Workload Characterization of a Large Systems Conference Web Server

Aniket Mahanti Carey Williamson Leanne Wu
University of Calgary, Canada

Abstract

This paper presents a workload characterization study of the WWW2007 conference Web site. We use both server-side and client-side measurement to collect data for our analysis. The collected datasets contain approximately 10 *million requests from about* 130, 000 *unique visitors generating* 215 *GB of traffic volume during a 1-year period. Our analysis focuses on usage behaviour, client errors, client network properties, bandwidth, and robot activity of the site. In addition to the workload characterization study itself, our work also offers a comparative discussion of different approaches to Web data analysis, and some insights for future organizers of large systems conferences with a significant Web presence.*

1 Introduction

Organizers of large systems conferences often confront numerous challenges [17]. Apart from handling the high load of paper submissions and the logistics of hosting the actual conference, organizers also have the job of creating and maintaining a conference Web site. Hosting a conference Web site stresses the existing infrastructure of a host institution because of additional overall traffic, different workload characteristics, and specific periods of extreme load. It also raises the possibility of malicious network attacks.

Modern Web sites have evolved significantly over the last decade [14]. In the past, a typical Web site was composed of pages that contained only text and HTML. Today, most Web sites are composed of multiple elements such as images, videos, external JavaScript, etc.

The evolution of the Web has also transformed Web analytics: the measurement and analysis of Web data. An indispensable tool for researchers and practitioners alike, Web analytics traditionally involves using server logs to assess the usage and performance of a Web site, such as the popularity based on the number of hits. Today, however, Web analytics often requires numerous metrics to understand usage patterns of a Web site, since a single pageview on a modern Web site generates multiple hits.

In this paper, we present a workload characterization study of the WWW2007 conference Web site [2], using data collected from both server-side and client-side. Our datasets were collected over a 1-year period in the form of server logs (server-side) and Google Analytics (client-side) reports. The datasets contain approximately 10 million requests from about 130, 000 unique visitors generating 215 GB of traffic volume. We use server-side and client-side measurement to characterize the usage behaviour of the Web site visitors. Our study comments on the accuracy of each measurement technique and reports upon interesting emerging trends in Web site usage.

Our paper makes contributions on three fronts. First, it provides another reference point for research on Web workload characterization, for a long-running Web server. Second, it provides side-by-side comparison of two different approaches to Web data analysis. Third, our experience from hosting the WWW2007 Web site provides some valuable insights for future organizers of large systems conferences with a significant Web presence.

The rest of the paper is organized as follows. Section 2 provides an overview of the WWW2007 conference Web site, providing context for our work. Section 3 describes our data collection methodology. Section 4 presents the results of our characterization study, while Section 5 discusses related work. Finally, Section 6 concludes the paper.

2 The WWW2007 Web Site

The World Wide Web (WWW) conference is a premier forum for academic and industry leaders to discuss and debate the latest ideas about the Web, its evolution, and its impact on society. The conference series is coordinated by the International World Wide Web Conferences Steering Committee (IW3C2). The WWW conference itself is an annual event, with a different location each year, and a different host institution as the local organizer.

The WWW2007 conference was organized by the University of Calgary. The conference took place in Banff, Alberta, Canada from May 8-12, 2007. The five-day event included 8 workshops, 12 tutorials, presentation of 111 refereed papers (15% acceptance rate) and 119 posters, 7 panel sessions, 4 plenary speakers, and 12 invited industry speak-

978-0-7695-3649-1/09 $25.00 © 2009 IEEE
DOI 10.1109/CNSR.2009.19

ers. The WWW2007 conference had 982 attendees from 40 different countries. Many of the attendees brought their laptops, and high-speed Wireless Internet connectivity was provided throughout the conference venue by the local organizers.

The WWW2007 conference Web site was launched on May 26, 2006, during the closing ceremony of the WWW2006 conference in Edinburgh. The Web server was located at the University of Calgary, and was running Apache 2.2.8. The initial Web site contained basic information about the WWW2007 conference, including the venue, dates, committee members, call for papers, and key deadlines related to the technical program. The basic design of the site (i.e., layout, colour scheme, navigation, menu items) was planned from the outset, with PHP (Hypertext Preprocessor) as the main language for the site content.

As the date of the conference drew closer, the content on the Web site grew in size. For example, travel and hotel information was added in September 2006, paper submission instructions in November 2006, registration information in December 2006, preliminary program in January 2007, and the full conference program in March 2007. Online proceedings were added in April 2007, and daily conference photo galleries in May 2007. Monthly newsletters were also issued for conference publicity, and sponsorship logos were added as they were received.

Table 1 summarizes the final content of the Web site. Approximately 6,000 files[1] are hosted on the server, consuming just over 1 GB of storage space.

Table 1. WWW2007 Web Site Content

File Type	Number	Avg Size	Total Size
PNG	2,727	4.6 KB	12.6 MB
JPG	905	370 KB	337 MB
GIF	655	4.5 KB	2.9 MB
PHP	533	3.7 KB	1.9 MB
PDF	352	415 KB	145 MB
Text	250	128 KB	32 MB
WMZ	133	609 B	81 KB
HTML	121	60 KB	7.2 MB
Directory	112	4.8 KB	0.5 MB
CSS	91	3.7 KB	0.3 MB
Perl	42	13 KB	0.5 MB
EMZ	33	460 KB	15 MB
PCZ	13	175 KB	2.3 MB
Artwork	10	39 MB	398 MB
DOC	4	234 KB	0.9 MB
AI	3	38 MB	116 MB
RAR	2	10.5 MB	21 MB
ZIP	1	440 KB	0.4 MB
Other	80	132 KB	10.7 MB
Total	6,062	182 KB	1.1 GB

3 Methodology

3.1 Web Data Collection

There are two well-known methods for data collection in Web analytics. The first method involves Web server log analysis. A Web server records all of its transactions and saves them to a log file. By analyzing these logs, we can understand the usage pattern of a Web site. This technique is known as server-side data collection. The second method, *page tagging*, uses the visitor's Web browser to collect data. This method requires placing a snippet of JavaScript code in every page of the Web site. Each time a page is requested, the embedded JavaScript is run, which in turn collects data that are sent to a remote server. Cookies are used to track user activity such as visit duration and return visits. This technique is known as client-side data collection.

In this study, we employ both server-side and client-side data collection. Our server logs, recorded in the Common Access Log format, were archived daily. Each line in the log files contains the IP address of the visitor, the date and time, URL, the status code, and the size of the response returned to the visitor. We used the Google Analytics [1] service for client-side data collection. This free service provides a multitude of customizable reports on visitor activity that can be accessed through a secure Web page. The reports provide detailed statistics about browsing patterns, geographic location, traffic sources, site referral, and other characteristics.

We separately analyze both the server logs and Google Analytics data. This approach provides a better understanding of the Web site usage when compared to using either one of the data collection techniques in isolation. Furthermore, this process will help us quantitatively assess the advantages and disadvantages of each approach.

Hybrid solutions such as cookie-fortified logs and server plug-ins appear to combine best of both server-side and client-side measurement. Furthermore, enabling better logging techniques (e.g., recording response times) could help in usability analysis of the Web site.

3.2 Terminology

The following terminology is used in this paper:

Hit - a resource request for a file from the Web server, as recorded in the server access log.

Pageview - a resource request for a file that is a Web page (e.g., `.php` or `.html` files).

Visitor - a unique client (IP) generating a hit or pageview.

Visit - a series of resource requests from a unique visitor that are temporally clustered. After 30 minutes of inactivity, a pageview by the same visitor is counted as a new visit. Visits are sometimes referred to as sessions.

Visit duration - the duration of a visit (i.e., amount of time a visitor spends browsing the site during a visit). Visit duration is also known as session duration.

Page depth - the number of unique pageviews during a visit.

[1] About half of these files (i.e., PNG, WMZ, EMZ, PCZ) are associated with the XHTML versions of several papers on the Web site, while several hundred more are conference photos (JPG).

Table 2. Summary of Data Sets

Characteristic	Server Logs	Google Analytics
Total Unique Visitors	129, 185	80, 554
One-time Visitors	99, 608(77%)	56% new visits
Total Visits	431, 698	143, 505
Avg. Visits per Day	1, 180	392
Avg. Visits per Visitor	3.34	1.78
Avg. Visit Duration (min:sec)	3:48	3:15
Total Pageviews	1, 578, 661	541, 639
Unique Pageviews	975, 895	391, 465
Avg. Pageviews per Visit	3.18	3.77

Traffic volume - the number of bytes transferred by the Web server. For example, traffic volume per day is the total bytes transferred to all visitors during a 24−hour period.

3.3 Trace Data Overview

Table 2 summarizes the two data sets used in this study. The data consists of access logs and Google Analytics reports collected between May 22, 2006 and May 22, 2007, a duration of 1 year. The server logs contained approximately 10 million hits. Approximately 215 GB of traffic volume was transferred during this period. On average the Web site received over 27, 000 hits per day and over 600 MB of data was transferred daily. Table 2 also shows the visit and visitor activity[2].

3.4 Differences between Server-side and Client-side Data Collection

Table 2 shows that there are many differences between client-side and server-side data collection. In general, Google Analytics seems to underestimate several of the traffic characteristics, sometimes by a factor of three or more.

There are several reasons for these observational differences. In our study, one reason is that the Google Analytics code was disabled for 5 days in early April while the conference proceedings were being produced (so that the Javascript code would not be included on the CD-ROM). The Web site was crawled multiple times during this period, generating many entries in the Web server access log.

Another reason for the discrepancy lies in the method employed to collect the data in each case. Google Analytics tracks pages (e.g., PHP, HTML, ASP), but does not record hits to the individual resources (e.g., images, documents) that are linked or embedded in the pages. In contrast, server logs record hits to every file on the server, providing a more complete record of Web resource access patterns.

A third reason is that Google Analytics does not record page errors, since JavaScript is only executed when a page load is successful. In contrast, server logs provide details about server and client errors. This is particularly helpful to fix broken links, or to detect server attacks. Furthermore, search engine spiders and bots do not execute JavaScript, and hence all such visits are ignored by Google Analytics. Thus, the page tagging method is better for measuring visits from humans rather than all site traffic.

In addition to the foregoing key differences, there are strengths and weaknesses to each data collection approach:

- Google Analytics cannot measure traffic volume of a Web site. The JavaScript tracker code runs under the scope of the Document Object Model tree and does not have access to size information of the objects embedded in a page. Furthermore, as mentioned earlier, non-embedded files such as Portable Document Format (PDF) files are difficult to track through page tagging. Server logs on the other hand record the size of each object provided by the Web server. When measuring bandwidth usage of a Web site, server logs are the only option.
- Google Analytics and other page tagging services track visitors and their visits by placing cookies based on session IDs in the visitors' browsers. This method allows for easy identification of returning visitors. Additionally, tracking the page traversal of a visitor (i.e., the series of pages visited) is easier using cookies. Page tagging requires the cooperation of the browser for proper data collection. Thus, if JavaScript is blocked or the user deletes the cookies, it could lead to inaccurate visit and visitor count. It is a non-trivial task to accurately measure visitors and visits using server log analysis. This task is further complicated due to the presence of Web proxies and visitors with dynamic IP addresses. Typically, in server log analysis, a unique visitor is identified by a unique IP address in the log.
- Visitor count using server log analysis can be affected by Web caching. Furthermore, it is hard to distinguish between a human visitor or a robot unless additional information is collected such as user agent. Page tagging can also provide information about the network properties, operating system, browser capabilities, traffic sources, site referrals, and search engine keywords of a visit that are not readily available from server logs.

We used the server logs for primary workload characterization, and augmented the results with reports from Google Analytics (specifically for characteristics that cannot be obtained from the server logs). Server logs have been long relied upon for Web workload characterization. Server logs are easily available and because they are saved in plain text format they are easy to process. Client-side data collection does not provide this feature; the raw data is not available. Client-side data collection is a convenient approach for people who want a third-party to collect data, analyze it, and manage the reports. Although the summary statistics for

[2] Note that the older Google Analytics tracker code that was embedded in the conference Web site did not count one-time visitors. Rather, it categorized visits as new or returning. The current Google Analytics code properly counts one-time and returning visitors.

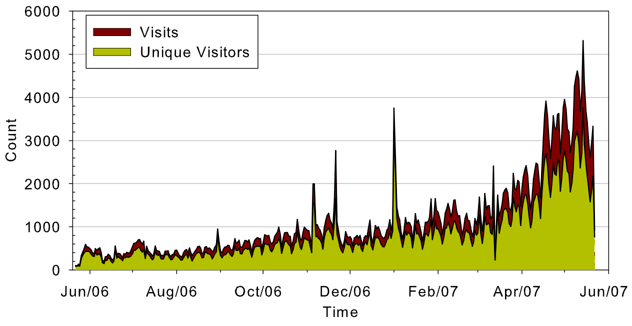


Figure 1. Visits and Visitors per Day

various characteristics differ in our data sets, we found that the trends (such as visits per day) are similar across both data sets. We believe that it is hard to get exact figures from either data source, but either one is suitable to understand the overall browsing patterns for a Web site.

4 Workload Characterization

We now provide multi-layered workload analysis of the WWW2007 conference Web site data. Section 4.1 discusses the traffic profile. Section 4.2 provides an insight into visitor activity on the site: number of visits, frequency of visits, visit duration, page depth, and errors. Section 4.3 analyzes traffic volume trends, network connection speeds, and geographic location of visitors. Section 4.4 discusses the various traffic sources for the site. Finally, Section 4.5 discusses robot activity on the site.

4.1 Traffic Profile and Trends

We begin our analysis of the WWW2007 conference Web site with a high-level look at the traffic profile. For this purpose we use the server logs.

1) Traffic Profile: Figure 1 shows the total number of visits and unique visitors per day for the 1-year trace period. The traffic volume was low for the first few months, until the first conference newsletter was sent on August 28. On this date the number of daily visits nearly doubled, indicating that the publicity attracted some attention from the target audience. The next spike is observed on October 25, which can be attributed to the third newsletter and the imminent paper submission deadline.

The most daily visits during the 2006 calendar year were observed on November 20 (paper submission deadline). The paper submission mechanism was being handled by a third-party conference management system and the conference site had a link to it. The visits count reached a peak, unexpectedly, between Dec 31, 2006 and Jan 2, 2007. (We explain the reason for this spike in Section 4.5.)

Traffic to the site slowly increased in early 2007, and continued to build until the conference in May. There are numerous peaks during the 5-month period, with most related to conference activities. Peaks are observed on January 30 (paper notifications), February 12 (poster submission), March 5 (early registration), March 12 (poster notification), and April 2 (normal registration). Between April 5 and 12 the Web site was overhauled, which included addition of the online proceedings. The entire site was crawled by a robot to fix any broken links. The site had its highest usage during the conference itself.

Table 3. File Type Distribution

File Type	Hits (%)	Traffic Vol. (%)
Image	81.8	63.1
Download	2.4	25.2
Page	13.4	9.1
Audio/Video	0.0	0.2
Other	2.4	2.4
Total	100	100

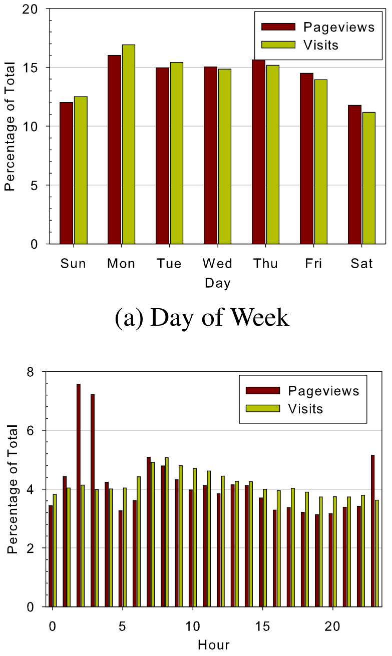
Table 4. Server Response Codes

Code	Hits (%)	Traffic Vol. (%)
200 (OK)	78.8	90.1
304 (Not Modified)	16.0	0.0
404 (Not Found)	3.2	0.0
206 (Partial Content)	1.9	9.9
301 (Moved Permanently)	0.1	0.0
302 (Found)	0.0	0.0
Total	100	100

2) File Types: Table 3 provides a breakdown of Web resources accessed, based on file type. Web pages and images account for approximately 75% of the total data transferred by the Web server. This is not surprising, since modern Web sites (including WWW2007) contain a lot of graphics, including photos, banners, logos, maps, and menus [4]. Caching can be helpful in such scenarios since static objects (i.e., images) are requested more often than other objects. About 25% of the traffic volume is attributable to visitors downloading `PDF` documents from the Web site.

3) Response Codes: Table 4 provides a breakdown of the server response codes. A majority (79%) of requests for objects were successful (status code 200). The successful requests accounted for 90% of the content transferred to the visitors, with most of the remaining 10% from partial transfers (status code 206). The second most common status code noticed was 304, indicating conditional GET requests (16% of the hits), wherein the Web browser validates its cached copy of a resource with the server.

Approximately 3% of the requests result in client errors. Our analysis found approximately 60% of these errors were related to two files, namely `robots.txt` and `favicon.ico`. `Robots.txt` is a default file read by a "polite" robot (such as a search engine spider), to know which directories or files are off limits. The WWW2007 conference site did not contain such a file. The `favicon.ico` file contains an icon that the Web browser



(a) Day of Week

(b) Hour of Day

Figure 2. Percentage of Total Visits or Pageviews

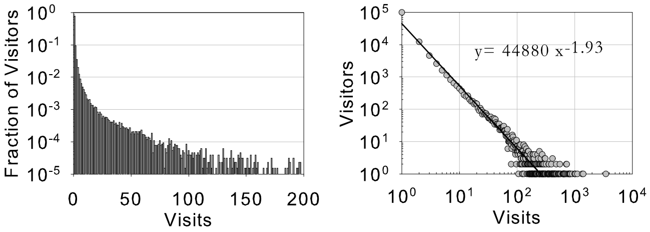
can display in the address bar and bookmark file. While the WWW2007 Web site has such an icon now, this icon was not added until April 14, 2007 (a month before the conference).

4.2 Visitor Trends

Our next analysis focuses on visitor trends. We are interested in understanding how the conference Web site was accessed on a daily and hourly basis. We also want to know how frequently visitors visited the site, the amount of time they spent browsing the site, and the number of pages viewed per visit. We use the server logs to answer these questions.

Figure 2 shows visitor activity based on day of the week and hour of the day. Figure 2(a) shows the weekly cycle typical to any workplace. We find that on average Monday is the busiest day of the week, while activity is almost constant over the rest of the work week. Visit counts tapered as the weekend approached. With most visitors to the Web site being from academia or industry, it is easy to comprehend this pattern.

Figure 2(b) shows the hourly usage pattern of the Web site. We do not observe a strong diurnal pattern. The Web site had a global reach with a good proportion of visitors who were located outside North America. The normal work hours account for almost 40% of the total visits.



(a) Truncated pdf

(b) Log Log Distribution

Figure 3. Visits per Visitor

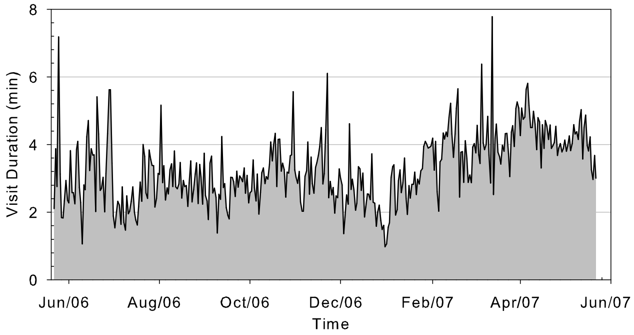
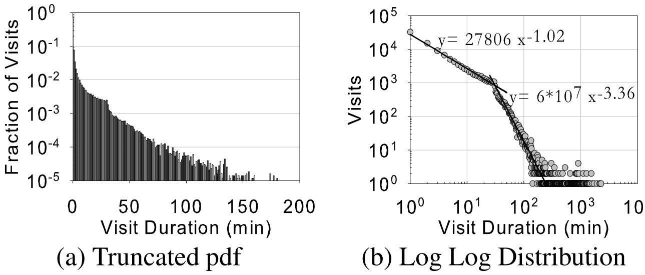


Figure 4. Average Visit Duration per Day

Surprisingly, there is a high percentage of page views between 2 AM and 4 AM, accounting for approximately 14% of the total. The proportion of visits for this time period is not as high as that for pageviews. Upon deeper inspection, we found that approximately 35% of the pageviews came from the University of Massachusetts, where one of the technical program chairs maintained a mirror site as a backup. Also, approximately 75% of these pageviews happened during March/April 2007, the period when the conference site was being updated. Furthermore, approximately 50% of the pageviews came from Asian and European countries. Because of the time difference between Calgary and Europe/Asia (6 hours or more), part of their work hours coincide with this time period.

2) Frequency of Visits: Figure 3 shows the frequency of visits to the Web site. Figure 3 shows a truncated probability density function (pdf) of visits per visitor. Over 75% of the visits were one-time visits. The return visits are dominated by visitors who were affiliated with the organization of the conference. Figure 3(b) shows the long-tailed nature of the distribution. A power-law relationship between visitors and number of visits is also observable. The exponent for the power-law function is 1.93, with $R^2 = 0.96$. see lecture

3) Visit Duration: We are interested in knowing how much time visitors spent browsing the Web site. We wanted to know how average visit duration varied over time and the



(a) Truncated pdf (b) Log Log Distribution

Figure 5. Visit Duration per Visit

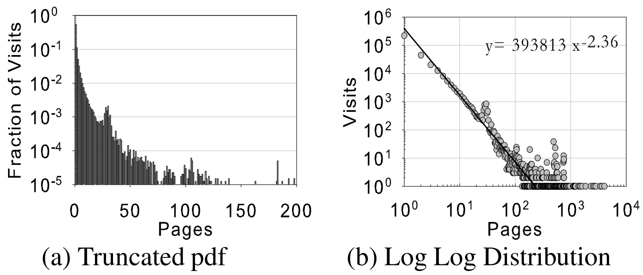
relationship between visit duration and number of visits.

Figure 4 shows the average visit duration of all visits per day for the trace period. The average visit duration varied between 2-4 minutes, except for the last three months leading up to the conference when the average visit duration varied between 4 minutes and 6 minutes. The spike on May 25, 2006 is due to some of the organizing members testing the newly launched conference site. Most of the spikes can be attributed to such activities by the organizing committee members. However, there are a few instances where spikes are caused due to other visitors. A spike on October 17, 2006 was caused due to someone browsing the site at Xerox Research with a total visit duration of 11 hours (perhaps due to the person not closing the browser). Other peaks in the graphs coincide with important dates of the conference such as submission deadline, etc.

Figure 5 shows the frequency of visits for a certain visit duration. From Figure 5(a) we notice that approximately 70% of the visits lasted less than 1 minute. These visits are mostly caused by conference participants and people who took a look at the site out of curiosity. The remaining visits are attributable to committee members and search engine spiders. Figure 5(b) shows the two-mode power law relation between visit duration and number of visits with exponents $1.02, 3.36$ and $R^2 = 0.95$. The dividing point is at 30 minutes, which is our chosen visit duration timeout period.

4) Page Depth: Figure 6 shows the number of unique pages browsed per visit. We observe that more than 40% of visits include browsing more than one page. We can again notice the long-tailed nature of distribution of page depth per visit in Figure 6(b). We also observe a power-law function with an exponent 2.36 and $R^2 = 0.95$. Most of the visits with page depth greater than 3 can be attributed to search engine spiders and conference organizers. Furthermore visits with single-page visits were mostly restricted to the homepage, the program page, the call for papers page, the important dates page, or specific paper downloads.

We found that the top 18 pages accounted for about 53% of the total pageviews. For obvious reasons the most viewed page was the homepage at 19%. The technical program pages (list of refereed papers and posters) accounted for about 11% of the pageviews. Other pages that were viewed highly were the important dates (3%), call for papers (3%), submissions (2%), and registration (2%) pages.



(a) Truncated pdf (b) Log Log Distribution

Figure 6. Page Depth per Visit

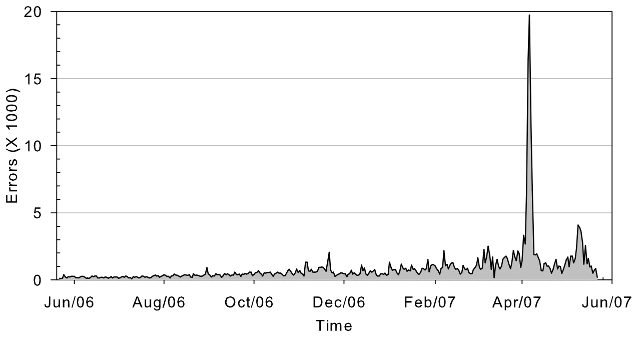


Figure 7. Client Errors per Day

5) Errors: The server logs recorded approximately $314,000$ client errors (status code 404 and 416), which accounted for 3% of the total hits. No server-side errors were noted, indicating that the Web site functioned properly with few outages during the 1-year period.

Figure 7 shows the daily error hits as seen by the server. As mentioned earlier, most of the errors were due to two missing files, namely `favicon.ico` and `robots.txt`. The `favicon.ico` file was added on April 14, 2007, while the `robots.txt` file was never added.

Figure 7 shows that the error rate was relatively low until the first week of April 2007, when a spike occurred. As part of the online conference proceedings production, the entire site was crawled for HTML validation and to check for broken links. The spike in errors represents a transient state when many of the new links were not working, though they were fixed later in the week. It should be noted that not all client errors are due to a missing file; sometimes visitors typed in a wrong URL.

4.3 Network Traffic Trends

This section presents traffic volume and network usage characteristics from our data sets. We utilize the server logs to understand the load patterns of the Web site and the lo-

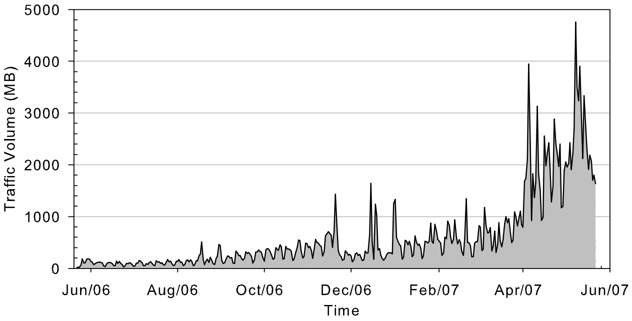


Figure 8. Bytes Transferred per Day

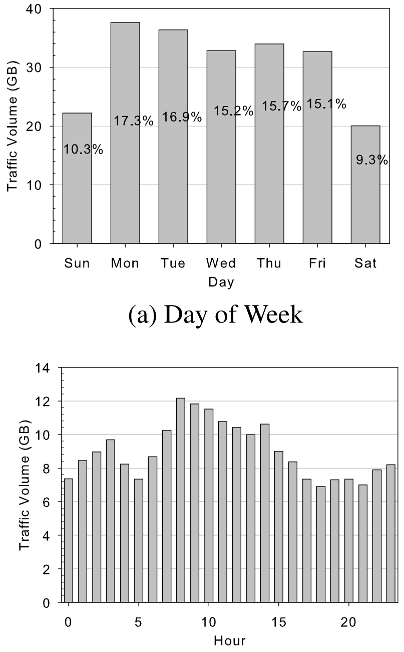
cality properties of the visitors. Using Google Analytics, we report on the network connection speeds, as well as the browsers and operating systems used by the visitors.

1) Traffic Volume: Figure 8 shows the total bytes transferred each day during the 1-year period. We observe that the traffic volume almost quadruples during April and May of 2007 compared to the preceding months. Up until April 2007, approximately 800 MB of data was transferred each day. During the first week of April the entire Web site was crawled and the online proceedings was added containing `PDF` files of all accepted papers and posters. May 7, 2007 was the busiest day when 4.7 GB of data was transferred by the server to its visitors. This increase in traffic volume was mostly due to visitors accessing `PDF` files of papers from the online proceedings and site updates. Approximately 55% of the total traffic volume transferred occurred during the last 60 days.

Figure 9 shows the total traffic volume transferred based on day of week and hour of the day. We find that the work week accounted for almost 80% of the total traffic volume. Each weekday has about $15 - 17\%$ of the total traffic volume, with Monday being slightly busier. The (local) work period is clearly visible in Figure 9(b) when almost 50% of the data is transferred. The remaining hours mostly represent the access patterns of overseas users, committee members, and search engine spiders.

2) Geographic Location: Figure 10 shows the breakdown of unique visitors and traffic volume based on geographic location. We observe that the G7 countries[3] account for the majority (60%) of visitors and traffic volume, with the U.S. alone accounting for approximately 40%. These results match well with country of affiliation for registered attendees. About 75% of the conference participants were from G7 countries, including 43% from the U.S. Participants from Canada accounted for 15% of the total.

We also observe that certain countries, namely U.K., China, and Canada, transfer a higher percentage of traffic as compared to percentage of unique visitors. The local organizing committee was based in Calgary and was involved in building and testing the Web site, which meant higher page depths per visits and this translated into higher traffic volume. The average visit duration of Canadian visitors was the highest at 5.4 minutes. Some of the site maintenance was done from U.K. The level of activity from China reflected some early planning for WWW2008 in Beijing. In addition to a large delegation of WWW2008 organizers at WWW2007, the conference drew significant attention from the Chinese government, academia, and industry.



(a) Day of Week

(b) Hour of Day

Figure 9. Total Bytes Transferred

3) Network Connection: Figure 11 shows the distribution of visits categorized by the network connection type. Of the 14,400 distinct network providers identified, approximately 57% of the visits were made using a broadband connection. With most of the visitors being from the G7 countries where broadband connectivity is pervasive, the results are easy to comprehend. However, since the conference Web site was targeted towards academic and industry members, who have easy access to T1/T3 network connections, we expected the percentage of T1 connections to be higher. We also found that a significant percentage of visits were categorized as unknown connection speed.

We conjecture that Google Analytics does not use empirical testing to determine the speed; rather, it resolves the IP address to determine the ISP, and uses heuristics to guess

[3] The G7 is a group of seven industrialized economies of the world, namely, U.S., Japan, Germany, France, U.K., Italy, and Canada.

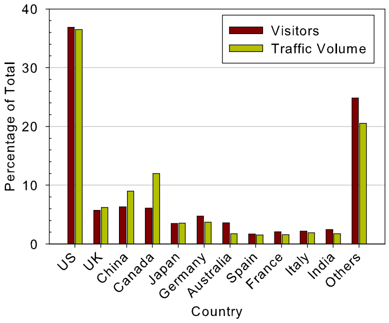


Figure 10. Geographical Analysis

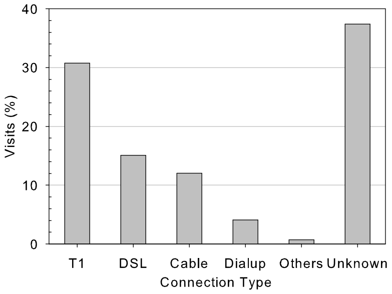


Figure 11. Network Connection Type

the speed of visitor's connection. To test our hypothesis, we looked at the connection speed of visits as reported by Google Analytics during the conference itself. Most of the visits were classified as unknown, even though the conference had provided high-speed WiFi (35 Mbps) connection to its attendees.

We looked at the top 10 ISPs using the server logs and found a mix of academic institutions (e.g., University of Calgary, University of Massachusetts, University of Southampton), commercial service providers (Shaw, Comcast, MyCanopy, Road Runner, Verizon, AT&T, Telus), search engines (Microsoft, Yahoo!, Google), and a research lab (Bell Labs). These ISPs accounted for about 20% of the total traffic volume. The universities and Bell Labs were associated with the conference organization. Canadian regional service providers such as Shaw and Telus were also used by the local organizers. MyCanopy was the Internet service provider during the conference.

4) Browser and Operating System: Google Analytics data can indicate the browser and operating system used by the client. Internet Explorer (IE) was prevalent among site visitors, accounting for 50% of all visits. Firefox accounted for 40%. For operating systems, most of the visits (83%) came from Windows, while 9% were from Mac OS and 7% from Linux. These statistics reflect the typical market share for browsers and operating systems [3]. In 2007, Windows and IE had market shares of approximately 90% and 50%, respectively.

4.4 Traffic Sources

In this section, we present an analysis of the traffic sources for the Web site using Google Analytics. Typically, Web accesses can be classified as direct (e.g., typing the URL in the browser, using a bookmark, or clicking a link to the site from an email client), referral page (e.g., a hyperlink from a related Web site), or search engine (e.g., a user clicked on a link in search results). We found that only 27% of the traffic was direct. The rest of the visits came via search engines (47%) and site referrals (26%). The only exception was during the conference period when visits from all traffic sources were almost equal.

These results indicate that users rely heavily on search engines as their entry points into the Web. Google was the overwhelming choice for search engine among the site visitors, accounting for 95% of the visits. Yahoo! accounted for 3% of the visits. While these two search engines are well known worldwide, we noticed some popular regional search engines as well. Baidu and Yandex, the biggest search engines in China and Russia, respectively, provided traffic to the site. Visitors also seem to use meta search engines (e.g., `mamma.com`). These search engines query multiple other search engines.

For site referrals, the WWW2006 Web site was the major single source, providing about 9% of the total visits. The organizers (University of Calgary, IW3C2, W3C) accounted for approximately 14% of the visits. Many academic institutions hosted workshops that were co-located with the main conference. These accounted for another 5% of the visits (e.g., Lehigh University hosted the AirWeb 2007 workshop page). Other sources of site referrals were social networking sites (e.g., Upcoming), blogs, Wikipedia, computing societies (e.g., ACM, International Association of Cryptologic Research), and industry (e.g., Google Labs). We also noticed site referrals from Web-based free email services such as Gmail, perhaps triggered by the newsletters sent via email to prospective participants.

4.5 Anomalous Activities

In this section we discuss anomalous patterns seen in the Web site traffic, whether generated by humans or robots.

1) Unusual Search Referrals: We were perplexed by the unusually high number of visits between December 31, 2006 and January 2, 2007 (a traditional holiday period). Fearing that this might be malicious activity, we analyzed both the server logs and Google Analytics for that time frame to uncover the reasons for this activity. We found that about 4,200 unique visitors made approximately 4,300 visits. This was the most unique visitors observed for any day in the 1-year period, and the only time when the number of visits almost matched the number of unique visitors. A majority (over 90%) of these visits lasted for less than 10 seconds (the lowest average visit duration in the 1-year

Table 5. Characteristics of Robot Visits

Characteristic	Maintenance	External
Total Hits	601, 916	283, 762
Unique Visitors	7	4, 263
Total Visits	400	39, 666
Avg. Visit Duration (min:sec)	41:13	9:45
Traffic Volume (GB)	7.7	7.2
Avg. Traffic Volume per Visit	19.6 MB	189.8 KB
Total Pageviews	161, 685	162, 255
Unique Pageviews	58, 159	110, 237
Avg. Pageviews per Visit	404	4

period). Furthermore, 97% of the visits were by first-time visitors and 95% came from via search engines.

These behaviours deviate significantly from the average characteristics for the site, indicating that a completely different set of users visited the site during this time. Since most of the traffic came via a search engine, we looked at the keywords used by the visitors. Over 94% of the visitors had typed the keyword `2007`. We believe that search engines would have returned the conference Web site as the top result (WWW2007 was still in the top 20 search results on Google as of August, 2008.). It appears that novice Web users were trying to find some interesting site, event, or activity related to the New Year. The search engine returned the conference Web site in the list, and people would click on it. Upon noticing that this was not the site they were looking for, they would exit the site.

2) Robot Visits: The conference Web site was regularly visited by search engine spiders and other robots. We use the server logs to study the activities of such visits. A robot can be identified by looking at its IP address and user agent in the log and matching with a list of known robots. Well behaved robots will request the `robots.txt` file from the server before searching the site. Since our server logs did not contain any user agent information, we used any request for the `robots.txt` file to identify a robot. Bad robots will often bypass this file and use fake user agent names. Detecting such robots is a non-trivial task. We used a list containing IP addresses of known bad robots and searched for them in the server logs [5, 11]. We report on these robots later in the section.

Table 5 shows the characteristics of two robot groups. Robots that were used by the conference organizers to update the conference Web site are labelled as Maintenance. All other robots are classified as External. Major updates to the Web site were done by visitors from the following domain names: University of Massachusetts, University of Southampton, Bell Labs, and Linuxfromscratch. The first three of these domain names are associated with conference organizers, while the fourth is not. However, Linuxforscratch provided the routers (including the on-site router) and mail servers for the ISP (MyCanopy) that provided Internet service to the attendees during the conference. All visits from this domain name were during the conference period, when the site administrators were updating the site at the end of each day. We observed that all hits to the Web site occurred between 2 AM and 4 AM. Robots from University of Massachusetts accounted for 55% of the hits and 18% of traffic volume. Linuxfromscratch robots caused 41% of the traffic volume and 26% of the hits. Maintenance robots accounted for over half of the total traffic volume transferred by the server and their average visit duration was five times that of External robots.

Among the external robots, search engine spiders accounted for over 75% of the visits. Almost 45% of the search engine spider visits were due to the Inktomi spider (Inktomi is used by Yahoo! search engine). Microsoft robots accounted for about 10% of the visits followed by Google at 7% of the visits. We also observed crawlers from image search engines such as Picsearch.

Other robot visits came from educational institutions, Web filtering companies, anti-virus companies, and individual agents. For example, we identified a crawler from the Database Group at University of Wisconsin that indexes relevant conference sites as part of the DBLife project. Examples of some other institutions sending robots were MIT, UCLA, UIUC, etc. We also noticed a robot from `Twtelecom.net`, a Web filtering services company. Their robots crawl sites and decide whether to block them or not for their customers. We also noticed robot visits from an anti-virus vendor, Symantec.

Finally, we did notice some malicious robot activity on the Web site. For example, we found robots from hosts on `svservers.com` and `interwave.ru` involved in spamming, particularly targeting forums and blogs. In our case they were accessing a file called `comment.php` intending to leave spam messages.

5 Related Work

Significant research has been conducted in characterizing Web workloads of clients [6], servers [8, 7, 18, 20], proxies [15], and national domains [9]. A survey of Web workload characterization studies prior to 1997 can be found in [19]. These studies have provided a better understanding of the Web, and their findings have resulted in improved caching policies, prefetching techniques, and load balancing strategies for servers and proxies.

Arlitt and Jin [7] were the first to characterize the workload of a large commercial Web site (World Cup 1998). Padmanabhan and Qiu [18] studied the dynamics of server content and client accesses of a large commercial news Web site. They found that file modifications were more frequent than file creations, file popularity followed a Zipf distribution, and popularity of documents tended to decrease as they became older. Shi *et al.* [20] analyzed the server workloads of a customizable university Web portal. They deduced that client-side latencies can be reduced by actively prefetching a few non-shared personalized channels.

Bent *et al.* [10] characterized a large server farm hosted by an ISP. They observed that a high percentage of the workload consisted of uncacheable content, there was widespread usage of cookies, and Web sites did not utilize the cache control features of the HTTP/1.1 standard. The prevalence of uncacheable Web content in Internet traffic was also reported in [22].

Menascé *et al.* [16] studied the characteristics of e-commerce workloads at the user, application and protocol levels. They found that session lengths are heavy-tailed, most requests are human-generated, product selection functions were used more frequently than product ordering functions, and the popularity of product search terms followed a Zipf distribution. Vallamsetty *et al.* [21] found that e-commerce workloads exhibit a higher degree of self-similarity than general Web workloads [12].

The evolution of the Web was studied by Fetterly *et al.* [13]. They observed that the frequency and degree of change of a Web page were strongly related to document size and past changes were a sound indicator of future changes. The results of the work have implications on improving the indexing strategies of search engine crawlers.

We believe that our experiences with WWW2007 and our characterization results will be useful for future conference organizers, in several ways. With knowledge of the frequency, duration, and source of client visits, organizers can perform appropriate search engine optimizations in order to promote the site more efficiently. Furthermore, familiarity with the access patterns of users could allow organizers to schedule site updates and maintenance such that it causes minimal disturbance to visitors.

Site administrators could improve the user experience on the site by designing the site with a more complete knowledge of system and network properties of potential visitors. The *flash-crowd* effect (i.e., sudden and unexpected spike in traffic) during the new year indicates that organizers should be prepared for the unexpected. We showed that robot loads should not be underestimated, and that an understanding of robot visit patterns may allow organizers to prepare security procedures to safeguard their Web site.

6 Conclusions

This paper provides a multi-layer workload characterization of a large conference Web site. Using server-side and client-side measurement we analyzed usage behaviour, client errors, client network properties, bandwidth, and robot activity of the site. We found that both methods have their strengths and weaknesses. By combining the two, we can get a more complete picture of the Web site operation.

Our analysis showed that approximately $130,000$ unique visitors visited the site, generating 10 million requests and about 215 GB of traffic volume during a 1-year period. The Web site traffic was non-stationary, with much of the Web site activity in the month just prior to the conference. Visitor activity showed no strong diurnal pattern, reflecting the international usage of the site. Almost half of all visits came via search engine queries. Robot visits were also prevalent on the site.

References

[1] Google Analytics. http://www.google.com/analytics/.

[2] WWW2007 Web Site. http://www2007.org/.

[3] Market Share for Browsers, Operating Systems and Search Engines, 2007. http://marketshare.hitslink.com/.

[4] Western Europe Leads North America in Broadband Growth, 2007. http://www.websiteoptimization.com/bw/0710/.

[5] List of Bad Bots, 2008. http://www.kloth.net/internet/badbots.php.

[6] A. Adya, P. Bahl, and L. Qiu. *Content Networking in the Mobile Internet*, chapter Characterizing Web Workload of Mobile Clients. Wiley, 2004.

[7] M. Arlitt and T. Jin. A Workload Characterization Study of the 1998 World Cup Web Site. *IEEE Network*, 14(3), 2000.

[8] M. Arlitt and C. Williamson. Internet Web Servers: Workload Characterization and Performance Implications. *IEEE/ACM Trans. Netw.*, 5(5):631–645, 1997.

[9] R. Baeza-Yates, C. Castillo, and E. Efthimiadis. Characterization of National Web Domains. *ACM Trans. Interet Technol.*, 7(2), 2007.

[10] L. Bent, M. Rabinovich, G. M. Voelker, and Z. Xiao. Characterization of a Large Web Site Population with Implications for Content Delivery. *WWW*, 9(4), 2006.

[11] C. Bomhardt, W. Gaul, and L. Schmidt-Thieme. Web Robot Detection - Preprocessing Web Logfiles for Robot Detection. In *Proc. SIS CLADAG*, 2003.

[12] M. Crovella and A. Bestavros. Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. *IEEE/ACM Trans. Netw.*, 5(6):835–846, 1997.

[13] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A Large-scale Study of the Evolution of Web Pages. In *Proc. WWW*, 2003.

[14] R. Levering and M. Cutler. The Portrait of a Common HTML Web Page. In *Proc. ACM DocEng*, 2006.

[15] A. Mahanti, C. Williamson, and D. Eager. Traffic Analysis of a Web Proxy Caching Hierarchy. *IEEE Network*, 14(3), 2000.

[16] D. Menascé, V. Almeida, R. Riedi, F. Ribeiro, R. Fonseca, and W. Meira. A Hierarchical and Multiscale Approach to Analyze E-business Workloads. *Perform. Eval.*, 54(1), 2003.

[17] J. Mogul and T. Anderson. Open Issues in Organizing Computer Systems Conferences. *ACM CCR*, 38(3), 2008.

[18] V. Padmanabhan and L. Qiu. The Content and Access Dynamics of a Busy Web Site: Findings and Implications. In *Proc. ACM SIGCOMM*, 2000.

[19] J. Pitkow. Summary of WWW Characterizations. *WWW*, 2(1-2), 1999.

[20] W. Shi, Y. Wright, E. Collins, and V. Karamcheti. Workload Characterization of a Personalized Web Site and its Implications for Dynamic Content Caching. In *Proc. WCW*, 2002.

[21] U. Vallamsetty, K. Kant, and P. Mohapatra. Characterization of E-Commerce Traffic. *Elect. Comm. Res.*, 3(1-2), 2003.

[22] Z. Zhu, Y. Mao, and W. Shi. Workload Characterization of Uncacheable HTTP Content. In *Proc. ICWE*, 2003.