Traffic Model and Performance Evaluation of Web Servers

Zhen Liu, Nicolas Niclausse, Cesar Jalpa-Villanueva, Sylvain Barbier

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Traffic Model and Performance Evaluation of Web Servers

Zhen Liu — Nicolas Niclausse — Cesar Jalpa-Villanueva — Sylvain Barbier

N° 3840

Décembre 1999

THÈME 1

ISSN 0249-6399 ISRN INRIA/RR--3840--FR+ENG

INRIA
SOPHIA ANTIPOLIS

Traffic Model and Performance Evaluation of Web Servers

Zhen Liu , Nicolas Niclausse , Cesar Jalpa-Villanueva , Sylvain Barbier

Thème 1 — Réseaux et systèmes

Projet Mistral

Rapport de recherche n° 3840 — Décembre 1999 — 28 pages

Abstract: In this paper we present a new model of Web traffic and its applications in the performance evaluation of Web servers. We consider typical behavior of a user's hypertext navigation within a Web server. We propose a traffic model at the session level, formulated as a stochastic marked point process, which describes when users arrive and how they browse the server. We provide results of statistical analyses and goodness-of-fit of various simple parametric distributions and of their mixtures. We developed a Web server benchmark: WAGON (Web trAffic GeneratOr and beNchmark), and we validated the traffic model by comparing various characteristics of the synthetic traffic generated by WAGON against measurements.

We then report benchmark and analysis results on the Apache server, the currently most used Web server software. We analyze the impact of the traffic parameters on the HTTP request arrival process and the packet arrival process. We also show that the aggregate traffic is self-similar in most cases, and that, more importantly, the Hurst parameter is increasing in the traffic intensity. We further provide performance comparison results between HTTP1.0 and HTTP1.1 and show that HTTP1.1 could be much worse for users as well as for servers if some control parameters of the server and of browsers are incorrectly set. Indeed, when the server load is relatively high, the page response time under HTTP1.1 increases exponentially fast with the number of parallel persistent connections and with the timeout value used in Apache for persistent connections. We investigate the impact of user network conditions on the server performance. We also propose a queueing model to analyze the workload of persistent connections on the Apache server, and we establish optimal solution of the timeout parameter for the minimization of workload.

Based on our analyses, we suggest the following practical guidelines. It is usually beneficial for both Web servers and Web clients to use HTTP1.1 instead of HTTP1.0. When HTTP1.1 is used, it should be used with pipeline. In terms of the management of persistent connections, it is useful for browsers to implement Early Close policy which combines the advantages of both HTTP1.0 and HTTP1.1. Browsers should in general, except for users with low bandwidth network connections (such as Modem), avoid establishing multiple parallel persistent connections from one browser window to the same Web server. On the server side, servers should set small timeout values for persistent connections if fixed timeout control mechanism is used (as in Apache) or if dynamic timeout control mechanism is used and the measured workload is high.

Key-words: Web server performance, Web traffic modeling, Apache server, server model, HTTP1.0, HTTP1.1, persistent connection, self-similarity, user perceived quality of service.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, B.P. 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : 04 92 38 77 77 - International : +33 4 92 38 77 77 —Fax : 04 92 38 77 65 - International : +33 4 92 38 77 65

Modélisation de trafic et évaluation de performances de Serveurs Web

Résumé : Dans ce papier, nous présentons un nouveau modèle de trafic Web et ses applications à l'évaluation de performance de serveurs Web. On considère le comportement typique des utilisateurs lors d'une navigation hypertexte sur un serveur Web. Nous proposons un modèle de trafic au niveau de la session, formulé comme un processus stochastique ponctuel marqué qui décrit quand un utilisateur arrive sur le serveur et comment il navigue sur celui-ci. Nous apportons des résultats d'analyse statistique et de tests d'ajustement de plusieurs distributions paramétriques simples, ainsi que de leur mélange. Nous avons développé un logiciel de «benchmark» de serveur Web, WAGON (Web trAffic GeneratOr and beNchmark) qui implémente ce modèle, et nous avons validé le modèle en comparant plusieurs caractérisations du trafic synthétique généré par WAGON avec des mesures.

Ensuite, nous décrivons les résultats d'expérimentations menées avec le serveur Apache (le serveur Web le plus utilisé dans le monde actuellement). Nous présentons une analyse de l'impact de différent paramètres du trafic sur les processus d'arrivée de requêtes HTTP et de paquets IP. Nous montrons également que le processus agrégé est auto-similaire dans la plupart des cas, et que le paramètre de Hurst croit avec l'intensité du trafic. De plus, nous comparons les résultats obtenus avec les protocoles HTTP/1.0 et HTTP/1.1 et nous montrons que HTTP/1.1 peut être beaucoup moins efficace pour les utilisateurs si certains paramètres du serveur et des clients ne sont pas correctement fixé. En effet, lorsque la charge est assez importante, le temps de réponse augmente exponentiellement avec le nombre de connexions persistantes parallèles et avec la valeur du *timeout* du serveur Apache. Nous analysons également l'influence de la connectivité réseau des clients sur les performances du serveur. Enfin, nous proposons un modèle de file d'attente pour analyser la charge des connexions persistantes pour le serveur Apache, et nous établissons une solution optimale de la valeur du *timeout* pour minimiser la charge.

Basé sur nos analyses, nous suggérons les pratiques suivantes: il est généralement avantageux pour les clients et les serveurs Web d'utiliser HTTP/1.1 à la place de HTTP/1.0; lorsque HTTP/1.1 est utilisé, il doit l'être avec le pipeline. En terme de gestion des connexions persistantes, les clients ont intérêt à implémenter la politique «Early Closing» qui combine les avantages de HTTP/1.0 et HTTP/1.1. En général, les navigateurs ne devraient pas, sauf pour les clients ayant un débit très faible (avec un modem, par exemple), utiliser plusieurs connexions persistantes en parallèle sur un même serveur Web. Du coté serveur, la valeur de *timeout* des connexions persistantes doit être peu élevée si le serveur utilise un mécanisme de *timeout* fixe (comme apache) ou s'il utilise un mécanisme de *timeout* dynamique, le *timeout* doit diminuer lorsque la charge augmente.

Mots-clés : Performance de serveurs Web, modélisation du trafic Web , serveur Apache, modélisation de serveur , HTTP1.0, HTTP1.1, connections persistantes , auto-similarité, qualité de service perçu par l'utilisateur

1 Introduction

The exponential increase of the number of servers and of the number of users causes performance problems of access to Web objects, due to the saturation of Web servers and of the communication network. One of the main preoccupations of Web server administrators is to propose a fast and reliable service to satisfy their users. The tremendous success of the Web makes this task difficult to accomplish. Indeed, not only is it necessary to provide an efficient service for a given time instant, but also it is required to anticipate traffic growth in order to maintain the quality of service for short or medium term. It is thus important to understand the statistical properties of the Web traffic and to develop appropriate traffic models for the performance evaluation of Web servers and Web applications.

Many studies have identified important Web traffic characteristics. Heavy tail distributions for document sizes, popularity, and requests have been reported in [1] and [7]. In [6] the self-similar nature of Web traffic is demonstrated and explained. In [2] six different data sets are used to identify some invariants in Web traffic. HTML and image files account for 90-100% of requests, mean transfer size is less than 21 kBytes, file size distribution is Pareto, 10% of accessed files account for 90% of server requests and 90% of bytes transferred. It was also found, in analyzing the extended logs of the 1994 Californian congressional elections Web server [19], that no correlation exists between file size and connection time for files under 30 kBytes, that the majority of traffic was generated by transfers of small images, and that request arrivals did not appear to follow a pure Poisson process. In [11, 24] a work on the probabilistic model of the number of pages that a user visits within a Web site was presented. It seems that inverse Gaussian distribution is appropriate. In [29, 28, 12], statistical characteristics of the Web request traffic patterns in dynamic and heavily-accessed Web server environments are analyzed. The authors developed traffic models at the request level using data from the official Web site during the 1998 Winter Olympic Games in Nagano, Japan. Their analysis of the traffic data illustrates traffic patterns that exhibit both light-tailed and heavy-tailed behaviors. They also feed these traffic processes into one of the Web server systems modeled as a general single-server queue and analyze the waiting-time process.

At the time HTTP1.1 was proposed, a comparison study between HTTP1.0 and HTTP1.1 was carried out [21]. The effect of persistent connections, pipelining, document compression, bandwidth, and latency, on the performance was investigated using a single Web page containing 42 embedded images. A work to characterize Web response time is presented in [14]. Four proxy logs files are used, they are replayed using a URL re-player multi-process program. The workload is replayed by reading a log file of URLs, sending HTTP requests, and timing the transfer. Ten experiments were done to investigate the effects of proxy caching, network bandwidth, traffic load, and persistent connections. The periodicity of the response times was also studied. More recently, two other studies on the performance comparisons between HTTP1.0 and HTTP1.1 were reported. One focuses on the effect of the CPU and memory capacities [4]. The other reports preliminary studies on the performance difference of the two versions of HTTP under different network connection conditions (bandwidth and propagation delay) [15].

There exist several Web benchmarks. WebSTONE [31] is a distributed, multi-process benchmark. The WebSTONE generates HTTP traffic that allows to stress an HTTP server. The load is generated by requesting pages and files from the server as fast as the server can send them. The existing servers currently in use on the Web are represented by four different files (small, large and mixed pages). SPECweb [27] works in the same way as WebSTONE (both two are based on LADDIS benchmark for file systems) with the difference that the workload is composed of a mixture of four file classes, where the weight of each class is obtained from the analysis of logs. Benchmark hbench:Web [17] follows the same paradigm as WebSTONE and SPECweb. hbench:Web derives its workload by analyzing existing Web server logs to determine the site's page set, a collection of user profiles, and the inter-arrival time between users. In particular, the workload is generated according to the empirical distributions of user inter-arrival times. httperf [20] permits generating HTTP workload in several ways: as HTTP calls generated deterministically and at a fixed rate, as sessions consisting of a number of HTTP call bursts spaced by a fixed user think time (also sessions are generated deterministically and at a fixed rate), as URL requests generated over and over again, and as URL request sessions generated at a given rate.

SURGE [3] is a benchmark tool that tries to imitate a stream of HTTP requests originating from a fixed population of web users. A user is modeled by an ON/OFF process (User Equivalent) who during the ON period makes requests for Web files and during the OFF period lies idle. Within an ON period there are active OFF time periods corresponding to the time between transfer of components of a page. Inactive OFF time periods correspond to the user think time. The workload is generated using an analytic approach to capture properties observed in real web workloads concerning file sizes (what is stored in the file system), request sizes

(what is transferred over the network from the server), file popularity, and temporal locality. Distributional models for the file size, the active OFF times, and the embedded references, were developed using a client trace data set. The most important advantage of SURGE compared to the other benchmarks is that SURGE allows to generate much more realistic traffic. Using this new model the authors of [4] analyzed the performance of Web servers (Apache and IIS) and studied the impact that server hardware (CPU, memory, etc.) has on the performance. They also compared performance of HTTP1.0 and HTTP1.1 and concluded the advantage of HTTP1.1.

In our work we propose a new model of Web traffic and use this model in the performance evaluation of Web servers. We consider typical behavior of a user's hypertext navigation within a Web server. We propose a traffic model at the session level, formulated as a stochastic marked point process, which describes when users arrive and how they browse the server. We provide results of statistical analyses and goodness-of-fit of various simple parametric distributions and of their mixtures. It turns out that mixture of simple parametric distributions significantly improve the goodness-of-fit metrics.

We have developed a benchmark tool of Web servers: WAGON (Web trAffic GeneratOr and beNchmark). It comprises a generator of Web traffic, a robot which sends and analyzes requests and a monitoring tool. It can generate different types of traffic requests, which, in turn, can be sent out to the server from different machines with different (simulated) delays and bandwidths. Using this tool we validated the traffic model by comparing various characteristics of the synthetic traffic generated by WAGON against measurements.

The basic advantage of our model compared to WebSTONE-like benchmark is its closeness to the real traffic. Compared to SURGE, one of the advantages of this new traffic model is its simplicity and its ease of parameterization. Indeed, it is at a slightly higher level (session level) and has a direct correspondence with user's hypertext navigation behavior. The random variables used in the model, such as the number of clicks in a session and interclick idle times (or user think time), are simple to interpret and are relatively easy to measure and their analyses have independent interest. The statistical analyses of the traffic parameters within this framework also reveal some interesting properties of the Web traffic. It turns out that the session arrivals are well described by a renewal process, and in many cases, they are simply a Poisson process. This nice property allows for a simple way to model arrivals and to parameterize the traffic intensity. This is in contrast with the HTTP request arrivals which usually form a long-range dependent process, and is thus more difficult to analyze. Another advantage is that this new traffic model allows the benchmarks to factorize users' navigation behavior from the implementation issues of the servers, browsers and protocols. This is in contrast with other models at the HTTP request level which are dependent on both the server (Apache, IIS, etc.) and the navigator (Netscape, IE, etc.) and the ways they implement the protocol (HTTP1.0 or 1.1, parallel connections, persistent connections, pipeline, etc.). Last, our model intrinsically describes the nature of dynamic changing user population of Web server.

With this traffic model and our benchmark WAGON, we carried out a series of experiments, most of which are difficult to accomplish with other existing benchmarks. We analyze the impact of the traffic parameters on the (Apache) server performance and the user perceived quality of service. We show that the aggregate traffic is self-similar in most cases, and that, more importantly, the Hurst parameter is increasing in the traffic intensity. We further provide performance comparison results between HTTP1.0 and HTTP1.1 and show that HTTP1.1 could be much worse for users as well as for servers if some control parameters of the server and of browsers are incorrectly set. We also investigate the impact of user network conditions on the server performance. Indeed, when the server load is relatively high, the page response time under HTTP1.1 increases exponentially fast with the number of parallel persistent connections and with the timeout value used in Apache for persistent connections.

These results indicate that it is harmful to establish multiple parallel persistent connections from one browser window to the same Web server. Such a practice, which seems to be the case in the current implementation of the two most popular browsers Netscape and IE, yields little improvement (only for users with poor network condition such as Modem) in the user perceived quality of service, and could be disastrous to Web servers, and in turn, to Web clients.

Based on this traffic model, we also propose a queueing model to analyze the workload of persistent connections on the Apache server. We mathematically establish optimal solution of the timeout parameter for the workload minimization. As a by-product of this analysis, we show that the Early Close policy proposed in [4] is near optimal in this regard.

This theoretical result as well as the previously mentioned experimental results imply that, in terms of the management of persistent connections, it is useful for browsers to implement Early Close policy which combines the advantages of both HTTP1.0 and HTTP1.1. On the server side, these results imply that servers should set small timeout value for persistent connections if fixed timeout control mechanism is used (as in Apache). If dynamic timeout control mechanism is used, however, then small timeout value should be used once the measured workload is high.

The paper is organized as follows. In section 2 we describe the traffic model and we report the statistical analysis results. In section 3 we present the benchmark tool WAGON and the setup of the experimentation environment. We also present the model validation results. In section 4 we present results pertaining to the impact of the distributions of the random variables of the traffic model. We discuss how the performance characteristics depend on these statistical laws. In section 5 we analyze the performance of the Apache server from various angles: the effect of transport protocols (HTTP1.0 vs. HTTP1.1), the impact of network conditions of the users, and the effect of the browsers' and server's control parameters of persistent connections. Finally in section 6 we conclude the presentations and provide some practical guidelines with regard to the implementation of HTTP1.1 in Web browsers and Web server softwares.

2 Traffic Model and Statistical Analysis

In this section we present our work on the HTTP traffic. We first propose a stochastic model of HTTP traffic at the session level. We then present statistical analysis results of some Web servers for such a model. We shall also provide validation results of the traffic model by comparing performance characteristics observed using synthetic traffic against those of real traffic.

2.1 Traffic Model

We propose a stochastic model of HTTP traffic at session level. We confine ourselves to HTTP requests to the same Web server. We are interested in the typical behavior of users traversing the hypertext of a Web server, namely, when they arrive and how requests are generated. This can actually be well described by the notion of session: a sequence of clicks of a user on the hyperlinks of the same server.

The traffic model we propose can be described by a stochastic marked point process. The arrival times correspond to the beginning times of sessions. Each arrival is associated with the following random variables (the marks):

- number of clicks during the session;
- interclick idle times (or user think times), i.e. time durations elapsed between the completion of the transfer of the previous requested page and the beginning of the transfer of the current page;
- transfer and CPU costs of the successive clicks.

Figure 1 illustrates a user session, where τ_i denotes the time taken to retrieve Web pages (including embedded documents), and γ_j represents the interarrival times of clicks.

Note that the random variables representing transfer and CPU costs of the successive clicks can take different forms depending on the abstraction level of the model and on the performance evaluation technique. When it is used by a simulation tool or an analytical approach, they can be the total sizes of the pages (including embedded objects), or, in a more detailed level, vectors of sizes of objects of the pages. When it is used in a benchmark, they can be addresses of the successively required pages, i.e., the addresses of the first page and the successively clicked hyperlinks. In such a case, embedded objects are not specified and are requested automatically by browsers. Note also that CPU cost of a static page is usually negligible. It could however be dominant for dynamic pages.

It turns out (see below) that the session arrivals are well described by a renewal process, and in many cases, they are simply a Poisson process. This nice property allows for a simple way to model arrivals and to parameterize the traffic intensity. The distribution function of the number of clicks of a session seems to depend heavily on the traces. It can be short-tailed (such as geometric distribution) or heavy-tailed (such as Pareto,

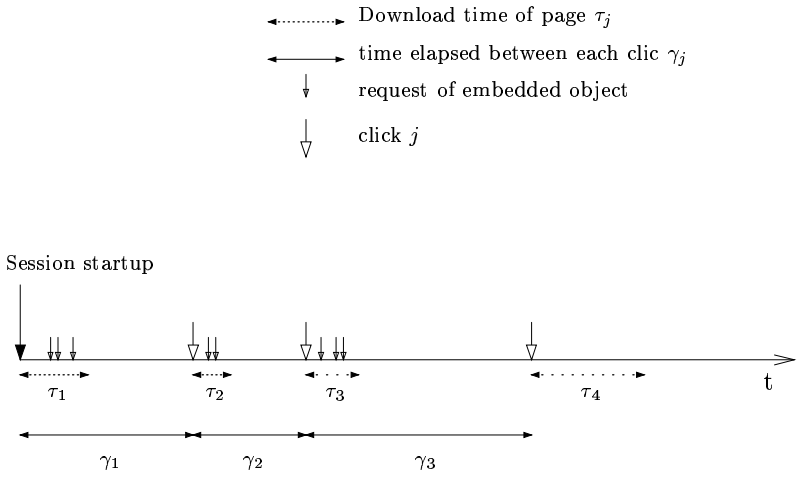

Figure 1: User Session

Lognormal and Inverse Gaussian). The marginal distribution of the interclick idle times most often belong to the class sub-exponential distribution functions including Pareto, Lognormal, Inverse Gaussian and Weibull distributions.

2.2 Statistical Analysis

We report now some of the statistical analyses we performed on traces (log files) of Web servers. They are concerned with the Web servers at W3C (http://www.w3.org/), INRIA (http://www.inria.fr/) and www.clark.net (http://ita.ee.lbl.gov/html/contrib/ClarkNet-HTTP.html). Some characteristics of these server traces are presented in Table 1.

	www.w3.org	www.inria.fr	www.clark.net	INRIA proxy
Time period:	Feb 97	Oct 96	Sep 95	Nov 99
Duration:	24h	24h	18h	24h
Total number of requests:	275000	50000	150000	125000
Total number of pages:	4726	7773	9294	9712
Average page size:	23kB	15kB	13kB	9kB

Table 1: Characteristics of the servers www.w3.org, www.inria.fr, www.clark.net, and the INRIA proxy cache

The goal of this part of the work is to identify the statistical laws of the above described random variables of the traffic model. For ease of use of the traffic model in performance evaluation, we consider only the class of parametric distributions. Moreover, we are interested in distribution functions with a small number of parameters.

Thus, we fix a small set of most commonly used exponential type and sub-exponential distribution functions including Exponential, Geometric, Weibull, Normal, Lognormal, Pareto and Inverse Gaussian distributions. These are the hypothetical distributions that samples are to be tested on. For each sample, we calculate the maximum likelihood estimators (except for some scale parameters) according to each of the hypothetical distributions, and compare the goodness-of-fit of these distributions.

In order to develop a tool that recognize the best fitting distribution function, we used, in conjunction with statistical test, a combination of three metrics: λ^2 ([23, 3]), Cramer-von Mises and Anderson-Darling [8]. The first metrics is based on splitting the values into bins (as in the χ^2 test) and count the number of samples in each bins, and compare it with the theoretical distribution to match (see [22]):

$$\lambda^2 = \frac{\sum_{i=1}^{N} \frac{(Y_i - np_i)^2}{(np_i)} - K - df}{n-1},$$

Where N is the number of bins, $p_1, p_2, \ldots, p_N$ is the probability that the matching distribution fits in the i-th bin. Y_i is the number of samples in bin i ($\Sigma_{i=1}^{N} Y_i = n$), $K = \sum_{i=1}^{N} \frac{Y_i - np_i}{np_i}$ and $df = N - 1 - Est$ is the "degrees of freedom" (Est is the number of estimated parameters). The other two metrics are based on empirical cumulative distribution functions,

$$Q = \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \psi(x) dF(x)).$$

If $\psi(x) = 1$, it is called *Cramer-von Mises* statistics. If $\psi(x) = [F(x)(1 - F(x))]^{-1}$, it is the *Anderson-Darling* statistics.

As we mentioned previously, sample data of each of the random variables are extracted from the log files. Due to the lack of information available in *Common Log Files*, we used heuristics to identify sessions. In particular, we use two thresholds $T_{\min}$ and $T_{\max}$ to identify clicks with requests: the interclick times should be larger than $T_{\min}$ (e.g. 2s) and smaller than $T_{\max}$ (e.g. 10mn). Such kind of heuristics is also used in [16] for identifying the number of files in a page and other data from traces. More theoretical investigations were reported in [24] where different path reconstruction methods were described which are based on the available information in log files (cookies, referrer). The authors propose a Markovian model for routing probability.

For the session arrival process, according to this combined metric, the Poisson process has the best fit in these servers, with rate $\lambda = 0.39$ for www.w3.org, $\lambda = 0.037$ for www.inria.fr and $\lambda = 0.21$ for www.clark.net. Previous results in the literature concerning the analysis of HTTP and IP traffic have rejected the Poisson hypothesis, see for example [22, 7]. However, if one considers the arrival process of sessions instead of the arrival process of requests or packets, the Poisson assumption turns out to be valid in these cases. It is worthwhile noticing that Poisson assumption does not hold for all session arrival processes. However, in all the statistical analyses we performed, the session arrivals do form a renewal process.

The numbers of clicks of sessions are independent. Their distributions are heavy-tailed for the above mentioned three traces: inverse Gaussian ($\mu = 5.96; \sigma = 3$ for www.inria.fr) Pareto ($a = 0.748; \beta = 0.807$ for server www.w3.org, $a = 0.94; \beta = 1.16$ for www.clark.net).

2.3 Refined Analyses with Mixture of Distributions

It turns out that in many cases statistical tests fail, even for the distributions with the best fit. One reason is the well-known problem of large sample, see discussions in [23]. Another reason is that the above mentioned simple parametric distributions are not rich enough to fit the real data of Web traffic. We thus propose to use (finite) mixture of probability densities to circumvent this last problem.

Let $f_1(x), \cdots, f_k(x)$ be k parametric probability density functions with parameters (possibly vectors) $\theta_1, \cdots, \theta_k$. We consider the mixture

$$p(x) = \pi_1 f_1(x) + \cdots + \pi_k f_k(x), \tag{1}$$

where the coefficients (or mixing weights) π_i are positive (in the sense of nonnegative) and sum to 1: $\pi_1 + \cdots + \pi_k = 1$.

Finite mixture densities can be interpreted as densities associated with a statistical population composed of k component populations with associated densities $f_1(x), \cdots, f_k(x)$ and mixing proportions $\pi_1, \cdots, \pi_k$. The reader is referred to [30] for discussions about such finite mixture of parametric (as well as non-parametric) distributions.

EM algorithm. We use the well-known EM (Expectation-Maximization) algorithm to compute the maximum-likelihood estimates of the parameters in mixture densities. The definite reference of this method is the paper by Dempster, Laird and Rubin [9]. A comprehensive treatment can be found in [25].

Let $x = (x_1, \cdots, x_n)$ be a sample of $p(y)$, and $\psi = (\pi, \theta)$ be the vector of parameters to be estimated, where $\pi = (\pi_1, \ldots, \pi_k)$ and $\theta = (\theta_1, \ldots, \theta_k)$. The EM algorithm generates, from some initial approximation $\psi^{(0)}$, a sequence of estimates $\{\psi^{(m)}\}_m$. Each iteration consists of the following double step:

E step : Compute $E[\log p(y|\psi)|x, \psi^{(m)}] =: Q(\psi, \psi^{(m)})$

M step : Find $\psi = \psi^{(m+1)}$ to maximize $Q(\psi, \psi^{(m)})$.

Let

$$\begin{aligned} w_{ij}^{(m)} &= \frac{\pi_j^{(m)} f_j(x_i|\theta_j^{(m)})}{p(x_i|\psi^{(m)})} \\ w_i^{(m)} &= \left(w_{i1}^{(m)}, \cdots, w_{ik}^{(m)}\right) \\ V(\pi) &= (\log \pi_1, \cdots, \log \pi_k) \\ U_i(\theta) &= (\log f_1(x_i|\theta_1), \cdots, \log f_k(x_i|\theta_k)) \end{aligned}$$

Then,

$$Q(\psi, \psi^{(m)}) = \sum_{i=1}^{n} \left(w_i^{(m)}\right)^T V(\pi) + \sum_{i=1}^{n} \left(w_i^{(m)}\right)^T U_i(\theta).$$

By differentiate $Q(\psi, \psi^{(m)})$ with respect to π_j we obtain the M step for π:

$$\pi_j^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} w_{ij}^{(m)} = \frac{1}{n} \sum_{i=1}^{n} \pi_j^{(m)} \frac{f_j(x_i|\theta_j^{(m)})}{p(x_i|\psi^{(m)})} \tag{2}$$

The M step for θ is problem specific. We derive below the solution for the density functions we are more interested of. Note that some of these results are available in the literature.

- Normal distribution: $f_j(x) = \frac{1}{\sigma_j \sqrt{2\pi}} e^{-\frac{(x-m_j)^2}{2\sigma_j^2}}$

$$\frac{\partial Q}{\partial m_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \pi_j^{(m)} \frac{f_j(x_i|\theta_j^{(m)})}{p(x_i|\psi^{(m)})} \frac{(x_i - m_j)}{\sigma_j^2} = 0 \Leftrightarrow m_j^{(m+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(m)} x_i}{\sum_{i=1}^{n} w_{ij}^{(m)}}.$$

$$\frac{\partial Q}{\partial \sigma_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \pi_j^{(m)} \frac{f_j(x_i|\theta_j^{(m)})}{p(x_i|\psi^{(m)})} \frac{1}{\sigma_j}(-1 + \frac{(x_i - m)^2}{\sigma_j^2}) = 0 \Leftrightarrow \sigma_j^{2(m+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(m)} (x_i - m^{(m+1)})^2}{\sum_{i=1}^{n} w_{ij}^{(m)}}.$$

- Exponential distribution: $f_j(x) = \lambda_j e^{-\lambda_j x}$

$$\frac{\partial Q}{\partial \lambda_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \pi_j^{(m)} \frac{f_j(x_i|\theta_j^{(m)})}{p(x_i|\psi^{(m)})} (\log \lambda_j - \lambda_j x_i) = 0 \Leftrightarrow \lambda_j^{(m+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(m)}}{\sum_{i=1}^{n} w_{ij}^{(m)} x_i}.$$

- LogNormal distribution: $f_j(x) = \frac{1}{x\sigma_j \sqrt{2\pi}} \exp\left(-\frac{1}{2}(\frac{\ln x - m_j}{\sigma_j})^2\right)$

$$\frac{\partial Q}{\partial m_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \pi_j^{(m)} \frac{f_j(x_i|\theta_j^{(m)})}{p(x_i|\psi^{(m)})} \frac{(\log x_i - m_j)}{\sigma_j^2} = 0 \Leftrightarrow m_j^{(m+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(m)} \log x_i}{\sum_{i=1}^{n} w_{ij}^{(m)}}.$$

$$\frac{\partial Q}{\partial \sigma_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \pi_j^{(m)} \frac{f_j(x_i|\theta_j^{(m)})}{p(x_i|\psi^{(m)})} \frac{1}{\sigma_j}(-1 + \frac{(\log x_i - m)^2}{\sigma_j^2}) = 0 \Leftrightarrow \sigma_j^{2(m+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(m)} (\log x_i - m^{(m+1)})^2}{\sum_{i=1}^{n} w_{ij}^{(m)}}.$$

- Inverse Gaussian distribution: $f_j(x) = \sqrt{\frac{\lambda_j}{2\pi x^3}} \exp\{-(\lambda(x-\mu_j)^2/2\mu_j^2 x)\}$

$$\frac{\partial Q}{\partial \mu_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \pi_j^{(m)} \frac{f_j(x_i|\theta_j^{(m)})}{p(x_i|\psi^{(m)})} \frac{\lambda_j}{\mu_j^2}\left(\frac{x_i-\mu_j}{x_i} + \frac{(x_i-\mu_j)^2}{\mu_j x_i}\right) = 0 \Rightarrow \mu_j^{(m+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(m)} x_i}{\sum_{i=1}^{n} w_{ij}^{(m)}}.$$

$$\frac{\partial Q}{\partial \lambda_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \pi_j^{(m)} \frac{f_j(x_i|\theta_j^{(m)})}{p(x_i|\psi^{(m)})} \frac{1}{2}\left(\frac{1}{\lambda_j} - \frac{(x_i-\mu_j)^2}{\mu_j^2 x_i}\right) = 0 \Leftrightarrow {\lambda_j}^{(m+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(m)}}{\sum_{i=1}^{n} w_{ij}^{(m)} \frac{(x_i-\mu_j)^2}{\mu_j^2 x_i}}.$$

- Pareto distribution: $f_j(x) = \beta_j a_j^{\beta_j} x^{-(\beta_j+1)}$.

$$\frac{\partial Q}{\partial \beta_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \pi_j^{(m)} \frac{f_j(x_i|\theta_j^{(m)})}{p(x_i|\psi^{(m)})}\left(\frac{1}{\beta_j} - \log x_i\right) = 0 \Leftrightarrow \beta_j^{(m+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(m)}}{\sum_{i=1}^{n} w_{ij}^{(m)} \log x_i}.$$

- Weibull distribution: $f_j(x) = (b_j x^{b_j-1}/a^{b_j}) e^{-(x/a)^{b_j}}$

$$\frac{\partial Q}{\partial b_j} = 0 \Leftrightarrow h(b_j, x) := \sum_{i=1}^{n} \pi_j^{(m)} \frac{f_j(x_i|\theta_j^{(m)})}{p(x_i|\psi^{(m)})}\left(\frac{1}{b_j} + \log x_i - b_j x^{b_j-1}\right) = 0.$$

The solution can be computed numerically using e.g. binary search. It is interesting to notice that the function $h(b_j, x)$ is monotonically decreasing if all elements of x are greater than or equal to 1. Indeed,

$$\frac{\partial h}{\partial b_j} = \sum_{i=1}^{n} w_{ij}^{(m)}\left(-\frac{1}{-b_j^2} - x_i^{b_j-1} - b_j x_i^{b_j-1} \log x_i\right).$$

- Geometric distribution: $P(X = n) = p_j(1-p_j)^n$.

$$\frac{\partial Q}{\partial p_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \pi_j^{(m)} \frac{f_j(x_i|\theta_j^{(m)})}{p(x_i|\psi^{(m)})}\left(\frac{1}{p_j} - \frac{x_i}{1-p_j}\right) = 0 \Leftrightarrow p_j = \frac{\sum_{i=1}^{n} w_{ij}}{\sum_{i=1}^{n} w_{ij}(x_i+1)}.$$

Improved statistical analysis results. In table 2 we summarize statistical analyses on single and mixture of distributions of INRIA proxy cache logs files. As we can see, the metrics for mixtures of distributions are much smaller than with single distributions. When we use the Anderson-Darling test for subsamples with mixture of distributions, we obtain positive results in most cases. Comparison of metrics between single and mixture of distributions for other log files is similar.

	mean	**Distribution**	**Parameters**	λ^2	AD	CVM
Click number	4.6	Inverse Gaussian	$\mu = 4.61$ $\lambda = 4.18$	0.106	90.36	13.17
		Exp-Normal	$\pi_1 = 0.571$	0.091	82.5	11.6
			$\lambda_1 = 0.154$ $\mu = 2.12$ $\sigma = 0.91$			
Idle Time	33.4s	Weibull	$a = 11.34$ $\beta = 0.33$	0.006	484.2	65.0
		Weibull-InvG	$\pi_1 = 0.442$	0.005	351.1	39.85
			$\beta = 0.15$; $\mu = 39.5$ $\lambda = 10.54$			
Transfer time	7.03s	LogNormal	$\mu = 0.77$ $\sigma = 1.53$	0.006	27.9	4.3
		InvG-InvG	$\pi_1 = 0.856$	0.016	7.79	1.28
			$\mu_2 = 12.47$ $\lambda_2 = 0.098$			
			$\mu_1 = 6.12$ $\lambda_1 = 2.12$			

Table 2: Statistical analyses on single and mixture of distributions of INRIA proxy cache logs files

3 Synthetic Traffic Generation and Traffic Model Validation

3.1 Synthetic Traffic Generation using WAGON

We have developed a benchmark tool of Web servers: WAGON (Web trAffic GeneratOr and beNchmark). It is composed of a generator of Web traffic, a robot which sends and analyzes requests and a monitoring tool. It can generate different types of traffic requests, which, in turn, can be sent out to the server from different machines with different (simulated) network delays and bandwidth constraints. It is written in Java for portability.

For this, user can use the graphic interface of WAGON and can specify

- server features: server name or IP address, home page address of the server, server log files, etc.
- client features: laws and parameters of the random variables in the traffic model; transport protocols and parameters used.
- experiment configuration: local client caching, sampling rate, bandwidth constraint and network delay of different types of clients, etc.

Experiments using WAGON are carried out on a set of machines connected through a LAN. One machine is used as the server machine with the Web server software (such as Apache, IIS and Jigsaw) and the Web contents (i.e. the Web hypertext) installed on it. Other machines are considered as client machines in charge of sending HTTP requests to the server machine. Each of these machines is assigned one or several types of clients.

URL addresses are generated by WAGON according to the popularity of the pages of the Web hypertext. More precisely, for each page we compute the probability of getting a session started from it. We also compute the routing probabilities for each page: given the page is visited, what is the probability that a page it refers to will be visited next.

During a benchmark, various performance measures can be obtained and observed on-line through the monitoring tool of WAGON. Of particular interest are request latency and user perceived throughput.

3.2 Experimentation Setup

In the sequel we shall present various experimentation results obtained using the traffic model and WAGON. Unless otherwise stated, the experimentation setup is the following.

For the experiments presented in the paper, we have chosen a subset of the Web server hypertext INRIA[1] as the Web server contents. It consists of the Web pages at the highest level of the INRIA server. The majority of the files are HTML files with in average five embedded images. There are 10000 documents in total, with the average size 7.5kB. Table 3 summarizes the first-order characteristics of the hypertext. Figure 2 illustrates the empirical distribution of page sizes.

	INRIA
Number of files	10627
Total size	255Mo
Number of Web pages (with popularity > 0)	1828
Number of HTML pages (id)	1024
Average document size	7.5Ko
Average Page size (except icons)	38.5Ko
Average HTML size	4.0Ko
Mean number of embedded documents per HTML page	5.0
Mean number of links per HTML page	8.9

Table 3: Server content

[1] http://www.inria.fr

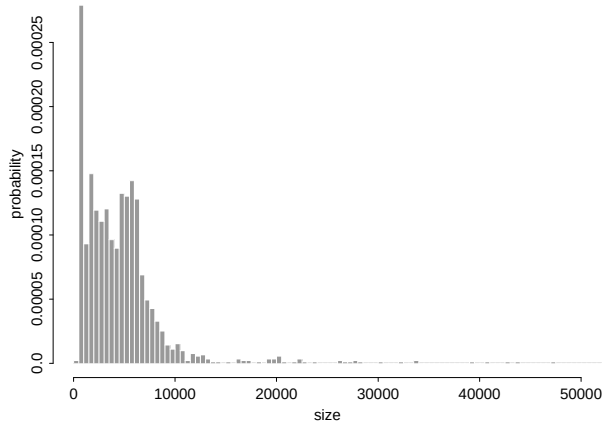

Figure 2: Page size empirical distribution of INRIA server

In all the experiments we shall report below, the Web traffic is generated using the above described traffic model under the additional assumption that the sequences of the inter-arrival times, of the number of clicks and of the interclick idle times are all renewal processes which are mutually independent.

Although the assumption of mutual independence among the random variables is not justified, it turns out that the resulting synthetic traffic has characteristics close to the real traffic, see discussions in section 3.3.

The parameters used for generating Web traffic are summarized in table 4.

Variable	**Law**	**Mean**	**Std deviation**
Session arrival	Poisson Process ($0.0005 \leq \lambda \leq 0.01$)		
Number of clicks	Inverse Gaussian ($\mu = 5; \lambda = 3$)	5	1.28
interclick idle time	LogNormal ($m = 3; \sigma = 1.1$)	36.8s	56.4s

Table 4: Client parameters used in the experimentation

This workload corresponds to approximately 10 requests per second for the case of the smallest load ($\lambda = 0.0005$), and to 180 requests per second for the case of the highest load ($\lambda = 0.01$) we use is these experiments.

The experiments have been carried out on seven client machines and one server machine connected through a LAN of 100Mb/s. The clients machines are PC Pentium II 450Mhz with 128MB main memory, under FreeBSD 3.2. The server machine is a PC Pentium Pro with 64MB main memory, under FreeBSD 3.2. For WAGON, we use the JDK version 1.1.8 with JIT (tya) enabled.

In all our experiments, we use only static documents. Therefore, all our conclusions are valid only for static page, and not for dynamic one. In the dynamic case indeed, the server is far more CPU intensively loaded, and the underlying dynamic processes (Perl CGI's, Database requests, Java Servlet, etc.) can play a major role on the performance of servers.

The Web server software we used in the experiments is Apache[2], the currently mostly used Web server software according to Netcraft's survey[3]: about 5 million sites and 55% market share in November 1999. There are two key parameters used in the Apache server that control the number of connections of the server: the maximum number of connections (MaxClients, default value 150) and the timeout beyond which an idle connection is cut off (KeepAliveTimeout, default value 15s). The maximum number of connections depends largely on the memory and CPU capacity of the Web server. The KeepAliveTimeout parameter, however, is to be set in accordance with the arrival traffic pattern. Unless otherwise specified, these default values are used in most experiments.

[2]http://www.apache.org
[3]http://www.netcraft.co.uk/survey/

For network constraints, bandwidth are limited by WAGON for each client, whereas delays are simulated at the OS level, using the Dummynet [26] module available in the FreeBSD kernel. Delays are therefore simulated at the packet level.

We investigate the following different types of connections such as Modem, T1 link, WAN, satellite connections, etc. In order to simulate such connections, we put the delay and bandwidth constraints as in table 5.

	delay	bandwidth
Modem1	300 ms	56kb/s
Modem2	300 ms	33kb/s
T1, DSL	20ms	1.5Mb/s
WAN	180ms	150kb/s
satellite	500ms	2Mb/s
Ethernet	0.1ms	100Mb/s

Table 5: Delay and bandwidth constraints

3.3 Traffic Model Validation

We now present some results of experimentation which aims at validating the traffic model. We are interested in different characteristics of the resulting synthetic traffic: the self-similarity, request arrival process, document popularity and the stack distance. The results presented in this subsection are obtained using parameters of table 4 with arrival rate $\lambda = 0.008$.

Self-similarity. According the various experiments we carried out, the synthetic traffic is in most cases self-similar (except for the degenerate case with deterministic input, see discussions in section 4.1). In figures 3, 4 and 5, we illustrate the curves of the number of packets observed at different time scales in the experimentation network. Figure 6 shows the estimated Hurst parameter $H = 0.73$.

HTTP request arrival process. We now investigate the the arrival process of HTTP requests. Due to the lack of precision of log files, we shall analyze the sequence of number of HTTP requests per second instead of inter-arrival times.

Figures 7 and 8 illustrate the qqplot (quantile to quantile plot) of both processes (WAGON and real traffic) with different distributions. One can see that in both cases, Weibull distribution fits the data (confirmed by the λ^2 metric).

In terms of the auto-correlation structure of this arrival process, we observe from the auto-correlation curves of figures 9 and 10 that the synthetic traffic and the real traffic are both highly auto-correlated with quite similar auto-correlation structure.

Document popularity. We next examine the page popularity resulting from the synthetic traffic. Figure 11 illustrates the comparisons of real data and WAGON popularity curves (in the same log-log coordinate). The real data comes from the log files of INRIA during September 1998. We can see that the two popularity curves are quite close to each other, except for the lowest ranked documents.

It is worthwhile noticing that in the literature (see e.g. [5]) results were reported indicating that the document popularity can be characterized by Zipf type distribution with a slope close to or larger than 1. It turns out that in our case, for both real data and the WAGON traffic, Zipf type distribution does not seem to be suitable. One reason may be that in our experimentation we have chosen a subset of the INRIA Web server. If we consider the whole Web site of INRIA, Zipf distribution does fits well, see figure 12 where the straight line has slope -1.

Temporal Locality. Another characteristic of interest is the so-called temporal locality of HTTP requests which plays an important role on the efficiency of caching algorithms. A usual way (cf. [1]) to measure temporal locality is to compute the "stack distance": the distances between requests to the same document (by counting

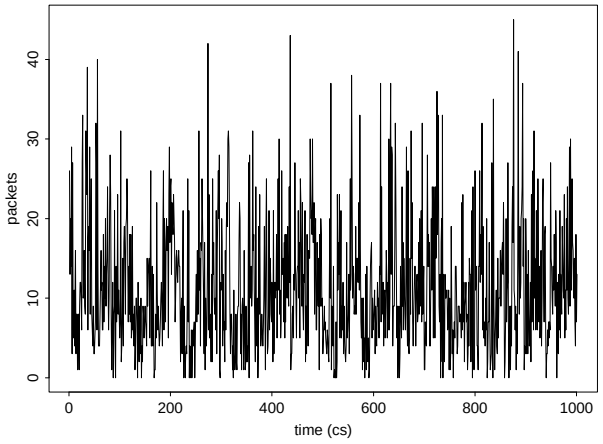

Figure 3: Number of packets, scale=10ms

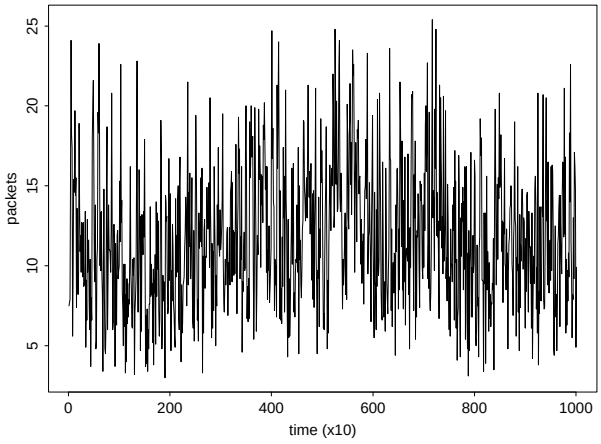

Figure 4: Number of packets, scale=100ms

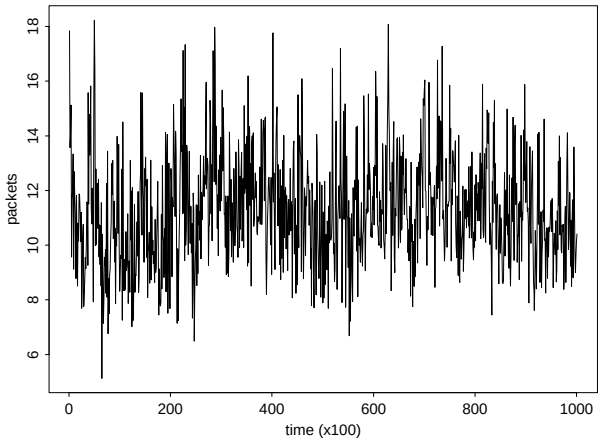

Figure 5: Number of packets, scale=1s

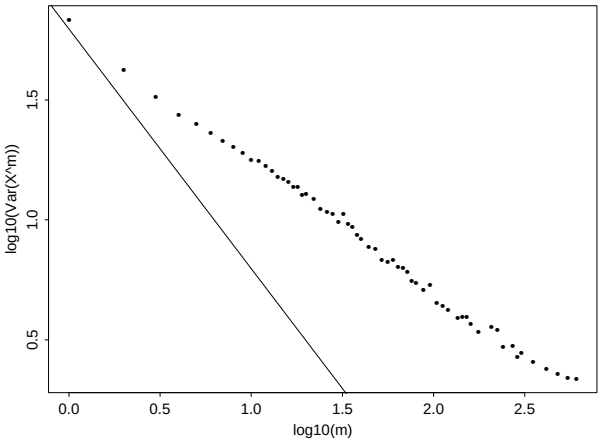

Figure 6: Hurst estimation using variance analysis

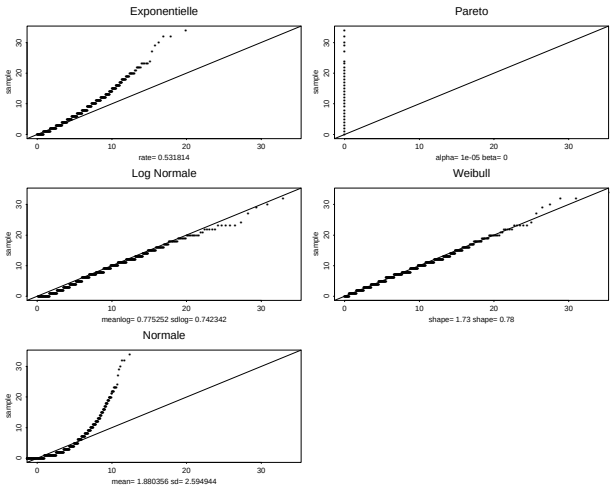

Figure 7: QQplot, HTTP requests/sec, INRIA server

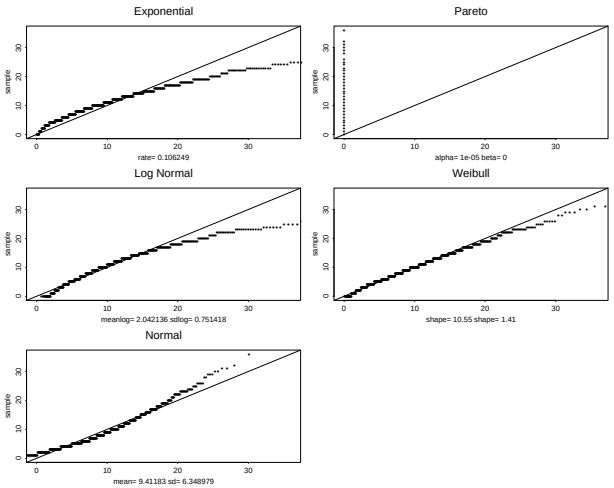

Figure 8: QQplot, HTTP requests/sec with WAGON

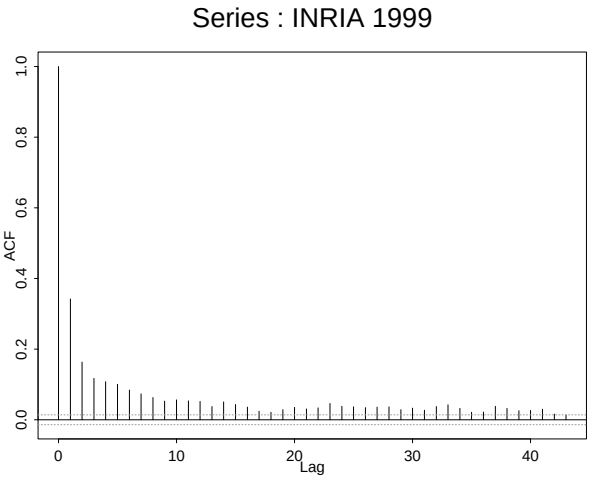

Figure 9: Auto-correlation (number of requests per seconds), INRIA server

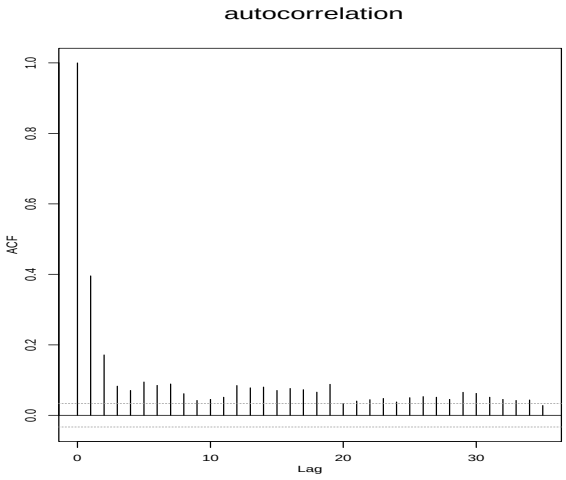

Figure 10: Auto-correlation: number or request per seconds obtained with WAGON

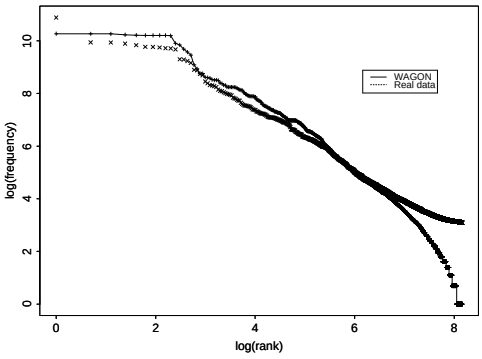

Figure 11: Document popularity in real and WAGON data

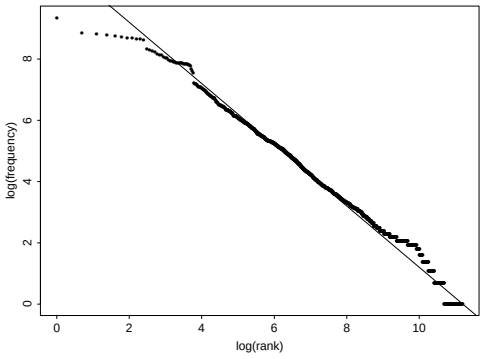

Figure 12: Document popularity for the whole Web site of INRIA

the number of different requests between them). It was observed in [1] that the marginal of the stack distance is Lognormal.

While using this measure on the our real data (computed with logs files) and the synthetic traffic data, we also observed that the Lognormal distribution yields the best fit to the marginal distribution of the stack distance for both cases.

4 Impact of Traffic Parameters

We investigate in this section the impact of the random variables of the traffic model on the performance characteristics.

4.1 Impact of the Randomness

We first consider the effect of the randomness of the key quantities of the traffic model, in particular, the inter-arrival times of the sessions, the number of clicks, the inter-click idle times. We use WAGON to simulate Web traffic in both the case where these quantities are random with the distributions described in table 4 and the case where they are all deterministic with the same averages as in the random case.

Consider the resulting HTTP request arrival processes. As we mentioned previously, the marginal distribution of the inter-arrival times of HTTP requests in real data as well as in the synthetic traffic generated according to table 4 is Weibull. In the case of deterministic input, however, the Lognormal distribution turns out to be the best fit (with positives tests on sub-samples). More importantly, the auto-correlation structure is completely different. As we already saw, in the random case, the auto-correlation and the Hurst parameter are close to those of real traces (figure 10. In the deterministic case, the auto-correlation function is close to zero (see figure 13) and the Hurst parameter is close to 0.5.

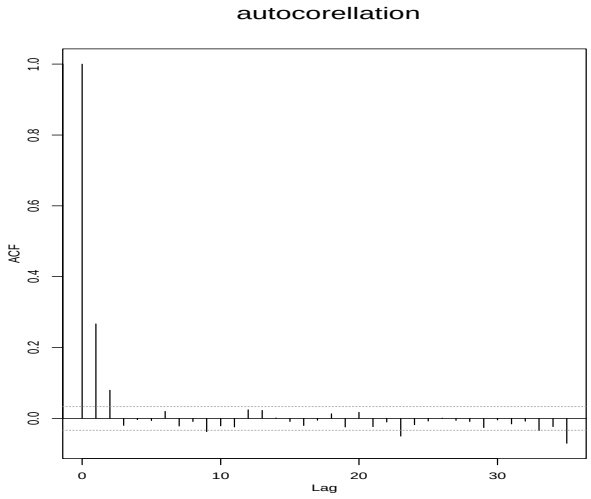

Figure 13: Auto-correlation: number or requests per second obtained with deterministic input.

Recall that in both cases, documents (size, number of embedded documents per HTML page, popularity) and the mean arrival rate are the same. However, the resulting performance characteristics differ significantly. As a consequence, it is important for benchmarks to generate realistic synthetic traffic. Benchmarks such as SPECWeb [27] or WebStone [31], which use deterministic arrival process have a severe drawback in this regard.

4.2 Impact of the Arrival Rate

As is reported often in the literature, Web traffic as well as other Internet traffic has the long-range dependence and the self-similarity. A key parameter measuring these properties is the Hurst parameter which varies inbetween 0.5 and 1. A detailed analysis on the way the self-similarity depends on the TCP is presented in [10].

We investigate here how the Hurst parameter depends on the traffic intensity. For this, we vary the session arrival rate and measure the resulting traffic. We consider the situation when the network inbetween the server and the clients becomes the bottleneck when the arrival rate increases. This situation is configured such that all traffic inbetween server and clients goes through a router of bandwidth 10Mb/s.

In figure 14 we provide the Hurst parameter estimations as a functions of the arrival rate. Each of these estimations is obtained through variance analysis. We observe that the Hurst parameter increases in the traffic intensity.

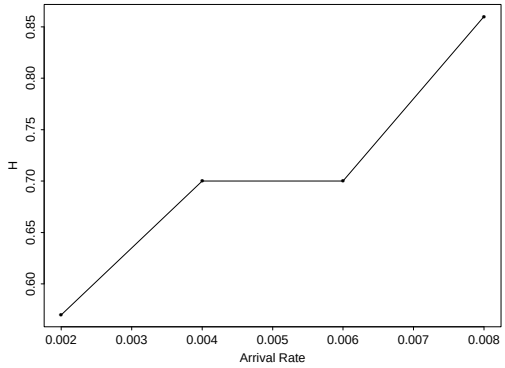

Figure 14: Hurst parameter estimation when arrival rate increases

5 Performance Evaluation of the Apache Server

In this section we analyze more specifically the performance of the Apache server. The previous section deals with the effect of the traffic model on the server performance. Here we are interested in how the server performance and the user perceived quality of service depend on issues related to transfer protocols, user network conditions, browser control parameters and server control parameters.

Results reported in [4] are concerned with the impact of server hardware (CPU, memory, etc.) on the performance of Web servers. We shall be more interested in the impact of network conditions. For this purpose, we configure the experimental network in such a way that all traffic inbetween server and clients goes through a router of bandwidth 10Mb/s. In addition, clients have simulated individual network bandwidth and delay constraints as specified in section 3.2, table 5.

In this section we will analyze two quality of service (QoS) measures:

Response Time of Web pages: the time it takes to retrieve an entire page (HTML plus embedded objects)

HTML latency (User perceived latency). For a user, the quality of service is perceived mainly by the speed of display of the pages. The quicker, the better. To display a page, the browser needs to have the information on the content (HTML) and also on the embedded objects. The first bytes of images can give information to the browser in order to display them (size). Therefore, it is important to get the first byte of each embedded objects as fast as possible. A way to measure this is to compute the time elapsed between the click (beginning of connection to the HTML page) and the time epoch where the first byte of data of the last image is available.

Most existing results of performance evaluation of HTTP1.0 and HTTP1.1 are concerned with the first measure, in addition to the network traffic, see e.g. comparisons between HTTP1.0 and HTTP1.1 [21, 4, 15].

Our benchmark WAGON allows us to measure detailed transaction information on the client side so that we will report comparison results on this user perceived QoS.

Note that in HTTP1.1, one can use the pipeline mechanism to send requests on the same connection without waiting for the completion of the previous transfer. Our extensive experiments show that HTTP1.1 with pipeline is uniformly better than HTTP1.1 without pipeline. Thus, in the following, as far as HTTP1.1 is concerned, we shall only report results of HTTP1.1 with pipeline.

However, it seems (cf. [33]) that request pipelining is not supported by the two most popular Web browsers (Netscape and IE). We think that according to our analyses as well as those in the literature, one of the ways for HTTP1.1 to achieve significantly its performance gains over HTTP1.0 is through request pipelining on a persistent connection. We thus suggest that this request pipelining be implemented in browsers.

5.1 How Persistent Should Persistent Connections Be

We first analyze the effect of the persistent connections of HTTP1.1. As is now well known, most Web documents are small objects. However, in HTTP1.0, every object transfer requires to establish a new TCP session and to go through the slow-start phase. Thus it is interesting to reuse previously established connections to transfer such objects. The persistent connection is proposed in HTTP1.1 so that TCP sessions are kept alive. However, this has a cost. Mogul [18] pointed out that persistent connection has a higher main memory cost.

In Apache, the control parameter of persistent connections is KeepAliveTimeout which controls the duration above which an idle session is disconnected. The default value is 15s. It is reported in [4] that when the disc system is the bottleneck, it is beneficial to keep connections open just for the transfers of a Web page.

In figure 15, we study the response time by varying the timeout parameter of the server. We consider only clients with bandwidth 150KB/s and network delay 80ms. The other configurations have similar behavior.

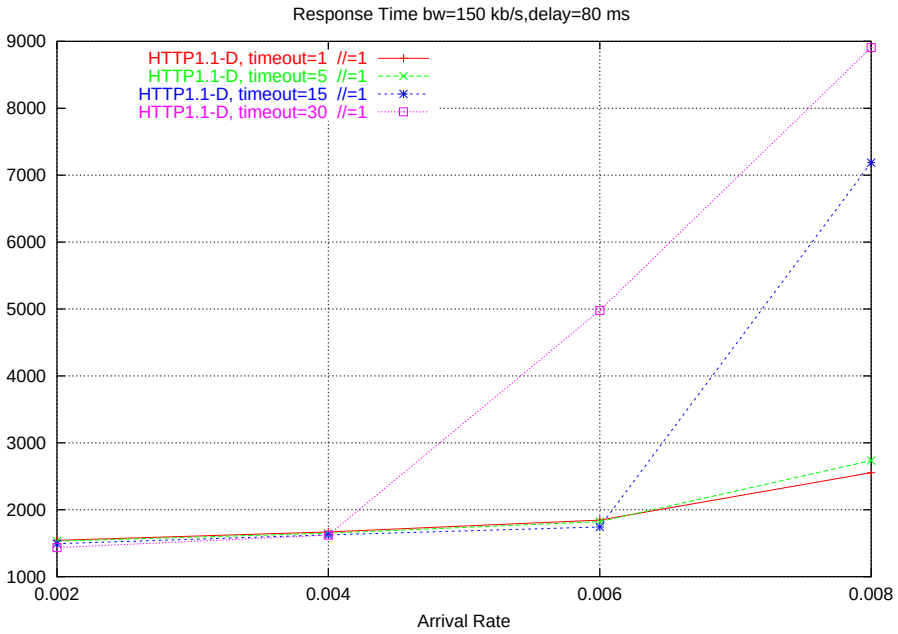

Figure 15: Response Time with Apache when the timeout increase

We observe that when the server load is small, it is beneficial to keep connections alive for long time. However, when the server load increases, this advantage vanishes and becomes a drawback due to its memory cost. The improvement of persistent connections (when there are) with large timeout values are rather small. This gain depends heavily on the RTT (Round Trip Time) between the client and the server. At low traffic intensity , the improvement due to persistent connections is at best around 10%. When the traffic intensity grows, the increase in the response time is exponentially fast in the timeout value. This phenomenon could be explained theoretically below as an issue of workload or stability region.

5.2 Queueing Model for the Apache server

Based on the traffic model presented in section 2, we analyze the workload brought in to the server by a session. We fix the parameter of the maximum number connections (MaxClients) and analyze the impact of the timeout parameter (KeepAliveTimeout). As this parameter has effect only on persistent connections, we confine ourselves to the HTTP1.1 sessions with persistent connections.

We propose to use a multi-server queueing system to model the Apache server. Consider a $G/G/m$ queueing system with repetitive customers, illustrated in figure 16. There are m servers and an infinite-size queue in the system. Customers arrive according to a general, say stationary and ergodic, process and get served according to FCFS (first come first serve). The servers represent the processes serving the persistent connections. Customer arrivals correspond to the session arrivals.

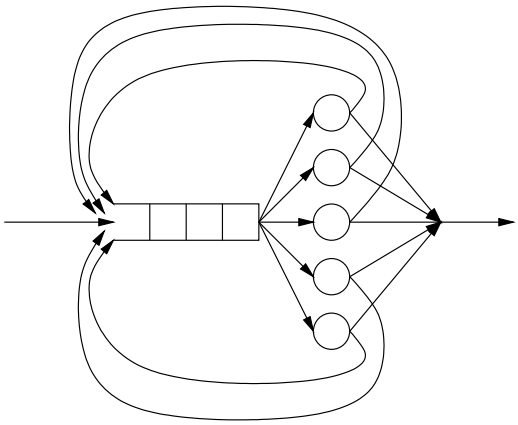

Figure 16: Apache server model

When a customer starts service on one of the servers, it iterates the service and idle phases. The service time corresponds to the transfer time of an entire Web page. The first service time also includes the connection establishing time and the slow-start phase in TCP. The first transfer is thus usually longer. The idle time corresponds to the click idle time. The customer stays on the server until either the session finishes or the connection is timed out while it is in the idle state. If the connection is timed out, the session goes back to the queue when the next click is done.

As with the synthetic traffic generation, we assume that the transfer times of Web pages (resp. interclick idle times, numbers of clicks) are independent and identically distributed random variables. Moreover, these random variables are assumed to be mutually independent.

Let X_1 be the (generic) random variable representing the transfer time of a Web page when using a new connection (it thus includes the connection establishing time and also the slow-start delay). Let X_2 be the (generic) random variable representing the transfer time of a Web page when using an existing connection. Let Y be the (generic) random variable representing the click idle time, and N the random variable representing the number of clicks in a session. Finally, let Δ denote the timeout duration.

Define $p = P(Y \geq \Delta)$ as the probability that the connection be timed out. It can be shown using Wald's identity (see e.g. [13, page 264]) that the expected number of times that a session is timed out is given by $E[Q] = E[N-1]p$. Indeed, if $Y_1, Y_2, \cdots, Y_n, \cdots$ is the sequence of interclick idle times, then, thanks to Wald's identity,

$$E[Q] = E\left[\sum_{n=1}^{N} \mathbf{1}(Y_n \geq \Delta)\right] = E[N]E\left[\mathbf{1}(Y_1 \geq \Delta)\right] = pE[N].$$

Similarly, using again Wald's identity we can show that total expected work brought in by a session is

$$\begin{aligned} E[W] &= E[X_1] + pE[N-1]E[X_1] + (1-p)E[N-1]E[X_2] + (1-p)E[N]E[Y|Y<\Delta] \\ &= E[X_1] + E[N-1]E[X_2] + pE[N-1]\left(E[X_1] - E[X_2]\right) + (1-p)E[N]E[Y|Y<\Delta]. \end{aligned}$$

Denote by $f(x)$ and $F(x)$ the density function and the cumulative distribution function of Y, respectively. Then,

$$\begin{aligned} E[Y|Y<\Delta] &= \int_0^\infty P(Y \geq x|Y<\Delta)dx = \int_0^\infty \frac{P(x \leq Y < \Delta)}{P(Y<\Delta)}dx \\ &= \frac{1}{1-p}\int_0^\Delta P(x \leq Y < \Delta)dx = \frac{1}{1-p}\int_0^\Delta \int_x^\Delta f(y)dydx \\ &= \frac{1}{1-p}\int_0^\Delta dy \int_0^y f(x)dx = \frac{1}{1-p}\int_0^\Delta F(y)dy. \end{aligned}$$

Hence,

$$E[W] = E[X_1] + E[N-1]E[X_2] + (1-F(\Delta))E[N-1]\,(E[X_1]-E[X_2]) + E[N]\int_0^\Delta F(x)dx,$$

so that the derivative of $E[W]$ with respect to Δ is

$$\frac{dE[W]}{d\Delta} = -f(\Delta)E[N-1]\,(E[X_1]-E[X_2]) + F(\Delta)E[N].$$

Thus, we obtain

Theorem 1 *The timeout value Δ that minimizes the workload is given as a solution of*

$$\frac{F(\Delta)}{f(\Delta)} = \frac{E[N-1]}{E[N]}\,(E[X_1]-E[X_2]) =: \delta \tag{3}$$

The difference $(E[X_1]-E[X_2])$ can be considered as the gain of transfer time by using persistent connections. According to our measurements, the average gain is usually upper bounded by twice of the RTT (round-trip time). We can assume without loss of generality that $\delta < 1s$. A more realistic assumption is actually $\delta < 400ms$.

When the interclick idle time has a Weibull distribution, $f(x) = (bx^{b-1}/a^b)e^{-(x/a)^b}$ and $F(x) = 1-e^{-(x/a)^b}$. As $e^{(x/a)^b} \simeq 1+(x/a)^b$ we obtain from (3) that the optimal timeout satisfies $\Delta \simeq \delta b < b$. If the shape parameter $b < 1$, in which case Y has heavy tail distribution, we conclude that Δ should be smaller than $1s$. The same conclusion holds for exponentially distributed interclick idle times.

If, however, the interclick time has a Lognormal distribution, $f(x) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}(\frac{\ln x - m}{\sigma})^2\right)$, then we need to numerically solve the equation (3). For the case where our traffic is generated, $m = 3; \sigma = 1.1$, which implies mean $36.8s$ and standard deviation $56.4s$, the optimal value turns out to bounded by 2s, even for δ close to 1, see figure 17. If we assume that $\delta < 400ms$, we then obtain $\Delta < 1s$.

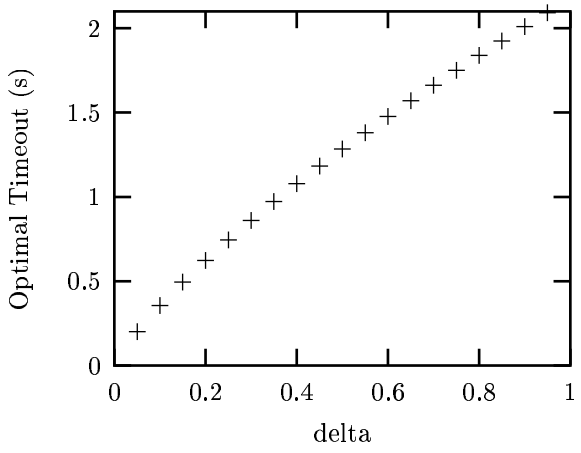

Figure 17: Optimal timeout value for LogNormal($m = 3; \sigma = 1.1$)

We conjecture that the optimal value of the timeout is increasing in δ and is usually small (within seconds). Since the interclick idle times are usually larger than $2s$, we see that Early Close is near optimal for the workload minimization. Under the Early Close scheme (see [4]), persistent connections are closed by clients after each Web page (*i.e.* after retrieving HTML document and all of its embedded objects).

Note that the total arrival traffic intensity is given by $\lambda E[W]$, where λ is the arrival rate of sessions. We conjecture that the usual stability condition $\lambda E[W] < m$ holds for this queueing system with m servers. In this case, the solution of theorem 1 also maximizes the stability region of the Web server

As a conclusion of the results of sections 5.1 and 5.2, we suggest that if HTTP1.1 is used, browsers implement Early Close policy and servers set small timeout values if fixed timeout control mechanism is used (as in Apache). If dynamic timeout control mechanism is used, however, then small timeout value should be used once the measured workload is high.

5.3 Comparison of HTTP/1.0 and HTTP/1.1

Previous comparison results on protocols HTTP/1.0 vs. HTTP/1.1 are mostly concerned with amount of traffic that HTTP/1.1 saves [21] and the effect of server hardware components [4]. We provide here comparison between the protocols HTTP/1.0 and HTTP/1.1 with different network conditions. These comparison results complement those of [21, 4].

For HTTP/1.0, we use 4 parallel connections. For HTTP/1.1, we use a single persistent connection using pipelined requests. The KeepAliveTimeout parameter is the default value (15s). We also look at the Early Close scheme where persistent connections are closed by clients after each Web page. In the sequel, HTTP/1.1 with default KeepAliveTimeout parameter is denoted HTTP/1.1-D, HTTP/1.1 with Early Close is denoted HTTP/1.1-EC.

The experimentation results are summarized in figure 18 for response times and figure 19 for latencies. It turns out that the performance comparison results depend heavily on both the traffic intensity and the type of client network conditions.

When the traffic intensity is small, HTTP/1.1-D provides smaller response time, whatever client network condition is. However, when the traffic grows, HTTP/1.1-D performs very badly. Indeed, the number of simultaneous connections reaches its maximum (MaxClients). Apache server in this case refuses new connections until an old one has finished (this happens when an existing connection times out or when a client closes its connection). However, with Early Close policy, HTTP/1.1 is still better than HTTP/1.0, even when the traffic is very high.

Consider now the user perceived QoS: HTML latency. The results are quite different. HTTP/1.1 no longer outperforms HTTP/1.0 in this regard. Indeed, with a single connection, HTTP/1.1 serializes its requests of all the objects of a page. Thus, multiple parallel connections of HTTP/1.0 can yield a smaller latency, even if all these connections have to go through the connection establishment phase and the TCP slow start phase.

HTTP1.0 uniformly outperforms HTTP1.1-D with respect to the HTML latency. The difference in latency between the two protocols increases as the traffic intensity grows. Compared to HTTP1.1-EC, the advantage of HTTP1.0 is particularly important with small-bandwidth clients (e.g. Modems), in which case, the latency of HTTP/1.0 is better off by about 25%. The difference in latency between HTTP1.0 and HTTP1.1-EC decreases as client bandwidth increases. At high speed (1.5Mbs), the performances are almost equal.

In light of this comparison of latency, one might be interested in implementing multiple parallel persistent connections with HTTP/1.1. This will be the subject of our discussions in the next subsection.

5.4 How Many Parallel Persistent Connections Should There Be

Since parallel HTTP1.0 requests could reduce the latency and HTTP1.1 persistent connections could reduce network traffic as well as page response time, one is tempted to implement parallel persistent connections in order to combine the advantages of both protocols. It seems to be indeed the case in the two most popular browsers [33]: Netscape seems to use 6 parallel persistent connections to the same Web server, and IE seems to use 2.

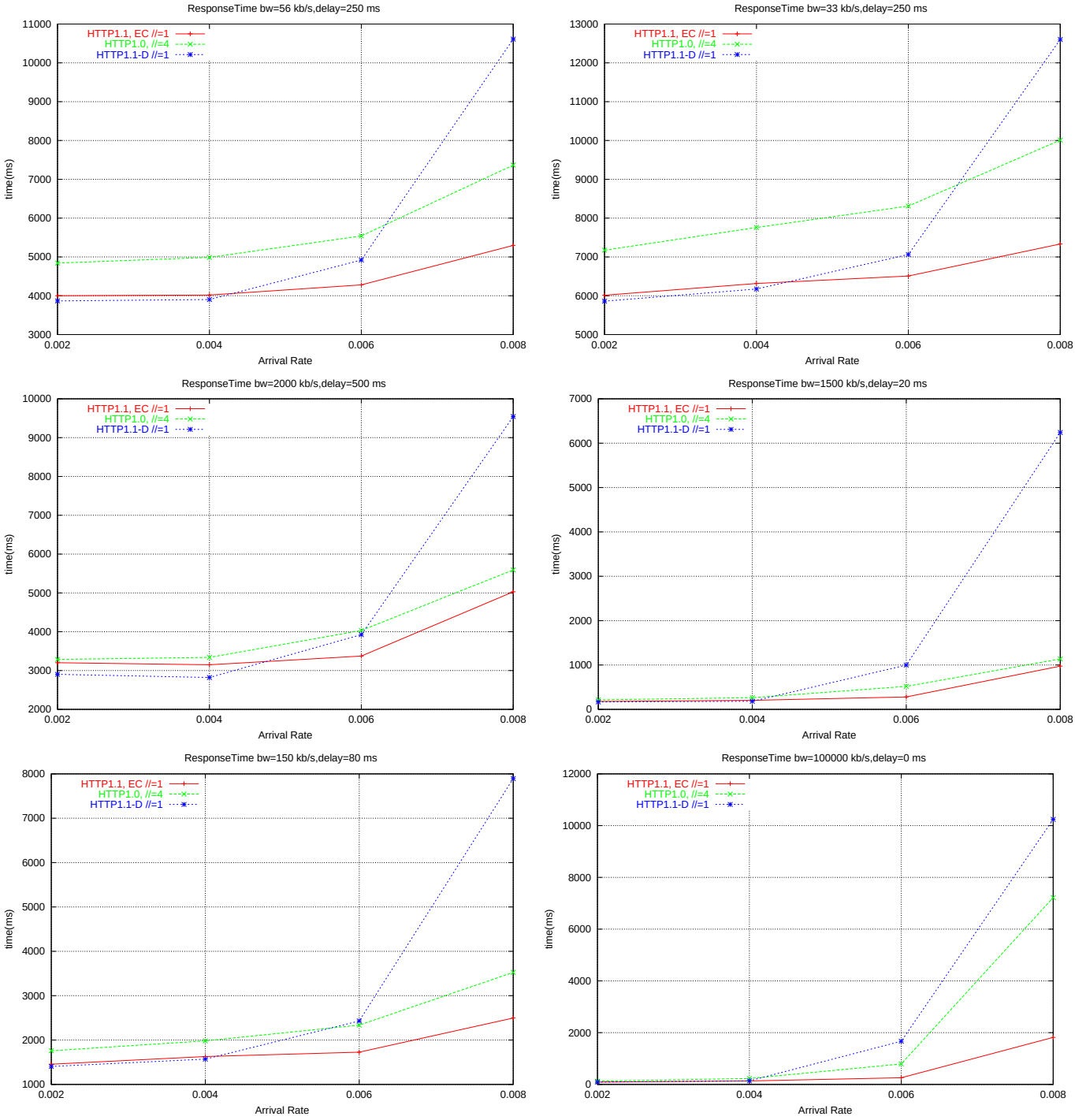

Figure 18: Comparison between HTTP1.0 and HTTP1.1: Mean Response time

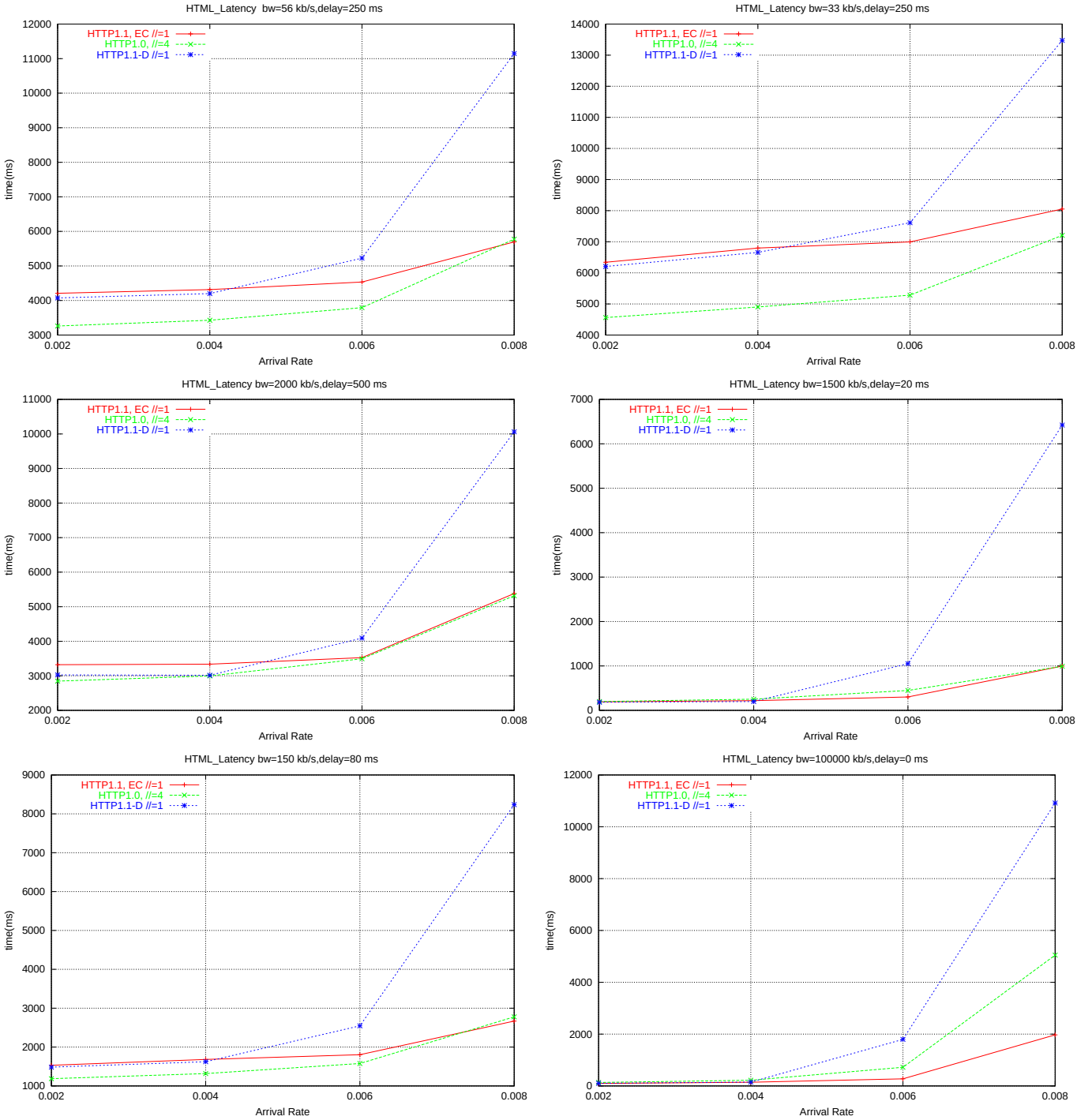

Figure 19: Comparison between HTTP1.0 and HTTP1.1: Mean HTML Latency

We show in figure 20 that under both HTTP1.1-D and HTTP1.1-EC, the response time increases exponentially with the number of parallel connections. With the same number of parallel persistent connections, the performance under HTTP1.1-EC is usually better than that under HTTP1.1-D, except for the cases of light loads. The gain of HTTP1.1-D over HTTP1.1-EC in light load, however, is much smaller (several orders of magnitude difference) than the gain of HTTP1.1-EC over HTTP1.1-D in medium or heavy load.

Concerning the latency, figure 21 shows that under HTTP1.1-D, the latency increases exponentially with the number of parallel connections. The behavior of HTTP1.1-EC, however, is quite similar to that of HTTP1.0. There is gain, especially for low bandwidth connections, in using multiple parallel HTTP1.1-EC persistent connections when the load is light. This gain vanishes when the network bandwidth increases. In the heavy load case, even under HTTP1.1-EC, it is harmful to use multiple parallel persistent connections: The latency increases exponentially with the number of parallel persistent connections.

In view of the results of sections 5.3 and 5.4, we conclude that HTTP1.1 with Early Close combines the advantages of both HTTP1.0 (latency minimization with parallel connections) and HTTP1.1. We think that browsers should, in general, implement HTTP/1.1 with Early Close (and pipelining). We suggest that if multiple parallel persistent HTTP1.1-EC connections are used in browsers, they be used only for low bandwidth network connections. In other words, except for low bandwidth network connections, browsers should open only one persistent connections to the same Web server from one browser window. Multiple parallel persistent connections are usually harmful.

6 Concluding Remarks

We have presented a new model of Web traffic and its applications in the performance evaluation of Web servers. We have proposed a traffic model at the session level, formulated as a stochastic marked point process, which describes when clients arrive and how they browse the server. We have provided results of statistical analyses and goodness-of-fit of various simple parametric distributions and of their mixtures. We have developed a Web server benchmark WAGON, and we have validated the traffic model by comparing various characteristics of the synthetic traffic generated by WAGON against measurements.

Using this benchmark tool we have performed extensive experiments, most of which are difficult to accomplish with other existing benchmarks. We carried out our experiments on the Apache server, the currently most used Web server software. We have analyzed the impact of the traffic parameters on the HTTP request arrival process and the packet arrival process. We have shown that the aggregate traffic is self-similar in most cases, and that, more importantly, the Hurst parameter is increasing in the traffic intensity.

We have provided further performance comparison results between HTTP1.0 and HTTP1.1, with emphasis on the effect of network conditions. We have shown that HTTP1.1 could be much worse for clients as well as for servers if the timeout value for persistent connections used by the server is too large or if browsers use multiple parallel persistent connections to the same server without Early Close.

With regard to HTTP1.1, we have shown that request pipelining is useful, and that response time depends tightly on the timeout value used by server for managing persistent connections. When the workload traffic is light, it is beneficial (with a small gain) to set a large timeout value. When the workload traffic is heavy, however, the response time grows exponentially in the timeout value. We have also proposed a queueing model to analyze the workload of persistent connections on the Apache server, and we have established optimal solution of the timeout parameter that minimizes the workload. HTTP1.1 with Early Close turns out to be near optimal. We have moreover shown that multiple parallel persistent connections are harmful: the response time increases exponentially in the number of parallel connections.

Based on our analyses, we make the following suggestions for the implementation and for the parameterization of Web server softwares and Web browsers. Browsers should in general support HTTP1.1 with request pipelining and implement HTTP/1.1 with Early Close. HTTP1.1 with Early Close combines the advantages of both HTTP1.0 and HTTP1.1. Browsers should in general avoid establishing multiple parallel persistent connections from one browser window to the same Web server. Multiple parallel connections could be beneficial only to clients with low bandwidth connections (such as Modem). On the server side, for the management of persistent connections, servers should set small timeout values if fixed timeout control mechanism is used (as in Apache) or if dynamic timeout control mechanism is used and the measured workload is high.

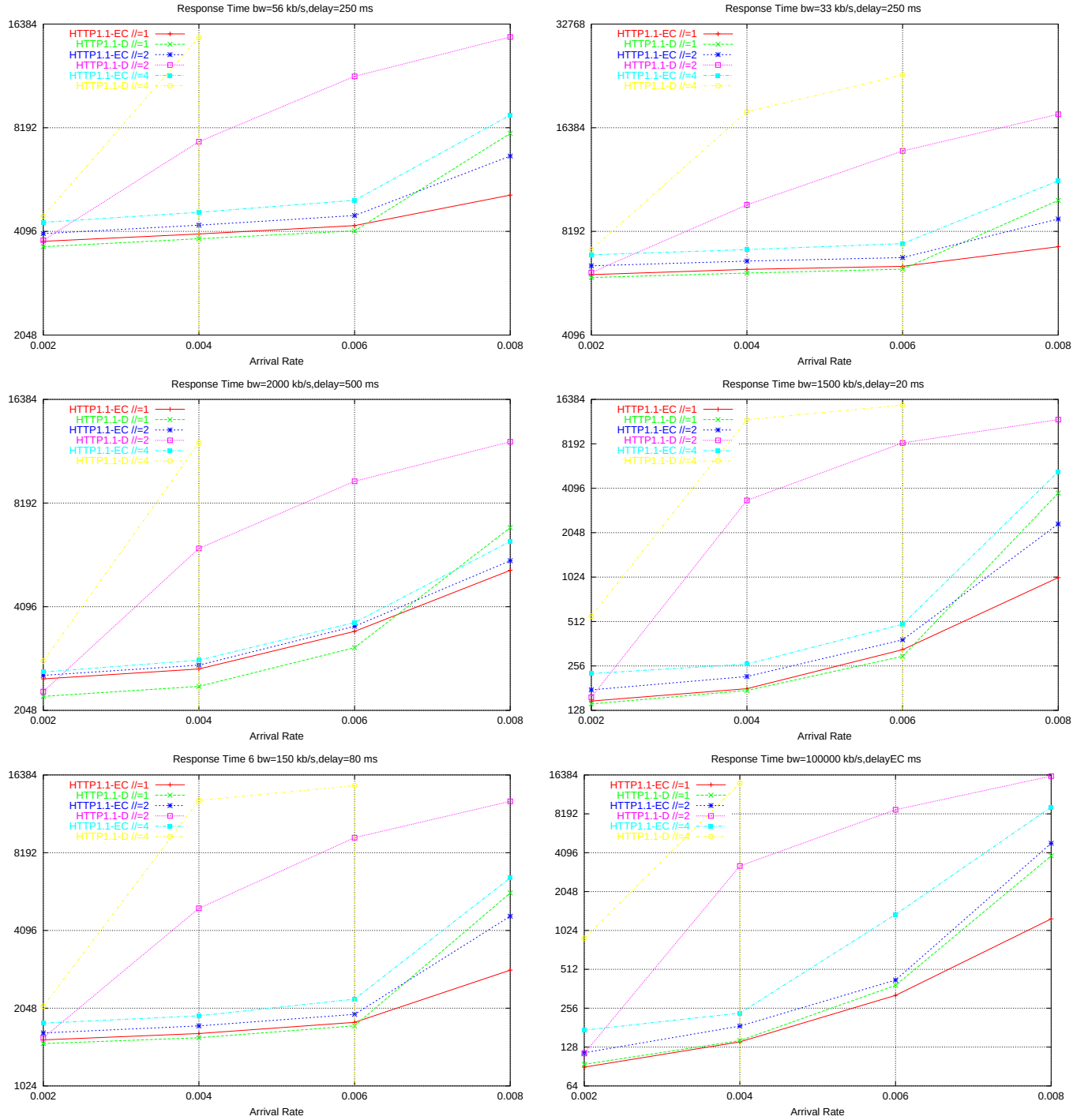

Figure 20: Effect of Multiple Parallel Persistent Connections: Mean Response time

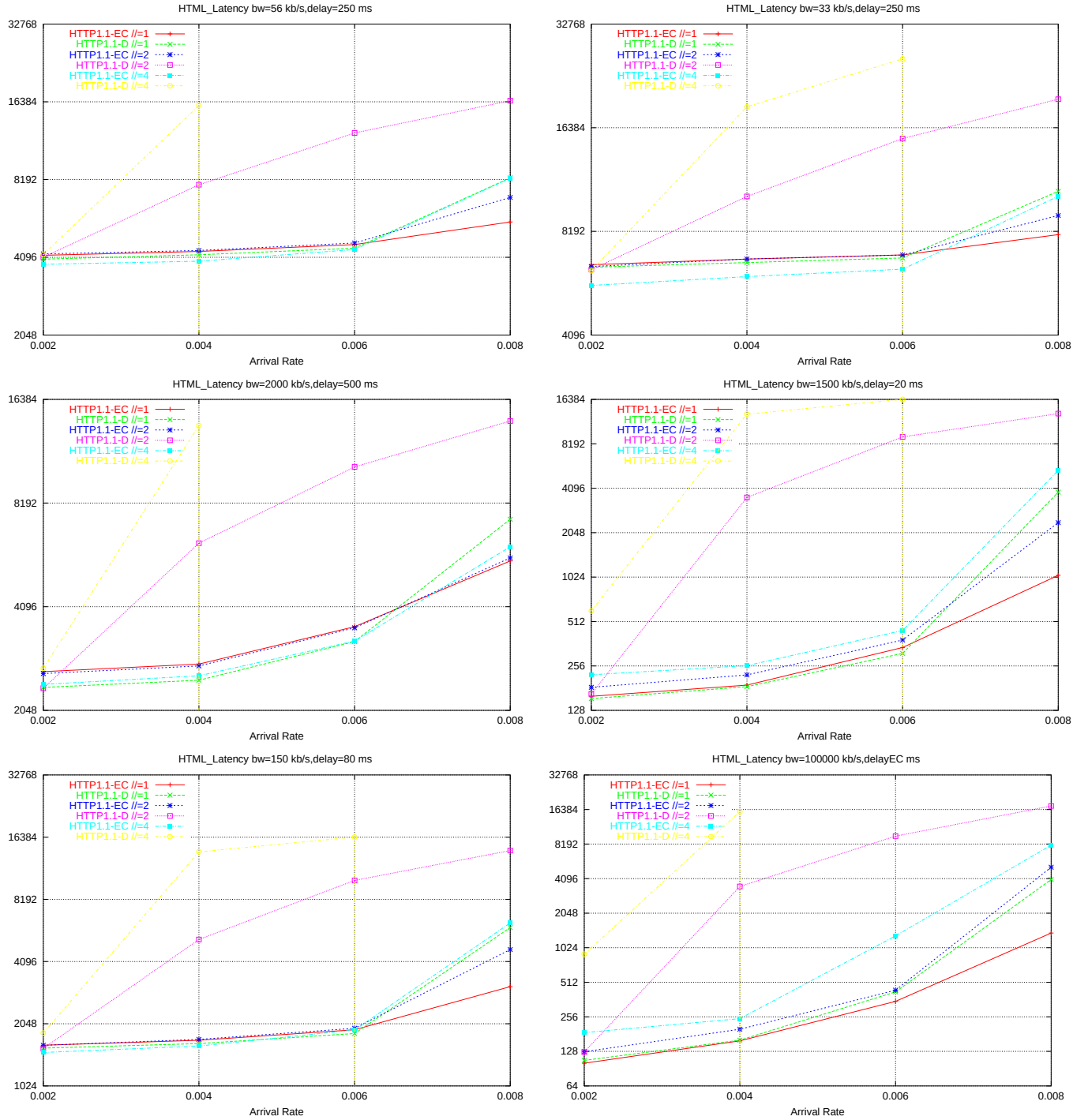

Figure 21: Effect of Multiple Parallel Persistent Connections: Mean HTML Latency

We are currently evaluating performances of Web server softwares other than Apache. We use the benchmark WAGON to compare these Web servers. We are also extending the traffic model and the benchmark tool for Web caches. More theoretical work is undergoing on multi-server queueing system proposed in section 5.1. We are analyzing formally the stability condition that allow for the weak convergence of response times. We are also computing approximate solutions of the response time and their minimization by an optimal solution of the timeout parameter. In parallel, we are pursuing the investigations on the monotonicity of the Hurst parameter in the traffic intensity. Finally, we are investigating the multiplexing scheme proposed in HTTP-NG [32]. In view of the results we obtained in this paper, we believe that a solution better than HTTP1.1-EC could be achieved by using sending (or multiplexing) parallel requests on the same connection.

Acknowledgments: The authors are grateful to Dr. Jean-François Abramatic for his support and various useful comments on this work. They are also grateful to Professor Eric Moulines for his suggestion of using the EM technique to estimate parameters of mixture of densities.

References

[1] V. Almeida, M.E. Crovella, A. Bestavros, and A. de Oliveira. Characterizing reference locality in the www. In *Proceedings of PDIS'96: The IEEE Conference on Parallel and Distributed Information Systems*, Miami Beach, Florida, December 1996.

[2] M. Arlitt and C. Williamson. Web server workload characterisation: The search for invariants. In *Proceedings of the 1996 ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, Philadelphia, May 1996.

[3] Paul Barford and Mark E. Crovella. Generating representative web workloads fot network and server performance evaluation. In *Proceedings of ACM Sigmetrics'98*, 1998.

[4] Paul Barford and Mark E. Crovella. A performance evaluation of hyper text transfer protocols. In *Proceedings of ACM Sigmetrics'99*, 1999.

[5] Lee Breslau, Pei Cao, Li Fan, Graham Phillips, and Scott Shenker. Web caching and zipf-like distributions: Evidence and implications. In *Proceedings of INFOCOM'99*, 1999.

[6] Mark E. Crovella and Azer Bestavros. Explaining world wide web traffic self-similarity. Tr-95-015, Computer Science Dept., Boston University, 1995.

[7] Mark E. Crovella, Murrad S. Taqqu, and Azer Bestavros. Heavy-tailed probability distributions in the world wide web. To appear in the book: A Practical Guide To Heavy Tails: Statistical Techniques for Analysing Heavy Tailed Distibutions., 1996.

[8] R. B. D'Agostino and M. A. Stephens, editors. *Goodness-of-Fit Techniques*. Marcel Dekker, Inc., 1986.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Statistic. Soc.*, 39(1):1–38, 1977.

[10] Anja Feldmann, Polly Huang, Anna C. Gilbert, and Walter Willinger. Dynamics of ip traffic: A study of the role of variability and the impact of control. In *Proceedings of the ACM/SIGCOMM'99*, 1999.

[11] B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, and R. Lukose. Strong regularities in world wide web surfing. *Science*, (280), 1998.

[12] A.K. Iyengar, Squillante, M.S., and L. Zhang. Analysis and characterization of large-scale web server access patterns and performance. *World Wide Web*, 2:85–100, 1999.

[13] S. Karlin and H. M. Taylor. *A first course in stochastic processes*. Academic Press, New York, 1975.

[14] Binzhang Liu. Characterizing web response time. Master's thesis, Virginia Polytechnic Institute and State University, April 1998.

[15] Zhen Liu, Nicolas Niclausse, and Cesar Jalpa-Villanueva. Web traffic modeling and performance comparison between http1.0 and http1.1. In E. Gelenbe, editor, *System Performance Evaluation: Methodologies and Applications*. CRC Press, August 1999.

[16] Bruce A. Mah. An empirical model oh HTTP network traffic. In *Proceedings of IEEE INFOCOM'97*, 1997.

[17] S. Manley, M. Courage, and M. Seltzer. A self-scaling and self-configuring benchmark for web servers. Technical report, Harvard University, 1998.

[18] Jeffrey C. Mogul. The case for persistent-connection HTTP. In *Proceedings of the SIGCOMM'95 conference*, Cambridge, MA, August 1995.

[19] Jeffrey C. Mogul. Network behavior of a busy web server and it's clients. Technical report, Digital, Western Research Laboratory, October 1995.

[20] David Mosberger and Tai Jin. httperf–a tool for measuring web server performance. In *Workshop on Internet Server Performance (WISP'98)*, Madison, WI, June 1998.

[21] H. Frystyk Nielsen, J. Gettys, A. Baird-Smith, E. Prud'hommeaux, H. Lie, and C. Lilley. Network performance effects of http/1.1, css1, and png. In *Proceedings of the ACM SIGCOMM '97*, Cannes, France, September 1997.

[22] V. Paxson and S. Floyd. Wide-area traffic: The failure of poisson modeling. In *Proceedings of ACM/Sigcomm'94*, pages 257–268, September 1994.

[23] Vern Paxson. Empirically-derived analytic models of wide-area tcp connections. *IEEE/ACM Transactions on Networking*, 2(4), 1994.

[24] Peter L.T. Pirolli and James E. Pitkow. Distributions of surfers' paths through the world wide web: Empirical characterizations. *World Wide Web*, 2(1-2):29–45, January 1999.

[25] R. A. Redner and H. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26(2):195–239, 1984.

[26] Luigi Rizzo. Dummynet: a simple approach to the evaluation of network protocols. *ACM Computer Communication Review*, January 1997.

[27] SPEC. An explanation of the specweb96 benchmark, 1996. the Standard Performance Evaluation Corporation.

[28] M.S. Squillante, D.D. Yao, and L. Zhang. Internet traffic: Periodicity, tail behavior and performance implications. In E. Gelenbe, editor, *System Performance Evaluation: Methodologies and Applications*. CRC Press, August 1999.

[29] M.S. Squillante, D.D. Yao, and L. Zhang. Web traffic modeling and server performance analysis. In *Proceedings of the IEEE Conference on Decision and Control (CDC)*, 1999.

[30] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley Series in Probability and Mathematical Statistics, 1985.

[31] Gene Trent and Mark Sake. Webstone: The first generation in http server benchmarking, 1995.

[32] Http-ng working group. http://www.w3.org/Protocols/HTTP-NG/.

[33] Zhe Wang and Pei Cao. Persistent connection behavior of popular browsers. http://www.cs.wisc.edu/ cao/papers/persistent-connection.html.

Contents

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - B.P. 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Lorraine : Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 Villers lès Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot St Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, B.P. 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399