRIBUTION OF WEB USERS DATA WITH SELF SIMILAR BEHAVIOR

Pushpalatha Sarla[1], D.Mallikarjuna Reddy[2], Manohar Dingari[3]

[1]*Department of Engg.Mathematics, School of Technology, GITAM University, Hyderabad, India 502329*
[2]*Department of Engg.Mathematics, School of Technology, GITAM University, Hyderabad, India 502329*
[3]*Department of Engg.Mathematics, School of Technology, GITAM University, Hyderabad, India 502329*

Abstract

It has been reported that Internet traffic exhibits self-similarity. Motivated by this fact, real time web users at different web centers can be treated as arrivals consider as traffic and has been examined by various techniques to test the self-similarity. The outcome from the experiments carried out and proven that arrival pattern has self similar behavior. In this paper, various techniques used to compute Hurst index which is a measure to know the intensity of self-similarity. In addition to this, the mean queue length and busy time distribution at a web center has been computed against traffic intensity. Numerical results clearly reveal that analysis accessible in this paper is very helpful in improvement of designs of web centers to give quality of service (QOS) guarantee.

Keywords—*self similarity, Hurst index, auto correlation function, queue length, traffic intensity.*

----------***----------

1. INTRODUCTION

Now a day's arrival pattern at different web centers is one of the fundamental issues in planning and designing. One of the major issues to know various traffic flows are in self similar nature to study and design some performance measures as that of Ethernet traffic, mobile networking etc. Until recently Poison approach has been used to model the road traffic irrespective of traffic intensity. This was similar to the practice in the cases of Ethernet, LAN, WAN, and WWW traffic. But seminal studies [1-3] reveal that IP packet traffic in supposed networks tends to be burst in nature on many time scales. This burstyness of traffic can be characterized mathematically as long-range dependence (LRD) or self-similar. It is clear from the work agreed [4] that Poisson process could not emulate the self-similar network traffic. Markovian arrival process (MAP) emulating self-similar traffic is fitted over desired time scales by equating descriptive statistics measures such second-order statistics of the counts [5-8]. The theme of the paper is, we examined the nature of real time web users traffic data is self-similar [20] and this is an enhancement to study the performance metrics such as mean length and busy time distribution against the traffic intensity. This kind of research is useful for future studies to know the performance analysis and continuous improvement of web centers. The rest of the paper has been organized as follows: mathematical definitions of self-similarity or long range dependence is given in Section II. Materials and methods is placed in Section III. In Section IV, Hurst index is computed using various techniques. Queue length and busy time distribution is discussed in Section V. Finally, conclusions are given in Section VI.

2. SELF-SIMILARITY AND LONG-RANGE DEPENDENCE

The definition of exact second-order self-similar process is given as follows. Arrival instants are modeled as point process. Divide the time axis into disjoint intervals of unit length and let $X = \{X_t : t = 1, 2, ...\}$ be the number of points (arrival) in the t^{th} interval. Let X be a second order stationary process with variance σ^2 and the autocorrelation function $\gamma(k), k \geq 0$ is given by

$$\gamma(k) = \frac{Cov(X_t, X_{t+k})}{Var(X_t)} \quad (1)$$

For each $m = 1,2,3,....,$ let a new time series $X_t^{(m)}$ is obtained averaging the original time series X over non-overlapping blocks of size m. That is

$$X_t^{(m)} = \frac{1}{m}\sum_{i=1}^{m} X_{(t-1)m+i}, \; t = 1, 2, ... \quad (2)$$

This new series $X_t^{(m)}$, for each m, is also a second order stationary process with autocorrelation function $\gamma^{(m)}(k)$.

Definition -1

The process $'X'$ is said to be exactly second order self-similar with Hurst parameter $H = 1 - \frac{\beta}{2}$ and variance σ^2 if

$$\gamma(k) = \frac{\sigma^2}{2}\left[(k+1)^{2H} - 2k^H + (k+1)^{2H}\right], \forall k \geq 1 \quad (3)$$

Definition-2

The process $'X'$ is said to be asymptotically second order self-similar with Hurst parameter $H = 1 - \frac{\beta}{2}$ and variance σ^2 if

$$\sum_{m\to\infty} \gamma^{(m)}(k) = \frac{\sigma^2}{2}\left[(k+1)^{2H} - 2k^H + (k+1)^{2H}\right], \forall k \geq 1 \quad (4)$$

In variance terms, self- similar process is defined as follows:

Definition-3

The process $'X'$ is said to be exactly second order self-similar with Hurst parameter $H = 1 - \frac{\beta}{2}$ and variance σ^2 if

$$Var(X^{(m)}) = \sigma^2 m^{-\beta}, \forall m \geq 1 \quad (5)$$

Now we shall differentiate long range dependence (LRD) and short range dependence (SRD) processes. For $H \neq 0.5$, from the Eq. (3), we can see that $\gamma(k) = H(2H-1)k^{2H-2}$ as $k \to \infty$, and we have

$$\sum_k \gamma(k) \sim c\sum_k k^{-\beta}, \; c = H(2H-1). \quad (6)$$

The series $c\sum_k k^{-\beta}$ is divergent if $0.5 < H < 1$ or $0 < \beta < 1$ otherwise they are convergent, being a positive term series. Accordingly the left hand series $\sum_k \gamma(k)$ is divergent if $0.5 < H < 1$ or $0 < \beta < 1$, otherwise they are convergent. That is, for $0.5 < H < 1$, the autocorrelation functions decays slowly, that is hyperbolically. In this case, the process X is called LRD. The process X is SRD if $0 < H < 0.5$ and the autocorrelation function is summable.

3. MATERIALS AND METHODS

As discussed in the introduction, we are primarily interested collecting data from various sources. Real time web users data has been considered. The sample number of users logged on to an Internet server each minute over 100-minutes . In the study web users data can be treated as traffic and verified that the said traffic is satisfying self similar characteristics. Here, we investigated few of performance metrics such as mean length and busy time distribution against the traffic intensity etc.

4. TECHNIQUES FOR MEASURING HURST INDEX OF SELF-SIMILAR PROCESS

The intensity of self-similarity is given by Hurst parameter H. The parameter H was named after the hydrologist H.E. Hurst who spent many years to investigate the problem of water storage and also to determine the level patterns of the Nile River. Hurst parameter is perfectly well defined mathematically, measuring if it is a problematic one.

The data must be measured at high lags or low frequencies where fewer readings are available. The parameter H has range $0.5 \leq H \leq 1$. Estimation of H is a difficult task. Several methods are available to estimate degree of self-similarity in a time-series. We also present the four techniques to calculate the Hurst index: Periodogram analysis, Correlogram method, Variance-time analysis and another method based on percentiles is applied and then compare with the said methods.

4.1 Periodogram Analysis

In the frequency domain, analysis of time series [10] is merely the analysis of a stationary process by means of its spectral representation. The periodogram is given by

$$I_N(\lambda) = \frac{1}{2\pi N}\left|\sum_{k=0}^{N-1} X_k e^{k\lambda}\right|^2 \quad (7)$$

where λ is the Fourier frequency, N is the number of terms in the time series and X_j is the data of the given series. To estimate H, first, one has to calculate this periodogram. Since $I_N(\lambda)$ is an good sample estimator of the spectral density, a series with long-range dependence should have a periodogram, which is proportional to $|\lambda|^{1-2H}$ close to the origin. Then a regression of the logarithm of the periodogram on the logarithm of the frequency λ should give a coefficient of $1-2H$. The slope of the fitted straight line is the estimate of $1-2H$. Using this method the H value is computed for the data given in section III. The obtained value of H in this case is 0.763.

4.2 Correlogram Method

In time series analysis, plot of ACF (autocorrelation function) is known as correlogram where the estimated correlation can be given in terms of auto-covariance function $\gamma(k)$

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} \quad (8)$$

It has already been observed that slow decay of correlation, which is proportional to K^{2H-2} for $\frac{1}{2} < H < 1$ indicates the long-memory process [11]. Therefore, the plot of the sample autocorrelation should exhibit this property. A much better plot for the handling of long-range dependence is the plot of ACF in logarithmic scale. If the asymptotic decay of the correlation is hyperbolic, then the points in the plot should be approximately scattered around a straight line with a negative slope of $2H-2$ for the long memory processes but for short memory, the points should tend to diverse to minus infinity at an exponential rate. If the time series is long enough or if the series has strong long-range dependence, then this log-log correlogram is useful. Correlogram is useful as a preliminary heuristic approach to the data. Some pitfalls of sample correlation which are less known can be found in Mandelbrot [12, 17]. Even though it is neither widely used nor attractive method for estimation, still H, the self-similarity parameter, can be estimated by this method deriving an equation of the form

$$\rho(k) = \hat{H}(2\hat{H}-1)K^{2\hat{H}-2} \quad (9)$$

Using this method, the obtained value of H in this case is 0.79.

4.3 Variance-Time Analysis

This method, variance time analysis [18] is very popular and is based on property of slowly decaying variance of self-similar processes undergoing aggregation. The m -averaged process $X^{(m)} = (X_1^{(m)}, X_2^{(m)}, ...)$ of a discrete-time stationary parent process $X_1, X_2, ...$ as:

$$X_j^{(m)} = \frac{1}{m}\sum_{(j-1)m+1}^{m} X_i, \quad j = 1, 2, ..., \frac{N}{m} \quad (10)$$

Where m and j are positive integers. The variance is defined as:

$$Var[X^{(m)}] = \frac{1}{N/m}\sum_{j=1}^{m}(X_j - \bar{X})^2 \quad (11)$$

The variances of the aggregated processes $X^{(m)}(m = 1, 2, 3, ...)$ decrease linearly (for large m):

$$Var[X^{(m)}] = Var[X]m^{\beta} \quad (12)$$

The variance-time plot is obtained by plotting $\log Var[X^{(m)}]$ against $\log m$ and by fitting a sample least squares line through the resulting points in the plane, ignoring the small values for m . Values of the estimate ß of the asymptotic slope between –1 and 0 suggest self-similarity and an estimate for the degree of self-similarity is given by

$$H = 1 - \frac{\beta}{2} \quad (13)$$

using this method, the obtained value of H in this case is 0.761.

4.4 Percentile Method

A percentile is the value of a variable below which a certain percent of observations fall, like partition values of a process such as quartiles and deciles. There is no exact definition for the percentile [13], however all definitions yield similar results when the number of observations is very large. One definition of percentile, often given in texts, is that the P^{th} percentile $(1 \le P \le 100)$ of N ordered values is obtained by first calculating the rank.

$$n = \frac{P * N}{100} + \frac{1}{2} \quad (14)$$

Given data set or time series $(t, Z_t)\ (t \ge 0)$. First we can find the percentiles $(P_i,\ i = 1, 2, ...100)$ for a given time series or real time data using

$$P_i = \frac{i * N}{100} + \frac{1}{2};\ i = 1, 2, ..., 100. \quad (15)$$

$P_i = i^{th}$ percentile, this a special type of average such as partition values in descriptive statistics like quartiles (Q_1, Q_2, Q_3). Draw a scattered Plot percentile number against percentiles on log scales. A linear equation $Z_t = \beta t + c$ (say) is obtained with the slope (β) . The Hurst parameter (H) is then computed by Eq. (13). Using this method, the H value is computed for the data. The pertaining scattered data and trend line with the slope $(\beta) = 0.476$. The obtained value of H in this case is 0.762. One paper [14], explained how the 95-percentile depends on the aggregation window size, and how this phenomenon justifies the mathematical definition of self similarity or LRD. The advantages of this method are: This method is matter of a simple empirical formula, unlike other two methods. Data however large it may be is divided into hundred parts (partition values) and the plotting involves only 100 points (percentile versus percentile number).

5. QUEUE LENGTH - BUSY TIME DISTRIBUTION

The section presents some practical formulae, mean queue length ($\bar{L}$) and time length (busy period) (T) distribution, by model, the web users traffic at web centers as queuing

system with self similar input traffic. Here the numerical results discussed with two performance measures. Suppose that queue Length is L and the average queue length of the distribution is $\bar{L}$ against traffic intensity. For this, we use the formula [14] given under:

$$\bar{L} = \frac{\rho^{\frac{0.5}{1-H}}}{(1-\rho)^{\frac{H}{1-H}}} \qquad (16)$$

In the eq. (13), ρ is traffic intensity. Results are depicted in Fig. 1. From this figure, we conclude that as ρ increases mean queue length increases which is expected. Further, as H increases, mean queue length increases. This tendency agrees with our intuition.

The busy period distribution for bulky queues [19] is approximated by using large deviation techniques and is given by

$$P(T > t) \approx \exp(-T^{(2-2H)/2}) \qquad (17)$$

The busy period distribution is worked out for four values of H and the results are depicted in Fig. 2. This figure clearly shows that when H is higher the busy period distribution will be higher.

Table 1

Minute	*Web Users*	*Minute*	*Web Users*	*Minute*	*Web Users*	*Minute*	*Web Users*
1	88	26	139	51	172	76	91
2	84	27	147	52	172	77	91
3	85	28	150	53	174	78	94
4	85	29	148	54	174	79	101
5	84	30	145	55	169	80	110
6	85	31	140	56	165	81	121
7	83	32	134	57	156	82	135
8	85	33	131	58	142	83	145
9	88	34	131	59	131	84	149
10	89	35	129	60	121	85	156
11	91	36	126	61	112	86	155
12	99	37	126	62	104	87	171
13	104	38	132	63	102	88	175
14	112	39	137	64	99	89	177
15	126	40	140	65	99	90	182
16	138	41	142	66	95	91	193
17	146	42	150	67	98	92	204
18	151	43	159	68	84	93	208
19	150	44	167	69	84	94	210
20	148	45	170	70	87	95	215
21	147	46	171	71	89	96	222
22	149	47	172	72	88	97	228
23	143	48	172	73	85	98	226
24	132	49	174	74	86	99	222
25	131	50	175	75	89	100	220

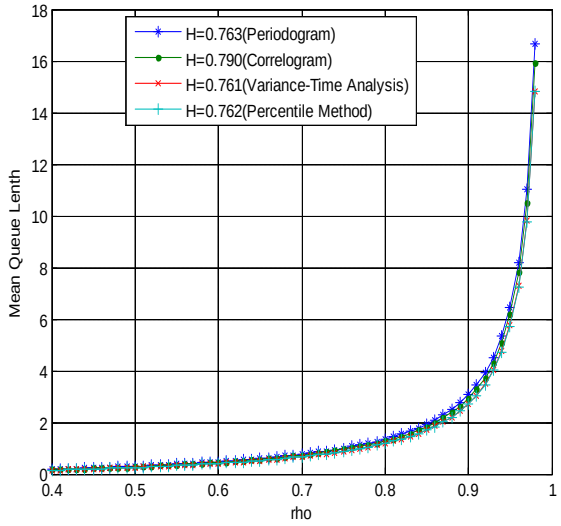

Fig 1: Mean Queue Length Vs Traffic Intensity (rho).

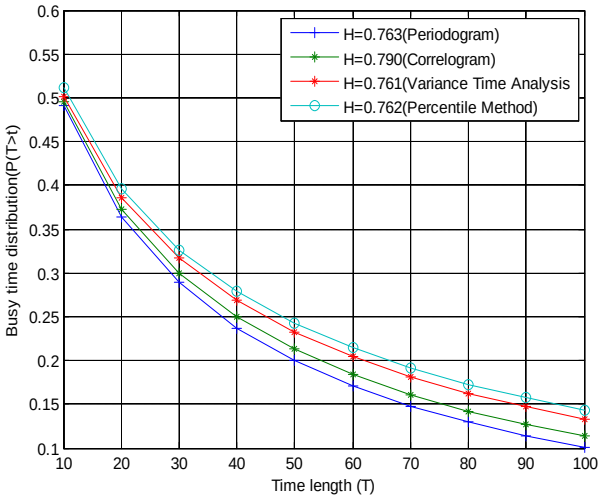

Fig 2: Mean Queue Length Vs Traffic Intensity (rho).

6. CONCLUSION

In this paper, real time web users at different web centers can be treated as arrivals and consider as traffic has been proven to be self-similar. Various techniques have been discussed and applied to test the self-similar behavior. The obtained values of the Hurst parameter are reasonably close to each other. Mean queue length and busy period distribution are computed. Numerical results reveal that the mean queue length increases as ρ and H increase, and the busy time distribution is found to be higher if value of H is high. The first measure, mean queue length against traffic intensity, can be used to determine the optimal number of web centers; whereas another metric, busy time distribution can be used to know the congestion problem like things. Based on this analysis, web users can change their decision over the busy period. This kind of analysis in computer science engineering is very useful to give a dimension for the improvement of and designing web centers.

REFERENCES

[1] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson On the Self-Similar Nature of Ethernet Traffic (Extended version) , IEEE / ACM Trans. Networking, 2, pp. 1-15, 1994.

[2] M. Crovella and A.Bestavros, Self-Similarity in World Wide Web traffic: evidence and possible causes, IEEE/ ACM Trans. Networking, pp.835-846, 1997.

[3] Vern Paxson, Sally Floyd, Wide Area Traffic: The Failure of Poisson Modeling, IEEE/ACM Trans. Networking, 3, pp.226-244, 1995.

[4] Allan T. Anderson, Bo Friis Nielsen , A Markovian Approach for Modeling Packet Traffic with Long Range Dependence, IEEE Journal on Selected Areas in Communications, Vol.16, No.5, pp.719-732, June 1998.

[5] T. Yoshihara, S.Kasahara, and Y. Takahashi, Practical time-scale fitting of self-similar traffic with Markov-modulated Poisson process, Telecommun. Syst., vol.17, pp.185-211, 2001.

[6] S.K. Shao, Malla Reddy Perati, M.G. Tsai, H.W. Tsao and J. Wu, Generalized variance-based Markovian fitting for self-similar traffic modeling, IEICE Trans. Commun., Vol.E88-B, no.12, pp.4659-4663, April 2005.

[7] D.Mallikarjuna Reddy "Second Order Statistics of Time Series of Various Real Time Problems in Conjunction with Periodogram Technique" International Journal of Latest Trends in Engineering and Technology (IJLTET) Vol. 3 Issue 1 September 2013.

[8] Nagatani, T. (2005). Self-similar behavior of a single vehicle through periodic traffic lights, Physica A, 347, 673–682.

[9] Qiang Meng and Hooi Ling Khoo, Self-similar characteristics of vehicle arrival pattern on Highways. Journal of Transportation Engineering, © ASCE / November 2009.

[10] P. J. Brockwell and R. A. Davis, \An introduction to time series and fore-casting", Springer - Verlag, New York (1996).

[11] M. M. A. Sarker, Estimation of the Self-similarity parameter in long memory processes, Journal of Mechanical Engineering, Vol. ME38, Dec. 2007, Transaction of the Mech. Eng. Div., The Institution of Engineers, Bangladesh.

[12] J.Beran, M. S. Taqqu and W. Willinger, Long- range dependence in variable bit rate traffic," IEEE Trans. on Communications, Vol. 43, pp.1566-1579

[13] Lane, David. "Percentiles". http://cnx.org/content/m10805/latest. Retrieved 2007-09-15.

[14] Web hosting talk Forum: 95th Percentile billing polling interval, http://www.webhostingtalk.com, Last accessed 09/23/2008.

[15] Beran J., Statistics for Long-Memory Processes, Chapman and Hall, 1994.

[16] Spyros Markidakis, steven C.Wheelwright, Rob J.Hydman "Forcating Methods and Applications" John Wiley & Sons.Inc. Third edition".
[17] Mallikarjuna Reddy Doodipala , Malla Reddy Perati K. Raghavendra; H. K. Reddy Koppula; Rajaiah Dasari 'Self-Similar Behavior of Highway Road Traffic and Performance Analysis at Toll Plazas" 1234 Journal Of Transportation Engineering © Asce October 2012.
[18] Mitko Gospodinov, Evgeniya Gospodinova "The graphical methods for estimating Hurst parameter of self-similar network traffic" International Conference on Computer Systems and Technologies - CompSysTech' 2005.
[19] K. Park, and W. Willinger, eds. (2000). Self-similar network traffic and
[20] performance evaluation, Wiley, New York.
[21] Pushpalatha Sarla, D.Mallikarjuna Reddy, Manohar Dingari, "A Study on Self Similarity Analysis of Web Users Data at Selected Web Centers" Proceedings of International conference on Mathematics ICM- 2015.