RC21562(97220)6SEPT1999 ComputerScience/Mathematics

Research Report

Web Traffic Modeling and Web Server Performance Analysis

Mark S. Squillante, Li Zhang
IBM T. J. Watson Research Center
P. O. Box 218
Yorktown Heights, NY 10598

David D. Yao
IEOR Department
Columbia University
New York, NY 10027

Research Division
Almaden - Austin - Beijing - Haifa - T. J. Watson - Tokyo - Zurich

Web Traffic Modeling and Web Server Performance Analysis

Mark S. Squillante*, David D. Yao†, and Li Zhang*

Abstract

The control and optimization of various performance measures in high-volume Web sites requires a fundamental understanding of the user request patterns and the performance impact of such traffic patterns. We present here a study of these key issues within the context of the official IBM Web site for the 1998 Nagano Olympic Games.

1 Introduction

A significant amount of research has considered models to characterize network traffic for different Web server environments; e.g., see [8, 10] and the references cited therein. Conversely, there has been very little research attempting to understand and model the request traffic of high-volume Web server environments, which are becoming increasingly common as the Web is used more and more often as a means to access the latest news, financial data and other information, and to support a wide range of Internet applications in areas such as e-business, education and entertainment. Many aspects of these Web server environments are very different from those that have been previously considered in the research literature.

In this paper we present a study to better understand the key characteristics of the request patterns in high-volume Web server environments, and to better understand the impact of such traffic patterns on Web server performance. Using the access logs from the official Web site for the 1998 Winter Olympic Games in Nagano, Japan, we develop traffic models to represent the user request process submitted to the geographically-distributed Web site. This is fundamental to gaining key insights into the impact of these traffic patterns on various measures related to performance, including quality of service (QoS), scalability, capacity planning, availability and reliability. Our analysis of the data illustrates traffic patterns that exhibit both light-tailed and heavy-tailed marginals together with relatively strong dependence structures and some seasonal effects. We then input these traffic processes to the set of Web server systems, each modeled as a general single-server queue, and analyze the tail behavior of the waiting-time process which models the latency encountered by user requests. We also study the performance impact of periodic traffic patterns in this setting.

We first summarize in §2 the IBM Web site used in our study. The development of traffic models is provided in §3. We then present in §4 the analysis of the $G/G/1$ queue under these traffic processes, followed by a study of periodic traffic in §5.

2 Web Server Environment

We consider the official Web site for the 1998 Nagano Olympic Games, which is representative of the class of high-volume Web server environments motivating our study. This system consisted of multiple SP2 machine frames at four different locations, where each SP2 frame consisted of 10 uniprocessor nodes that serve Web requests and 1 multiprocessor node that handles all updates to the underlying data. Incoming user requests are routed by a set of Network Dispatchers (ND) to specific nodes of a certain SP2 machine. Each ND follows a weighted round-robin policy, where the weight for a node is a function of its current load. This approach balances the load of the server across the set of SP2 machines and their nodes, and has the effect of smoothing out and equalizing (in a statistical sense) the user request process among the SP2 nodes. The interested reader is referred to [5] for additional details.

A high percentage of the pages at the Web site were created dynamically, since the content of these pages were constantly changing. Embedded image files, on the other hand, comprised the majority of the static page requests. The time to serve a dynamic page request is often significantly longer than the time required to satisfy a static page request, by as much as several orders of magnitude or more. Since the serving of dynamic pages can dominate performance in the Web servers under consideration, it is very important to understand the request patterns for dynamic pages in this class of Web environments and to understand the impact of these traffic patterns on various aspects

*IBM Research Division, T.J. Watson Research Center, Yorktown Heights, NY 10598; {mss,zhangl}@watson.ibm.com.

†IEOR Department, Columbia University, New York, NY 10027; yao@ieor.columbia.edu. Research undertaken while an academic visitor at the IBM T.J. Watson Research Center; supported in part by NSF Grant ECS-9705392.

of Web server performance.

We therefore focus on the requests for dynamic pages originating from Asia and Europe, noting that the traffic patterns from Europe are also representative of those originating from the Americas. The number of requests for dynamic pages were analyzed at different time scales, a representative example of which is provided in the graphs of Fig. 1 showing the aggregate number of requests received every 300 seconds (5 minutes) from both Asia and Europe.

We observe that the traffic from Asia contains very large bursts, whereas the traffic from Europe is much less bursty. The large traffic bursts from Asia are primarily due to a strong interest in Japan for events related to ski jumping; note that most of the traffic from Asia originated from Japan, and this traffic was often concentrated around the time when popular ski jumping events were taking place. In contrast, due in part to the time differences and broader interests in events, the traffic from Europe is more scattered and does not contain the very large spikes found in the traffic from Asia.

To better understand these traffic patterns, we plot in Fig. 2 the tail distribution of the batch processes corresponding to the set of request traffic in Fig. 1. The leftmost graph in Fig. 2 plots $-\log(\mathsf{P}[\text{Batch} > x])$ as a function of x for the traffic from Europe. This quadratic curve suggests that the distribution is light-tailed, where a distribution $F(x)$ is called *light-tailed* if its tail $\bar{F}(x) := 1 - F(x)$ decays at least exponentially fast. In our case for Europe we have $\mathsf{P}[\text{Batch} > x] \sim e^{-\alpha x^2}$. The rightmost graph in Fig. 2 plots $-\log(\mathsf{P}[\text{Batch} > x])$ against $\log x$ for the traffic from Asia. This result is also a quadratic curve, but in terms of $\log x$, suggesting a *heavy-tailed* distribution in which the tail decays in a sub-exponential fashion. (More precisely, this belongs to the *sub-exponential class* of distributions; see [7].) When the log-tail is a quadratic function of $\log x$, as in our case for Asia where $\mathsf{P}[\text{Batch} > x] \sim e^{-\beta(\log x)^2}$, a good model to capture this type of burstiness is the log-normal distribution.

Note that the light-tailed and heavy-tailed distributions only characterize the marginals of the traffic patterns in terms of the batch size per time unit. Hence, we also study the dependence structure of the batch process over time, and we find the autoregressive models recently developed in [5] quite suitable for our purposes. Although these models are for traffic processes with light-tailed marginals, they are readily adapted to handle heavy-tailed marginals as well, which we consider in the next section.

3 Web Traffic Models

A stationary time series $\{Z_n\}$ is said to satisfy an order (p, q) *autoregressive moving average* model, denoted by ARMA(p, q), if it can be represented as

$$Z_n - \phi_1 Z_{n-1} - \ldots - \phi_p Z_{n-p} = \epsilon_n - \theta_1 \epsilon_{n-1} - \ldots - \theta_q \epsilon_{n-q},$$

where p and q are respectively the orders of the AR and MA processes. A time series is an ARIMA(p, d, q) process if $(1 - B)^d Z_n$ is an ARMA(p, q) process, where $BZ_n = Z_{n-1}$ defines the backwards shift operator. Seasonal patterns can be additionally captured by extending this class of statistical models to the so-called *seasonal ARIMA* models.

Let $\{X_n\}$ denote the user request process submitted to a Web server, with X_n denoting the number of requests that arrive in the n^{th} time period, a pre-specified time unit. We suppose $\{X_n\}$ is a stationary sequence, and denote by $X \stackrel{\mathrm{d}}{=} X_n$ the generic r.v. that follows the (common) marginal distribution $F(\cdot)$.

For the type of less bursty traffic from Europe, X follows a standard light-tailed distribution. Specifically, we model X as a normal variate, i.e., $X = \mu + \sigma Z$ where μ and σ are positive real values, and Z denotes the standard normal variate (with zero mean and unit variance). Notice that

$$-\log \mathsf{P}(X > t) \sim \frac{(t-\mu)^2}{2\sigma^2},$$

which is consistent with our observations of the quadratic curves fitting the traffic from Europe in Fig. 2.

To model the bursty traffic from Asia, we let X follow a lognormal distribution, i.e., $X = e^{\mu + \sigma Z}$. The n^{th} moment of X is given by

$$\mathsf{E}(X^n) = e^{n\mu} \mathsf{E}(e^{n\sigma Z}) = e^{n\mu + n^2\sigma^2/2},$$

and in particular

$$\mathsf{E}(X) = e^{\mu + \sigma^2/2}, \qquad \mathsf{Var}(X) = \mathsf{E}(X)(e^{\sigma^2} - 1).$$

Let $\Phi(x)$ and $\phi(x)$ respectively denote the distribution function and the density function of Z. Define $\bar{\Phi}(x) := 1 - \Phi(x)$. We then have

$$\bar{F}(x) = \mathsf{P}(X \geq x) = \mathsf{P}(\mu + \sigma Z \geq \log x) = \bar{\Phi}(z_x), \quad (1)$$

where

$$z_x := (\log x - \mu)/\sigma.$$

It is easy to verify that $\bar{\Phi}(z) \sim \phi(z)/z$ when $z \to \infty$. Hence, when $x \to \infty$, we have $z_x \to \infty$ and

$$-\log \bar{F}(x) \sim \frac{1}{2\sigma^2}(\log x - \mu)^2 + \log(\log x - \mu). \quad (2)$$

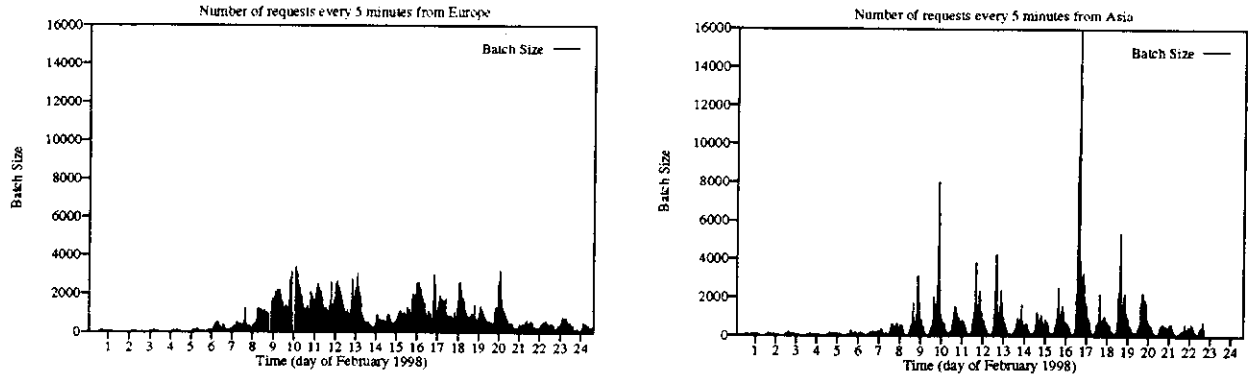

Figure 1: Traffic Patterns from Europe and Asia at Time Scale of 300 seconds

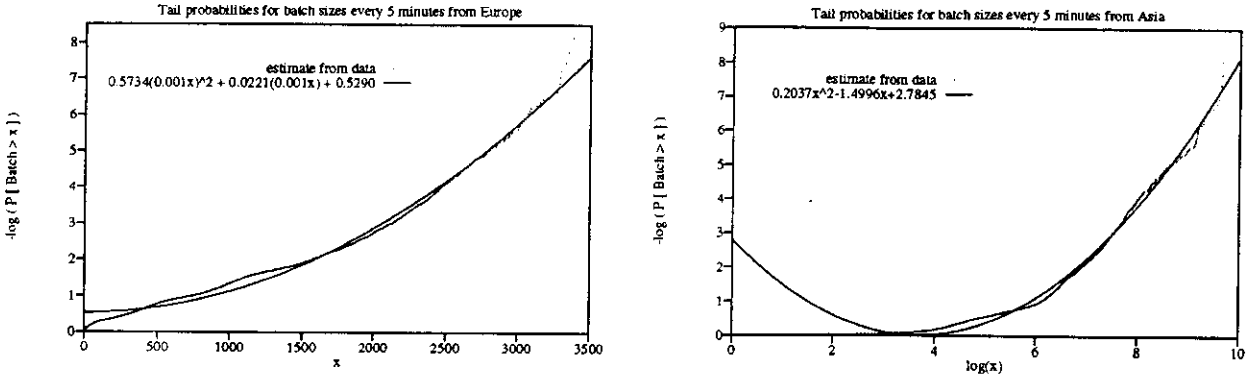

Figure 2: Tail Distribution of the Request Batch Process at 300 Second Intervals

That is, the log-tail distribution behaves as a quadratic function of $\log x$, and in this sense, X follows a heavy-tailed distribution.

We next consider the dependence structure of the sequence $\{X_n\}$. To start, consider the following simple model for the case of a light-tailed marginal distribution. Let

$$X_n = \mu + Y_n \tag{3}$$

for each n, with $\{Y_n\}$ represented by the AR(1) process

$$Y_n = \epsilon_n + \phi_1 Y_{n-1}, \tag{4}$$

where ϕ_1 is a real parameter and $\{\epsilon_1, \epsilon_2, \ldots\}$ is a sequence of i.i.d. normal r.v.s with mean 0 and variance σ_ϵ^2. Since $\{X_n\}$ is assumed to be a stationary sequence, $\{Y_n\}$ must also be a stationary sequence. Hence, for all n, Y_n is also a normal r.v. with mean zero and variance σ_y^2 that must satisfy

$$\sigma_y^2 = \sigma_\epsilon^2 + \phi_1^2 \sigma_y^2,$$

or equivalently

$$\sigma_y^2 = \frac{\sigma_\epsilon^2}{1 - \phi_1^2}. \tag{5}$$

From the recursion in (4), we have

$$Y_n = \epsilon_n + \phi_1 \epsilon_{n-1} + \phi_1^2 \epsilon_{n-2} + \cdots.$$

The autocorrelation function at lag k is then

$$\rho_k := \mathsf{Cov}(Y_n, Y_{n+k})/\mathsf{Var}(Y_n) = \phi_1^k,$$

and thus $\{X_n\}$ has a short-range dependent structure, and so does $\{Y_n\}$. In practice, the parameters of the model are obtained by fitting the data using standard time-series analysis. The data may exhibit more complex dependency structures, in which case a higher order ARIMA process might be needed for a good fit of the data.

For the case of heavy-tailed marginals, we replace (3) by

$$\log X_n = \mu + Y_n. \tag{6}$$

As before, it follows from the stationarity of $\{X_n\}$ that $\{Y_n\}$ is also a stationary process. The relation in (5) therefore still applies, and we have

$$X_n = e^{(\mu + \epsilon_n + \phi_1 \epsilon_{n-1} + \phi_1^2 \epsilon_{n-2} + \cdots)}.$$

The k^{th} moment of X_n is given by

$$\mathsf{E}[X_n^k] = \mathsf{E}[e^{k\mu + kY_n}] = e^{k\mu + k^2\sigma_y^2/2},$$

and we can then derive the autocorrelation function for $\{X_n\}$ to be

$$\rho_k \sim \frac{1}{e^{\sigma_y^2} - 1} \sigma_y^2 \phi_1^k \quad \text{as } k \to \infty.$$

Hence, $\{X_n\}$ continues to be a process with short-range dependence. We can therefore take the log of the data and apply standard time-series analysis to obtain the parameters for the model, using a higher order ARIMA process if the data possesses more complex dependency structures.

Following the above approach for the request traffic from Asia (see Fig. 1), we perform the log transformation on the original times series data starting after February 7 when the Olympic games started. The shape of the autocorrelation function for the transformed time series very closely resembles the autocorrelation function of a seasonal AR(1) model. We therefore performed another transformation to remove the seasonality, namely $Z_n = Y_n - 0.9Y_{n-288}$ where the period is 288 (representing the daily cycle) and the coefficient 0.9 is estimated from the data. The transformed series Z_n is plotted in Fig. 3.

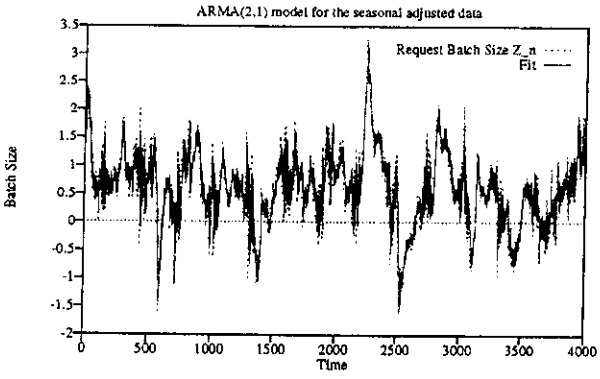

Figure 3: Time-Series Analysis of the Transformed Traffic from Asia

Our time-series analysis shows that the following ARMA(2,1) model provides a good fit for Z_n:

$$Z_n = \mu + H_n,\ H_n = \epsilon_n + \phi_1 H_{n-1} + \phi_2 H_{n-2} + \theta_1 \epsilon_{n-1},$$

where $\mu = 0.62091$, $\phi_1 = 1.23131$, $\phi_2 = -0.24350$, $\theta_1 = 0.59744$ and $\{\epsilon_n\}$ is a sequence of i.i.d. normal r.v.s with mean 0 and variance 0.04427; this is also plotted in Fig. 3. Moreover, the following simpler AR(1) model is also found to provide a reasonably good fit for Z_n:

$$Z_n = \mu + H_n,\ H_n = \epsilon_n + \phi_1 H_{n-1},$$

where $\mu = 0.62313$, $\phi_1 = 0.88332$ and ϵ_n has mean 0 and variance 0.12545.

We constructed a sample trace from the seasonal model by first generating a sample path from the ARMA series $\{H_n\}$, and then calculating $Z_n = \mu + H_n$. From this we obtain Y_n as

$$Y_n = Z_n + 0.9Y_{n-288},$$

and then set $X_n = e^{Y_n} - 3$, which we call an *eARMA*, or *logARMA*, process. The resulting traffic is plotted in Fig. 4. Note that the generated sample path contains some of the key characteristics of the traffic from Asia in Fig. 1.

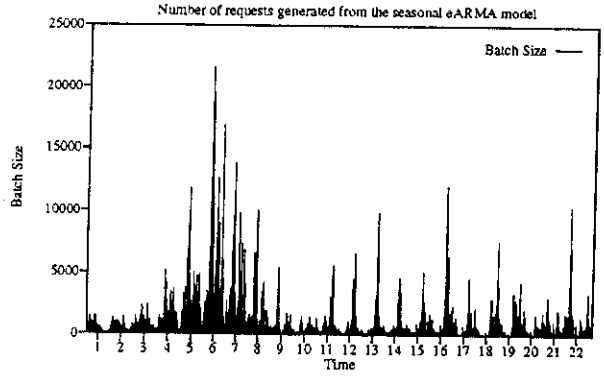

Figure 4: Generated Seasonal eARMA Process

4 Web Server Performance

We now turn to modeling the Web server of interest using a $G/G/1$ queue that takes as input the sequence $\{X_n\}$, which follows the traffic models developed in §3. When considering the waiting-time process, we focus on an individual SP2 node as a G/G/1 queue based on our statistical analysis showing that the set of ND routers has the effect of smoothing and equalizing the statistical properties of the arrival processes for each individual node. With respect to the workload process, however, the G/G/1 queue equally models the entire Web server system.

Suppose the server depletes c units of requests every period, where c is a deterministic constant. In order for the queue to be stable, we assume $c > \mathsf{E}(X)$. Let W_n denote the amount of work (i.e., the number of requests) in the system at time n (i.e., in the n^{th} period). Then, W_n follows the well-known Lindley recursion:

$$W_n = [W_{n-1} + X_n - c]^+. \tag{7}$$

Since $c > \mathsf{E}(X)$, we know that, as $n \to \infty$, W_n will converge (weakly) to a finite r.v., denoted by W_∞.

Consider first the case of a light-tailed input sequence. Iterating on the Lindley recursion in (7) and making use of the stationarity of $\{X_n\}$, we know that W_n is equal in distribution to

$$\max\{0, X_1 - c, X_1 + X_2 - 2c, \ldots, X_1 + X_2 + \cdots + X_n - nc\}.$$

Making use of the traffic model in the previous section, we have

$$S_n := \sum_{i=1}^{n} (X_i - c)$$

$$= n(\mu - c) + \frac{1}{1-\phi_1}[(1-\phi_1)\epsilon_n + (1-\phi_1^2)\epsilon_{n-1} + \cdots + \phi_1(1-\phi_1^n)\epsilon_0 + \phi_1^2(1-\phi_1^n)\epsilon_{-1} + \cdots].$$

Hence, S_n is a normal r.v. with mean $m_n = n(\mu - c)$ and variance (with algebra):

$$\sigma_n^2 = \frac{1+\phi_1}{1-\phi_1}\sigma_y^2\left\{n - \frac{2\phi_1(1-\phi_1^n)}{1-\phi_1^2}\right\}.$$

Following the large deviations theory for the steady-state tail probabilities in a single-server queue [4], we know that as $W_n \to W_\infty$ $w.p.1$, we have

$$x^{-1}\log \mathsf{P}(W_\infty > x) \longrightarrow -\theta^* \tag{8}$$

as $x \to \infty$, where

$$\theta^* = \frac{2(c-\mu)(1-\phi_1)}{\sigma_y^2(1+\phi_1)}. \tag{9}$$

That is, the tail probability of the workload process decreases (as x increases) exponentially fast with rate θ^*. Note that ϕ_1 is the autocorrelation of lag 1 for the input process. Therefore, the higher the correlation of the input process, the slower the decay rate of the tail probability of the workload process, which is very much in line with our intuition.

Suppose the QoS requirement is $\mathsf{P}(W_\infty > x) \le \alpha$, where both x and α are given. Then, making use of (8) and (9), we can derive

$$c = \mu + \frac{\sigma_y^2(1+\phi_1)}{2(1-\phi_1)} \cdot \frac{-\log\alpha}{x}. \tag{10}$$

That is, the required capacity c is equal to the input rate μ, plus a "risk premium" term that is increasing in the burstiness (in terms of both variance and correlation) of the traffic, as well as in the QoS requirement.

We next consider a heavy-tailed input sequence. In this case, the generating function for X does not exist, and thus the large deviations result above no longer applies. To start, consider the case of a renewal input, with the same (heavy-tailed) marginals. Specifically, let $\{X_n'\}$ denote an i.i.d. sequence with the same marginal distribution as $\{X_n\}$, i.e.,

$$X_n' \stackrel{\mathrm{d}}{=} X_n \stackrel{\mathrm{d}}{=} X = e^{\mu+\sigma Z},$$

for all n. Let W_n' and W_∞' respectively denote the corresponding workloads (via the Lindley recursion) at time n and in the limit (i.e., steady state). Then, from standard queueing theory ([1, 3, 6]), we have the following tail distribution for W_∞'

$$\mathsf{P}(W_\infty' \ge x) = [c - \mathsf{E}(X)]^{-1}\int_x^\infty \bar{F}(y+c)dy.$$

Notice that $\bar{F}(y+c) = \mathsf{P}(X_n - c \ge x)$ is the tail distribution of the displacement, $X_n - c$, in the random walk associated with $\{W_n\}$. With a change of variable, $y \leftarrow y + c$, we can rewrite the above tail distribution as

$$\mathsf{P}(W_\infty' \ge x - c) = \frac{\mathsf{E}(X)}{c - \mathsf{E}(X)}\bar{F}_0(x), \tag{11}$$

where

$$\bar{F}_0(x) := \frac{\int_x^\infty \bar{F}(y)dy}{\mathsf{E}(X)}$$

is a familiar object in renewal theory: the stationary excess-life distribution associated with $F(x)$. In the literature on heavy-tailed distributions (e.g., [7]), it also has a prominent status, and its numerator $\int_x^\infty \bar{F}(y)dy$ is usually referred to as the "integrated tail". For instance, the result in (11) simply says that the tail behavior of the delay in a single-server queue with *renewal* input is essentially the same (up to a multiplicative constant) as the *integrated* tail behavior of the input distribution.

Making use of the properties of the log-normal distribution, in particular (1), we can derive the result

$$\bar{F}_0(x) = \frac{\sigma\phi(z_x - \sigma)}{z_x(z_x - \sigma)}. \tag{12}$$

Upon substituting the above into (11), we obtain

$$\mathsf{P}(W_\infty' \ge x - c) \sim \frac{\rho}{1-\rho} \cdot \frac{\sigma\phi(z_x - \sigma)}{z_x(z_x - \sigma)}, \tag{13}$$

where $\rho := \mathsf{E}(X)/c$ is the traffic intensity. Furthermore,

$$-\log \mathsf{P}(W_\infty' \ge x - c) \sim (z_x - \sigma)^2/2 + \log z_x + \log(z_x - \sigma). \tag{14}$$

Comparing the above with (2), we obtain

$$\frac{-\log \mathsf{P}(W_\infty' \ge x - c)}{-\log \mathsf{P}(X' \ge x)} \sim O(1). \tag{15}$$

On the other hand, from (7), we have

$$\mathsf{P}(W_n \ge x - c) \ge \mathsf{P}(X_n \ge x).$$

Hence,

$$\mathsf{P}(W_\infty \ge x - c) \ge \mathsf{P}(X \ge x) = \bar{F}(x), \tag{16}$$

or equivalently

$$-\log \mathsf{P}(W_\infty \ge x - c) \le -\log \bar{F}(x). \tag{17}$$

In a special case considered in [2], where the input process is a Markov modulated Poisson process,

it is shown that the relation in (13) also holds for W. Specifically,

$$\mathsf{P}(W_\infty \geq x - c) \sim C \cdot \bar{F}_0(x), \tag{18}$$

where the constant multiplier C, which depends on the Markov chain that modulates the arrival rate, will of course be different in value from the constant in (13), i.e., $\rho/(1-\rho)$. In other words, when the marginals are heavy-tailed, the dependence structure of the input process does not have much impact on the tail distribution of the delay; the tail behavior is essentially the same, up to a multiplicative constant, as the tail behavior in the case of renewal inputs. Therefore, in view of the above analysis, it appears that in our case, we can reasonably expect the relation in (18) to hold as well.

In summary, we have established that for both light-tailed and heavy-tailed input traffic, the tail behavior of the delay distribution is qualitatively the same as the tail behavior of the input. The parameters involved are, of course, different. For light-tailed input, the delay distribution has an exponential tail, with the decay rate following (9). For heavy-tailed input, with the batch size following a lognormal distribution, it appears that there is enough evidence for us to expect the tail of the delay distribution to also follow a lognormal distribution, which is essentially the integrated tail of the input distribution; refer to (13), (14), (18).

To further support this result, we simulated the above G/G/1 queue under the heavy-tailed input process from §3 and obtained the tail of the delay distribution. We observe from these results that the shapes of the tails of the delay distributions are indeed quadratic, i.e., they follow lognormal distributions and thus are qualitatively the same as the tail behavior of the input processes, which is consistent with the corresponding result established above by our analysis. We also observe from our simulation results that the delay distribution is always worse when the arrival patterns are positively correlated (i.e., $0 < \phi_1 < 1$) than when the arrival patterns contain no correlations (i.e., $\phi_1 = 0$). In fact, the delay improves as the positive correlation decreases from $\phi_1 = 0.9$ toward $\phi_1 = 0$. Interestingly, we further observe that this improvement in the delay distribution continues as ϕ_1 decreases from 0 to -0.4, beyond which the delay distribution continues to worsen as ϕ_1 decreases from -0.4 to -0.9.

5 Periodic Traffic

The results established in the previous section characterize the impact of the marginal distribution and the dependence structure of the user request process on queueing system performance. Another important aspect of the traffic patterns observed in §3 concerns the regular periodic behavior found in the traffic from Asia, where there are obvious peaks around noon and valleys during the night. The lower traffic intensity at night is mainly due to limited Internet access from homes in Japan, whereas the generally high volume of traffic around noon is primarily due to public interest in an ongoing sports event for which most people check on the latest results during their lunch break.

In this section we consider two basic aspects of such periodic traffic patterns. First we use an experiment to simply demonstrate that strong periodic trends in the user request process (including corresponding changes in the dependence structure) can have an important impact on the workload and waiting-time processes. Then we examine an approach to model periodic traffic patterns in the request process using a deterministic function to capture these trends.

Recall that X_n is the number of requests at time n. Consider a sample path of the following eARMA process:

$$X_n = e^{\mu + Z_n}, \tag{19}$$

$$Z_n = \epsilon_n + \phi_1 Z_{n-1}, \tag{20}$$

where $\mu = 4.8, \phi_1 = 0.5$ and $\epsilon_n \sim N(0, \sqrt{2.1(1-\phi_1^2)})$. The upper graph in Fig. 5 plots such a sample path from time 0 to 3600. We then partition this process into 6 equal-length intervals, sorting the 1st, 3rd and 5th intervals in increasing order of the X_n values, and sorting the 2nd, 4th and 6th intervals in decreasing order of the X_n values. Upon concatenating the sorted intervals back together, we obtain a new process that converts the original process into a strongly periodic process, with 3 periods, while preserving the empirical distribution of the original process. Note that this procedure also changes the dependence structure of the original process. The resulting *shuffled process* is plotted in the lower graph of Fig. 5.

We next feed these two request processes into a G/G/1 queue with service capacity $c = 800$ to obtain the corresponding workload processes. These four processes are plotted in Fig. 6. The key point is to observe that the average workload measure for the shuffled process (i.e., 31936) is much larger than that for the original process (i.e., 1500). Moreover, the workload process associated with the shuffled input is much more regular and much less bursty than the workload process for the original input. Such large workload buildups in between relatively long idle periods under the shuffled process in Fig. 6 are due to the large concentration of arrivals occurring at around the same time. In these cases, to support QoS guarantees, one would have to design the system to

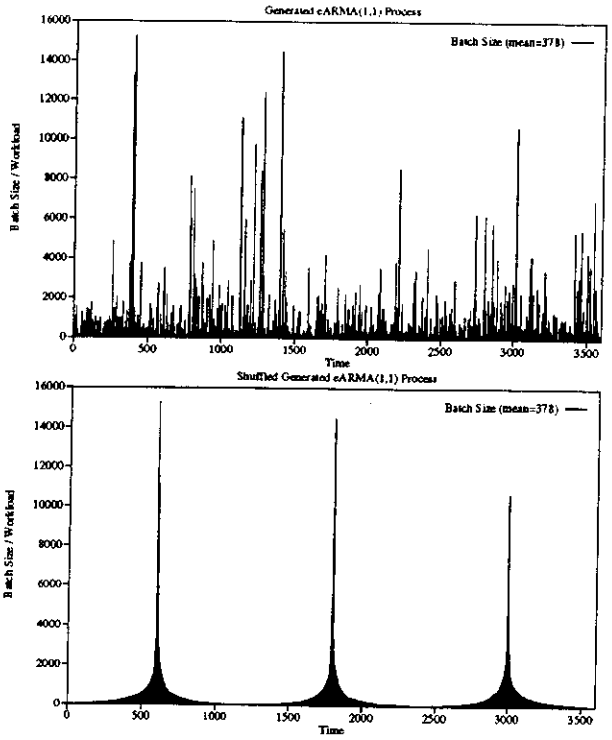

Figure 5: Generated eARMA Process and the Shuffled Process

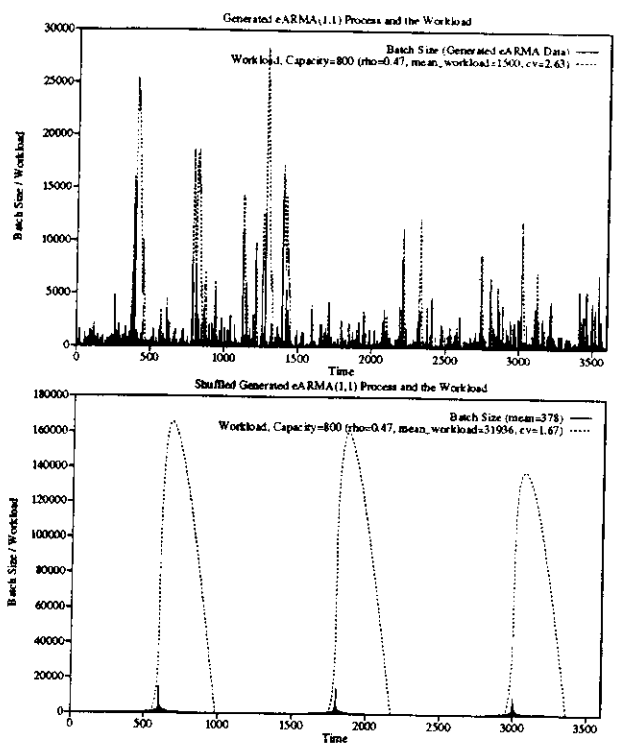

Figure 6: The eARMA and Shuffled Processes and their Workload Processes

have a capacity that is very close to the expected peak load.

We now turn to consider the problem of modeling periodic patterns of this type in the user request traffic. One approach is to first capture the deterministic trend, and then characterize the random variations of the actual traffic about this trend. Intuitively, one would expect that system performance is primarily determined, to first order, by this deterministic trend when the traffic contains strong periodic patterns.

To examine this idea in more detail, let D_n denote the deterministic trend and N_n denote the random fluctuations about this trend. We therefore have

$$X_n = D_n + N_n. \tag{21}$$

Then consider two versions of a G/G/1 queue with fixed capacity c, one with input sequence $\{X_n\}$ and the other with input sequence $\{D_n\}$. We want to compare the workload processes of these two queueing systems. In particular, if $\{D_n\}$, through its time-dependent trend, is the dominant factor that contributes to most of the workload in the system, then we should expect the performance of the two systems to be quite close.

To extract the deterministic trend D_n from the data, we compute the average level of traffic over a moving window of length one hour and use this average for the middle of the time interval D_n; i.e., $D_n = \sum_{i=n-6}^{n+6} X_i/13$. As shown in Fig. 7, the plots of X_n and D_n for the traffic from Asia indicate that they indeed both follow the same pattern. Furthermore, D_n is relatively smoother, and statistically unbiased as well. Upon feeding both of these processes into the same single-server queue with capacity $c = 1300$, we obtain the workload processes plotted in Fig. 8, and observe that the two workload processes virtually coincide with each other. The corresponding noise process N_n is also plotted in Fig. 9, where it is shown to be quite close to following a normal distribution. These results suggest that (21) may be a good (first-order) traffic model for user request processes containing strong periodic trends. We plan to further investigate this as part of our ongoing research.

References

[1] ASMUSSEN, S., Rare Events in the Presence of Heavy Tails. In: *Stochastic Networks: Stability and Rare Events*, P. Glasserman, K. Sigman, and D.D. Yao (eds.), Springer-Verlag, New York, 1996, 197-214.

[2] ASMUSSEN, S., HENRIKSEN, L.F., AND KLÜPPELBERG, C., Large Claims Approximations for Risk Processes in a Markovian

Environment. *Stoch. Proc. Appl.*, **54** (1994), 29-43.

[3] EMBRECHTS, P. AND VERAVERBEKE, N., Estimates for the Probability of Ruin with Special Emphasis on the Possibility of Large Claims, *Insurance: Mathematics and Economics*, **1** (1982), 55-72.

[4] GLYNN, P.W. AND WHITT, W., Logarithmic Asymptotics for Steady-State Tail Probabilities in a Single-Server Queue. In: *Studies in Applied Probability*, J. Galambos and J. Gani (eds.), *J. Appl. Prob.*, **31A** (1994), 131-156.

[5] IYENGAR, A.K., SQUILLANTE, M.S., AND ZHANG, L., Analysis and Characterization of Large-Scale Web Server Access Patterns and Performance. *World Wide Web*, **2** (1999), 85-100.

[6] JELENKOVIĆ, P. AND LAZAR, A.A., A Network Multiplexer with Multiple Time Scale and Subexponential Arrivals. In: *Stochastic Networks: Stability and Rare Events*, P. Glasserman, K. Sigman, and D.D. Yao (eds.), Springer-Verlag, New York, 1996, 215-235.

[7] KLÜPPELBERG, C., Subexponential Distributions and Integrated Tails. *J. Appl. Prob.*, **25** (1988), 132-141.

[8] LELAND, W.E., TAQQU, M.S., WILLINGER, W. AND WILSON, D.V., On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Trans. on Networking*, **2(1)**, (1994), 1-15.

[9] SHAKED, M. AND SHANTHIKUMAR, J.G., *Stochastic Orders and Their Applications*. Academic Press, New York, 1994.

[10] WILLINGER, W., TAQQU, M.S., SHERMAN, R. AND WILSON, D.V., Self-Similarity through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level. *IEEE/ACM Trans. on Networking*, **5(1)**, (1997), 71-86.

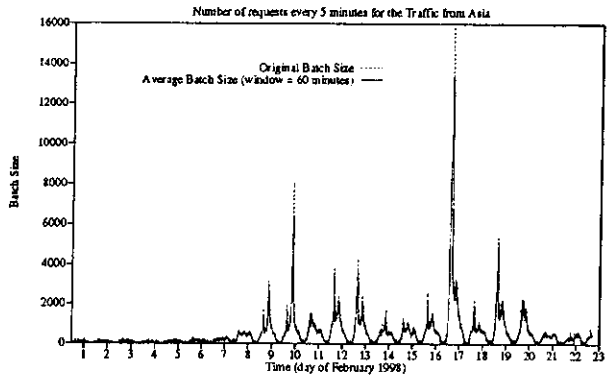

Figure 7: The Original Traffic from Asia and the corresponding Deterministic Trend

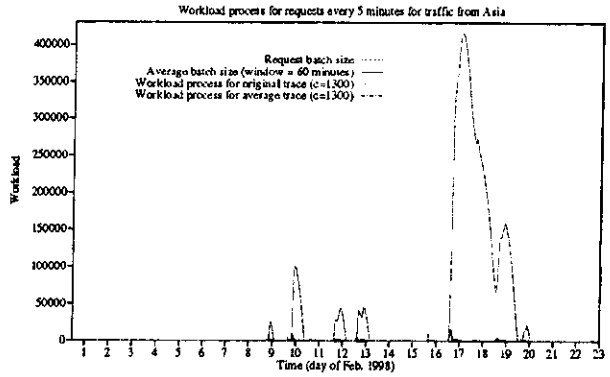

Figure 8: The Workload Processes for the G/G/1 Queue with the Original Traffic and its Deterministic Trend as Input

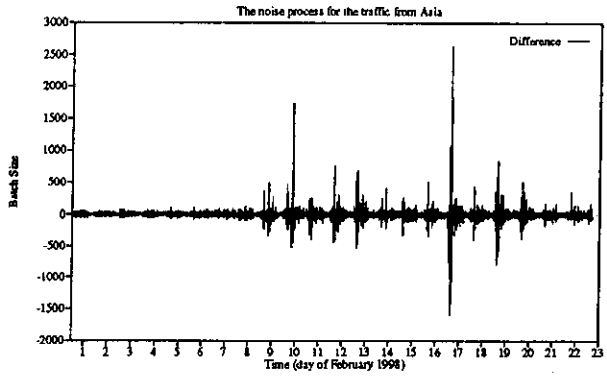

Figure 9: The Difference between the Original Traffic and its Deterministic Trend