Dynamics and Evolution of Web Sites: Analysis, Metrics and Design Issues

Ludmila Cherkasova and Magnus Karlsson
Hewlett-Packard Laboratories
1501 Page Mill Road, Palo Alto, CA 94303
E-mail: {cherkasova,karlsson}@hpl.hp.com

Abstract

Our goal is to develop a web server log analysis tool that produces a web site profile and its system resource usage in a way useful to service providers. Understanding the nature of traffic to the web site is crucial in properly designing site support infrastructure, especially for large, busy sites. The main questions we address in this paper are the new access patterns of today's WWW, how to characterize dynamics or evolution of web sites, and how to measure the rate of changes. We propose a set of new metrics to characterize the site dynamics, and we illustrate them with analysis of three different web sites.

1 Introduction

Design, management and support of large web sites can be a challenging task for service providers. Far from all sites are large from the very beginning, and these sites tend to grow at a certain rate and at one point in time some major efforts and decisions have to be made to create a scalable and highly available solution. This might include some caching and load balancing infrastructure. Understanding the nature of web server workloads is crucial to properly designing and provisioning current and future services. Issues of workload analysis, performance modeling, and capacity planning play essential roles during the design stages.

There are several tools freely available, e.g. Analog [13], Webalizer [14], WebTrends [15], that give detailed data analysis useful for business sites to understand customers and their interests. However, these tools lack the information which is of interest to system administrators and service providers; the information which provides insight into the system's resource requirements and traffic access patterns.

We have instead developed a tool called **WebMetrix** that characterizes a web site profile and its system resource usage in both a quantitative and qualitative way. It extracts and reports information that could be used by service providers to evaluate current solutions and to improve and optimize relevant future components. **WebMetrix** performs an analysis which is entirely based on web server access logs, which can be from one or multiple servers in a cluster. The tool is written in Perl for the Common Log Format, which is the most popular default for web server access logs.

In this paper, we discuss only part of the statistics available from our tool. We introduce a set of new metrics and observations which help to characterize web site access patterns and the dynamics and evolution of these over time. We also show how these metrics can make various design choices easier.

A number of studies [2, 3, 4, 5, 6, 10] have analyzed web-servers in order to identify a number of invariants. In Section 3, we revisit these invariants and show new trends in modern web server workloads. We also complement these invariants via characterization of the web server *working set locality* in Section 4. Our results show that typically in current web server workloads the files responsible for 90% of the requests or more account for about 5% or even less of the total working set. This metric can be used to characterize memory requirements of the underlying workloads, and could be used for simple capacity planning of web servers. One example of the use of this information would be for locality-aware load balancing strategies [8, 12].

The main question we address in this paper in Section 5 is how to characterize the dynamics and evolution of web sites. This has previously been studied by Manley *et al.*[10] that considered evolution of web sites over time, but mostly concentrated on web sites growth. A more recent study by Padmanabhan *et al.*[11] examined in more detail the dynamics of a busy news site. While they try to answer some of the questions we pose, they have developed a different set of metrics to answer these questions. Moreover, our study reports its results from business and academic sites instead of a news site.

Most web sites are adding new content and removing some of the old one. However, these changes to the content only partially characterize what we call dynamics or evolution of the site. These content changes contribute to the dynamics of the site only when the new content is accessed. The first natural step is to observe the introduction of new files in the logs, and to analyze the portion of all requests destined for those files. The metric that aims to characterize the site evolution due to new content, is performed by computing the ratio of the accesses targeting these new files over time.

Some research has already been performed to characterize the rate of changes of modified documents: what percentage of documents that get modified and how often this happens. This metric helps to optimize web cache consistency protocols [9], and to decide on how often web search engines (crawlers) have to re-visit pages and re-index them to stay current [7]. Our tool collects this statistics too. However, the main goal of this paper is to introduce more general metrics to capture the evolution and dynamics of the web sites due to overall content evolution and corresponding clients access pattern changes.

Clients' interest in various data is another critical variable defining the dynamics of the site: some topics and documents will with time loose their popularity, and that changes their access pattern. Can we propose a metric characterizing dynamics of the changes for web sites?

Our approach, is to define such a metric by examining the properties and changes of a file set, called the *core*, defined by 90% of the web site accesses. This choice is justified by the very nature of the web site access pattern as 90% of the accesses usually defines the core files of clients current interests. The dynamics of the core reflect changes in the traffic access patterns on the web site. To quantify the dynamics of the site, we observe the core files, and measure the duration they belonged to a core. For visualization purposes, we partition the files in three groups: *stable*, *long-lived* and *short-lived* that reflect the duration a file stays in the core. To reflect the dynamics of access patterns and the clients interests, we compute the percent of the accesses targeting these groups of files. For the sites under study, the stable and long-lived files get up to 98% of all the accesses to the core. We believe that these metrics will be useful to service providers and help them to uncover the dynamics and evolution rate of their web sites.

Dynamics of the site can be taken into account (in addition to load information) when making a decision about different load balancing solutions, caching or content distribution systems. For example, if the site is very dynamic, i.e. a large portion of the daily client requests are accessing new content, news sites being a prime example, then Akamai [1] approach might be a good choice to handle the load. Frequently accessed documents will be replicated closer to clients on Akamai servers, as this will improve user quality of service. However, if the site's traffic pattern shows consistently that clients access a slowly changing subset of documents, then currently existing Internet caches might be a useful solution at no cost for the service provider.

Another set of data reported in Section 6 that **WebMetrix** provides is related to quality of service for web servers. Aborted connections often reflect unsatisfactory level of service, typically due to high response time, however they are not easily recognizable. From web server logs information, we identify the requests which are most likely due to aborted connections. Our results show two different access patterns for aborted connections, which hint whether the network or the server is responsible for the low quality of service. This profiling technique can be useful as a first warning sign for system administrators about poor quality of service on their sites.

2 Data Collection Sites

In our case study, we use three access logs from different servers:

- OpenView (www.openview.hp.com) provides the complete coverage on OpenView solutions from HP. It contains the product descriptions, white papers, demos illustrating the products usage, software packages, business related events, conferences on the topic, etc. The log covers a duration of 5 months, from the end of November, 1999 to the end of April, 2000.
- HP Labs (www.hpl.hp.com) provides information about HP Laboratories, its current projects and research directions, lists current job openings, provides access to an archive of published HP Labs research reports, hosts a collection of personal web pages. The access log was collected during 5 months, from January 1 to May 31, 2000.
- HP (www.hp.com) provides diverse information about HP: business news, major events, detailed coverage of the most software and hardware products, and the press related news. The access log covers a month's duration and was collected on one of the servers in the cluster supporting the HP.com site.

The web-access logs record information about all the requests and responses processed by server. Each line from the access log provides a description on a single request for a document or file. A typical entry contains the following fields:

```
hostname - - [dd/mm/yyyy:hh:mm:ss tz]
request status bytes
```

Each log entry specifies the name of the host machine making the request, the time stamp the request was made, the filename of the requested document and the number of bytes transferred in the reply. The entry also provides the information about the server's response to this request: successful requests (code 200), so called conditional get requests (code 304) and errors which are the rest of the codes. Since the successful responses with code 200 are responsible for all of the files transferred by the server, we will concentrate our analysis only on those responses for the rest of the paper. The three access logs provide information on web servers with different number of requests. OpenView, and HP Labs servers had somewhat comparable number of requests, if normalized per month. HP.com had three orders of magnitude heavier traffic.

3 New Trends

In this section we will revisit some previously identified invariants. For all our sites under study, 90% of the server requests target only 2%-4% of the files instead of the previous 10% [4], showing high locality of references, as previously reported in [11]. The bytes transferred due to these requests vary in much broader range: from 26% percent for HP Labs site to 49% percent for HP.com site. These two facts reflects a new trend in current web server workloads. We speculate that this is due to four main factors. First and second, improved web server side performance and available Internet bandwidth. Third, that current web sites use more graphical content as the ratio of graphical content per html page has risen from 1.66 reported in [4] to 8.6 on the average for our sites in about four years. Finally, that the mean transfer/request size has increased. Arlitt *et al.* [4] reported a mean transfer size between 6 KBytes and 21 KBytes while we observe a mean transfer size between 4 KBytes and 324 KBytes.

4 Working Set Locality

For each access log, we build a site profile by evaluating the following characteristics: 1) *WS* - the *working set* of the site which is the combined size of all the accessed files in bytes during the observed period; 2) *BT* - the number of bytes transferred from the site during the observed period.

It is well known that web server performance greatly depends on efficient RAM usage. A web server works faster when it transfers pages from RAM, and its throughput is higher too. The working set of the site characterizes the site's memory requirements. If the working set fits in RAM, all files will be served from RAM after the first access to it, and this leads to the best server performance. The bytes transferred metric gives an approximation of the load to a server provided by the traffic to the site. These parameters outline a high-level characterization of web sites and their system resource requirements. Both of these characteristics can be used for first step capacity planning of the underlying system. Table 1 shows the working sets and access rates of the sites. From Table 1, high-level, "at-a-glance" site specifics can be observed.

Table 1. Basic site measurements.

	OpenView	HP Labs	HP.com
Working Set Size (MB)	8,553	2,701	7,883
Bytes Transferred (GB)	1,709	208	428
Accessed Files	14,249	44,598	204,336

In order to analyze the density of references against the working set size, we introduce a metric called *working-set locality*. We define this as the percentage of the working set that the most frequently accessed files occupy, that contribute to a specific percentage of the total number of requests.

For all the logs in our study, we observed a high working set locality: 87% to 96% of all requests are to files that constitute only 5% of the total access log working set as can be seen in Figure 1a). HP.com has the highest working set locality across all the points, while HP Labs site shows least working set locality.

Two major factors impact web servers performance: the number of requests the server must process and the amount of corresponding response bytes the server must transfer (from disk versus memory, to the network). We use the *bytes-transferred locality* to characterize the amount of bytes transferred because of the requests to the most frequently accessed files. Figure 1b), shows this locality for our collection of access logs. These metrics normalized with respect to the site working set, make it possible to easily compare different workloads, identify similarity and differences in web server workloads. Both graphs in Figure 1 (complemented with absolute values also provided by **WebMetrix**) could be used for high-level capacity planning (amount of memory, I/O capacity) when choosing web servers support for targeted sites.

5 Web Site Dynamics

Our **WebMetrix** tool attempts to characterize the dynamics of web-sites, in order for this information to be used when making performance critical decisions on design choices for the underlying infrastructure supporting the site.

5.1 Changes in Absolute Statistics over Time

In this section we will show that the volume of web server traffic can be quite predictable for some sites when measured on a large enough time scale. Figure 2 shows the absolute number of requests over time for the three sites under study. In the case of HP.com there is a steady, predictable weekly access pattern. Every weekend has lower traffic as expected, but more interestingly we can also see a consistent, predictable increase in the number of requests over the week days. HP Labs and Open View show less consistent behavior. OpenView has short traffic bursts that

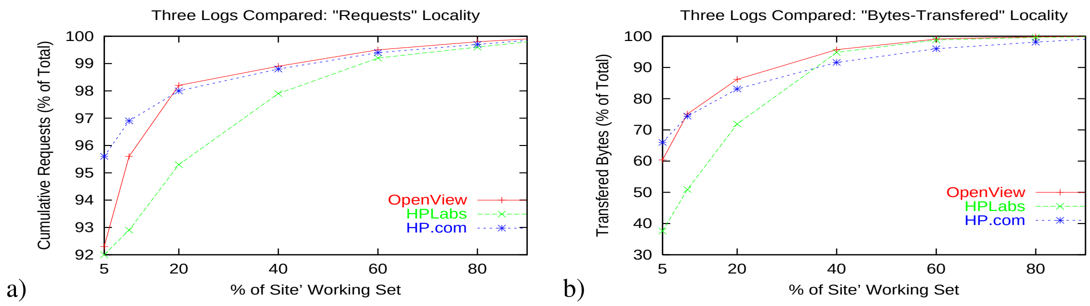

Figure 1. Three traces compared: (a) working-set locality and (b) bytes-transferred locality.

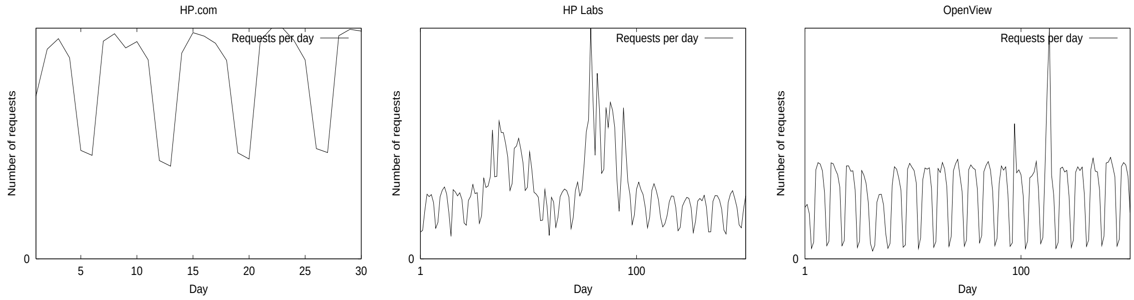

Figure 2. The absolute number of requests over time.

only lasts for a day or two, while HP Labs have traffic bursts that stretches for about a month in duration.

If instead the number of requests were measured per hour or minute this shows a much less predictable behavior. Table 2 shows the minimum and maximum number of request relative to the mean value of requests when this is measured using four different time intervals. A value of e.g. 10 under the max and 1 hour column means that the hour with the highest number of requests had 10 times as many requests as the average hour. The difference between the minimum minute and the maximum minute can be as high as 25586 times for HP.com when measured on a per minute basis. However, if measured daily or even hourly, all the sites show much less variation.

Table 2. The relative minimum and maximum number of requests to the mean number of requests, measured over 4 time intervals.

Site	1 minute		10 minutes		1 hour		1 day	
	min	max	min	max	min	max	min	max
OV	.024	11.2	.003	6.8	.015	6.3	.12	3.6
HPL	.011	9.5	.001	1.7	.022	6.3	.32	3.3
HP	.0001	3.2	.53	1.6	.55	1.5	.51	1.3

5.2 The Dynamics of New Files

One of many important decisions that a web-service provider of a busy site has to make is if the documents on the server should be replicated in several places over the world in order to off-load the servers in the original location. The proxy caches found around many clients sites are one approach that off-loads the original servers. While this approach is at no cost for the service provider, it does not work efficiently for new data that is immediately popular as it takes some time for the new files to propagate to Internet proxy caches. A novel approach by Akamai [1] allows to replicate frequently accessed site content across many servers around the world soon after it is created and accessed. However, it is not a free solution for web-service providers. Characteristics of the access patterns to new documents makes this choice easier.

Figure 3 shows two curves for each site: the percentage of new files introduced each day relative to the number of files accessed that day, and the percentage of requests to these files on the introductory day relative to the total number of request that day. A file is considered new if the file name has not been encountered before during the measurement interval. There is no statistics for the first week as it is used as a warm-up period. These diagrams show that for these sites even if the relative amount of new files introduced on a certain day can be high, especially for HP Labs, the percentage of the requests due to these new files is not high on the introduction day.

To understand the traffic contribution of new files over time, Figure 4 shows the percentage of requests due to all new files as a function of time. It also shows the percentage of new files accessed each day. Figure 4 shows how much of the old content that has been replaced by new content from a request point of view.

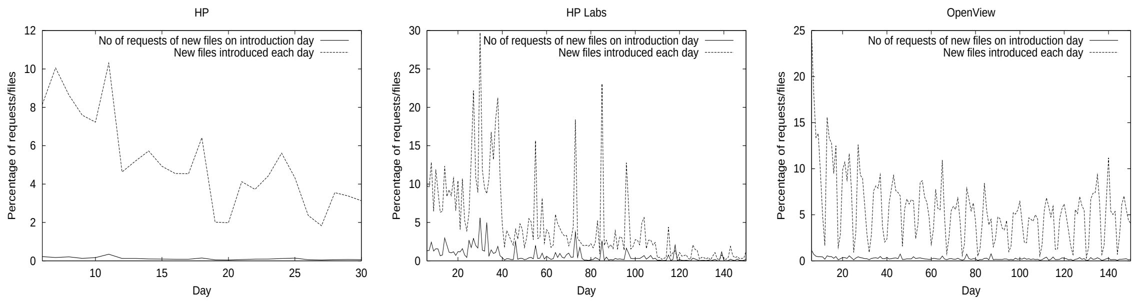

Figure 3. The percentage of new files introduced each day relative to the number of files accessed that day, and the percentage of requests to the new files on the introduction day.

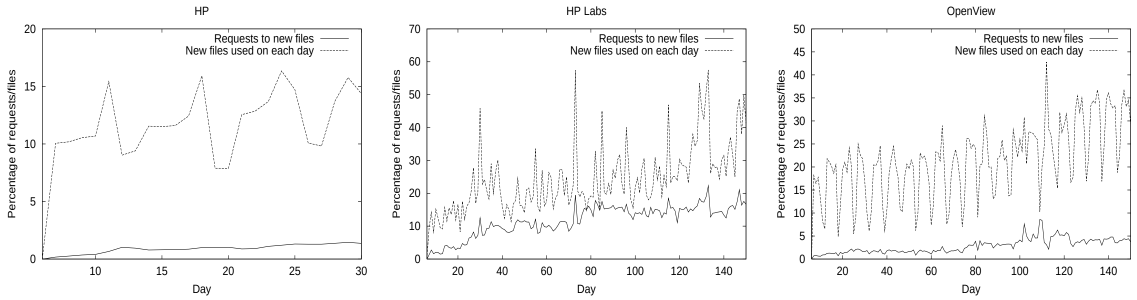

Figure 4. The percentage of files used that are new as a function of time, and the percentage of requests that are to these new files.

For HP Labs up to 50% of the files used and around 19% of the requests are due to new content that has replaced the previous content over the five months the measurements are for. Similarly to HP Labs, OpenView users access a lot of new content but the frequencies of these accesses are low. With OpenView the users mostly access the old content even though new content is added. For HP.com we only have one month of data, but by the end of this month only around 2% of the requests are to new content. Padmanabhan *et al.*showed in [11] that 10% of the accessed documents during one week were actually created during that time period. That this number is higher than our numbers is intuitive as they studied a news site and we study more static academic and business sites.

5.3 The Core

Web server workloads exhibit high locality of references. In [11] it was observed that 90% of the server requests come to only 2%-4% of the files. Thus, this small set of files has the strong impact on the web server performance.

The most performance critical part of the working set is what we call the *core*. We define the core as the set of most frequently accessed files that makes up for 90% of the requests. Going back to Figure 1, 90% of the requests for our sites constitute less than 5% of their working sets. From a performance point of view it is these core files we should concentrate on to obtain good performance as most accesses are to them.

One potential performance problem can occur if the core changes frequently. If this is the case the core files might not be in memory but instead on disk which degrades performance. Figure 5 depicts how the core changes over days and weeks. More specifically, it shows the relative number of requests that are to previous core files each day. For HP Labs and OpenView the core is measured over a week instead of a day as there is not enough traffic to get statistically valid data for the core files with the lowest access frequency if it was measured on a daily basis. From the figure we can see that less than 6% of the requests are due to accesses to new core files. Thus the most performance critical working set is quite steady over time even though the rest of the site changes more as seen in Figures 3 and 4.

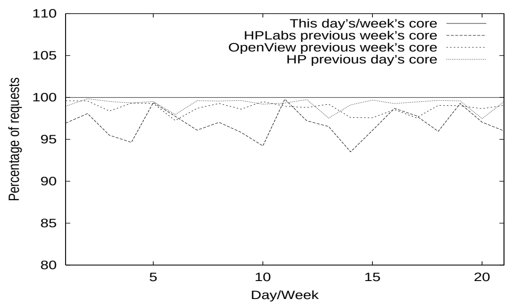

Figure 5. Relative number of requests that are to previous day's core files each day.

What the previous figure does not show is how long the files stay in the core and what the rate of changes is. It also does not distinguish whether the changes are due to the certain files which leave and enter the core frequently or the core slowly goes through the evolution process and the new files replace the old core files. To measure the dynamics of the core, we accumulate statistics about all the files from the core: for how many days or weeks they did belong to the core during the observed period of time.

Figure 6 shows how long the files stay in the core relative to the total amount of files that were in the core during the measurement period. Each pie chart is divided into three categories: *stable* files that stay in the core during at least 90% of the time period, *short-lived* files that stay 10% or less, and *long-lived* files that stay between 10% and 90% of the time period in the core.

These pie-charts reflect the dynamics of the file set in the core, or rather the stability for the three sites studied here: HP Labs has the most stable core file set: 77% of the files have been found in the core during most of the measurement period, and for Open View site stable files constitute 48% of the core. According to this measurement of stability, Open-View's core file set is the most dynamic one: the stable files are 30% of core and the percentage of short-lived files are 26%.

The above metric describes the dynamics of the core file set. To understand the site dynamics and the access patterns characterization over time it is not enough. We also need to understand how the client accesses are distributed across the files in the core accordingly to this classification.

If we look at the number of requests that are due to stable, long-lived, and short-lived files, respectively, the picture is quite different as seen in Figure 7. The number of requests due to the stable files in the core outnumber the other categories by far. The percentage of requests due to short-lived files in the core is now small, between 0% and 2%. It was somewhat unexpected that for the sites under study, the stable and long-lived files get at least 98% of all the accesses to the core.

The introduced new metrics demonstrate stability of the access patterns as well as the content stability for all the three sites under study. It suggests that network caches might be an efficient way to offset a sizeable portion of web server traffic. Indeed, it is clearly seen from the percentage of 304 type requests found in the logs. The status code of 304 relates to the documents cached somewhere in the Internet (or by proxy caches) which send a 'request-validation' whether the document was modified since the last requested time. No data bytes need to be transferred in this case. For the logs we considered, the percentage of *conditional-get* requests (code 304) varied from 19% for HP.com, to 33% for Open View which shows significantly increased efficiency of caching in the Internet nowadays compared with the 4%-13% reported in [4].

By using **WebMetrix**, service providers will gain more information and better understanding of the access patterns and content dynamics of their sites. This will help them to justify the decision whether to subscribe to the Akamai service or rely on the infrastructure of network caches.

Our case study is limited by the access logs from three sites, two of them being business sites and one a research organization. Padmanabhan *et al.*[11] studied the dynamics of a busy news site and showed that this was a highly dynamic site compared to our sites.

Our proposed metrics give insight into site dynamics, and how traffic access patterns can be characterized over time due to change of the content or its popularity on a site. Since the core and its traffic patterns are performance critical, we believe that these observations are important to consider when making design choices for the site.

6 Aborted Connections: Quality of Service

User perceived *quality of service (QoS)* is also an important metric to consider for a web service provider. One way to measure the QoS of a web server from a performance point of view is to measure the amount of aborted connections, the logic behind this being that if the site is not fast enough a user will get impatient and hit the stop button, thus aborting the connection. The response time that the user reacts to has two components: the network transfer time and the web server processing time. Thus it is important to be able to distinguish between the two in a tool for service providers.

Unfortunately most access logs do not give any explicit statistics about aborted connections. In our tool, to calculate the number of aborted connection, we count the number of times the file size in the log was less than the real size of that file. As we do not know the real size of the file we have to estimate this size in what we call the *perceived size*. The perceived size is set whenever a file has the same size two references in a row. We define a connection as aborted if the following holds: The size of a file in the log is less than the perceived size of that file, there is a perceived size set, and the file is not dynamically generated.

OpenView and HP.com has few connections, thus in this Section we will focus on HP Labs and a site called `ESN-Europe`. ESN stands for "Electronic Solutions Now" and is a site for some of HP's corporate customers.

For ESN-Europe there is a strong correlation between the number of aborted connections per day and the number of requests the web server receives per day as seen in Figure 8a. Since increased number of requests leads to higher server load, there is a strong correlation between number of aborted connections and server load. This suggests that, server performance and quality of service significantly degrade with the load and more clients chose to abort their

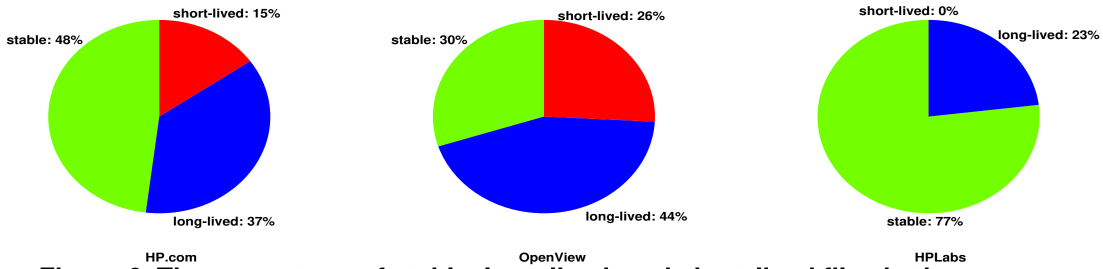

Figure 6. The percentage of stable, long-lived, and short-lived files in the core.

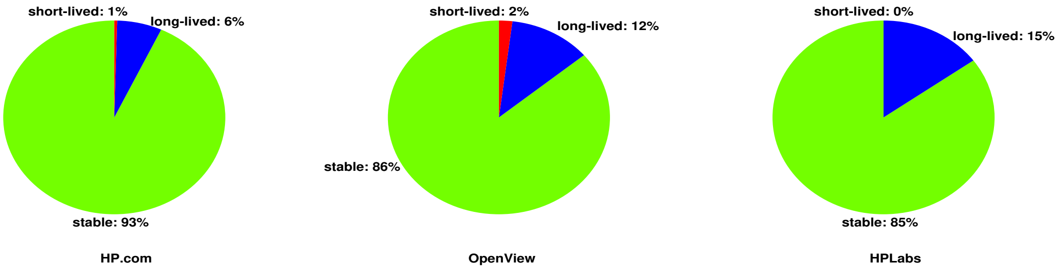

Figure 7. The percentage of clients requests that are targeting stable, long-lived, and short-lived files.

requests.

We verified this observation by comparing load data gathered from the server over the time period in question, we found that the server was heavily loaded when the number of requests were high. We can also see in Figure 8b that the larger files tend to be aborted. This figure shows the cumulative percentage of requests that has a file size smaller than a specific file size. The result is intuitive as we would expect small files to load relatively quickly even though the site is under heavy load. For ESN-Europe the server is clearly to blame and should be dealt with accordingly, and it actually was shortly after these statistics were gathered.

The behavior of HP Labs is quite different. Looking at Figure 8c we can see that there is nearly no correlation between the number of requests per day and the number of aborted connections. There is more or less a constant number of aborted connections over the days observed. The explanation to this observed pattern is due to network or client side performance problem, and is not related to a server performance.

We verified this observation by translating the IP addresses of clients that aborted connections we found that the large majority of these clients could be divided into two categories. First, users using a dial-up service to connect to the Internet with a modem, and second, users surfing the Internet from a number of developing countries with generally slow Internet connections even through LAN accesses. As seen in Figure 8d the file sizes of the aborted connections are generally larger than for all the completed connections. However comparing the request-size distribution of aborted connections of ESN-Europe with the one from HP Labs, we see that there are relatively more larger files aborted for ESN-Europe than for HP Labs. ESN-Europe has very few modem users as it is a site mainly for corporations. We speculate that modem users tend to hit interesting links as soon as they appear instead of waiting for all the pictures on the page to load, thus even small files get aborted.

We have shown how to differentiate between aborted connections due to network delays, and those due to web-server delays. This could be useful as a early warning sign that the QoS on a site is getting poor.

7 Conclusions

Web server access logs are invaluable sources of information not only to extract business related information, but also for understanding traffic access patterns and system resource requirements of different web sites. Our tool **WebMetrix** is specially designed for system administrators and service providers to understand the nature of traffic to their web sites. Issues of workload analysis, performance modeling, and capacity planning are crucial to properly designing the site support infrastructure, especially for large, busy web sites.

In this paper, we analyze some of the new access patterns and trends specific to new, mature web sites. We observe that the bytes transferred due to 90% of the server requests is much lower: from 26% to 49% percent for the sites under study. Another related observations are that the ratio of graphical content per html page has risen from 1.66 to 8.6 on the average for our sites, and the mean transfer size has increased. This reflects new trends in current web server workloads.

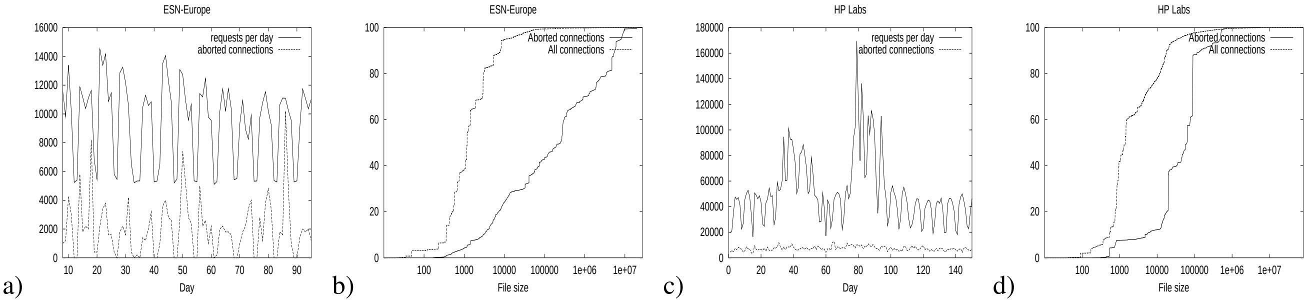

Figure 8. a,c) Aborted connections over time scaled up 100 times for easier reference. b,d) Request size distribution for aborted connections and for all the requests.

The main contribution of the paper is a set of new, novel metrics to characterize the dynamics and evolution of the web site and to measure the rate of changes. First we analyze the properties and changes of a file set, called the core, defined by 90% of the web site accesses. We partitioned the core files into three groups: stable, long-lived, and short-lived files. We measured these files distribution in the core, as well as the percent of the accesses targeting each group of these files. For the sites under study, the stable and long-lived files get up to 98% of all the accesses to the core. Our study was limited by the three sites we had access logs from. However, we believe that these metrics will be generally useful to service providers to understand the dynamics of their web sites. Capturing the dynamics and evolution of the site might be useful to the site designers as well: whether the rate of changes in the client's access patterns is what the designer had expected and desired, or not?

We also propose a simple metric based on aborted connections and their pattern to characterize the quality of service for web servers. To sum it up, **WebMetrix**'s analysis helps to observe specific site access patterns in order to predict the changes and efficiently provision for them well in time.

8 Acknowledgments

Both the tool and the study would not have been possible without the help provided by Guy Mathews, Dean Baender, Wai Lam, Len Weisberg, and Mike Rodriquez.

References

[1] Akamai, Cambridge, MA, USA. *FreeFlow: How it Works*. http://www.akamai.com.

[2] V. Almeida, A. Bestavros, M. Crovella, and A. Oliviera. Characterizing reference locality in the www. In *Proceedings of the 4th Int. Conf. Parallel and Distributed Information Systems (PFIS)*, pages 92–106, 21, 1996.

[3] M. Arlitt and T. Jin. Workload characterization of the 1998 world cup web site. Technical Report HPL-1999-35R1, HP Laboratories, 1998.

[4] M. Arlitt and C. Williamson. Web server workload characterization: the search for invariants. In *Proceedings of the ACM SIGMETRICS Conference*, pages 126–137, May 1996.

[5] P. Barford, A. Bestavros, A. Bradley, and M. Crovella. Changes in web client access patterns: Characteristics and caching implications. Technical Report TR-1998-023, Department of Computer Science, Boston Univeristy, 1998.

[6] A. Bestavros, A. Carter, M. Crovella, C. Cunha, A. Heddaya, and S. Mirdad. Application-level document caching in the internet. Technical Report 1995-002, Department of Computer Science, Boston Univeristy, 1995.

[7] B. Brewington and G. Cubenko. How dynamic is the web? In *Proceedings of the 9th International World Wide Web Conference*, pages 257–276, May 2000.

[8] L. Cherkasova. Flex: Load balancing and management strategy for scalable web hosting service. In *Proceedings of the Fifth International Symposium on Computers and Communications (ISCC'00)*, pages 8–13, July 2000.

[9] F. Douglis and A. Feldmann. Rate of change and other metrics: a live study of the world wide web. In *USENIX Symposium on Internet Technologies and Systems*, pages 147–158, December 1997.

[10] S. Manley and M. Seltzer. Web facts and fantasy. In *USENIX Symposium on Internet Technologies and Systems*, pages 125–134, December 1997.

[11] V. Padmanabhan and L. Qiu. The content and access dynamics of a busy web site: findings and implicatins. In *Proceedings of SIGCOMM*, pages 111–123, August 2000.

[12] V. Pai, M. Aron, M. Svendsen, P. Drushel, W. Zwaenepoel, and E. Nahum. Locality-aware request distribution in cluster-based network servers. In *Proceedings of the 8th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS VIII)*, pages 205–216, October 1998.

[13] Statistical Laboratory, Univeristy of Cambridge, UK. *Analog*. http://www.statslab.cam.ac.uk/ sret1/analog.

[14] *Webalizer*. http://www.mrunix.net/webalizer/.

[15] WebTrends Corporation. *WebTrends*. http://webtrends.com/.