A Contribution Towards Solving the Web Workload Puzzle

Katerina Goševa-Popstojanova, Fengbin Li, Xuan Wang, and Amit Sangle
Lane Department of Computer Science and Electrical Engineering
West Virginia University, Morgantown, WV 26506-6109
{katerina, fengbinl, xwang, sangle}@csee.wvu.edu

Abstract

World Wide Web, the biggest distributed system ever built, experiences tremendous growth and change in Web sites, users, and technology. A realistic and accurate characterization of Web workload is the first, fundamental step in areas such as performance analysis and prediction, capacity planning, and admission control. Compared to the previous work, in this paper we present more detailed and rigorous statistical analysis of both request and session level characteristics of Web workload based on empirical data extracted from actual logs of four Web servers. Our analysis is focused on exploring phenomena such as self-similarity, long-range dependence, and heavy-tailed distributions. Identification of these phenomena in real data is a challenging task since the existing methods may perform erratically in practice and produce misleading results. We provide more accurate analysis of long-range dependence of the request and session arrival processes by removing the trend and periodicity. In addition to the session arrival process (i.e., inter-session characteristics), we study several intra-session characteristics using several different methods to test the existence of heavy-tailed behavior and cross validate the results. Finally, we point out specific problems associated with the methods used for establishing long-range dependence and heavy-tailed behavior of Web workloads. We believe that the comprehensive model presented in this paper is a step towards solving the Web workload puzzle.

1 Introduction

The growing availability of Internet access has led to enormous increase in the use of World Wide Web which has become the biggest distributed system ever built. Users increasingly see large–scale Web services as essential to the world's communication infrastructure and demand 24/7 availability and response time within seconds. With the tremendous growth and change in Web sites, users, and technology, expanding usage in different application domains, and high consequences of failures and unsatisfactory performance, a comprehensive analysis and prediction of Web quality attributes is essential.

Understanding the nature and characteristics of Web workload is a precondition for proper design, implementation, and tuning of Web–based systems which lead to improved quality of service offered to users. Therefore, in the last decade a considerable amount of research work was focused on studying the network traffic in general, and Web traffic in particular. In their pioneering work, Leland, Taqqu, Willinger and Wilson [18] established that Ethernet LAN traffic is self–similar in nature and showed that the degree of self–similarity increases with the traffic intensity. Following these breakthrough results, the WAN traffic was studied in [22]. In this study, the authors presented a complete model for TELNET traffic (FULL-TEL), which uses Poisson connection arrivals, log–normal connection sizes, and Tcplib packet inter–arrivals. In [28] the authors suggested that the superposition of many ON/OFF sources whose ON and OFF periods are modelled with heavy–tailed distributions produce aggregate network traffic which is self–similar or long–range dependant in nature.

The analysis of the Web traffic at request level presented in in [7] showed that the busiest hours are well described as self–similar, while many less busy hours do not show self–similar characteristics. Another study of Web traffic at request level [2], based on the access logs from six Web servers, showed that the file size and transfer size distributions are heavy–tailed.

A unique characteristic of Web workload is the concept of session which is defined as a sequence of requests from the same user during a single visit to the Web site; session boundaries are delimited by a period of inactivity by a user. In [5] authors proposed a session–based admission control aimed at increasing the chances that longer sessions will be completed. In [3] authors studied how the threshold value affects the number of sessions and focused on other session characteristics such as the number of requests per session, session length, and inter–session arrival times. The work presented in [19] used Customer Behavior Model Graph (CBMG) to represent Web sessions. As a continuation of this work, priority–based resource management policies based on CBMG representation and simulated workload were proposed in [20]. The work presented in [21] studied the request, function, and session characteristics of two weeks of data from two actual e–commerce sites.

Next, we summarize a few papers which raised interesting questions about the methods used to establish the existence of self–similarity, long–range dependence, and heavy–tailed distributions applied to Web and other types of network traffic. In [13] it was shown that the methods for estimating self–similarity and long–range dependence could give conflicting results. Furthermore, it was shown that the trend, periodicity, and noise may affect the accuracy and consistency of the Hurst exponent estimations. In a closely related papers [15], [16], the study of a network backbone traffic showed that the packet arrivals appear to be Poisson at sub-second time scales, non–stationary at multi–second time scales, and exhibit long–range dependence at scales

Proceedings of the 2006 International Conference on Dependable Systems and Networks (DSN'06)
0-7695-2607-1/06 $20.00 © 2006 IEEE

IEEE COMPUTER SOCIETY

of seconds and above. In [9] it was suggested that the methods employed for estimating the index of heavy–tailed distributions could produce misleading results. Thus, it was shown that the lognormal distribution, which is not heavy–tailed, may result in a log-log complimentary distribution (LLCD) plot which appears to be heavy–tailed.

In our earlier work [11], [12] we introduced several inter–session and intra–session characteristics which collectively describe Web workload in terms of sessions. In this paper we present more detailed and rigorous statistical analysis of Web workloads based on empirical data extracted from the actual logs of four Web servers. Similarly to the work presented in [22], which proposed so called FULL-TEL model for describing the TELNET traffic, we conduct statistical analysis of the request–based and session–based attributes of Web traffic aimed at building FULL-Web model. Specifically, for typical intervals with low, medium, and high workload and for one week, we analyze the following characteristics of the Web workloads:

- *request–based analysis*: number of requests per unit of time and request inter–arrival time
- *session–based analysis*:
  - *inter–session characteristics*: sessions initiated per unit of time and time between sessions initiated
  - *intra–session characteristics*: session length in time, number of request per session, and number of bytes transferred per session.

Our analysis is focused on exploring important phenomena, such as self–similarity, long–range dependence, and heavy–tailed distributions. As it has been observed recently [9], [13], [16], despite almost ten years of history of using these phenomena to model the Internet traffic, their identification in real data is a challenging task since the existing methods may perform erratically in practice and produce misleading results. Therefore,

- For establishing self–similarity and long–range dependence on request and session level we test the stationarity of the request and session-based time series and remove the trend and periodicity. Then, we use several methods for estimating the Hurst exponent.
- For intra–session characteristics we use several different methods to test the existence of heavy–tailed behavior and cross validate the results.
- We point out specific problems associated with the methods used for establishing long–range dependence and heavy–tailed behavior of Web workload.

It should be emphasized that the previous research on statistical characterization of Web workloads was focused only on request level [7] or very limited non-rigorous analysis of only one session characteristic - session length in number of requests [21]. We believe that the comprehensive model presented in this paper contributes towards better understanding and more formal statistical description of the Web workloads, which is a fundamental step necessary for performance modelling and prediction, capacity planning, and admission control.

The rest of the paper is organized as follows. The data extraction and analysis process is briefly described in section 2, while the background on self–similarity, long–range dependence, and heavy–tailed distributions is summarized in section 3. We present the analysis of Web workload at request and session level in sections 4 and 5, respectively. Finally, the concluding remarks are given in section 6.

2 Data extraction and analysis process

The Web logs used in this paper were obtained from four Web servers: university wide Web server at West Virginia University (WVU), Web server of the Lane Department of Computer Science and Electrical Engineering (CSEE), Web server of the commercial Internet provider ClarkNet, and Web server at the NASA Independent Verification and Validation Facility (NASA-Pub2)[1].

In this paper, for practical reasons, we define a session as a sequence of requests issued from the same IP address with the time between requests less than some threshold value. As in all other research papers that considered sessions, we consider each unique IP address in the access log to be a distinct user. Clearly, this is not always true [3]. However, in spite of the inaccuracies, we believe that using the IP address provides a reasonable approximation of the number of distinct users. Based on the study of the effect of different threshold values on the total number of session presented in [12], we adopt a 30 minute time interval as the threshold value.

The data collection and analysis process is summarized in Figure 1. (For more detailed information, the reader is referred to [11], [12].) After merging the access and error logs for architectures that employ redundant Web servers (i.e., WVU and CSEE), we include the log entries from the access and error logs as records in the corresponding database tables, which allows more flexible and customized analysis. In our earlier work [11], [12] we presented detailed error and reliability analysis and introduced several intra–session and inter–session attributes which collectively describe Web server sessions. In this paper we provide more detailed and statistically rigorous analysis of both request–based and session–based workload characteristics (the bottom right part in Figure 1).

Table 1 summarizes the raw data for one week period for the Web servers analyzed in this paper. It should be noted that the workload on different servers varies by three orders of magnitude. Also, the servers are from different domains: two from educational institutions, one from research institution, and one from a commercial Web site. In addition to the analysis of one week of data, our goal is to study the effect of the workload intensity on the request–based and session–based characteristics. For this purpose, we divided the one week period into 42 intervals of 4 hours and for each data set selected typical low (Low), medium (Med), and high (High) intervals using the total number of requests as a criterium. Although we select the intervals accordingly to the total number of requests, the total number of sessions and total number of bytes transferred within these intervals adhere the same trend.

[1] The Web logs of the NASA IV&V server were sanitized, that is, IP addresses were replaced with unique identifiers.

IEEE COMPUTER SOCIETY

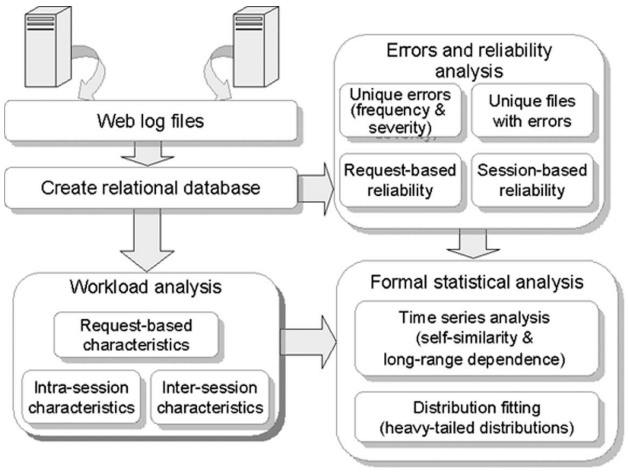

Figure 1. Data extraction and analysis process

Data set	Start date	Requests	Sessions	MB transf.
WVU	12-Jan-04	15,785,164	188,213	34,485
ClarkNet	28-Aug-95	1,654,882	139,745	13,785
CSEE	12-Apr-04	396,743	34,343	10,138
NASA-Pub2	12-Apr-04	39,137	3,723	311

Table 1. Summary of the raw data

3 Background on used statistical methods

3.1 Self–similarity and long–range dependence

Since in Web workload context we deal with time series, the self–similarity is defined as follows. Let $X = \{X_i, i \geq 1\}$ be a stationary sequence. Let

$$X_k^{(m)} = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i, \quad k = 1, 2, \ldots \tag{1}$$

be the corresponding aggregated sequence with level of aggregation m obtained by averaging non-overlapping blocks of size m. Then, for all integers m, the following holds for a self–similar process

$$X(t) \stackrel{d}{=} m^{1-H} X^{(m)}. \tag{2}$$

A stationary sequence is said to be exactly second–order self–similar if $m^{1-H}X^{(m)}$ has the same variance and autocorrelation as X for all m. A stationary sequence is said to be asymptotically second–order self–similar if $m^{1-H}X^{(m)}$ has the same variance and autocorrelation as X as $m \to \infty$. Asymptotically second–order self–similar processes are also called long–range dependant processes. Long–range dependent processes are characterized by hyperbolically decaying autocorrelation function, that is, $r(k) \sim k^{-\beta}$ as $k \to \infty$, where $0 < \beta < 1$. Since $\beta < 1$, the sum of the absolute values of the autocorrelation function approaches infinity, that is, the autocorrelation function is non-summable. Simply put, long–range dependence describes the property that the correlation structure of a time series is preserved irrespective of time aggregation, that is, the autocorrelation function (ACF) is the same in either course or fine time scales.

A predominant way to quantify the self–similarity and long–range dependence is through the Hurst exponent H. For a self–similar process $0.5 < H < 1.0$; as H increases from 0.5 to 1.0, the degree of self–similarity increases. Calculating this exponent, however, is not straightforward due to following reasons [13], [16]. (1) It cannot be calculated definitely, only estimated. (2) No estimator is robust in every case and it is not clear which estimator provides the most accurate estimation; estimators can hide long–range dependence or report it erroneously. (3) Long–range dependence may exists, even if the estimators have different estimates in value, provided that the estimates show $0.5 < H < 1$. (4) For accurate characterization it may be necessary to process and decompose the signal since the trend and periodicity can obscure the analysis.

In general, Hurst exponent estimators can be classified into two categories [27]: those operating in time–domain and those operating in frequency– or wavelet–domain. In this paper we use the Variance and R/S from the time–domain estimators, and Periodogram, Whittle and Abry-Veitch from frequency- and wavelet–domain estimators. Whittle and Abry-Veitch methods, in addition to the estimate of H, provide confidence intervals. For detailed description of the Hurst exponent estimators the reader is referred to [1], [18], and [27].

3.2 Heavy–tailed distributions

The random variable X, with cumulative distribution function $F(x)$, is said to be heavy–tailed if

$$1 - F(x) = P[X > x] = x^{-\alpha} L(x) \tag{3}$$

where $L(x)$ is slowly varying as $x \to \infty$, i.e., $\lim_{x\to\infty} L(ax)/L(x) = 1$ for $a > 0$ [24]. That is, regardless of the behavior for small values of the random variable, if the asymptotic shape of the distribution is hyperbolic, it is heavy–tailed. The simplest heavy–tailed distribution is the Pareto distribution which is hyperbolic over its entire range. The classical Pareto distribution with shape parameter α and location parameter k has the cumulative distribution function

$$F(x) = P[X \leq x] = 1 - (k/x)^{\alpha}. \tag{4}$$

There is an important qualitative property of the moments of heavy–tailed distributions. If X is heavy–tailed with parameter α then its first $m < \alpha$ moments $E[X^m]$ are finite and its all higher moments are infinite. Thus, if $1 < \alpha \leq 2$ the distribution has a finite mean and infinite variance; if $\alpha \leq 1$ the distribution has infinite mean and variance. As α decreases an arbitrary large portion of the probability mass may be present in the tail of the distribution. In practical terms, a random variable that follows a heavy–tailed distribution can give rise to extremely large values with non–negligible probability.

To estimate the tail index α of a Pareto distribution we employ the *log-log complementary distribution (LLCD) plots* [2], [3], [7]. These are plots of the complementary cumulative distribution function (CCDF) $P[X > x] = 1 - F(x) = \bar{F}(x)$ on log-log axes. Plotted this way, heavy–tailed distributions have the property that

$$\frac{d \log \bar{F}(x)}{d \log x} = -\alpha, \quad x > \theta$$

for some θ. In practice, we select a value for θ from the LLCD plot above which the plot appears to be linear. Then, we estimate the slope, which is equal to $-\alpha$, using least–square regression.

IEEE COMPUTER SOCIETY

Hill estimator [24] is an alternative approach for estimating the tail index α of a semiparametric Pareto type model given by (3). For the discussion that follows, let $X_1, X_2, \ldots, X_n$ denote observed values of the random variable X and let $X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$ be the ordered statistics of the data set. The idea behind the Hill estimator is to use only k upper-order statistics, that is, to sample from the part of the distribution which looks most Pareto-like. Therefore, we pick $k < n$ and compute the Hill estimator

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log X_{(i)} - \log X_{(k+1)}. \quad (5)$$

Thus, for each value of k we obtain an estimate of the tail index parameter $\alpha_{k,n} = 1/H_{k,n}$. In practice, the estimates of the tail index $\alpha_{k,n}$ are plotted as a function of k, for the range of k-values. A typical Hill plot varies considerably for small values of k, but becomes more stable as more and more data points in the tail of the distribution are included (often up to a cut-off value, to the left of which (3) no longer holds). If the plot stabilizes to a constant value one can infer the value of the tail index α. The absence of such straight line behavior is a strong indication that the data are not consistent with the heavy–tailed distribution (3).

4 Request–based analysis

In this section we first examine whether the long–range dependence (i.e., asymptotically second–order self–similarity) applies to the request arrival process and then formally test the assumption for Poisson arrivals.

4.1 Number of requests per unit of time

Figure 2 shows the time series plot of the number of requests per second for one week period for the WVU data set. As it can be seen from Figure 3, the autocorrelation function is slowly decaying which indicates long–range dependence. Next, we estimate the values of the Hurst exponent using the SELFIS tool [14]. These values are presented in Figure 4, with Web sites sorted by the total number of requests in descending order.

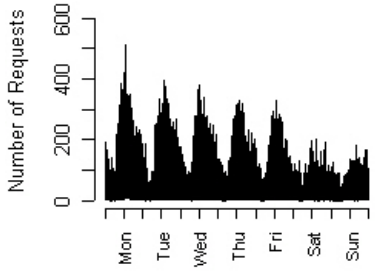

Figure 2. Number of requests per second - WVU

As described in section 3.1, all Hurst exponent estimators assume stationary time series, that is, the trend and periodicity can obscure the analysis based on Hurst exponent. However, related papers that studied self–similarity and long–range dependence of Web traffic either avoided dealing with non–stationarity of the time series or ignored it. Thus, in [7] the authors concentrated on individual hours from the request–based time series in order to provide as nearly a stationary dataset as possible, thus avoiding to deal with non-stationarity of the traffic. A period of two

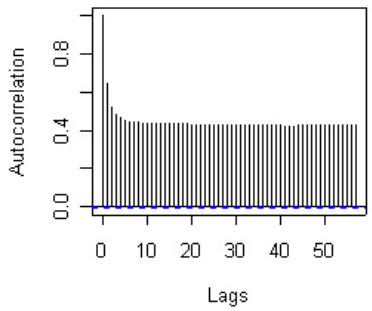

Figure 3. ACF for number of requests per second - WVU

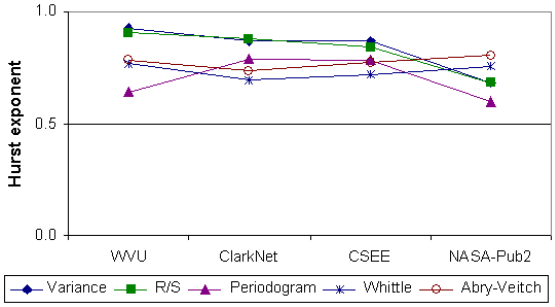

Figure 4. Hurst exponent for request per second based on the raw data

weeks for two e–commerce sites was considered for the request–based analysis presented in [21]. The existence of long–range dependence was suggested based on the variance time plot, without testing the stationarity of the time series, that is, ignoring the trend and periodicity of the signal.

One of our goals is to study how non–stationarity of the traffic affects the estimates of the Hurst exponent, and consequently the conclusions drawn about long–range dependence. We use the Kwiatkowski-Phillips-Schmidt-Shin test [17] to test the null hypothesis of stationarity against an alternative of a unit root which means that time series is non-stationarity. According to this test the request arrival processes (i.e., number of requests per second) for all Web servers considered in this paper are non–stationary. Therefore, we estimate and eliminate the trend and periodicity from the request–based time series using the least square estimation of trend, peridogram for finding the periodicity, and differencing method for removing the seasonal component [4]. All datasets considered in this paper had a slight trend component and a 24 hour period corresponding to day/night change of traffic intensity. After removal of the trend and the seasonal component, the Kwiatkowski-Phillips-Schmidt-Shin test [17] proved that the time series is stationary. The autocorrelation function of the stationary time series shown in Figure 5 still seams to be non–summable, which is an indication of long–range dependence. However, its value is lower than for the original (non-stationary) time series, which indicates that not accounting for the trend and periodicity leads to overestimating the long–range dependence. To confirm these findings formally, we estimate the Hurst exponent for the stationary request–based time series. The

IEEE COMPUTER SOCIETY

results are presented in Figure 6. The following observations can be drawn from the estimates of the Hurst exponent based on the raw data and stationary data: (1) The values of the Hurst exponent based on the raw data, with a few exceptions, are higher than the values based on the stationary time series. This proves the fact that for accurate estimates of self–similarity and long–range dependence the analysis must account for phenomena such as trend and periodicity. (2) The values of the Hurst exponent for all Web sites are higher than 0.5, which indicates that the request arrival processes on a second time scale is asymptotically second–order self–similar (i.e., long–range dependent); the degree of self–similarity increases with the workload intensity, which is consistent with the observations made for the LAN traffic in [18] and for the Web traffic in [7]. (3) The Hurst estimators provide consistent estimates for all four Web servers, which is not necessarily always the case [13]. (4) In most cases Abry-Veitch method provides slightly higher value of H than Whittle method which is consistent with the results presented in [13].

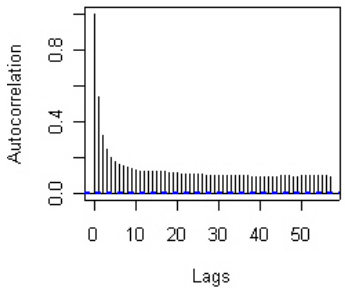

Figure 5. ACF for number of requests per second after removing the trend and periodicity - WVU

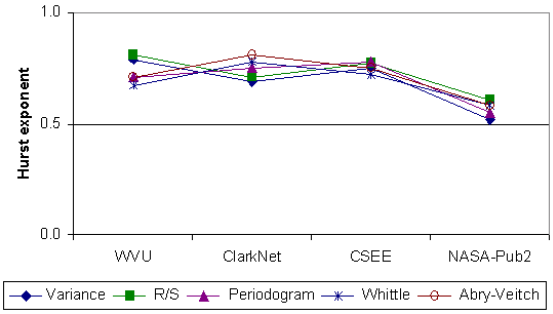

Figure 6. Hurst exponent for request per second based on the stationary data

Since the mathematical definition of the long–range dependence is asymptotic in nature, we next employ the Hurst exponent estimators on aggregated time series [7], [18]. Each one week dataset is aggregated at increasing levels m as described with equation (1), and the estimators are applied to each m–aggregated dataset[2]. As m increases, short–range dependencies are averaged out of the dataset; if the value of H remains relatively constant, we can be confident that it measures a true underlying level of (asymptotic second–order) self–similarity. Figures 7 and 8 show the estimates $\hat{H}^{(m)}$ of the Hurst exponent obtained from the aggregated series $X^{(m)}$ using Whittle and Abry-Veitch methods for the stationary request–based time series of the WVU server. The upper and lower dotted lines are the limits of the 95% confidence intervals on H. These Figures show that for WVU dataset, the values of $\hat{H}^{(m)}$ are relatively consistent as the aggregation level is increased (i.e., $\hat{H}^{(m)} \in [0.768, 0.986]$ for Whittle method, and $\hat{H}^{(m)} \in [0.748, 0.925]$ for Abry-Veitch method). The same holds for the 95% confidence interval bands, indicating a statistical evidence for long–range dependence of the request arrival process. The estimates of H for other sites, such as for example NASA-Pub2, are even more stable and fluctuate only slightly throughout the aggregation levels (i.e., $\hat{H}^{(m)} \in [0.534, 0.606]$ for Whittle method, and $\hat{H}^{(m)} \in [0.533, 0.688]$ for Abry-Veitch method).

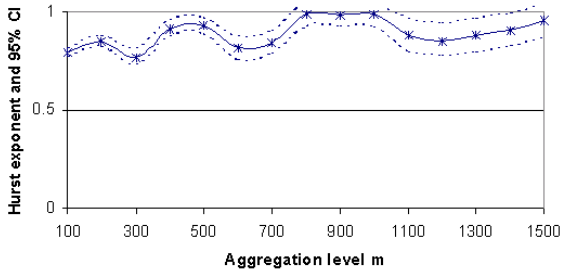

Figure 7. Whittle estimates for stationary request–based time series - WVU

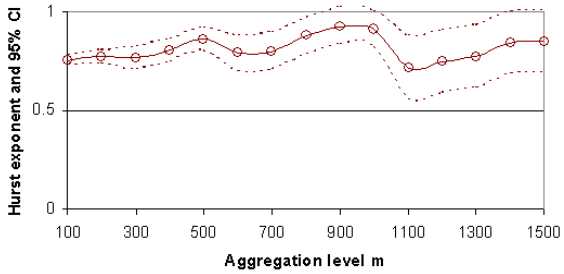

Figure 8. Abry-Veitch estimates for stationary request–based time series - WVU

4.2 Testing for Poisson arrivals at request level

In this subsection, we formally test whether the request arrivals can be modelled with Poisson process for each of the Low, Med, and High intervals. To test the two main characteristics of the Poisson process – request inter–arrival times are independent random variables which follow the exponential distribution - we use the method proposed in [22].

Before the test for Poisson arrivals can be applied, the original signal has to be processed due to the following reasons. (1) The Web servers considered in this study have timestamps with granularity of one second, which leads to multiple requests with the same timestamp. Assumptions about how these requests are distributed within a one second interval have to be made before we can apply the test for Poisson arrivals. Since different assumptions may lead to different results [29], we use two distributions for the request arrivals over the one second interval: uniform and

[2]As the aggregation level m increases the confidence intervals tend to widen since for larger m there are fewer observations in $X^{(m)}$.

IEEE COMPUTER SOCIETY

deterministic (i.e., requests evenly spread out over the one second interval). (2) Since the request arrival rate varies during the four hours intervals, testing for homogeneous Poisson model with a fixed rate is not appropriate. Therefore, we divide each of the Low, Med and High four hour intervals into four 1-hour intervals with approximately constant arrival rate. Then, we test each 1-hour interval for independent and exponentially distributed inter–arrival times.

Test for independent request inter–arrival times
For each 1-hour interval i ($i = 1, 2, 3, 4$), we compute its lag one autocorrelation ρ_i. Let S be the random variable of number of intervals having ρ_i less than $1.96/\sqrt{n_i}$, where n_i is number of samples in the ith interval. Then S follows the binomial distribution $B(4, 0.95)$. Suppose s is the observed value of S, then if $P(S = s) < 0.05$, we conclude with 95% confidence that the inter-arrivals are not independent. We also apply one further test for independent request inter–arrivals. If the inter–arrivals are truly independent, then their autocorrelation would be negative with probability 0.5 and positive with probability 0.5. Let X be the random variable of number of positive ρ_i's, then X follows the binomial distribution $B(4, 0.95)$. Suppose x is the observed value of X, then if $P(X = x) < 2.5\%$, the inter–arrivals are significantly positively correlated. Similarly, let Y be the random variable of number of negative ρ_i's, and y is the observed value of Y. Then if $P(Y = y) < 2.5\%$, the inter–arrivals are significantly negatively correlated.

Test for exponentially distributed request inter–arrival times
Let the null hypothesis be $H_0 : F(x) = 1-e^{-\hat{\lambda}x}$ where $\hat{\lambda} = 1/\bar{X}$ is estimated from the sample. To test the goodness of fit for each 1-hour interval, we use the Anderson–Darling (A^2) test [26] because it is generally much more powerful than either of better known Kolmogorov–Smirnov or χ^2 tests. A^2 is an empirical distribution test which looks at the entire observed distribution and it is particularly good for detecting deviations in the tail of a distribution.

The null hypothesis is rejected on an interval if the modified test statistic $A^2(1+0.6/n)$ is greater than the critical value 1.341. Let Z be the random variable of total number of intervals having test statistic less than 1.341, then Z follows the binomial distribution $B(4, 0.95)$. Suppose z is the observed value of Z, then if $P(Z = z) < 0.05$, we conclude with 95% confidence that the inter–arrivals are not exponential.

We repeat the same methods to test for independent and exponentially distributed request inter–arrival times by dividing each four hour period in 10-minute intervals. The results show that the request arrivals do not follow the Poisson process with fixed 1-hour or 10-minute rates for any of the considered Web sites. These results are valid regardless of the assumption made about the distribution of the request arrivals over one second (i.e., uniform and deterministic). Our results are in agreement with the recent study which showed that the backbone Internet traffic exhibits long–range dependence at scales of seconds and above [15]. The same study showed that the Internet traffic can be well represented by the Poisson model for sub–seconds time scales. The granularity of the measurements in our datasets is one second, which does not allow testing the Poisson assumption on the finer time scales.

In summary, the results presented in this section show that Web workload at request level, similarly to LAN and WAN workload, is long–range dependant. These results are consistent with earlier result for Web traffic presented in [7]. In addition, we have explicitly shown that the assumption that the request arrivals can be modelled with Poisson process is not valid. This means that several Web performance models which used queuing networks [23], [25], [30] or layered queuing networks [8] are based on incorrect assumptions and most likely provide misleading results.

5 Session–based analysis

In this section we study, in a rigorous statistical manner, the session arrival process (i.e, inter–session characteristics) and intra–session characteristics introduced in our earlier work [11]. It should be emphasized that the empirical studies which addressed Web sessions in the past were mainly focused on simple analysis and did not explore the long–range dependence and heavy–tailed behavior.

5.1 Inter–session characteristics

The analysis of inter–session characteristics is based on the same methods used in section 4 for analysis of request level workload.

5.1.1 Number of sessions initiated per unit of time

The values of the Hurst exponent for the raw data of the sessions initiated per second times series are presented in Figure 9 with Web servers sorted by the total number of sessions initiated within a week in descending order. As in case of the request–based time series, we test whether the session–based time series is stationarity using the Kwiatkowski-Phillips-Schmidt-Shin test [17]. The results show that WVU, ClarkNet, and CSEE Web servers have a slight trend and 24 hour period. The NASA-Pub2 session–based time series is stationary. Similarly to the request–based time series, removing the trend and periodicity leads to smaller values of the autocorrelation function.

The more formal analysis of the long–range dependence of the stationary session–based time series, based on the estimates of the Hurst exponent presented in Figure 10, leads to the following conclusions: (1) The values of the Hurst exponent based on the raw data are higher than the values based on the stationary time series in most of the cases. (2) The values of the Hurst exponent for all Web servers are higher than 0.5. These results indicate that the session arrival process on a second time scale is long–range dependant. (3) The long–range dependence of the sessions initiated per second times series seems to be less influenced by the workload intensity than the request–based time series. (4) The Hurst estimators provide consistent estimates, which is not necessarily always the case [13]. (5) Abry-Veitch method provides slightly higher value of H than Whittle method, which is consistent with the results presented in [13].

Again, as with request arrival process, we study the estimates $\hat{H}^{(m)}$ of the Hurst exponent obtained from the aggregated series $X^{(m)}$ for increasing level of aggregation m. The values of $\hat{H}^{(m)}$ for all datasets are quite stable and fluctuate slightly. The same holds for the 95% confidence interval bands, indicating a statis-

IEEE COMPUTER SOCIETY

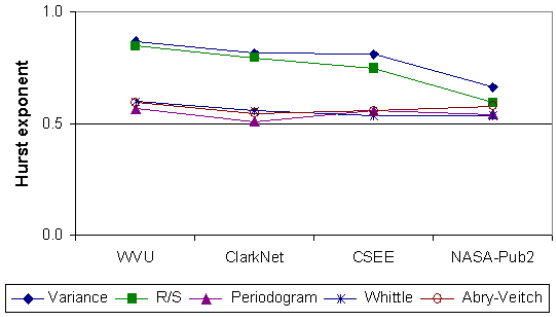

Figure 9. Hurst exponent for sessions initiated per second based on raw data

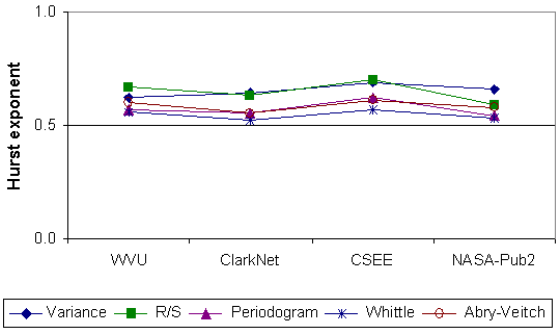

Figure 10. Hurst exponent for sessions initiated per second based on stationary data

tical evidence for a long–range dependence of the session arrival process.

5.1.2 Testing for Poisson arrivals at session level

Next, we test whether the time series of sessions initiated per second can be modelled with Poisson process, using the same methods as in section 4.2. It should be noted that for NASA-Pub2 server the number of sessions in Low, Med, and High four hour intervals are not sufficient to conduct the test. The results of the statistical tests for four hour intervals divided into four 1-hour intervals show that only in CSEE Low and Med intervals session arrivals are indistinguishable from the Poisson process. Thus, unlike TELNET connection arrivals and FTP session arrivals which were well modelled as Poisson process with fixed hourly rates [22], Web session arrivals are Poisson only when the workload is low (less than 1,000 sessions in a four hour period for our datasets). As in the case of the request arrival process, the assumption made about the distribution of sessions initiated within a second (i.e., uniform and deterministic) does not affect the results.

5.2 Intra–session characteristics

In this section we analyze the session length, number of request per session, and number of bytes transferred per session. In particular, we use the statistical methods described in section 3.2 to examine whether intra–session characteristics can be modelled with heavy–tailed distributions.

5.2.1 Session length in time units

The first intra–session characteristic is the sessions length in units of time. The LLCD plot of the session length of WVU server for all 10,287 sessions that occurred during the High four hour interval is presented in Figure 11. For sessions longer than about 1000 seconds, the plot is nearly linear, which indicates a hyperbolic upper tail. The least square regression estimate of the heavy tail index is $\alpha_{LLCD} = 1.67$ with standard error $\sigma_\alpha = 0.004$. The coefficient of determination (R^2) is 0.993, which indicates a very good fit between the empirical and mathematical distribution.

To further confirm the observation that session length of WVU server can be described with Pareto distribution with finite mean and infinite variance, we also estimate the tail index α_{Hill} using the Hill plot. The value of Hill estimator for varying k restricted to the upper 14% tail is shown in Figure 12. The Hill estimator seems to settle to a relatively constant estimate $\alpha_{Hill} \approx 1.58$ which is consistent with the estimate obtained by the LLCD method.

Table 2 summarizes the values of α_{Hill} estimated using the Hill estimate, and α_{LLCD} and R^2 estimated using LLCD plot for each Low, Med, and High four hour interval and one week period for each Web server. As it can be seen, in most cases Hill estimator provides estimates of the tail index α close to the estimates obtained using the LLCD method. However, in a few cases Hill plots did not stabilize, which is annotated with NS in Table 2. For NASA-Pub2 server, which has low workload intensity, the number of sessions in the Low four hour interval were not sufficient to estimate α with either method (annotated with NA).

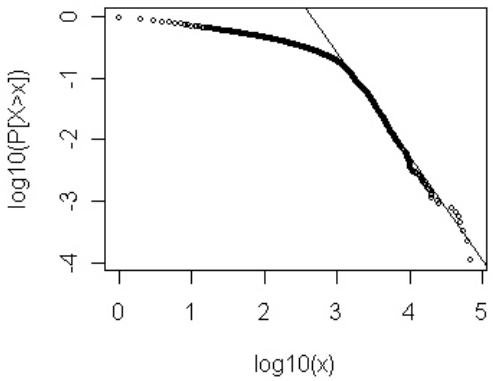

Figure 11. LLCD plot for WVU session length, High

The results for one week show that the session length is reasonably well modelled by a Pareto distribution with $1.723 \leq \alpha_{LLCD} \leq 2.329$. The session length of WVU and ClarkNet servers is heavy–tailed (with finite mean and infinite variance) for lengths longer than 21 minutes in both cases. It also can be observed that the session length for these two servers is heavy-tailed ($1 < \alpha < 2$) regardless of the workload intensity. The session length for the CSEE and NASA-Pub2 servers on one week of data has finite mean and variance (i.e., $\alpha > 2$). However, there are intervals (i.e., Med for CSEE and Med and High for NASA-Pub2) which have session length consistent with heavy–tailed distributions.

Since there is a group of researchers who advocate lognormal

IEEE COMPUTER SOCIETY

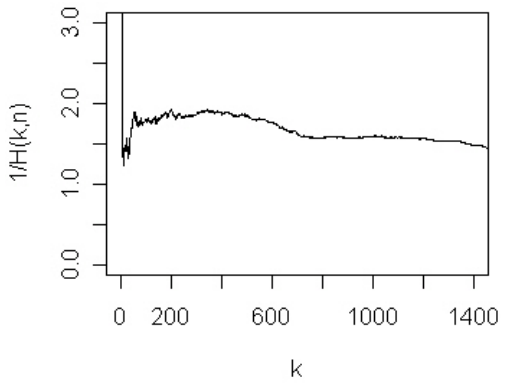

Figure 12. Hill plot for WVU session length, High

		WVU	ClarkNet	CSEE	NASA-Pub2
Low	α_{Hill}	1.02	0.8	NS	NA
	α_{LLCD}	1.044	1.03	2.172	NA
	R^2	0.941	0.982	0.937	NA
Med	α_{Hill}	1.55	1.27	1.73	NS
	α_{LLCD}	1.609	1.273	1.888	1.840
	R^2	0.990	0.981	0.976	0.977
High	α_{Hill}	1.58	1.5	NS	1.39
	α_{LLCD}	1.670	1.832	3.103	1.422
	R^2	0.993	0.966	0.981	0.857
Week	α_{Hill}	1.8	1.8	2.2	2.2
	α_{LLCD}	1.803	1.723	2.329	2.286
	R^2	0.994	0.994	0.987	0.976

Table 2. α_{Hill}, α_{LLCD}, and R^2 for session length

rather than Pareto distribution as correct description of the data (see for example [9]), we incorporate one more test in our analysis. Thus, it is known that when the variance is large, a lognormal CCDF is very close to a straight line in the log-log plot, that is, it appears long–tailed, at least to a point [9], [10]. The dramatic difference between lognormal and Pareto distributions lies in the extreme tail of their CCDFs. In a LLCD plot, the CCDF of a Pareto distribution decays with constant slope, while the CCDF of a lognormal distribution shows increasing slope in the extreme tail. To explore how good Pareto and lognormal models match our empirical data we applied the curvature test proposed in [9] on all datasets. The p-value both under Pareto and lognormal models for all intervals shown in Table 2 is greater than 0.05, which means that with 95% confidence we cannot reject the hypothesis that the sample comes from Pareto or lognormal distributions. According to the curvature test, for some intervals lognormal is better fit than Pareto distribution. Although in [9] the author claimed that the curvature test is insensitive to the estimated value of the parameter α, on our datasets different estimates of α led to different p–values for Pareto distribution. Furthermore, the same estimates for the tail index α with different random samples from Pareto distribution used as a part of the test, yielded different p–values. We believe that the reason for the sensitivity of the curvature test to the estimated value of α and the random sample, as well as the difficulty to distinguish Pareto and lognormal distributions, is the fact that very often there are very few sample observations in the extreme tail. In that case, as shown in [10], the 95% confidence intervals of Pareto and lognormal distributions have a large overlap at the extreme tail, which makes it hard to distinguish them.

The following example illustrates the importance of rigorous statistical analysis and the implications of our results. The simulation of the session–based admission control used for peak load management presented in [5], [6] was based on the assumption that the session length is exponentially distributed, which as our results show is an incorrect assumption.

5.2.2 Number of requests per session

Another intra–session characteristics is the number of requests per session (i.e., session length in number of requests). The results of the curvature test [9] for the number of request per session were similar as for the session length in time - neither Paerto nor lognormal models can be rejected for any interval. For example, note that although the LLCD plot of the session length in number of request for one week of data for ClarkNet server presented in Figure 13 shows increasing slope in the extreme tail, Pareto distribution provides better fit than the lognormal distribution.

As it can be seen from Table 3, the tail index of Pareto model for the distribution of session length in number of requests for one week of data is in the range $1.615 \leq \alpha_{LLCD} \leq 2.586$. Under the Pareto model, the session length in number of request shows clear heavy–tailed behavior with tail index α significantly smaller than 2 only for NASA-Pub2 server. For this server over 84% of requests belong to sessions in the 75 percentile tail. The other three servers have tail index around 2, that is, have session length in number of requests on the boarder line between finite and infinite variance. It should be emphasized that for all servers many long sessions in time units do not have many requests. That is, many sessions in the tail of session length distribution are completely different from the sessions in the tail of the number of request per session distribution.

The session length in number of requests is the only intra–session characteristic studied earlier. In [21], based on the LLCD plot, it was suggested that the tail of the distribution for the auction site falls abruptly, while for the bookstore site it remains close to the straight line plot of a Pareto-like distribution with $\alpha = 1$. However, the value of the tail parameter was not estimated and no evidence was presented that the Pareto model fits the data. Another important observation is that in cases when the session length in number of request is modelled with distributions with large variance, it does not make sense to derive and report metrics such as average session length in number of requests, as it was done in [19], [20].

5.2.3 Bytes transferred per session

For our last intra–session characteristic, total number of bytes transferred per session, we count the bytes transferred for both completed and partial transfers. Similarly to the other two intra–session characteristics, based on the curvature test [9], neither Pareto nor lognormal distribution can be rejected as models for the bytes transferred per session. Again, as in the other cases, the p–value for Pareto distribution was sensitive to the estimated value of α and generated random sample.

If the Pareto distribution is used for modelling the bytes transferred per session, as it can be seen from Table 4, $0.954 \leq \alpha_{LLCD} \leq 1.842$ for one week of data. This means that all Web servers have heavy–tails (with infinite variance) for the number of

IEEE COMPUTER SOCIETY

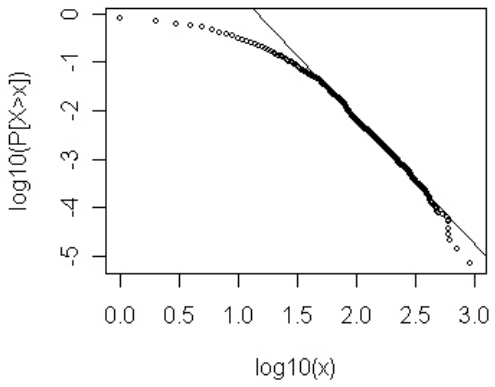

Figure 13. LLCD plot of session length in number of requests for ClarkNet, one week

		WVU	ClarkNet	CSEE	NASA-Pub2
Low	α_{Hill}	1.7	2.32	2.0	NA
	α_{LLCD}	1.965	2.218	2.047	NA
	R^2	0.986	0.975	0.976	NA
Med	α_{Hill}	2.0	1.8	1.93	1.9
	α_{LLCD}	2.055	1.724	1.931	1.948
	R^2	0.996	0.987	0.987	0.903
High	α_{Hill}	1.9	1.9	2.33	1.62
	α_{LLCD}	1.965	1.928	2.167	1.437
	R^2	0.993	0.979	0.981	0.971
Week	α_{Hill}	2.1	2.6	2.0	1.6
	α_{LLCD}	2.151	2.586	1.932	1.615
	R^2	0.995	0.996	0.989	0.967

Table 3. α_{Hill}, α_{LLCD}, and R^2 for session length in number of requests

bytes transferred per session, including the Low, Med, and High intervals. Even more, the values of α for CSEE server are around 1 or even below 1 (implying infinite mean). It is obvious that under the Pareto model, the number of bytes transferred per session has the heaviest tail compared to the other two intra–session characteristics. One obvious reason for the heavy–tailed behavior of the number of bytes transferred per session is due to the fact that the distributions of files sizes and files transferred are heavy–tailed [2], [3], [7].

6 Concluding remarks

In this paper we have presented a rigorous statistical analysis of request level and session level Web workload based on data extracted from four real Web servers. Our goals included development of a FULL-Web model which provides comprehensive view on Web workload and clear identification of the specific limitations associated with methods used for establishing long–range dependence and heavy–tailed behavior.

Our results show that all Web servers considered in this study have long–range dependant request arrival process. Unlike the related work on Web workload characterization which either avoided non–stationarity by focusing on one hour intervals or ignored it completely, we test the stationarity of the request–based time series and eliminate the trend and periodicity before studying the long–range dependence phenomenon. We show that not accounting for the trend and periodicity leads to overestimating the level of long–range dependence. Furthermore, we show that the piecewise Poisson process with fixed 1-hour or 10-minute rates cannot be used to model the request arrival process.

		WVU	ClarkNet	CSEE	NASA-Pub2
Low	α_{Hill}	1.1	1.7	0.8	NA
	α_{LLCD}	1.168	1.786	0.788	NA
	R^2	0.998	0.978	0.935	NA
Med	α_{Hill}	1.32	1.89	0.84	NS
	α_{LLCD}	1.371	1.799	0.898	1.676
	R^2	0.996	0.991	0.974	0.949
High	α_{Hill}	1.63	1.86	1.06	1.78
	α_{LLCD}	1.418	1.754	1.026	1.641
	R^2	0.993	0.993	0.989	0.949
Week	α_{Hill}	1.4	2.0	0.95	1.1
	α_{LLCD}	1.454	1.842	0.954	1.424
	R^2	0.995	0.990	0.998	0.960

Table 4. α_{Hill}, α_{LLCD}, and R^2 for bytes transferred per session

In addition to the analysis of the request–based Web workload, we provide a comprehensive model of session–based Web workload which has not been considered earlier. Thus, we study the Web session arrival process and show that, unlike TELNET and FTP traffic, it is long–range dependant for all servers considered in this paper. Even though piecewise Poisson process with fixed hourly rates models well some four hour intervals under low to moderate workload, it fails on longer periods (e.g., one week).

We also study several intra–session characteristics, such as session length in time, number of request per session, and bytes transferred per session. Since long–range dependence is usually accompanied with heavy–tailed distributions, for each intra–session characteristic we examine whether it follows heavy–tailed distribution. For this purpose, we use several different methods (i.e., LLCD plot, Hill plot, and curvature test for the extreme tail). Some highlights of this analysis include: (1) In most cases LLCD plot and Hill estimator give consistent results. (2) Based on the curvature test, the intra-session parameters are modelled well with both Pareto and lognormal distributions. (3) The results of the curvature test for Pareo distribution are somewhat sensitive to the estimated values of the tail index and simulated sample of Pareto distribution. (4) The reason behind the difficulty to statistically distinguish between Pareto and lognormal distribution is the small number of observations in the extreme tail. (5) Under the Pareto model, intra–session characteristics for some intervals exhibit heavy-tailed behavior.

In summary, in this paper we presented a comprehensive model which contributes towards better understanding of Web workloads. We also showed that, despite of almost ten years of research efforts in this area, a number of challenges remain to be addressed in the future work.

Acknowledgements

This work is funded by the National Science Foundation under CAREER grant CNS-0447715 and by the NASA OSMA SARP under grant managed through NASA IV&V Facility in

IEEE COMPUTER SOCIETY

Fairmont. The authors thank David Krovich and David Olsen of West Virginia University and Brian Kesecker of NASA IV&V Facility for making the Web logs available.

References

[1] P. Abry and D. Veitch, "Wavelet Analysis of Long–Range-Dependent Traffic", *IEEE Trans. Information Theory*, Vol.44, No.1, Jan. 1998, pp. 2-15.

[2] M. Arlitt and C. Williamson, "Internet Web Servers: Workload Characterization and Performance Implications", *IEEE/ACM Trans. Networking*, Vol.5, No.5, Oct. 1997, pp. 631-645.

[3] M. Arlitt and T. Jin, "Workload Characterization of the 1998 World Cup Web Site", *Hewlett-Packard Technical Report*, HPL-1999-35(R.1), Sep. 1999.

[4] G. E. P. Box, G. M. Jenkins and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, Third Edition, Prentice-Hall, 1994.

[5] L. Cherkasova and P. Phaal, "Session Based Admission Control: a Mechanism for Improving the Performance of an Overloaded Web Servers", *HP Labs Technical Reports*, HPL-98-119, 1998.

[6] L. Cherkasova and P. Phaal, "Session–Based Admission Control: A Mechanism for Peak Load Management of Commercial Web Sites", *IEEE Trans. Computers*, Vol.51, No.6, June 2002, pp. 669-685.

[7] M. E. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes", *IEEE/ACM Trans. Networking*, Vol.5, No.6, Dec.1997, pp. 835-846.

[8] J. Dilley, R. Friedrich, T. Jin and J. Rolia, "Measurement Tools and Modeling Techniques for Evaluation Web Server Performance", *Proc. 9th Int'l Conf. Computer Performance Evaluation, LNCS 1245*, June 1997.

[9] A. B. Downey, "Evidence for Long–tailed Distributions in the Internet", *Proc. 1st ACM SIGCOMM Workshop on Internet Measurement*, Nov. 2001, pp. 229–241.

[10] W. Gong, Y. Liu, V. Misra and D. Towsey, "Self–Similarity and Long Range Dependence on the Internet: A Second Look at the Evidence, Origins and Impications", *Computer Networks: The Interantional Journal of Computer and Telecommuncatoion Networking*, Vol.48, No.3, 2005, pp. 377-399.

[11] K. Goševa-Popstojanova, S. Mazimdar and A. D. Singh, "Empirical Study of Session–based Workload and Reliability for Web Servers", *Proc. 15th IEEE Int'l Symp. Software Reliability Engineering*, Nov. 2004, pp. 403-414.

[12] K. Goševa–Popstojanova, A. Singh, S. Mazimdar and F. Li, "Empirical Characterization of Session-based Workload and Reliability for Web Servers", *Empirical Software Engineering Journal*, Vol.11, No.1, Jan. 2006, pp. 71-117.

[13] T. Karagiannis, M. Faloutsos and R. H. Riedi, "Long-Range Dependence: Now You See It, Now You Don't!", *Proc. GLOBECOM*, Nov. 2002, pp. 2165–2169.

[14] T. Karagiannis, M. Faloutsos and M. Molle, "A User-Friendly Self-Similarity Analysis Tool", *ACM SIGCOMM Computer Communication Review*, 2003.

[15] T. Karagiannis, M. Molle, M. Faloutsos and A. Broido, "A Nonstationary Poisson View of Internet Traffic", *23rd Annual Joint Conf. of IEEE Computer and Communications Societies*, Vol.3, 2004, pp. 1558-1569.

[16] T. Karagiannis, M. Molle and M. Faloutsos, "Long–range Dependence: Ten Years of Internet Traffic Modeling", *IEEE Internet Computing*, Vol.8, No.5, 2004, pp. 57–64.

[17] D. Kwiatkowski, P. Phillips, P. Schmidt and Y. Shin, "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure are We that Enconomic Time Series have a Unit Root?", *Journal of Econometrics*, Vol.54, Oct/Dec 1992, pp.159-178.

[18] W. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic", *IEEE/ACM Trans. Networking*, Vol.2, No.1, Feb. 1994, pp.1-15.

[19] D. Menasce, V. Almeida, R. Fonseca and M. Mendes, "A Methodology for Workload Characterization of E-commerce Sites", *Proc. ACM Conf. Electronic Commerce*, Nov. 1999, pp. 119-128.

[20] D. A. Menasce, V. A. F. Almeida, R. Foneca and M. A. Mendes, "Business–oriented Resource Management Policies for E-commerce Servers", *Performance Evaluation*, Vol.42, No.2-3, 2000, pp. 223-239.

[21] D. Menasce, V. Almeida and R. Riedi, "In Search of Invariants for E-Business Workloads", *Proc. 2nd ACM Conf. Electronic Commerce*, Oct. 2000, pp. 56-65.

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

[23] P. Reeser and R. Hariharan, "Analytic Model of Web Servers in Distributed Environments", *Proc. 2nd Int'l Workshop on Software and Performance*, Sep. 2000, pp. 158-167.

[24] S. I. Resnick, "Heavy Tail Modeling of Teletraffic Data", *The Annals of Statistics*, Vol.25, No.5, Oct. 1997, pp. 1805-1849.

[25] C. U. Smith and L. G. Williams, *Performance Solutions: A Practical Guide to Creating Responsive, Scalable Software*, Addison-Wesley, 2001.

[26] M. A. Stephens, "EDF Statistics for Goodness of Fit and Some Comparisons" *Journal of the American Statistical Association*, Issue 347, 1967.

[27] M. S. Taqqu and V. Teverovsky, "On Estimating the Intensity of Long–range Dependence in Finite and Infinite Variance Time Series", in R. J. Alder, R. E. FEldman and M. S. Taqqu (Editors) *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, Birkhauser, Boston, 1998, pp. 177-217.

[28] W. Willinger, M. S. Taqqu, R. Sherman and D. V. Wilson, "Self–similarity through High Variability: Statistical Analysis of Ethernet LAN traffic at Source Level", *IEEE/ACM Trans. Networking*, Vol.5, No.1, Feb. 1997, pp. 71-86.

[29] C. H. Xia, Z. Liu, M. S. Squillante, L. Zhang and N. Malouch, "Traffic Modeling and Performance Analysis of Commercial Web Sites" *ACM SIGMETRICS Performance Evaluation Review*, Vol.30, Issue 3, Dec. 2002.

[30] Y. Zhu and K. J. Lu, "Performance Modeling and Metrics of Database-backed Web Sites", *Proc. 11th Int'l Workshop Database and Expert Systems Applications*, Sep 2000, pp. 494-498.

IEEE COMPUTER SOCIETY