Evaluating Self-Similar Processes for Modeling Web Servers *

Ronit Nossenson
Dept. of Computer Science
Technion, Haifa, 32000, Israel
e-mail: ronitt@cs.technion.ac.il

Hagit Attiya
Dept. of Computer Science
Technion, Haifa, 32000, Israel
e-mail: hagit@cs.technion.ac.il

Keywords: model-verification, simulation, statistical-analysis, self-similar processes, heavy-tailed distributions, Web server performance.

Abstract

The accuracy of self-similar processes that are widely used to model Web server systems is evaluated using simulation. Specifically, we consider two processes with self-similarity from a *single* origin: either with a *realistic* self-similar arrival process or with a *realistic* service-time distribution, but not both. A detailed examination and comparison of these processes is presented, together with conclusions regarding the scenarios in which one process outperforms the other.

The main results of the simulation are that when the system has medium or low utilization levels, both processes fail to estimate the realistic *maximum number of clients in the system* and the realistic *average response time*.

1 INTRODUCTION

The performance of a given Client/Server system is affected mainly by its statistical characteristics: the arrival process of clients to the server and the distribution of the service time.

The request arrival process in Web servers is an ON/OFF process [5, 7], with several requests during an ON period followed by an OFF period that is significantly longer than the interarrival times during the ON period. The ON, the OFF times and the inter-arrival times within the ON-times have *heavy-tailed* distributions. Such distributions are characterized by extremely high variability, and may have infinite variance and infinite mean. Figure 1 depicts a realistic arrival process together with a Poisson arrival process.

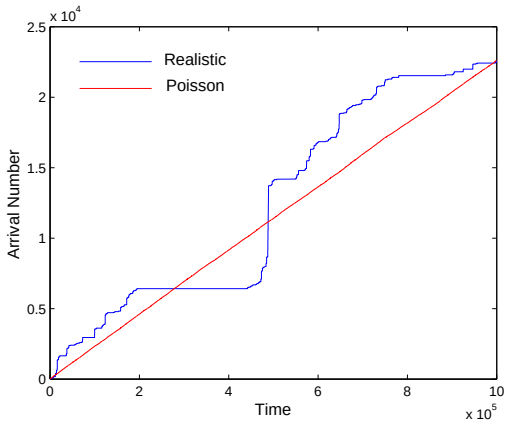

Figure 1: Realistic and Poisson Arrival Processes

When considering the service-time properties of Web servers, we restrict our attention to service in which clients (browsers) send requests for files, and servers reply with the requested files. In such systems, the service-time distribution is characterized by the files' *transmission-duration distribution* (TDD). This distribution was shown to be heavy-tailed as well [6, 13]. In fact, there is evidence that the TDD of the same file from the same server to the same client is heavy-tailed [13]. Thus, every Web server, including servers that serve a single static page, suffers from bursty service-time.

*This research was supported by the *Israel Science Foundation* (grant number 105/01).

Either a heavy-tailed arrival process [10] or a heavy-tailed service-time distribution [2] suffices to create *self-similar* behavior, that is, a process which is bursty over a wide range of time-scales. Practically, a self-similar process does not become smoother and closer to its average as the time scale grows. Parameters such as the average rate, or other moments, do not give valuable information about a self-similar process. Indeed, studies show that performance figures are radically different when simulations use data that incorporates long-range dependence and when simulations use data without long-range dependence [8, 15].

It is now widely accepted that when evaluating Web servers' applications and policies, e.g., for load balancing or task scheduling, the assumption that the process is self-similar is important to accomplish accurate analysis results (for example, [4, 11]).

Using simulation, we show in this study that an accurate queueing model for a Web server system should consider the *two* origins of self-similarity. Specifically, the model should incorporate both the *Realistic* arrival process and the *Realistic* service-time distribution. We present evidence that when the system has low to medium utilization, processes with self-similarity from a *single* origin (that is, either a realistic arrival process or a realistic service-time distribution, but not both) do not estimate the performance parameters in a reliable manner.

So far, models of a Web server system considered processes with self-similarity from single origin only (for example, [9]). A process with self-similarity from two origins is hard to analyze since it is not a Markovian process and it does not fit into any known solution of a queue with general arrival and general service-time distribution (G/G/1 queue). Boxma and Cohen [3] obtain *heavy-traffic* results for the *waiting time* of single server queue with heavy-tailed distributions, but there are no theoretical results for low to medium utilization.

Boxma and Cohen [2] consider a realistic heavy-tailed service-time distribution with a Poisson arrival process that is not self-similar. We refer to such single origin self-similar process with Poisson arrivals and Realistic service as the M/R/1 model. Respectively, other models [8, 17] consider an ON/OFF arrival process but assume Exponential service-time distribution, which is not heavy-tailed. Such single origin self-similar process with Realistic arrivals and Exponential service is referred to here as the R/M/1 model.

This study evaluates the accuracy of each single origin self-similar process, M/R/1 and R/M/1, compared to the process with Realistic arrivals and Realistic service, called R/R/1. We estimate the error factor of analyzing a Web server system using each single origin self-similar model instead of a realistic one, to evaluate which of these models is better and in which scenario.

We also study the influence of each parameter of the realistic arrival process and service distribution on the reliability of each single origin self-similar model. Understanding the impact of each parameter is helpful for designing more realistic models.

The first finding of our study is that when the system has low to medium utilization, both the R/M/1 model and the M/R/1 model fail to estimate the realistic system performance parameters in a reliable way. The error factor is very large in both the *maximum client in the system* and the *average response time* performance parameters. For example, under medium utilization (see Section **??**), the R/M/1 model's estimation of the maximum client in the system has an error factor of 153.39 and its estimation of the average response time has an error factor of 932.64.

When the utilization is high, both models perform well. Unfortunately, scenarios of high utilization, in which tens-thousands of clients are simultaneously in the system, are less interesting in the Web context since most likely, clients will drop out from the system. For example, under high utilization (see Section 4), the R/M/1 model's estimation of the maximum client in the system of has an error factor of 1.03 and its esti-

mation of the average response time has an error factor smaller than 2, but the system was over-loaded for long time periods, in which it has more than 58500 clients simultaneously inside it.

We found that system performance is greatly influenced by the ratio between the *location* parameters of the ON-times and OFF-times and the ratio between the *shape* parameters of the ON-times and OFF-times. In particular, the system is over-loaded for long periods when the ratio of the location parameters of the ON-times and the OFF-times is larger than 1 : 20. This also happens when the shape parameter of the ON-times is smaller than the shape parameter of the OFF-times. In other cases, the system is hardly over-loaded.

Another interesting discovery of this research is that the shape parameter of the inter-arrival times within the ON-times have an almost negligible impact on the system performance. Changing this shape parameter from 2 to 0.75 has almost no effect on the system. These results indicate that it is reasonable to assume that the inter-arrival times are not heavy-tailed distributed (as done, for example, in [17]).

2 DEFINITIONS

The Exponential distribution is *light-tailed*, in contrast to a *heavy-tailed* distribution, formally defined as follows.

Definition 2.1 *A random variable X has a* heavy-tailed distribution *with tail index α, $0 < \alpha < 2$, if*

$$P[X > x] \sim x^{-\alpha}, \quad as \quad x \to \infty$$

Where $a \sim b$ means that $\lim_{x\to\infty} a/b = c$ for some constant c.

Heavy-tailed distributions are characterized by extremely high variability, which increases sharply as α decreases. Such a distribution has infinite variance; if $\alpha \leq 1$, then it also has infinite mean.

A simple heavy-tailed distribution is the *Pareto* distribution with *shape* parameter α and *location* parameter k. It has the following cumulative distribution function.

$$F(x) = P[X \leq x] = 1-(k/x)^{\alpha}, \quad \alpha,\, k > 0,\, x \geq k,$$

3 SIMULATION SETUP

We use the *Client in the System* process along the time axis, that is, the number of clients in the queue plus one (the client that is currently serviced). Other performance measurement parameters that we use are the *maximum number of clients* that are simultaneously in the system and the *average response time* experienced by the clients. These two parameters determine resources required by the system, such as buffer-size and computing power, and the system's Quality of Service.

Recall that the system utilization, ρ, is the arrival rate divided by the service rate and in a stable system $0 < \rho < 1$. Our evaluation is performed under three levels of system utilization: *low* in which $\rho < 0.3$, *medium* in which $0.3 < \rho < 0.5$, and *high* in which $\rho > 0.5$. Due to the high variability of the distributions, the system is already over-loaded when $\rho \geq 0.5$.

We study the *potential* of the R/M/1 and the M/R/1 models in estimating the realistic model performance. To provide the R/M/1 and the M/R/1 models the best possible comparison results, we use *actual* parameters fitting. That is, each simulation is divided into three stages. First, we simulate the realistic model and calculate its actual arrival and service rates. Second, we simulate the R/M/1 model with the same realistic arrival process and with service-rate parameter equal to the actual realistic service rate. Respectively, in the third stage, we simulate the M/R/1 model with arrival-rate parameter equal to the actual realistic arrival rate and with the same realistic service process.

Following the characteristic of [1], our *base configuration* has the following parameters. For the arrival process: the ON-times location pa-

rameter is 10 and the shape parameter is 1.0; the OFF-times location parameter is 1000 and the shape parameter is 1.0; the inter-arrivals location parameter is 0.2 and the shape parameter is 2.0. For the service-time distribution the location parameter is 0.000001 and the shape parameter 0.5.

4 THE ARRIVAL PROCESS

We start with simulation series that analyze the general effect of the arrival rate on the reliability of the R/M/1 and the M/R/1 models.

The simulation results are presented for low (Figure 2 and Figure 3), medium (Figure 4 and Figure 5) and high arrival rates and utilizations (Figure 6 and Figure 7). The results of the comparison between the model's maximum number of client simultaneously in the system and their average response time appear in Table 1.

The immediate conclusion from this comparison is that under low and medium utilization, both the R/M/1 model and the M/R/1 model fail to estimate the performance parameters of the R/R/1 model. Furthermore, as can easily be seen from the above figures, although these models describe self-similar processes, they do not predict the bursty nature of the realistic model. Under low and medium utilization level, the M/R/1 model outperforms the R/M/1 model, while under high utilization this direction change and the R/M/1 model outperforms the M/R/1 model. Note that under high utilization the R/M/1 model is very accurate with error factor smaller than 2 for both performance parameters.

Clearly, when the system is empty or overloaded all models are similar. The interesting scenarios which emphasize the differences between the models are in fact in the middle: when the system is not almost empty and not overloaded. It can be seen from the above figures that under medium utilization, the gap between either the R/M/1 model or the M/R/1 model and the realistic model, is the highest.

Figures 6 and 7 graphically illustrate the essential difference in the behavior of the R/R/1, R/M/1 and M/R/1 models. In R/R/1 the typical shape of the client in the system graph is oblong like. The bursty arrivals, and in particular, arrival peaks, cause the sharp ascent of the lines. The bursty service has two effects: very long service-times cause long horizontal lines, while very short service-times result in sharp declines.

In contrast, the typical shape of both the R/M/1 and the M/R/1 client in the system graphs are right-angle triangles like. The R/M/1 model's triangles are caused by the bursty arrivals and smooth service and thus the right angles are on the left. The M/R/1 model's triangles are caused by smooth arrivals and bursty service, and thus the right angles are on the right.

Note that the comparison between R/M/1 model and the R/R/1 is sensitive to changes in the arrival process—the gap between the is very large at low and medium utilization and is reduced at high utilization—although they have the same arrival process.

Each of the ON-times, the OFF-times and the inter-arrivals times (within the ON-times) distributions has location and shape parameters. In the full version of the paper [14] we describe the results of several simulation series, each analyzing the effect of one parameter of the arrival process on the performance of the realistic model, and its effect on the reliability of the R/M/1 and the M/R/1 models.

The conclusions from these simulations are that, again, under low and medium utilization, both models fail to estimate the parameters of the realistic model, and they do not predict its bursty nature. However, under high utilization, both models perform well.

5 THE SERVICE PROCESS

In this section we analyze the general effect of the service rate on the reliability of the R/M/1 and the M/R/1 models. In addition, the full version of the paper [14] include the results of

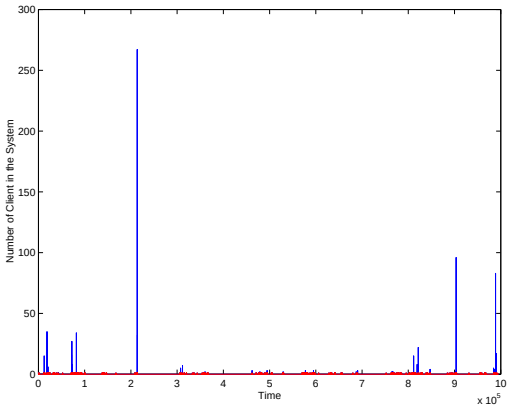

Figure 2: Client in the System Processes (R/R/1 is the blue line and R/M/1 is the red line) with low arrival rate and utilization.

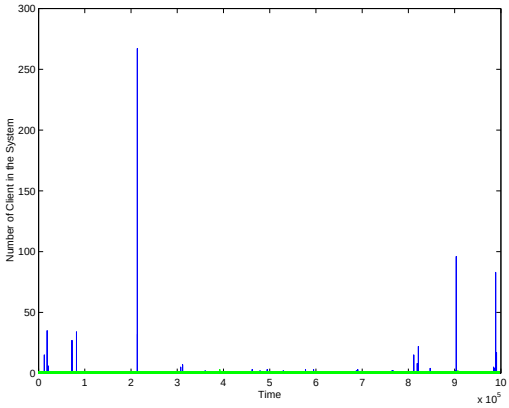

Figure 3: Client in the System Processes (R/R/1 is the blue line and M/R/1 is the green line) with low arrival rate and utilization.

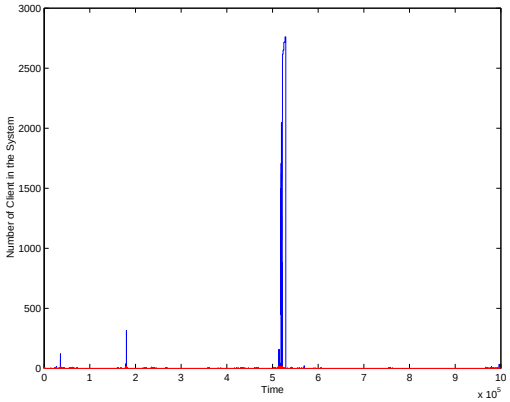

Figure 4: Client in the System Processes (R/R/1 is the blue line and R/M/1 is the red line) with medium arrival rate and utilization.

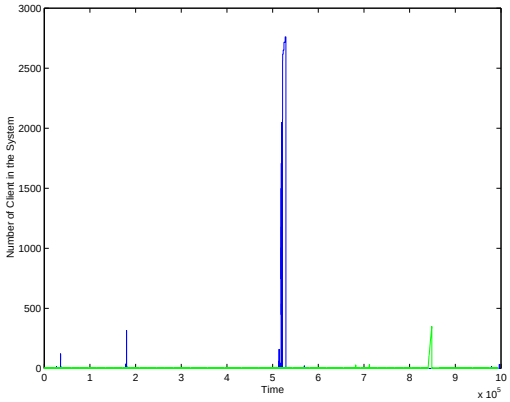

Figure 5: Client in the System Processes (R/R/1 is the blue line and M/R/1 is the green line) with medium arrival rate and utilization.

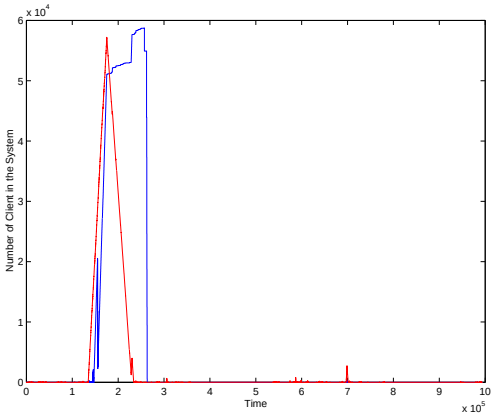

Figure 6: Client in the System Processes (R/R/1 is the blue line and R/M/1 is the red line) with high arrival rate and utilization.

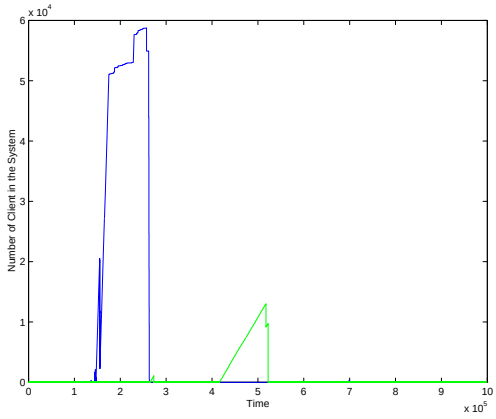

Figure 7: Client in the System Processes (R/R/1 is the blue line and M/R/1 is the green line) with high arrival rate and utilization.

Utilization level	Model	Maximum clients in the system	Error factor	Average response time	Error factor
Low	R/R/1	267		1.080	
	R/M/1	1	267.00	0.019	56.84
	M/R/1	3	89.00	0.024	45.00
Medium	R/R/1	2761		552.125	
	R/M/1	18	153.39	0.592	932.64
	M/R/1	348	7.93	37.187	14.85
High	R/R/1	58771		41519.760	
	R/M/1	57197	1.03	21232.644	1.96
	M/R/1	13054	4.50	5527.202	7.51

Table 1: Maximum number of clients in the system and average response time

two simulation series, analyzing the effect of the location and the shape parameters of the service-times distribution on the reliability of the R/M/1 and the M/R/1 models.

The simulation results are presented for high (Figure 8 and Figure 9), medium (Figure 10 and Figure 11) and low service rate (Figure 12 and Figure 13). Table 2 presents the results of the comparison between the model's maximum number of client simultaneously in the system and their average response time. The results of the comparison between the model's maximum number of client simultaneously in the system and their average response time appear in Table 2.

The immediate conclusions from this comparison are that under low utilization, both models fail to estimate the performance parameters of the realistic model. Under low and high utilization levels, the M/R/1 model provides more accurate estimations for both performance parameters than the R/M/1 model. This direction change under medium utilization and the R/M/1 model provides more accurate estimations. Note that under medium utilization, the R/M/1 model perform well with error factors smaller than 2, while the M/R/1 model fails to estimate the parameters. Under high utilization, both models perform well and provide good estimates for the parameters.

Note that the comparison between M/R/1 model and the R/R/1 is sensitive to changes in the service process—the gap between them changes in different utilization levels—although they have the same service process.

6 DISCUSSION

This research indicates that models using a process with self-similarity from a single origin are often unsuitable to describe a Web server system. It would be interesting to investigate the behavior of these processes resources are restricted, for example, bounded buffer size and limited client patience. We expect that in this more realistic environment, the estimations of the single origin self-similar processes will be even worse.

Clearly, new models with self-similarity from two origins should be developed. Recently, some theoretical studies concerning such models were proposed (for example, [12, 16, 18]), but they are not commonly used to evaluate Web servers. To understand the accuracy of these models, it would be useful to perform simulations similar to those performed in this paper.

References

[1] P. Barford and M. Crovella. Generating representative Web workloads for network and server performance evaluation. *SIGMETRICS '98/ PERFORMANCE '98*, pages 151–160.

[2] O. J. Boxma and J. W. Cohen. The M/G/1 queue with heavy-tailed service time distribution. *IEEE Journal on Selected Areas in Communications*, 16(5):749–763, June 1998.

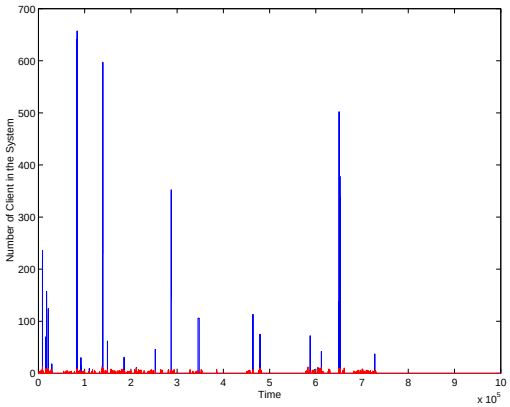

Figure 8: Client in the System Processes (R/R/1 is the blue line and R/M/1 is the red line) with high service rate processes

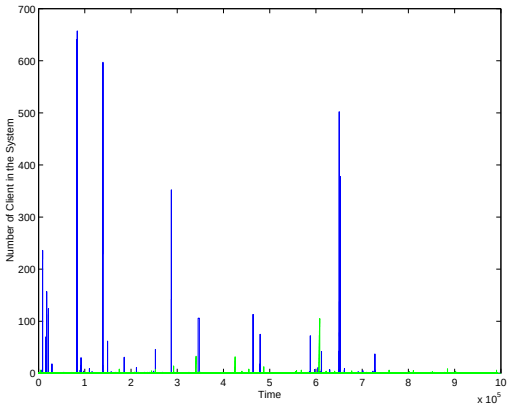

Figure 9: Client in the System Processes (R/R/1 is the blue line and M/R/1 is the green line) with high service rate processes

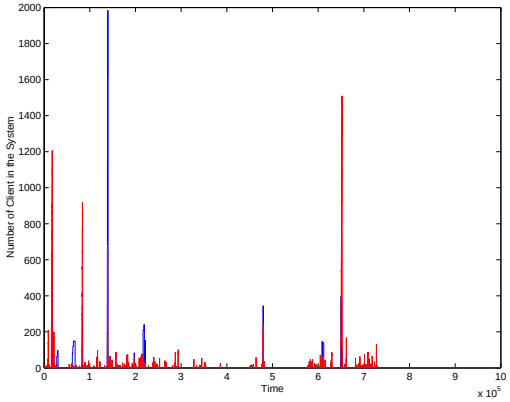

Figure 10: Client in the System Processes (R/R/1 is the blue line and R/M/1 is the red line) with medium service rate processes

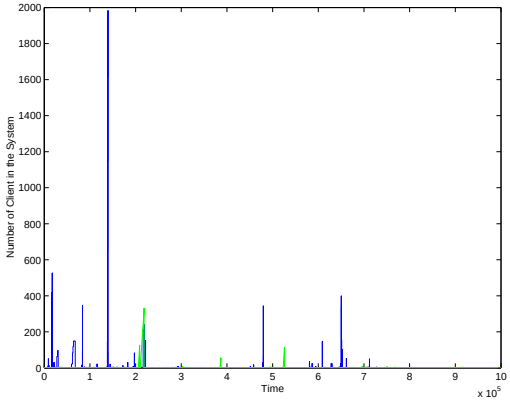

Figure 11: Client in the System Processes (R/R/1 is the blue line and M/R/1 is the green line) with medium service rate processes

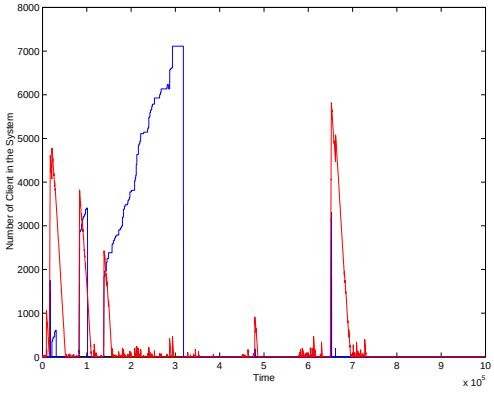

Figure 12: Client in the System Processes (R/R/1 is the blue line, R/M/1 is the red line and M/R/1 is the green line) with low service rate processes

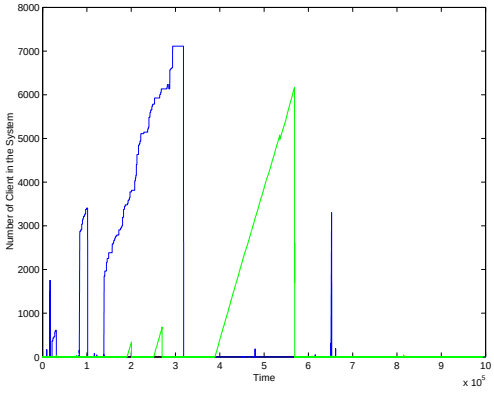

Figure 13: Client in the System Processes (R/R/1 is the blue line, R/M/1 is the red line and M/R/1 is the green line) with low service rate processes

Utilization level	Model	Maximum clients in the system	Error factor	Average response time	Error factor
Low	R/R/1	657		36.134	
	R/M/1	12	54.75	0.423	85.42
	M/R/1	105	6.26	5.836	6.19
Medium	R/R/1	1983		142.316	
	R/M/1	1508	1.31	166.831	0.85
	M/R/1	331	5.99	51.439	2.77
High	R/R/1	7113		26305.072	
	R/M/1	5812	1.22	9386.372	2.80
	M/R/1	6173	1.15	16151.190	1.63

Table 2: Maximum number of clients in the system and average response time

[3] O. J. Boxma and J. W. Cohen. Heavy-traffic analysis for the GI/G/1 queue with heavy-tailed distributions. *Queueing Systems*, 33(1):177–204, Jan. 1999.

[4] G. Ciardo, A. Riska, and E. Smirni. Equiload: a load balancing policy for clustered web servers. *Performance Evaluation*, 46, 2001.

[5] M. Crovella and A. Bestavros. Self-similarity in world wide web traffic evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, Dec. 1999.

[6] M. Crovella, M. Taqqu, and A. Bestavros. Heavy-tailed probability distributions in the world wide web. In *A practical Guide To Heavy Tails: Statistical Techniques and Application*, chapter 1, pages 3–26. Birkhauser, Boston, 1998.

[7] S. Deng. Empirical model of www document arrivals access link. *ICC'96*, pages 17–23.

[8] A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Transactions on Networking*, 4(2):209–223, 1996.

[9] M. Harchol-Balter, B. Schroeder, N. Bansal, and M. Agrawal. Size-based scheduling to improve web performance. *ACM Transactions on Computer Systems (TOCS)*, 21(2):207–233, 2003.

[10] K. Kant and Y. Won. Server capacity planning for web traffic workload. *IEEE transaction on knowledge and data engineering*, 11(5):731–747, Oct. 1999.

[11] E. M. Nahum, T. P. Barzilai, and D. D. Kandlur. Performance issues in WWW servers. *IEEE/ACM Transactions on Networking*, 10(1), Feb. 2002.

[12] R. Nossenson and H. Attiya. Evaluating web server performance with an extension of the N-burst model. Technical Report CS-2002-10, Department of Computer Science, Technion, 2002.

[13] R. Nossenson and H. Attiya. The distribution of file transmission duration in the web. *SPECTS 2003*, pages 647–654.

[14] R. Nossenson and H. Attiya. Evaluating self-similar processes for modeling web servers. Technical Report CS-2003-11, Department of Computer Science, Technion, 2003.

[15] K. Park, G. Kim, and M. Crovella. On the effect of traffic self-similarity on network performance. In *Proceedings of the SPIE International Conference on Performance and Control of Network Systems*, Nov. 1997.

[16] A. Riska, M. Squillante, S.-Z. Yu, Z. Liu, and L. Zhang. Matrix-analytic analysis of a map/ph/1 queue fitted to web server data. In G. Latouche and P. Taylor, editors, *Matrix-Analytic Methods: Theory and Applications*, pages 335–356. World Scientific, 2002.

[17] H. Schwefel and L. Lipsky. Impact of aggregated, self-similar on/off traffic on delay in stationary queueing models. In *Performance Evaluation 43*, pages 203–221, 2001.

[18] C. H. Xia and Z. Liu. Queueing systems with long-range dependent input process and subexponential service times. *SIGMETRICS 2003*, pages 25–36.