A Hierarchical and Multiscale Analysis of E-Business Workloads

Daniel A. Menascé • Virgílio A. F. Almeida • Rudolf Riedi • Flávia Ribeiro • Rodrigo Fonseca • Wagner Meira Jr.

Dept. of Computer Science, George Mason University, MS 4A5, Fairfax, VA 22030, USA

Dept. of Computer Science, Univ. Fed. Minas Gerais, Belo Horizonte, MG 31270, Brazil

Dept. of Electrical and Comp. Engineering, Rice University, Houston TX 77251, USA

Dept. of Computer Science, Univ. Fed. Minas Gerais, Belo Horizonte, MG 31270, Brazil

Dept. of Computer Science, Univ. Fed. Minas Gerais, Belo Horizonte, MG 31270, Brazil

Dept. of Computer Science, Univ. Fed. Minas Gerais, Belo Horizonte, MG 31270, Brazil

menasce@cs.gmu.edu • virgilio@dcc.ufmg.br • riedi@rice.edu • flavia@dcc.ufmg.br • rfonseca@dcc.ufmg.br • meira@dcc.ufmg.br

Understanding the nature and characteristics of E-business workloads is a crucial step to improve the quality of service offered to customers in electronic business environments. Using a multi-layer hierarchical model, this paper presents a detailed multiscale characterization of the workload of two actual E-business sites: an online bookstore and an electronic auction site. Our analysis of the workloads showed that the session length, measured in number of requests to execute E-business functions, is heavy-tailed, especially for sites subject to requests generated by robots. An overwhelming majority of the sessions consists of only a handful requests. This seems to suggest that most customers are human (as opposed to robots). A significant fraction of the functions requested by customers were found to be product selection functions as opposed to product ordering. An analysis of the popularity of search terms revealed that it follows a Zipf distribution. However, Zipf's law as applied to E-business is time scale dependent due to the shift in popularity of search terms. We also found that requests to execute frequent E-business functions exhibit a similar pattern of behavior as observed for the total number of HTTP requests. Finally, our analysis demonstrated that there is a strong correlation in the arrival process at the HTTP request level. These correlations are particularly stronger at intermediate time scales of a few minutes.

(*E-business, WWW, workload characterization, performance modeling, heavy-tailed distribution*)

1. Introduction

E-business sites are very complex, composed of several tiers of servers of different types (e.g., web servers, application servers, and database servers), and are subject to workloads that vary in ways hard to predict. The quality of service requirements for E-business sites are strict since customers demand fast response time and high availability or else they turn to competitors.

Understanding the nature and characteristics of E-business workloads is a crucial step to improve the quality of service offered to customers in electronic business environments. E-business workload characterization can lead to a better understanding of the interaction between customers and Web sites and can also help design systems with better performance and availability.

There are very few published studies of E-business workloads because of the difficulty in obtaining actual logs from electronic companies. Most companies consider Web logs to be very sensitive data. Past studies of WWW workloads concentrated on information provider sites and found several characteristics common to them. Some of the characteristics considered deal with file size distributions, file popularity distribution, self-similarity in Web traffic, reference locality, and user request patterns. A number of studies of different Web sites found file sizes to exhibit heavy-tailed distributions and object popularity to be Zipf-like. Other studies of different Web site environments demonstrated long-range dependencies in the user request process, in other words, strong correlations in the user requests. In particular, [4] identified ten workload properties, called *invariants*, across six different data sets, which included different types of information provider Web sites. Some of the most relevant invariants are: i) images and HTML files account for 90-100% of the files transferred; ii) 10% of the documents account for 90% of all requests and bytes transferred; iii) file sizes follow the Pareto distribution, and iv) the file inter-reference times are independent and exponentially distributed. Shortly after, [3] discovered that the popularity of documents served by Web sites dedicated to information dissemination follows a Zipf's Law. In [7], the authors pointed to the self-similar nature of Web server traffic. All these studies were performed almost five years ago. Since then, several major changes have been observed in the WWW. The most important are: clients now have much larger bandwidth, the number of users has grown exponentially, and E-business became one of the major applications on the Web. A question that naturally arises is: are the characteristics and invariants found in information provider Web sites still valid for E-business workloads?

To answer this question, we define a hierarchical structure to characterize the workload and apply this structure to two different types of actual E-business sites: an online bookstore and an electronic auction site. This paper extends our previous work [12] and examines statistical and distributional properties of the E-business workloads and compare these properties across the two datasets. As much as possible, we compare the features of these workloads with the invariants that were discovered for information dissemination Web sites and provide an extended multiscale analysis of the workload.

The rest of the paper is organized as follows. Section two shows the approach used to characterize E-business workloads. The next section describes the data collection process. Section four analyzes two logs from actual E-business sites and characterizes the workload at the HTTP request level. Characterizations at the E-business function and session levels are provided in sections five and six, respectively. Section seven briefly describes related work. Finally, section eight presents concluding remarks.

2. Approach

E-business workloads are composed of sessions. A *session* is a sequence of requests of different types made by a single customer during a single visit to a site. During a session, a customer requests the execution of various E-business functions such as browse, search, select, add to the shopping cart, register, and pay. A request to execute an E-business function may generate many HTTP requests to the site. For example, several images may have to be retrieved to display the page that contains the results of the execution of an E-business function.

Workload characterization can be accomplished at many levels: user level, application level, protocol level, and network level. An E-business workload can be viewed in a multilayer hierarchical way, as shown in Fig. 1. This paper focuses on the characterization of three levels, represented by the request layer (protocol level), function layer (application level), and session layer (user level). This hierarchy can be used to capture changes in user behavior and map the effects of these changes to the lower layers of the model.

Our approach is to analyze each layer individually in order to obtain a characterization of the arrival process and usage statistics. We perform multiscale statistical analysis, study long range dependence (LRD), and burstiness. Our analysis covers information such as: session

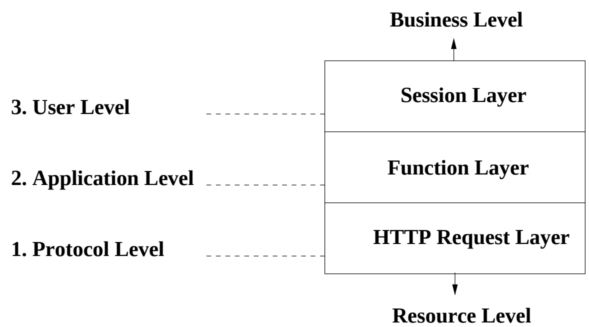


Figure 1: A hierarchical workload model.

inter-arrival times, inter-arrival times for specific E-business function requests, session length distribution, E-business function distribution per session, and number of active sessions and initiated sessions.

3. Data Collection

The data used for the workload characterization came from two actual E-businesses, an e-tailer and an auction server. The e-tailer is a bookstore that sells exclusively on the Internet. The auction site sells Internet domains. In both cases, the data consist of access logs recorded by the WWW server of each E-business.

The data comprises two weeks of accesses to each of these sites. The bookstore logs were collected from August 1st to August 15th, 1999, while the auction server logs are from March, 28th to April 11th, 2000.

During these two weeks, the bookstore handled 3,630,964 requests (242,064 daily requests on average), transferring a total of 13,711 megabytes of data (914 MB/day on average). The auction server has a smaller load, and answered 466,058 requests (31,071 requests/day) which amounts to 1,863 megabytes of data (124 MB/day). Most of these requests are for embedded images in the response pages. In the case of the bookstore, images account for 71% of the requests, while in the auction workload they represent 85.3% of the requests.

E-business function-related requests amounted to 26.3% and 14.7% of the requests received by the bookstore and auction sites, respectively. The difference in percentage between the two sites is explained by the larger number of images used by the auction site. Thus, the bookstore executed 63,711 E-business functions per day, and each service response had 12,618 bytes, on average. We should note that service-related requests are responsible for most of the network traffic, comprising 84.6% of the data sent by the bookstore server and

92.2% of the data sent by the auction server. This is explained through the fact that most of the gif files embedded in pages are usually already cached and are not transmitted back to the client. Although the auction pages contain more images than the bookstore pages, the auction site employs a smaller array of images, typically banner advertisements and page layout. The bookstore uses a larger number of different images, such as book covers, and can therefore benefit less from the advantages of caching.

4. Request-layer Characterization

In this section, we study the statistical nature of the arrival process of HTTP requests with two main goals in mind: i) the extraction of statistically significant features towards classification, understanding and modeling of request workload, and ii) the prediction of workload intensity, allowing for an adaptive provisioning of resources towards optimal performance.

It is now a well accepted fact that strong correlations are present in various aspects of the World Wide Web, from request arrivals on servers to packet arrivals on the network. These correlations lead to "burstiness" or high variability, which may degrade performance and throughput if not accounted for. We carry out a statistical analysis across all time scales to detect correlations and assess their strength.

The fact that statistical analysis and modeling has to incorporate different methods according to time scale is most apparent as we attempt to accommodate various trends. On the largest time scale of days (in our study), the weekend produces somewhat less volume, while on the scale of hours the presence of a periodic sleep-wake pattern per day is visually obvious. It is not our intention to explore these patterns. On finer time scales, structure is much less obvious and it is our goal to present a simple analytical tool that distinguishes scales of "noisy (non-predictable) oscillations" from scales with strong correlations (which foster prediction). Thereby, care is needed to avoid bias from large scale trends and non-stationarities which could manifest in both, small scale analysis and prediction.

A visual inspection of the number of requests arriving at the bookstore on different time scales, i.e., in time intervals of varying length (see Fig. 2) reveals, even to the inexperienced eye, an apparent strong dependence that shows long sequences of increase or decrease of volume (trends), particularly pointed at intermediate time scales. The purpose of our analysis is to decide whether these trends are purely due to changes in traffic volume during the day and week or whether there is predictable behavior beyond these cycles.

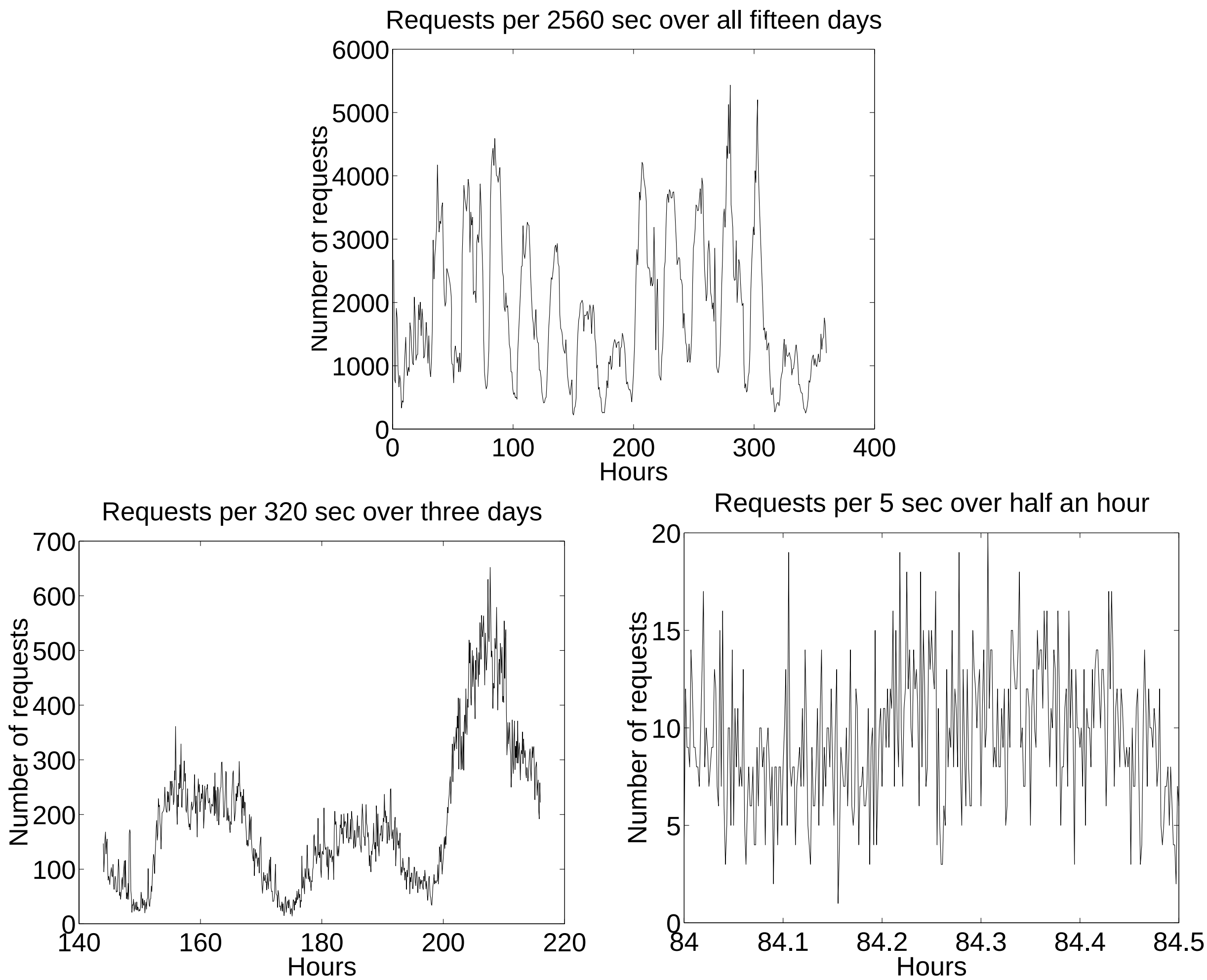


Figure 2: Total number of requests arriving at the bookstore site at various time scales. Top: Arrivals per 2560 secs over the full fifteen days captured; Bottom left: at a resolution of 320 secs over three days; Bottom right: at full resolution (5 secs), over half an hour.

A simple quantification of dependence over various scales is achieved by computing the sample variance by time scale: if arrivals occur independently of each other, the sample variance doubles if the length of the interval doubles. The variance exceeds twice the variance of the original interval if the arrivals are positively correlated and will not reach twice the variance if the arrivals are negatively correlated. Indeed, $\mathrm{var}(A+B) = \mathrm{var}(A)+2\mathrm{cov}(A,\ B)+\mathrm{var}(B)$. More specifically, if X_k denotes the number of arrivals in time interval $[k\ \delta, (k+1)\ \delta]$, where δ is the finest time resolution one is interested in, then $X_k^{(n)} = 2^{-n} \times (X_{k2^n} + X_{k2^n+1} + \ldots + X_{k2^n+2^n-1})$ averages the arrivals in $[k2^n\delta,\ (k+1)2^n\delta]$ and can be computed efficiently through the recursion $X_k^{(n)} = (X_{2k}^{(n-1)} + X_{2k+1}^{(n-1)})/2$. The log-log plot of the variance against scale, i.e., $\log_2\ \mathrm{var}\ X^{(n)}$ versus n, is called variance time plot (VTP). This plot has the slope

-1 for independent data (recall the normalization factor $1/2$ necessary to provide averages instead of total counts) and a different behavior for dependent data: The slower the VTP decays at a certain scale, the stronger the next-neighbor correlation within that scale.

The extreme case of positive correlation is a constant series X_k with a flat (horizontal) VTP. A more interesting case of dependent behavior constitutes the so-called "statistical self-similarity," which is defined by the requirement that var $X^{(n)} = \sigma^2\ 2^{n(2H-2)}$. Here H denotes the Hurst parameter and lies between 0 and 1. This case is of interest due to the existence of appealing, simple, Gaussian processes with such properties, such as the fractional Gaussian noise and the auto-regressive FARIMA processes [19]. For $H = 1/2$ we find ourselves back in the case of independent data where var $X^{(n)} = \sigma^2\ 2^{-n}$ for all time scales n. On the other hand, if the VTP decays at a slower rate, i.e., with slope $2H - 2$ where $H > 1/2$, then we have positive correlations.

It is important to be able to detect strong dependencies because they degrade estimation by increasing the variance of the estimation error. On the other hand, by detecting strong dependencies, one can foresee not only mean behavior but also temporary phases of increase or decrease in volume. The VTP is a crude measure of the correlation structure with known bias and poor performance as an estimator of the LRD parameter H and is particularly sensitive to non-stationarities such as changing mean. However, when properly applied, the VTP is completely valid as a tool for a first look (see [1, 2, 19] and references therein).

The VTP plot of the number of arrivals at the online bookstore (see Fig. 3) shows a decay of particular strength corresponding to $H = 0.98$ at intermediate time scales from 80 to 5120 sec, corresponding to aggregation 4-10 in Fig. 3 (a) (there δ corresponds to 5 sec). Due to the presence of large scale trends (or non-stationarity) this number has to be considered with caution since the estimate could be highly biased. Indeed, the scaling is not optimal, and it is wise to perform local scaling tests over regions where the data shows stationarity. Indeed, over periods of several hours (see Fig. 3 (b) for an analysis of six hours in the afternoon of the sixth day) the VTP becomes more straight and the measured Hurst exponent falls into the region generally observed in natural phenomena (.7 to .85). The scaling we found—typically noon to evening—over the fifteen days averaged to about .73. Also, in this local analysis, the dependence seems to be strongest at intermediate time scales.

This strong dependence on intermediate scales is further confirmed by the "cross-correlation" plots of the arrival process on various time scales: In Fig. 4 we display graphs of $X_{k-1}^{(n)}$ versus $X_k^{(n)}$, for three fixed values of n, where $X_k^{(n)}$ denotes the total number of requests

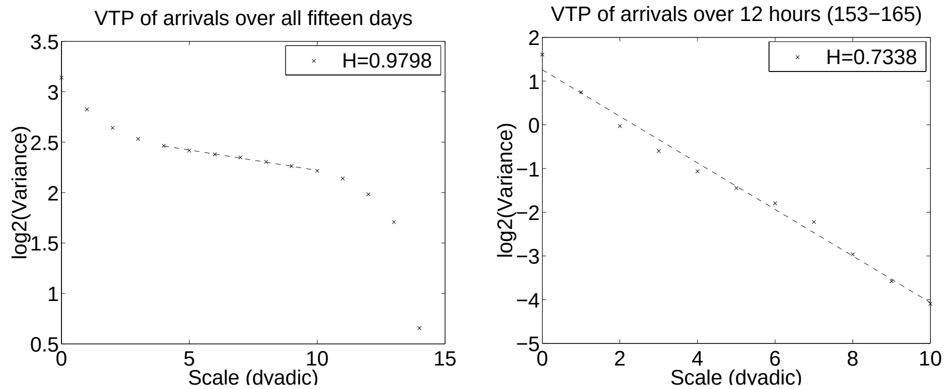


Figure 3: Variance time plot for the online bookstore. Left: over the full fifteen days (H = 0.9798). Right: over six hours of approximate stationarity of the mean arrival rate (H = 0.7338).

arriving at the online bookstore in the time interval $[k2^n\ \delta, (k+1)2^n\ \delta]$, δ being 5 seconds in our data. These plots give an idea of the next-neighbor dependence on the time scales of Fig. 3. Note that the more the data is clustered along the diagonal, the higher is the predictability: large values are most likely followed by large values, small values by small values. For illustration purposes, we also show in Fig. 4 the "correlation" plot of a series of independent random variables. In this case, no structure and no clear clustering is visible.

On an intermediate time scale we find close clustering along the diagonal while there is more spread on the fine scale displayed. This indicates superior predictability on intermediate scales from many minutes to maybe a couple of hours. On the coarsest scale, the data looks random again.

The difficulty in interpreting these "cross-correlation" plots resides in the presence of the predominant cyclic trends on the largest time scales. A critical observer could rightfully claim that the concentration along the diagonal is purely caused by these cycles, meaning that the data could be well approximated by a quasi-periodic (cyclic) mean superimposed with independent random fluctuations. Whether such an interpretation is valid cannot be decided from Fig. 4 because the amplitude of the fluctuations around the mean are considerably smaller than that very mean itself.

In order to get clear on this issue a wavelet analysis could be beneficial, as wavelets allow an analysis insensitive to trends (due to vanishing moments [8]) and provide an un-biased estimation of the Hurst parameter H [1, 2]. Such an approach, however, is beyond the scope

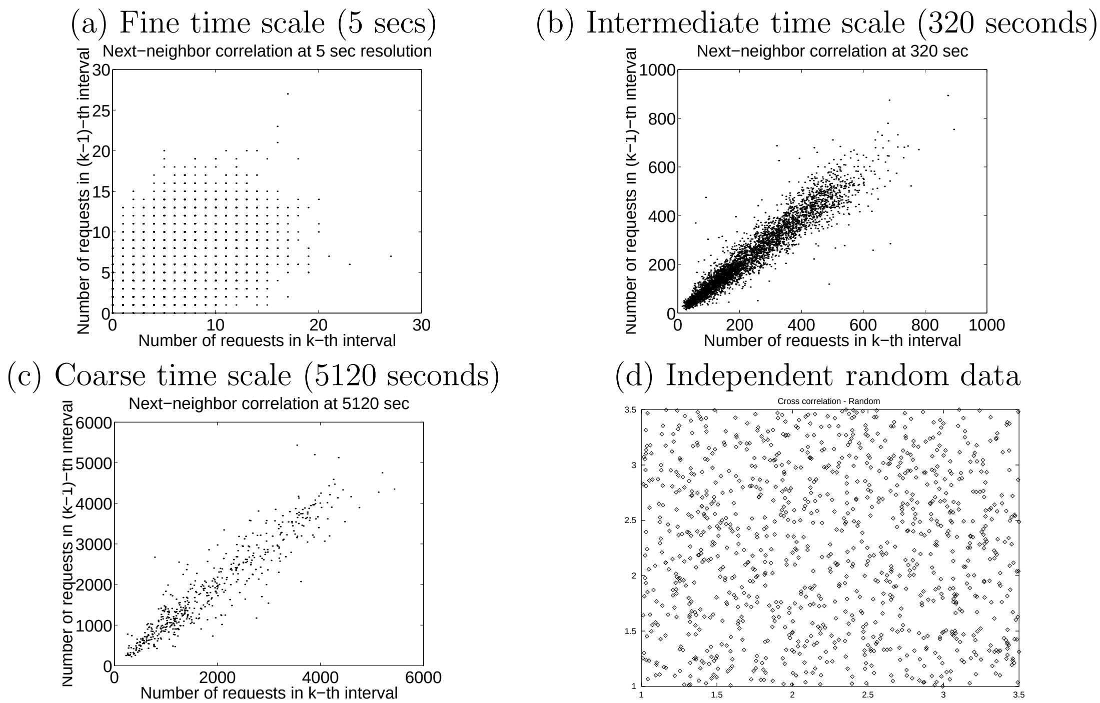


Figure 4: Cross correlation plots for the number of requests arriving at the online bookstore.

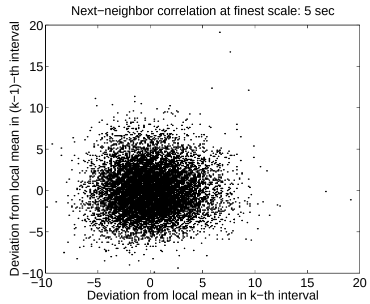
of this paper and we favor a more direct and simple approach. To test for predictability and at the same time remove bias from the changing arrival rate we study

$$Z_k^{(n)} = X_k^{(n)} - (1/8) * (X_{k-8}^{(n)} + \ldots + X_{k-1}^{(n)})$$

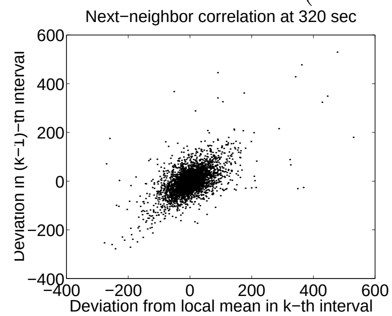
which is, in fact, the difference between the current number of arrivals and the average of the last eight arrivals at time scale $2^n * \delta$. The choice of averaging eight is arbitrary. Again, we display the next-neighbor correlation at three fixed values of n in Fig. 5, i.e. $Z_{k-1}^{(n)}$ versus $Z_k^{(n)}$. Again, we note the presence of correlations at intermediate time scales of hundreds of seconds, indicating that an increase in volume against past average volume is likely to be followed by yet another increase.

A possible explanation for particularly strong correlations on the time scale of few hundred seconds may be human think time and human distractedness. The overall self-similarity, at least "asymptotically," may be argued for by invoking the well known on-off process which was crucial in explaining self-similarity in network traffic loads [9]: The number of requests per session follows a heavy-tailed distribution. Since the number of requests sent per time unit is limited, sessions are thus sending requests over on-times which are heavy-tailed. Numerical support for this claim comes from our analysis of session duration in Sec. 6, which

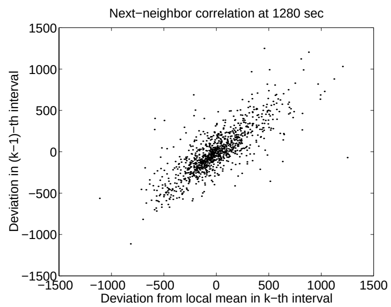
(a) Fine time scale (5 secs)



(b) Intermediate time scale (320 seconds)



(c) Coarse time scale (1280 seconds)



(d) The deviation itself at resolution 320 secs

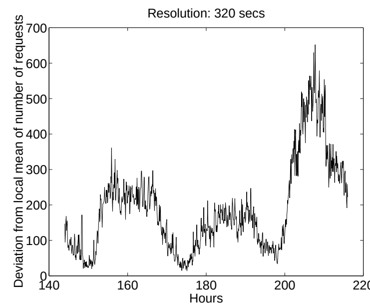


Figure 5: Cross correlation plots for the deviation from the local mean of the number of requests arriving at the E-business site.

shows that the distribution of session length follows a power law. The on-off model then relates the exponent of this fat-tailed distribution directly with H.

Having studied the data from the online bookstore in detail, let us now compare our two data sets. Figure 6 presents the average number of requests per day in our two-week logs. We can clearly notice the traffic reduction during the weekends, which are the days 0, 6, 7, 13, and 14. We can confirm this behavior by checking the graph in Fig. 7, where each point

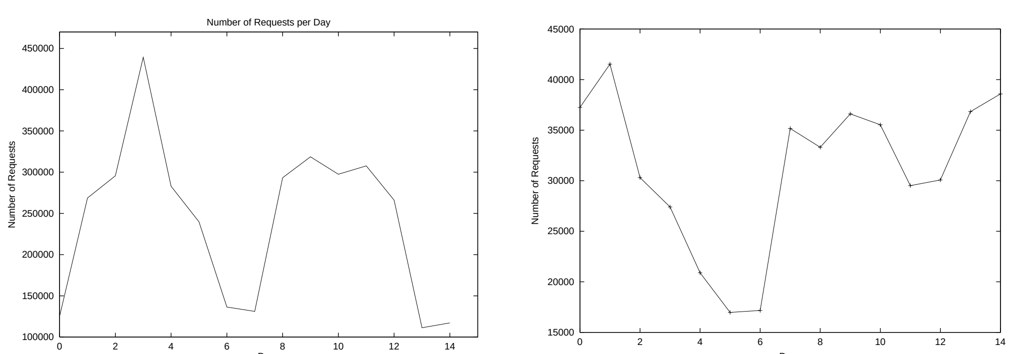


Figure 6: Number of requests per day for the bookstore (left) and auction sites

represents the hourly request average. As we can see, there are fourteen peaks, almost one per log day.

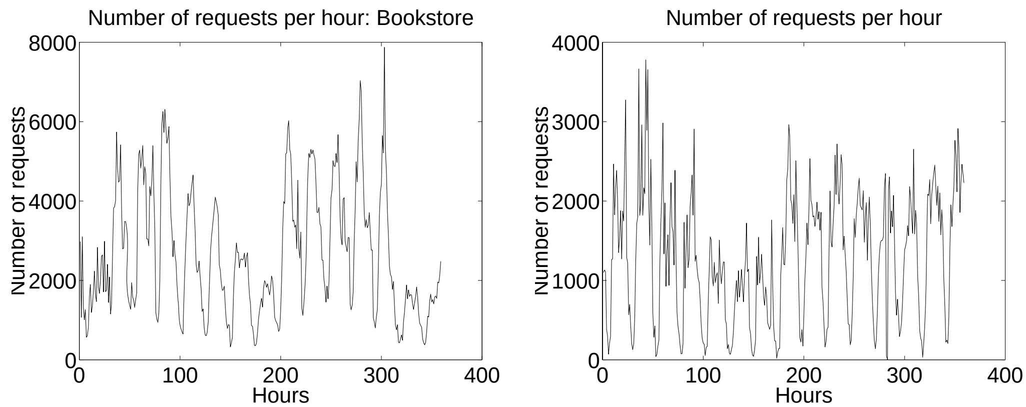


Figure 7: Number of requests per hour for the bookstore (left) and auction sites

Another representation of the request arrival process is given by the inter-arrival times graph (IAT) of Fig. 8 that plots a spike for each request in the log. The height of the spike is proportional to the time interval between a request and its predecessor. The lighter weekend traffic may also be observed in the graph by looking at the highest peaks.

Figure 9 shows the distribution of references per domain. If we compare these results with the ones presented in [4], we can see that the popularity of hosts is much less concentrated on few domains since E-business sites tend to have a much broader audience than content dissemination Web sites.

5. Function Characterization

In this section, we characterize the workload at the level of E-business functions. Our first criterion is the nature of the function. When considering an online store, we may divide the functions into four groups: static, product selection, purchase, and other. Static functions comprise the home and informational pages about the store. Product selection includes all functions that allow a client to find and verify a product they are looking for: browse, search, and view. Purchase functions indicate a desire to buy, either by selecting a product for later acquisition (e.g., add to cart) or by ordering it (e.g., pay). One interesting invariant in the logs we analyzed is that more than 70% of the functions performed are product selection functions. Table 1 presents a distribution of E-business function requests for both sites.

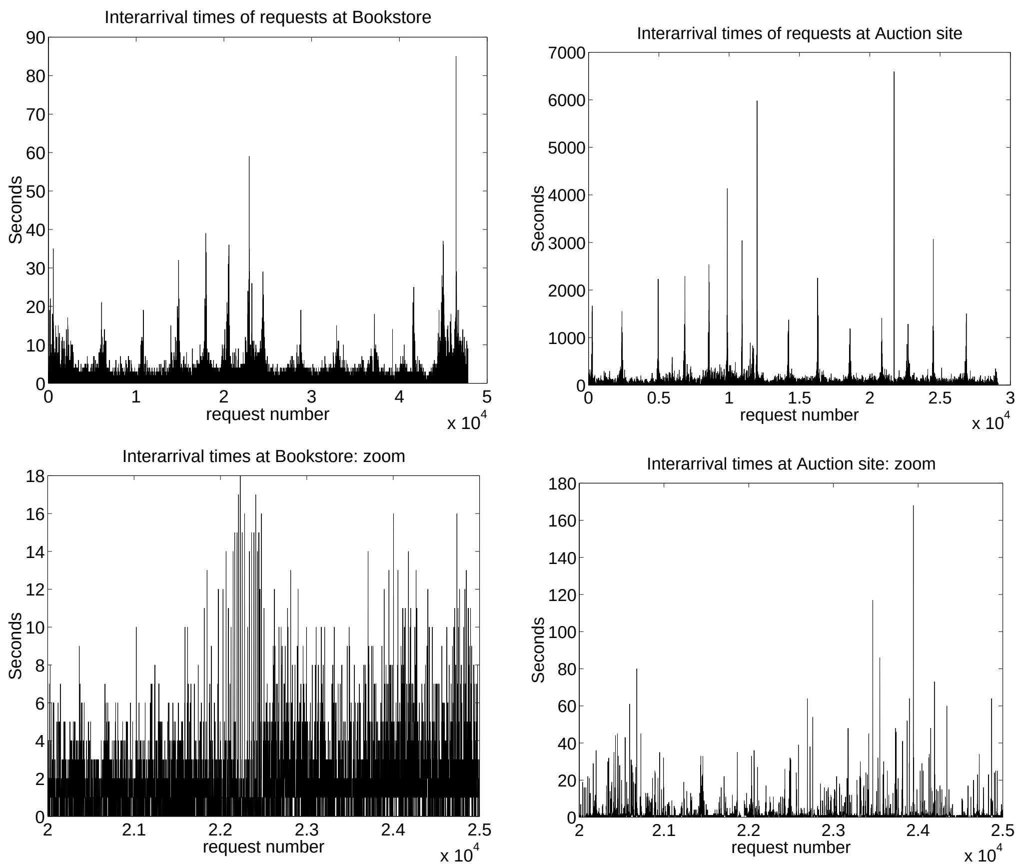


Figure 8: Inter-arrival times (IAT) of requests at bookstore (left) and auction site (right) for the full trace (top) and a selection of 5,000 requests (bottom). The IAT at the auction site are more 'explosive' indicating sudden unexpected long periods of silence while the bookstore seems to see long periods of silence which are less abrupt and which could be anticipated.

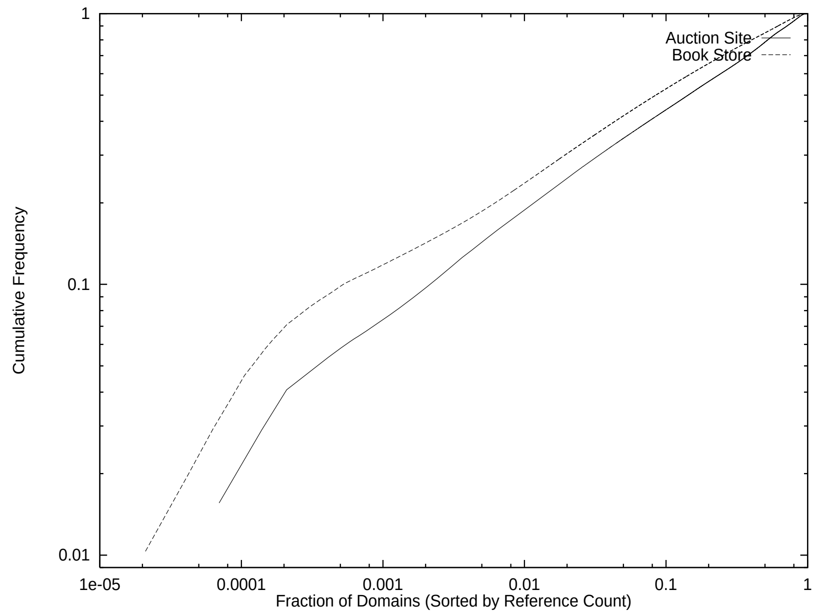


Figure 9: Distribution of References by Domain

Bookstore		Auction	
Function	Frequency	Function	Frequency
Home	11.92%	Home	20.70%
Browse	17.72%	Browse	14.66%
Search	36.30%	Search	16.74%
View	19.99%	View	4.87%
Add	5.44%	Bid	0.08%
Pay	1.19%	Sell	7.99%
Account	2.44%	Account	5.99%
Robot	0.04%	Robot	0.06%
Info	3.66%	Info	9.44%
Other	1.31%	Other	2.31%
		Auth	9.18%
		Register	7.29%
		Admin	0.71%

Table 1: Distribution of E-business functions.

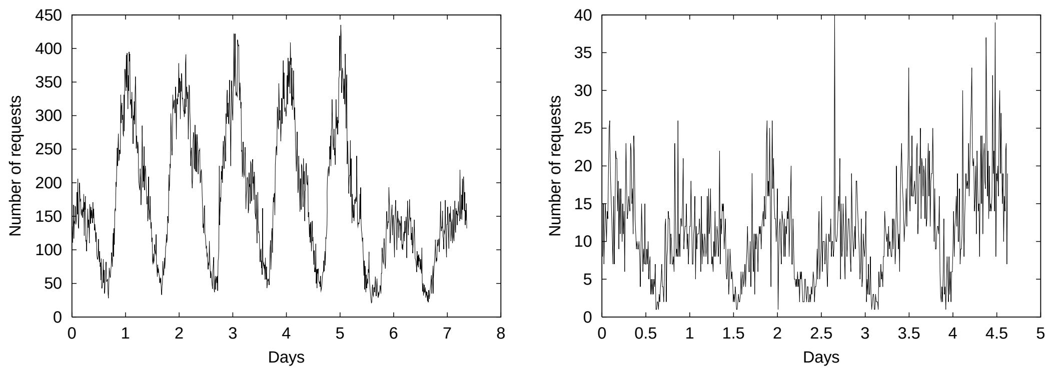


Figure 10: Number of arriving requests to execute frequent E-business functions. The time resolution is similar to that of Fig. 7. (Left) Bookstore: search, (Right) Auction site: home.

In the auction site, there are functions that relate to the process of posting items for sale. Similarly to the bookstore, though not as large in percentage, the majority of requests at the auction site concerns selection of products. On both sites, the functions directly related to spending money have a very low frequency.

When we split requests according to the E-business functions they invoke, i.e., search, browse, add, and pay, we find two clearly distinct classes. While the behavior on large time scales of hours and days of all functions follow the already observed human behavior, their small scale behavior is quite different. For example, Fig. 10 shows the number of requests to execute searches at the bookstore and to retrieve the home page of the auction site for several days for a time resolution similar to that of Fig. 7, i.e., in the order of ten minutes. If we compare Fig. 10 to Fig. 7, we see a similar pattern. This indicates that requests to execute frequent E-business functions exhibit a similar pattern of behavior as observed for the total number of HTTP requests.

The same is not true for less frequent functions such as pay and view, as indicated in Fig. 11. This figure shows clear bursts and a very different behavior from Fig. 7. Here, a more advanced statistical analysis revealing the multifractal scaling would be in place [18] and prediction is harder. In contrast, the more frequent functions such as “search” and “home” show statistics similar to the overall load of requests and are—appart from the cyclic trends—well described by Gaussian LRD processes.

This difference in small scale behavior is similar to the one we saw in the IAT process at the bookstore and the auction site (see Fig. 8). It is best understood when thinking in terms of doubly stochastic Poisson processes where Poisson arrivals are driven by a *varying*

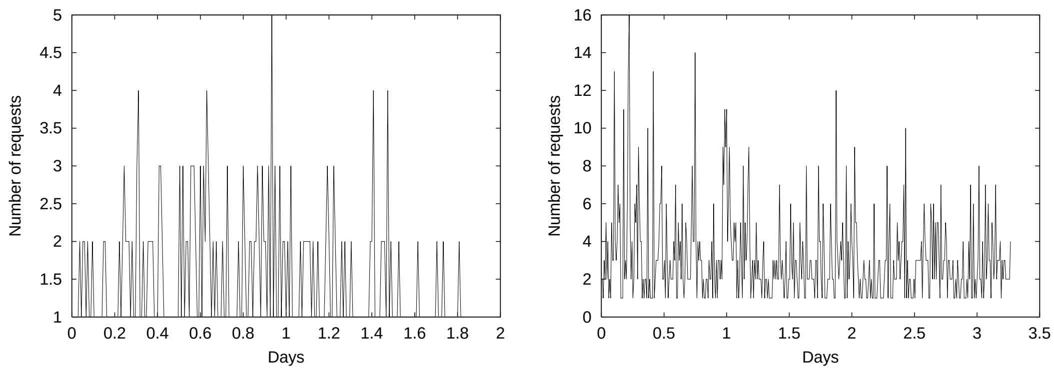


Figure 11: Number of arriving requests to execute less frequent E-business functions, The time resolution is similar to that of Fig. 7. (Left) Bookstore: pay, (Right) Auction site: view.

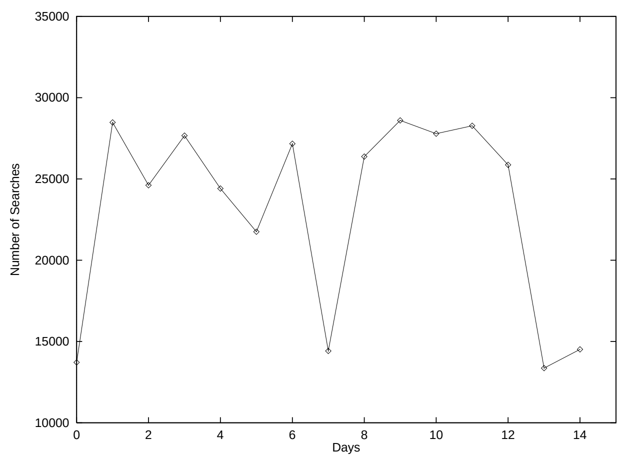


Figure 12: Number of searches arriving per day at the bookstore over two weeks.

intensity which is itself random. As intensities are low, the spikyness of Poisson arrivals are apparent; as intensities grow, the Poisson distributions are well approximated by the Gaussian. In a unifying approach one would aim at measuring the “hidden” intensity, thus capturing the driving stochastics of request arrivals and allowing for a deeper understanding and more control. This is left for future investigation.

Figure 12 shows the number of search requests for the bookstore on a daily basis. We can see that this graph exhibits, in the first week, a behavior different from the overall number of request per day (Figure 6). We attribute the difference to the fact that the search function is used by robots, which behave differently from human users. For instance, the spike observed in day 3 results from an unexpected number of requests for the home page. Such behavior could also be indicative of a denial of service attack and understanding such dynamics could be of advantage for security purposes.

5.1 Popularity of Search Terms

Prior studies of Web traffic have found that the popularity of static pages (i.e., documents) served by information provider web sites follows Zipf's law [3, 4]. In E-business sites, customers look for product information instead of documents or static pages. Product information is usually generated by dynamic pages, based on keywords provided by customers. A common way of finding product information in an online store is through query-based search functions, which are the central part of product seeking in E-business sites. Customers use keyword search functions to find out about products and services. To improve the efficiency of search functions it is important to understand the behavior of customers when they are looking for information. So, we want to examine the frequency of specific queries and find out the underlying distribution of these queries. We conjecture that a small set of queries, that refer to popular items of the store, are repeated many times over the course of a day.

Reference [11] shows that surfing patterns on the Web display strong statistical regularities, that can be described by universal laws. Zipf's law has been extensively used to explain the patterns of access to web servers and proxies. We investigate this issue further by studying the patterns of keywords used by customers during their interaction with an E-business site.

Zipf's law [20] is a relationship between the frequency of occurrence of an event and its rank, when the events are ranked with respect to the frequency of occurrence. Zipf's law [20] was originally applied to the relationship between words in a text and their frequency of use. It states that if one ranks the popularity of words used in a given text (denoted by r) by their frequency of use (denoted by $f(r)$) then

$$f(r) \sim 1/r. \tag{1}$$

This expression can be generalized as

$$f(r) = C/r^{\alpha}, \tag{2}$$

where C is a constant and α a positive parameter equal to one. This law describes phenomena where large events are rare, but small ones are quite common. Relationships such as Zipf's law can be used to facilitate both cache resource planning and strategies for distributing E-business functions.

In Fig. 13, we plot the relative percent frequency of a given query term versus its popularity rank. It shows that Zipf's law applies quite strongly to *the terms used for search functions*. This result is similar to the one found in [3], which showed that Web documents returned by Web servers also follow a Zipf's law. The figure displays three curves: one for the bookstore, one for the auction site, and other for Zipf's law. As it can be seen, there is a good match with Zipf's law over an extremely wide range of popularity, except for the most popular keywords. This fact is represented by the relatively flat part of the bookstore curve for small values of the term rank (i.e., popular search terms). Let us examine this fact in more detail.

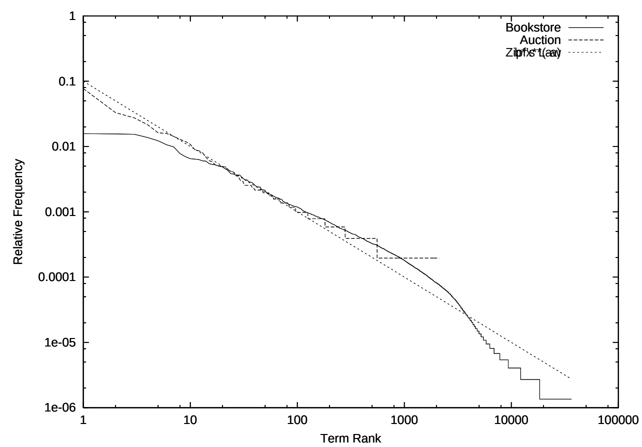


Figure 13: Popularity of search terms.

At first sight, it appears that the most popular keywords do not follow Zipf's Law. Our conjecture is different. We believe that Zipf's Law is time scale-dependent. In other words, the period used for popularity analysis should be a typical and representative period for accesses to the site. For example, in our analysis we used a two-week period of logs. In these two weeks, there is a kind of "shifting" in popularity. The most popular keyword in the first week may be different from the most popular one in the second week. However, cumulatively both keywords may get the same number of requests. If this phenomenon happens for several keywords, the result can be seen as a flat region in the leftmost part of the frequency versus rank plot. In order to verify this conjecture, we plot the popularity graphs for different time scales, as shown in Fig. 14.

The top leftmost graph of Fig. 14 corresponds to logs of one-day period of time. In this case, the flat part of the curve was clearly reduced. Our explanation is that one day is too short a period for a significant shift in popularity to occur. As we increase the period of

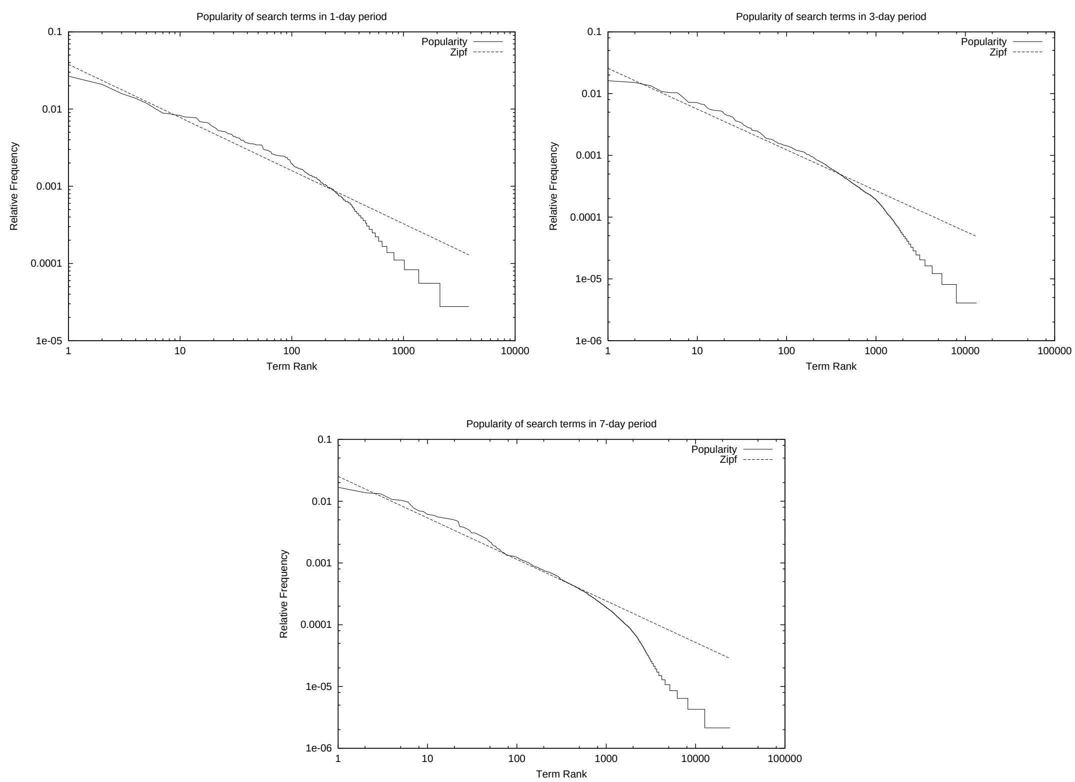


Figure 14: Popularity of search terms for different time scales.

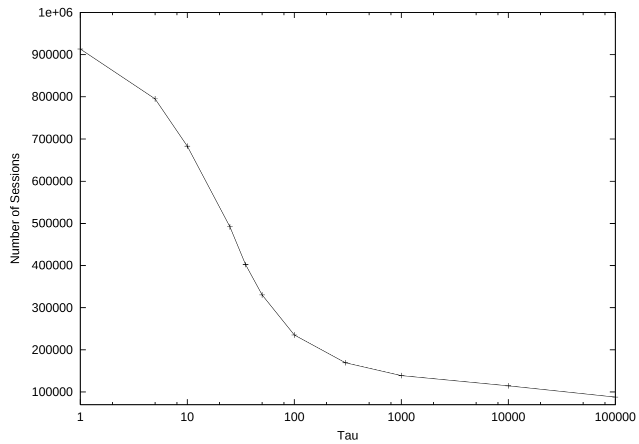


Figure 15: Influence of the threshold τ on the total number of sessions initiated.

analysis, we notice that the flat part of the curve increases accordingly as seen in the two remaining graphs of Fig. 14 and on Fig. 13. On the other hand, measuring Zipf's law over too short time intervals could bias the powerlaw since the most popular keywords could have not enough 'opportunity' to be requested in order to follow the exact power law.

6. Session Characterization

Session boundaries are delimited by a period of inactivity by a customer. In other words, if a customer has not issued any request for a period longer than a threshold τ, his session is considered finished. Usually, sites enforce this threshold and close inactive sessions to save resources allocated to these sessions. For the auction site, we know that the HTTP server enforced a threshold of twenty minutes. Since we do not have this information for the bookstore site, we had to estimate the threshold from the log. The value of τ has an influence on the number of sessions being handled by the site.

We discuss the effect of τ in what follows. Figure 15 shows the effect of the value of τ in the total number of sessions initiated for the bookstore site. As we can see, as the threshold increases from 1 to 100 sec, the number of sessions initiated decreases very rapidly. From 1000 sec on, the decrease is very small. This indicates that most sessions last less than 1,000 sec. A de facto industry-standard has been that 30 minutes (i.e., 1,800 sec) should be used to delimit sessions.

Figure 16 shows the distribution of session lengths, measured in number of requests to execute E-business functions, for both sites. The threshold τ used for the bookstore is 1,800

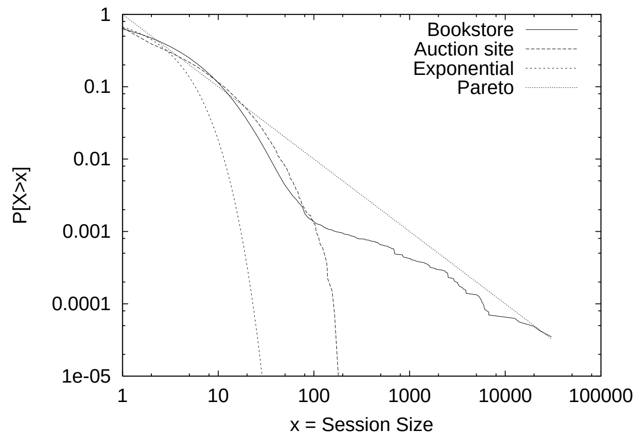


Figure 16: Session size distribution.

seconds while there is no threshold for delimiting the sessions at the auction site, since it implements timeouts for its sessions. The graphs of Fig. 16 show the empirical tail of the distribution of the session length X, i.e., $P\{X > x\}$ for the bookstore and auction sites, as well as the tail of the exponential and Pareto distributions. A random variable X, such as Pareto, that has a heavy-tailed distribution is characterized by $P\{X > x\} \sim x^{-a}, 0 < a < 2$. Among other implications, a heavy-tailed distribution presents a great degree of variability, and a non-negligible probability of high sample values. The exponential distribution decays much faster than a heavy-tailed distribution. In a log-log plot, x^{-a} is a straight line with inclination $-a$. We can distinguish two regions in the plot of Fig. 16. The first one comprises session sizes up to 100 requests, in which the curves for both sites are similar. In particular, in the region from about 5 to 100, they are fit by a straight line (not shown for clarity) with inclination ~ -2.05. For sessions longer than 100, the behavior changes. We can see that for the auction site, the probability for longer sessions falls abruptly, whereas for the bookstore it remains close to the straight-line plot of a Pareto-like distribution with $a = 1$. This "very" heavy tail is most likely due to the accesses by robots, which tend to exhibit long sessions. The auction site was not accessed by any detectable robot, and this explains why one does not see sessions much longer than 100 requests.

We can also notice that most sessions are small (about 90% of the sessions for both workloads have less than 10 requests).

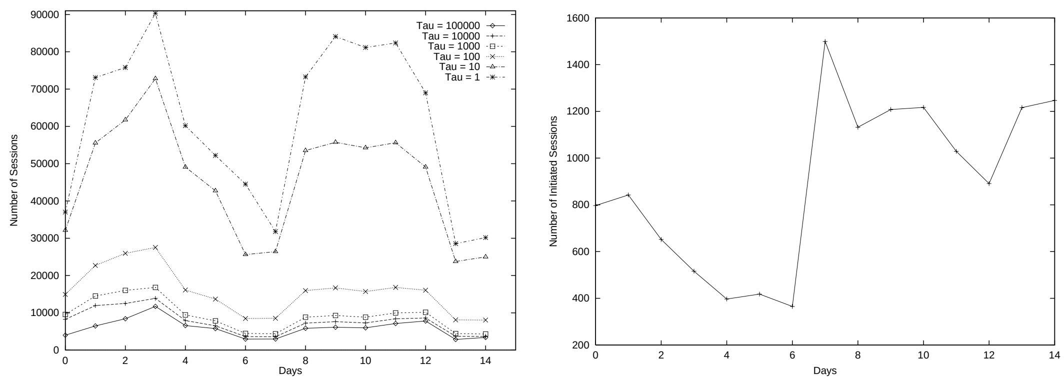


Figure 17: Number of initiated sessions per day for the bookstore (left) and the auction site.

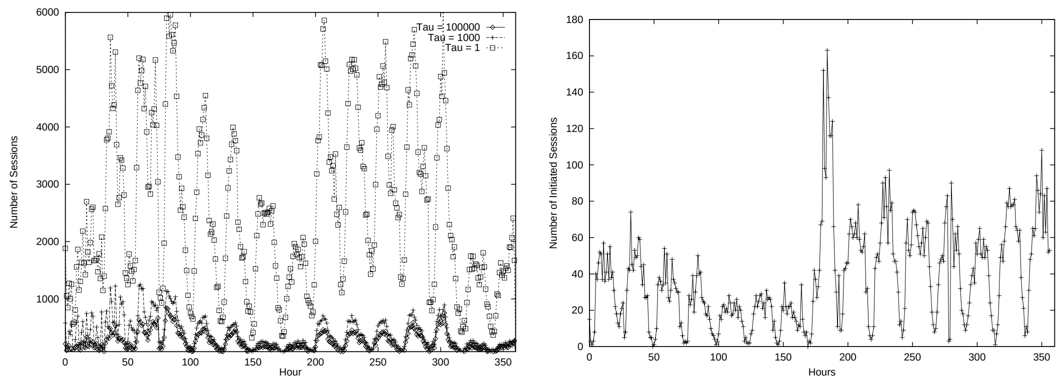


Figure 18: Number of initiated sessions per hour for the bookstore (left) and the auction site.

6.1 Usage Analysis

The left part of Fig. 17 shows the number of sessions initiated per day at the bookstore site for various values of the threshold τ. A small value of τ corresponds to the extreme case of considering each request as a session. The picture clearly shows that there is very little difference in the number of sessions as τ is increased from 1,000 sec to 10,000 sec. This is a strong argument in favor of the 30-minute standard. A similar behavior is seen in the left part of Fig. 18. The right part of Figs. 17 and 18 indicate the number of sessions initiated per day and per hour for the auction site. If we compare the shape of the graph of the number of initiated sessions for the bookstore site for $\tau = 1000$ and for the number of initiated sessions for the auction site with the corresponding graphs of Fig. 6, for number of arriving requests, we see some degree of similarity.

Figure 19 displays the number of active sessions on an hourly basis for various values of

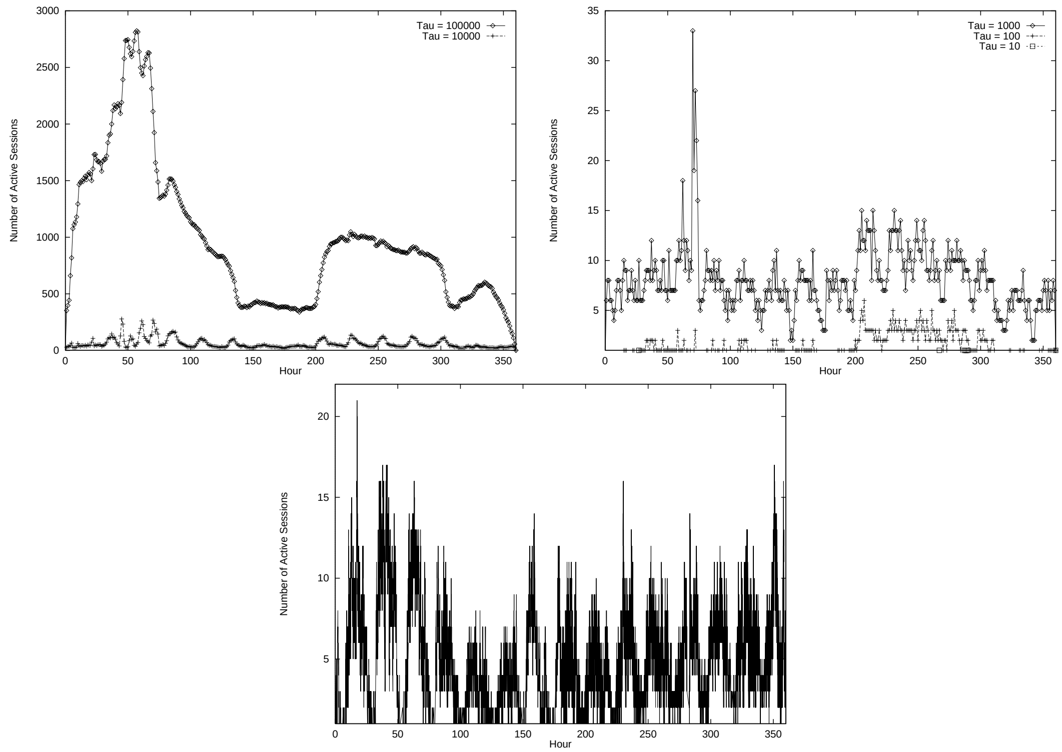


Figure 19: Number of active sessions. Top two figures are for the bookstore.

the threshold τ. Again, very little variation is seen for $\tau > 1000$ sec. At a time scale of one hour, we observe a high variability in the number of active sessions per hour since the session timeout for the auction site or the threshold of 1,000 sec for the bookstore are of the same order of magnitude as the time scale.

7. Related Work

To evaluate the performance of an E-business site, one needs a solid understanding of its workloads. Most of the existing references on workload characterization in the WWW focus only on information provider Web sites [16]. A few references have addressed the problem of workload characterization for E-business sites. Existing work concentrates on characterizing Web workloads composed of sequences of file requests [4, 7]. The characteristics and statistical properties of workloads on the Web have been studied by many [3, 4, 7, 16]. A number of studies of different sites identified WWW workload properties and invariants. For instance, file sizes have heavy-tailed distributions (e.g., Pareto distribution), object popularity follows Zipf's Law and WWW traffic is bursty across several time scales [7].

In [10], the authors introduce the notion of session, consisting of many individual HTTP requests. However, they do not characterize the workload of E-business sites, which is composed of typical requests such as browse, search, select, add, and pay. The analysis focuses only on the throughput gains obtained by an admission control mechanism that aims at guaranteeing the completion of any accepted session. The work in [17] proposes a workload characterization for E-business servers, where customers follow typical sequences of URLs as they move towards the completion of transactions. The authors though do not present any characterization or properties of actual E-business workloads. In [15], the authors propose a graph-based methodology for characterizing E-business workloads and apply it to an actual workload to obtain metrics related to the interaction of customers with a site. For example, the paper shows how to obtain information such as the number of sessions, average session length, and buy-to-visit ratio. Reference [13] presents several models (e.g., customer behavior model graph and customer visit model) for workload characterization. It also shows how models can be obtained from HTTP logs.

8. Concluding Remarks

Several studies have been published regarding the workload of information provider sites. However, very few studies are available for E-business sites. This paper used a hierarchical approach for workload characterization of E-business sites. The characterization was done at the session, E-business function, and request levels. The paper shows a large number of graphs containing a detailed characterization of the two sites analyzed: a bookstore and an auction site.

Some of the findings are: i) most sessions last less than 1,000 sec. ii) 88% of the sessions have less than 10 requests. iii) the session length, measured in number of requests to execute E-business functions, is heavy-tailed, especially for sites subject to requests generated by robots. iv) more than 70% of the functions performed are product selection functions as opposed to product ordering functions. v) requests to execute frequent E-business functions exhibit a similar pattern of behavior as observed for the total number of HTTP requests. vi) the popularity of search terms follows a Zipf distribution. However, Zipf's law as applied to E-business is time scale dependent due to the shift in popularity of search terms. vii) there is a strong correlation in the arrival process at the request level. This correlation is given by an average Hurst parameter value of 0.73. viii) correlations in the arrival process are particularly stronger at intermediate time scales of a few minutes. ix) the inter-arrival time pattern at the auction site exhibits sudden unexpected periods of inactivity while the bookstore seems to see long periods of silence which are less abrupt and which could be anticipated.

It is recognized by many that one of the major challenges in carrying out experimental work in E-business is the lack of data. Most companies regard their logs as sensitive information that should not be made public. The methodology presented in this paper can certainly be applied to logs of other E-business sites and constitutes an important step for capacity planning and performance tuning. The type of workload statistics one may find when studying other E-business sites may vary as a function of the types of products and services offered and as a function of the business model implemented by the site.

References

[1] P. Abry, P. Flandrin, M. Taqqu, and D. Veitch. Wavelets for the analysis, estimation and synthesis of scaling data. In *Self-similar Network Traffic and Performance Evaluation.* Wiley, Spring 2000.

[2] P. Abry, P. Gonçalvès, and P. Flandrin. Wavelets, spectrum analysis and $1/f$ processes. In A. Antoniadis and G. Oppenheim, editors, *Lecture Notes in Statistics: Wavelets and Statistics*, volume 103, pages 15–29, 1995.

[3] V. Almeida, M. Crovella, A. Bestavros, and A. Oliveira "Characterizing Reference Locality in the WWW," *Proc. IEEE/ACM International Conference on Parallel and Distributed System (PDIS)*, Dec. 1996.

[4] M. Arlitt and C. Williamson, "Web Server Workload Characterization," *Proc. 1996 SIGMETRICS Conference on Measurement of Computer Systems*, ACM, May 1996.

[5] J. Brown and P. Duguid, *The Social Life of Information*, Harvard Business School Press, 2000.

[6] M. Calzarossa and G. Serazzi, "Workload Characterization: A Survey," *Proc. of the IEEE*, Vol. 81, No. 8, August 1993.

[7] M. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes," *IEEE/ACM Transactions on Networking*, 5(6), pp. 835–846, December 1997.

[8] I. Daubechies, "Ten Lectures on Wavelets," SIAM, New York, NY, 1992.

[9] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, pp. 1–15, 1994.

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

[11] B. Huberman, P. L. T. Pirolli, J. E. Pitkow, and Rajan M. Lukose, "Strong Regularities in World Wide Web Surfing," *Science*, Vol. 280, April 3, 1998.

[12] D. A. Menascé, V. A. F. Almeida, R. Riedi, F. Pelegrinelli, R. Fonseca, and W. Meira Jr., "In Search of Invariants for E-Business Workloads," *Proc. 2000 ACM Conf. in E-commerce*, Minneapolis, MN, October 17-20, 2000.

[13] D. A. Menascé and V. A. F. Almeida, *Scaling for E-Business: technologies, models, performance and capacity planning*, Prentice Hall, Upper Saddle River, NJ, May 2000.

[14] D. A. Menascé, V. A. F. Almeida, R. Fonseca, and M. A. Mendes, "Business-oriented Resource Management Policies for E-Commerce Servers," *Performance Evaluation*, Vol. 42, September 2000, pp. 223–239.

[15] D. A. Menascé, V. Almeida, R. Fonseca, and M. Mendes, "A Methodology for Workload Characterization for E-Commerce Servers," *Proc. 1999 ACM Conference in Electronic Commerce*, Denver, CO, Nov. 3-5, pp. 119-128.

[16] Pitkow, J., Summary of WWW characterizations, *World Wide Web*, No. 2, 1999.

[17] D. Krishnamurthy and J. Rolia, "Predicting the Performance of an E-Commerce Server: Those Mean Percentiles," in *Proc. First Workshop on Internet Server Performance*, ACM SIGMETRICS 98, June 1998.

[18] R. Riedi, M. S. Crouse, V. Ribeiro, and R. G. Baraniuk. "A multifractal wavelet model with application to TCP network traffic," *IEEE Trans. Info. Theory, Special issue on multiscale statistical signal analysis and its applications*, Vol. 45, pp. 992–1018, April 1999.

[19] M. Taqqu, V. Teverovsky, and W. Willinger. "Estimators for long-range dependence: An empirical study," *Fractals*, Vol. 3, pp. 785–798, 1995.

[20] G. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA, 1949.