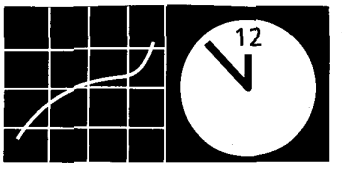


A Workload Characterization Study of the 1998 World Cup Web Site

Martin Arlitt and Tai Jin, Hewlett-Packard Laboratories

Abstract

This article presents a detailed workload characterization study of the 1998 World Cup Web site. Measurements from this site were collected over a three-month period. During this time the site received 1.35 billion requests, making this the largest Web workload analyzed to date. By examining this extremely busy site and through comparison with existing characterization studies, we are able to determine how Web server workloads are evolving. We find that improvements in the caching architecture of the World Wide Web are changing the workloads of Web servers, but major improvements to that architecture are still necessary. In particular, we uncover evidence that a better consistency mechanism is required for World Wide Web caches.

The 16th Federation Internationale de Football Association (FIFA) World Cup was held in France from June 10 through July 12, 1998. France '98, as the 16th FIFA World Cup was commonly called, was the most widely covered media event in history. An estimated cumulative television audience of 40 billion watched the 64 matches of France '98, more than twice the cumulative television audience of the 1996 Summer Olympic Games in Atlanta. http://www.france98.com, the Web site for France '98, was also popular, receiving more than 1 billion client requests during the tournament.

This article presents a detailed workload characterization of the France '98 Web site (more information is available in [1]). Workload characterization plays an important role in systems design. It allows us to understand the current state of the system. By characterizing the system over time we can learn what effects changes to the system have had. Workload characterization is also crucial to the design of new system components. In this article we focus on the characterization of a Web server workload. We compare our results to those from previous studies (e.g., [2]) to determine how Web server workloads have changed over time. Furthermore, the extremely heavy workload of the World Cup site allows us to predict what the workloads of future Web servers may look like so that we may plan accordingly.

Some of the more significant characteristics we observed in the World Cup workload and the performance implications of these characteristics include:

- HTTP/1.1 clients are becoming more prevalent, accounting for 21 percent of all requests. Widespread deployment of HTTP/1.1-compliant clients and servers is necessary for the functionality of HTTP/1.1 to be fully utilized.
- Eighty-eight percent of all requests were for image files; an additional 10 percent were for HTML files, indicating that most user interest was in relatively static (i.e., cachable) files.
- Almost 19 percent of all responses were "Not Modified," indicating that cache consistency traffic had a greater impact on the World Cup workload than on previous Web server workloads [2].
- The workload was quite bursty, although over longer time scales (e.g., hours or more) the arrival of these bursts was quite predictable.
- During periods of peak user interest in the World Cup site the volume of cache consistency traffic increased dramatically. This indicates that current consistency mechanisms do not allow Web caches to eliminate "flash-crowds" (i.e., sudden, unexpected increases in traffic to a site) in the network and at the servers, which is supposed to be one of the main benefits of Web caching.

Web server workload characterization is only one of the necessary steps for understanding the changes occurring in Web traffic. Research efforts on Web client workloads [3], Web proxy workloads [4], network traffic characterizations [5] as well as HTTP analyses [6] are all required in order better understand the Web.

The remainder of the article is organized as follows. We provide background information on the 1998 World Cup tournament, and introduce the France '98 Web site. Next, we discuss the data set used in the workload characterization study and present the results of the study. Then, we analyze a particularly busy segment of the World Cup workload and compare the results to the overall study previously discussed. We will describe in more detail the performance implications of the results from the previous sections. Finally, we summarize the contributions of our article and list future work.

0890-8044/00/$10.00 © 2000 IEEE

Background

The 1998 World Cup

In order to better understand the nature of the workload from the France '98 Web site knowledge of the tournament itself is required. This information is particularly useful for understanding the usage of the site.

The FIFA World Cup is a tournament held once every four years to determine the best football (soccer) team in the world. Due to the large number of teams interested in participating, a qualifying round is used to select the teams that will play in the World Cup tournament. Of the 172 countries that entered the qualifying round for France '98, 30 were selected to compete for the World Cup, along with the host country, France, and the reigning champions, Brazil.

France '98 began on June 10, 1998 and ended on July 12, 1998. The tournament consisted of several rounds of play. The opening round lasted from June 10 until June 26. The 32 participating teams were divided into eight groups. Each team then played one 90-min match against each of the other teams in its group. The top two finishers from each group qualified for the playoffs. During the playoffs only the winning team of each match advanced to the next round. If a playoff game was tied after 90 min, a 30-min sudden death overtime period was played. If the overtime failed to determine a winner, penalty kicks were used to decide which team would advance. The first round of the playoffs, known as the "Round of 16," lasted from June 27 through July 1. The remaining rounds of the tournament were the Quarter Finals, held on July 3 and 4; the Semi Finals, held on July 7 and 8; and the Final, held on July 12. A match to determine the third place finisher was held on July 11.

The 1998 World Cup Web Site

The Web site of the 1998 World Cup, http://www.france98.com, provided Internet-savvy football fans around the world with a wide range of information. Besides being able to access the current scores of the football matches in real time, fans could also access previous match results, player statistics, player biographies, team histories, information on the stadiums, facts about local attractions and festivities, as well as a wide range of photos from the matches and interviews with players and coaches. Fans could also download free software, such as World Cup screensavers and wallpapers from the France '98 Web site. All of the information on the site was available in English and French.

The France '98 Web site went online May 6, 1997. In anticipation of significant interest from the Internet community, emphasis was placed on developing an available, reliable, and low-latency platform to power the Web site. During the tournament 30 servers were used, distributed across four locations: Paris, France; Herndon, Virginia; Plano, Texas; and Santa Clara, California. All Web page creation and modifications occurred in France, and were distributed to the U.S. locations. A number of load balancers were used to distribute the requests across these four locations and among the servers at each location. In this article we examine only the aggregate workload. Information on the workload at each of these locations is available in [1].

Workload Characterization

This section presents the results of our workload characterization. The next section provides information on the collection and reduction of the data set used in our study. Since 99.88 percent of all requests contained the GET method, throughout the remainder of the article we analyze only these GET requests. We then discuss the protocol version, response status code and file type distributions. We also analyze the usage of the World Cup site. We describe the unique file size distribution while later we look at the file referencing patterns.

Log entry	Description
remotehost	The IP address of the client issuing the request
rfc931	The remote login name of the user
authuser	The username by which the user has authenticated himself
[date]	The date and time of the request (1 s resolution)
request	The request line exactly as it came from the client
status	The HTTP response status code returned to the client
bytes	The content length of the document transferred

■ Table 1. *A description of Common Log Format entries.*

Duration	May 1–July 23, 1998
Total requests	1,352,804,107
Average requests/min	10,796
Total bytes transferred (Gbytes)	4991
Average bytes transferred/min (Mbytes)	40.8

■ Table 2. *A summary of access log characteristics (raw data).*

Collection and Reduction of Data

The data set used in this workload characterization study is composed of the access logs collected from each of the servers used in the World Cup Web site. The access logs from each server were archived on a daily basis. For this study all of the access logs from May 1 through July 23, 1998 were analyzed.

Each access log is in the Common Log Format. For every request received by the Web server, the information described in Table 1 is stored. The `request` line from the client includes the method (e.g., GET) to be applied to the requested resource [7], the name of the resource (e.g., /index.html), and the protocol version in use (e.g., HTTP/1.0).

Table 2 summarizes the access logs that we acquired from the World Cup site. Our first concern was with the size of the raw access logs: 125 Gbytes in total, 14 Gbytes when compressed. In order to make our workload analyses more efficient, we chose to convert the logs to a more compact binary format. We reduced the storage requirements in two ways. One approach removed unnecessary data. For example, we deleted the `rfc931` and `authuser` fields since they were not used by the servers and thus provided no information. The second tactic we used to reduce the size of the data set was to represent the remaining fields in more efficient ways when possible. For example, we mapped each distinct URL to a unique integer identifier. Finally, we collated the access logs of all the servers by request time. The resulting binary log file was 25 Gbytes in size, 9 Gbytes when compressed. Furthermore, each request is now in a fixed size structure, which also helps to improve the efficiency of our analyses.

Despite the vast amount of data that was collected by each of the servers, a lot of interesting information is still not available. For example, although the logs do have a timestamp that records when the request was received by the server, it has a one second resolution which is too coarse-grained to be of use for numerous analyses (e.g., interrequest times). This is just

Response code	% of requests	% of data transferred
200 (Successful)	80.52	97.86
206 (Partial Content)	0.09	2.08
304 (Not Modified)	18.75	0.00
4xx, 5xx (Errors)	0.64	0.06
Total	100.00	100.00

■ Table 3. *Breakdown of server response codes.*

File type	% of requests	% of data transferred
HTML	9.85	38.60
Images	88.16	35.02
Compressed	0.08	20.33
Dynamic	0.02	0.38
Other types	1.89	5.67
Total	100.00	100.00

■ Table 4. *Breakdown by type.*

one example of useful information that could be added to a revised Web server log file format.

Statistical Characteristics

Our first analysis examined the version of Hypertext Transfer Protocol (HTTP) supported by the client issuing the request. As expected, we found that HTTP/1.0 is still the version used by most (78.7 percent) of the clients. However, over 20 percent of the traffic came from clients that support HTTP/1.1. This suggests that browsers supporting HTTP/1.1 are slowly replacing browsers that do not. These results do not indicate what percentage of the requests to the World Cup site, if any, actually used HTTP/1.1 functionality.

Table 3 shows the breakdown of server response codes. Table 3 reveals that the majority of requests resulted in the successful transfer of an object (response status 200). The successful transfers account for almost all of the content data (97.86 percent) transferred from the Web site back to clients. The second most common status code was the Not Modified (304) response, which accounted for almost 19 percent of all responses to client requests. This represents a substantial increase over the fraction of Not Modified responses seen in earlier server workloads [2]. The reason for this increase can be attributed to the improved caching architecture in the Web, including persistent caches in browsers, and more recently in proxies and networks (e.g., transparent caches). This type of response indicates that the client issued a conditional GET request to verify that its cached copy of the file is consistent with the version being served at the Web site.

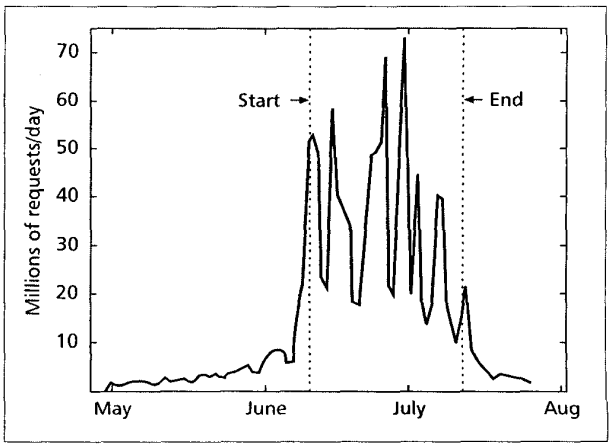
Table 4 shows the breakdown of responses by the type of file requested by the client. For the majority of the responses the file extension was used to determine the file type. For example, files ending with `.jpg` or `.gif` were placed in the Images category. We considered any URL that included a `cgi-bin` substring or a parameter list (e.g., `/home.htm?parameter_list`) to be a dynamic file. In some cases dynamic files can also be identified by file type (e.g., `.asp`). For all remaining (unique) requests we issued HEAD requests to the Web site and used the `Content-type:` response header to classify the file.

Table 4 reveals that almost all client requests (98.01 percent) were for either HTML (9.85 percent) or image (88.16 percent) files. A similar characteristic was observed in earlier Web server workloads [2]. HTML files had more impact than image files on the volume of data transferred from the Web site (38.60 percent for HTML compared with 35.02 percent for images). Most image requests were for small inline graphics, while the HTML requests were for substantially larger files. Compressed files (e.g., screensavers), which accounted for only 0.08 percent of all requests, were responsible for over 20 percent of the data traffic. The huge discrepancy between the percentages of requests and data transferred for the corresponding transfers is an indication of the effects large files can have on the workload of a Web server and of the network.

Usage Analysis

Figure 1 shows the daily traffic volume handled by the World Cup Web site. From the beginning of May until the start of the World Cup on June 10, the traffic volume is quite light, although clearly building in anticipation of the start of the event. Beginning on June 10, the volume of traffic grows enormously. This marks the beginning of a prolonged *flash-crowd*. That is, the site suddenly became very popular, remained popular for a period of time, and then quickly faded back into obscurity. Although the daily traffic volume is quite bursty during the World Cup, the traffic volume remains higher than it was at any time prior to the start of the event. The busiest day for the site was June 30, when over 73 million requests were handled by the France '98 site.

In order to better understand the causes of this burstiness we analyzed the traffic in more detail. Figure 2 shows the hourly traffic volume of the World Cup Web site. The figure consists of six bar graphs, one for each week of the World Cup tournament. The solid black curve in each graph represents the hourly volume of requests (y-axis) for the given time (x-axis, normalized to local time in France). The scale of both the x- and y-axes are kept constant across all bar graphs to facilitate comparisons in traffic volume over time and by day of the week. The dashed vertical lines indicate the starting time of a World Cup football match. The teams involved in each match are also listed (the abbreviations are defined in Table 5). For example, at 5:30 p.m. (in France) on Wednesday June 10, the first match of the 1998 World Cup was played between Brazil (BRA) and Scotland (SCO). At this time requests were arriving at a rate of 6 million/hr.



■ Figure 1. *Daily traffic volume to the World Cup Web site.*

Abbreviation	Team	Abbreviation	Team	Abbreviation	Team	Abbreviation	Team
ARG	Argentina	DEN	Denmark	ITA	Italy	NOR	Norway
AUT	Austria	ENG	England	JAM	Jamaica	PAR	Paraguay
BEL	Belgium	ESP	Spain	JPN	Japan	ROM	Romania
BGR	Bulgaria	FRA	France	KOR	South Korea	RSA	South Africa
BRA	Brazil	GER	Germany	KSA	Saudi Arabia	SCO	Scotland
CHI	Chile	HOL	The Netherlands	MEX	Mexico	TUN	Tunisia
CMR	Cameroon	HRV	Croatia	MOR	Morocco	USA	United States
COL	Columbia	IRN	Iran	NGA	Nigeria	YUG	Yugoslavia

■ Table 5. *Team name abbreviations.*

The bar graphs also indicate the days on which each round of the tournament began (e.g., the Round of 16 began on Saturday June 27), as well as those matches that required penalty kicks to decide a victor, indicated by a (P) following the names of the teams.

Figure 2 reveals that there were many variables that affected the hourly traffic volume at the World Cup Web site. For example, the volume of traffic increased when matches were in progress and decreased once they had finished. These bursts represent flash-crowds on a smaller scale. The traffic volume was also affected by the teams involved in the matches (e.g., traditional football powers like Brazil and Germany are of interest to football fans everywhere, not just Brazilians and Germans), the number of matches in progress (e.g., from June 23 through June 26 matches were played in parallel), and the playoff implications of the match. The time differences between the user's geographic location and France also contributed to the usage of the World Cup site.

One interesting observation to be made from Fig. 2 is that the volume of traffic to the World Cup Web site was quite low on weekends, even though a higher percentage of matches were played on Saturday and Sunday than on weekdays. The obvious reason for this reduction in traffic volume is that people preferred to watch the matches on television. When these fans were unable to watch the matches on television, such as when they were at work or school, or certain matches were not televised in their area, they relied on the Web to provide them with progress reports on the matches.

Unique File Size Distribution

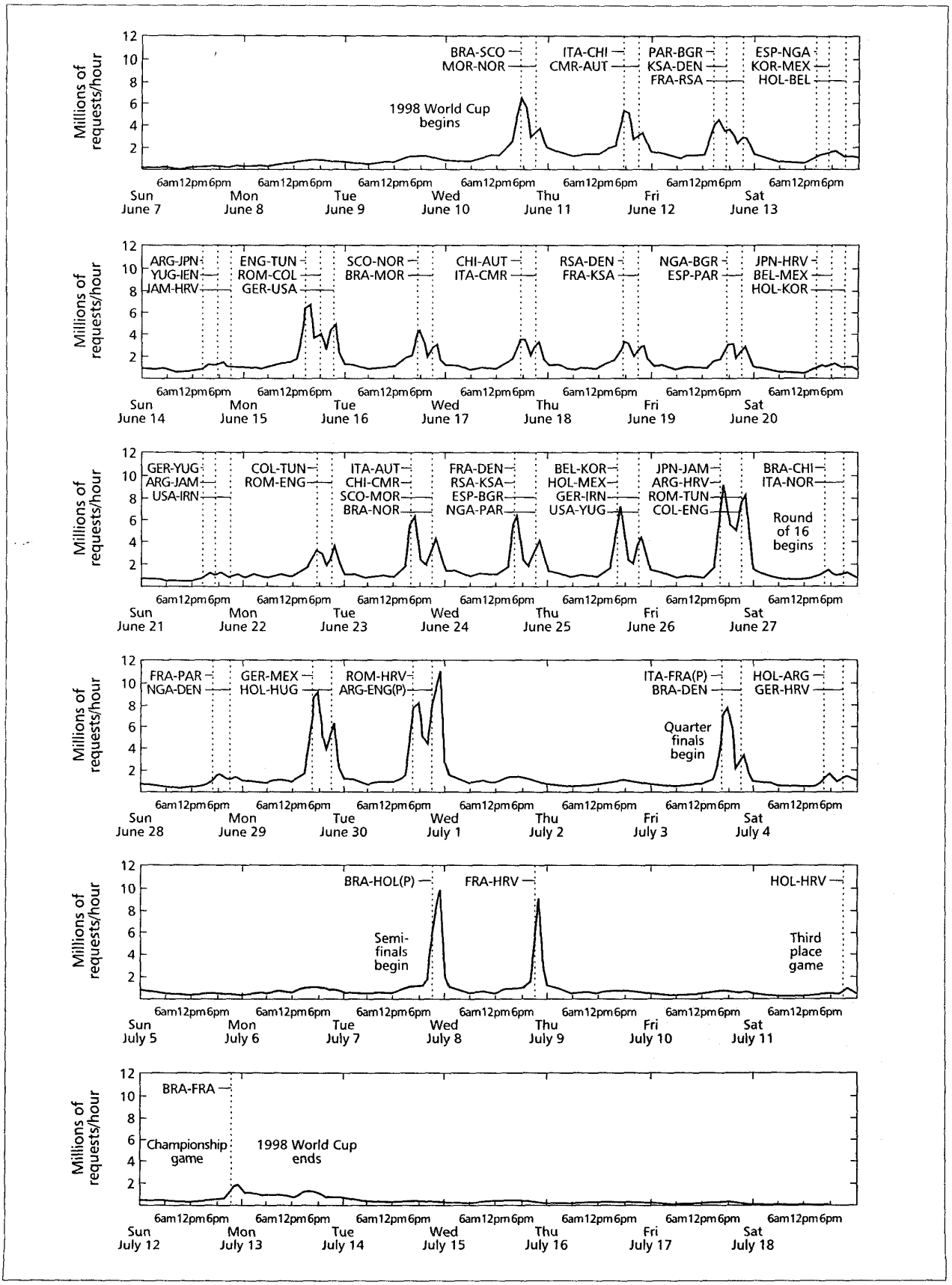
In this section we analyze the distribution of sizes for all unique files requested from the World Cup Web site. Information on the successful transfer size and overall transfer size distributions is available in [1].

Our first analysis looks at the sizes of each of the unique files requested and successfully transferred at least once in the access log. For the purpose of this study we utilize the initial nonzero size recorded for each unique file. Twenty thousand, seven hundred twenty-eight unique files were requested (and successfully transferred) from the World Cup site during the measurement period. The total combined size of these files was 307 MB. The mean size of these files was 15,524 bytes, the median size 4674 bytes, and the maximum size 61.2 Mbytes. The mean, median, and maximum values are consistent with those observed in other Web server workloads [2].

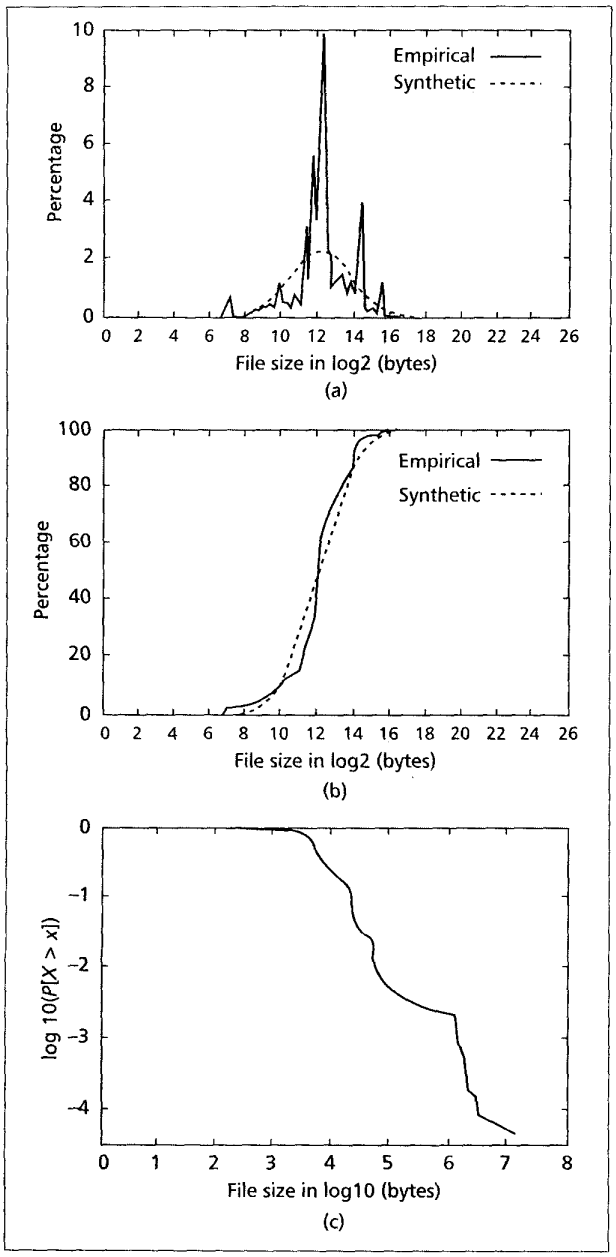
Figure 3a presents the frequency histogram; Fig. 3b provides the cumulative frequency histogram of the unique file sizes. We have applied a logarithmic transformation to the file sizes to enable us to identify patterns across the wide range of values [8]. For a $\log_2$ transformation, bin_i includes values in the range $2^i \le x < 2^{i+1} - 1$. Similarly, for a $\log_{10}$ transformation, bin_i includes values in the range $10^i \le x < 10^{i+1} - 1$. Figure 3a indicates that most unique files have sizes in the 256-byte to 64-kbyte range (2^8–2^{16} bytes). In other Web workload characterizations the file size distribution has been found to be *lognormal* [3, 4]. That is, after applying a logarithmic transformation to the data, the data appears to be normally distributed. We compare the unique file size distribution (the *empirical data*) to a synthetic lognormal distribution with parameters $\mu = 12.14$ and $\sigma = 1.73$. From Fig. 3a we can see that the empirical data deviates quite substantially from the synthetic model. These differences are due to the distinct nature of the World Cup site. For example, in Fig. 3a about 10 percent of all unique files were around 4 kbytes (212 bytes) in size. Sixty-five percent of these files are HTML objects that provided profiles on the individual players who participated in the World Cup tournament. The other large spikes in Fig. 3a are also the result of groups of related objects having very similar sizes. On "typical" Web sites we would not expect to see such large clusters of related objects making up a substantial percentage of all files on the Web site. Despite the number of spikes seen in Fig. 3a, the cumulative frequency histogram (shown in Fig. 3b) indicates that the lognormal distribution still provides a reasonable estimate for the body of the unique file size distribution.

While most of the unique files are less than 64 kbytes in size, a few are substantially larger. Our next analysis examined the tail of the unique file size distribution to determine if it is heavy-tailed. A distribution is considered heavy-tailed if $P[X > x] \sim x^{-\alpha}$, $x \to \infty$, $0 < \alpha < 2$. This means that if the asymptotic shape of the distribution is hyperbolic, it is heavy-tailed, regardless of the behavior of the distributions for small values [9]. To determine if the unique file size distribution from the World Cup Web site is heavy-tailed, we plotted the complementary distribution (CD) function on log-log axes and examined the results for linear behavior on the upper tail. This method of analysis is described in [9]. The results of this analysis for the World Cup data are shown in Fig. 3c. The tail of the distribution does exhibit some linear behavior, which suggests that the distribution is indeed heavy-tailed. However, this linearity does not exist throughout the entire tail. Specifically, a spike exists in the 1–4-Mbyte range. This spike is caused by the existence of 44 files whose sizes are in the 1–4-Mbyte range. These files include 13 uncompressed high-resolution images, four audio clips, 15 screen savers (i.e., downloadable software), and 12 video clips.

To verify that the unique file size distribution is indeed heavy-tailed, we utilized the scaling estimator tool `aest` created by Crovella and Taqqu [9]. This tool aggregates the data points in the distribution and then plots the CD of the aggre-



■ Figure 2. *Hourly traffic volume to the World Cup Web site.*



■ Figure 3. *The size distribution of unique files: a) frequency; b) cumulative frequency; and c) tail.*

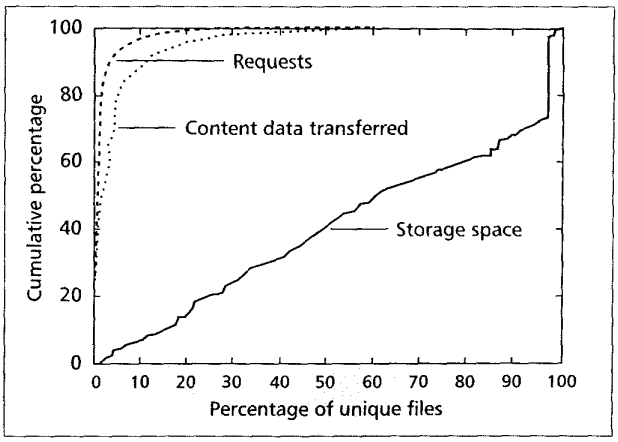
gated data set. If the distribution is heavy-tailed, the tails of each successive aggregated data set will be approximately parallel with slope approximately $-\alpha$ [9]. Using this tool we concluded that the unique file size distribution is heavy-tailed, and estimated α to be 1.37 [1].

File Referencing Behavior

In this section we analyze the World Cup workload for the presence of two important file referencing characteristics: temporal locality and concentration of references.

Temporal Locality — Temporal locality means that a recently referenced file is likely to be referenced again in the near future [2]. To measure the temporal locality we utilize the standard Least Recently Used (LRU) stack-depth analysis. This analysis works in the following manner. When a file is initially referenced it is added to the top of the LRU stack (position 1). All files currently in the stack are pushed down by one position. When a file is referenced again its current depth (i.e., position) in the stack is recorded, and then the file is moved back to the top of the stack. The other files in the stack are pushed down as necessary. Once the entire log has been analyzed, the record of the depths at which rereferences occurred is examined. Logs which exhibit a high degree of temporal locality will have a small average (or median) stack depth relative to the maximum stack depth (this is affected by the number of unique files accessed). Conversely, logs with a low degree of temporal locality will have a large mean (or median) stack depth.

For the World Cup workload the median stack depth was 106, which is significantly smaller than the number of unique files accessed (20,728). This indicates that the degree of temporal locality in the workload is quite strong. The degree of temporal locality is even stronger over shorter periods of time when user interest is focused on a specific subject (e.g., the current score of the match in progress).



■ Figure 4. *Concentration of references (cumulative distribution).*

Concentration of References — The second file referencing characteristic on which we focus is concentration of references. Many studies, including [2], have found that a nonuniform referencing pattern exists for files on the Web. This means that a small number of files on a Web site are extremely popular and receive most of the requests arriving at the site, while many unique files on a Web site are very unpopular.

Figure 4 shows the distribution of all client requests across the set of unique files available at the World Cup Web site. In this figure all of the unique files (x-axis) have been sorted in decreasing order by the number of references each received. The volume of content data each of these files generated in network traffic along with each file's storage requirements at the site were also computed. Figure 4 clearly shows that there is a concentration of references among a small subset of the unique files. For example, the top 10 percent (i.e., the most popular) unique files received 97 percent of all requests and generated 89 percent of the network traffic (i.e., content data) while occupying less than 7 percent of the storage space on the Web site. The top 1 percent of files received 75 percent of all requests, generated 46 percent of the network traffic, and consumed a mere 0.12 percent of the required storage space.

While a number of the files on the World Cup site were extremely popular, many were relatively unpopular. In fact, 9.2 percent of the unique files were requested only a single time. We refer to these files as *one-timers* [2]. The combined size of these one-timers was 98 Mbytes, or 31.8 percent of the com-

Duration	11:30pm–11:45pm, June 30th, 1998
Total requests	3,135,993
Avg requests/min	209,066
Total bytes transferred (Gbytes)	8.5
Avg bytes transferred/min (Mbytes)	580

■ Table 6. *A summary of access log characteristics (A-E workload).*

Response code	% of requests	% of data transferred
200 (Successful)	62.63	99.94
206 (Partial Content)	0.01	0.04
304 (Not Modified)	37.18	0.00
4xx (Client Error)	0.18	0.02
5xx (Server Error)	0.00	0.00
Other Codes	0.00	0.00
Total	100.00	100.00

■ Table 7. *A breakdown of server response codes (A-E workload).*

bined size of all unique files. This characteristic is of interest because of its obvious effect on caching; even over a long period of time and with an exceptionally heavy workload, some files on a Web site will not be referenced more than once. Thus, there is no benefit in caching these files. The spike in the storage space curve shown in Fig. 4 is due to the presence of a 61.2-Mbyte file (a `.zip` file). This file was also a one-timer and accounted for 19.8 percent of the storage space.

Analysis of a Peak Workload

Earlier we noted that much of the traffic to the World Cup site came in large bursts that occurred while football matches were in progress. In this section we analyze the busiest 15-min period from the overall World Cup workload. This period occurred from 11:30 p.m. until 11:45 p.m., June 30th, 1998. During this time penalty kicks were being used to determine the victor in a playoff match between Argentina and England. For the remainder of this article we shall refer to this subset of the overall World Cup workload as the A-E workload.

Table 6 reports some overall statistics on the A-E workload. Over 3 million requests were received by the World Cup site during the 15-min period. The average number of requests received per minute was over 19 times the average rate for the overall workload (Table 2). The average rate of data transfer per minute for the A-E workload was 13 times that of the overall workload. The busiest 1-min interval contained 215,241 requests.

The A-E workload differs from the overall workload in a number of ways. As one would expect, user interest in the A-E workload is focused on a much smaller set of files (i.e., those pertaining to the match between Argentina and England). This resulted in fewer unique files being accessed, a higher degree of temporal locality in the reference stream, and an increased concentration in the references to the most popular files. All of these characteristics suggest that caching should be able to reduce the site workload considerably.

Despite these characteristics, browser, proxy, and network caching did not significantly reduce the site workload (at least not to the degree we would expect). To understand why, we examine another characteristic of the A-E workload. Table 7 shows the breakdown of the server response codes from the A-E workload. This breakdown is quite different from the overall distribution provided in Table 3. The most significant change between the workloads is the percentage of Not Modified responses. In the A-E workload over 37 percent of all server responses were Not Modified, twice the percentage seen in the overall workload. This characteristic indicates that many of the clients are simply performing consistency checks to ensure that the World Cup files they have stored in their caches are still up to date. This is likely the result of users hitting the reload button on their browsers to check whether there has been a change in the status of the match. The large percentage of Not Modified responses (which do not contain any data) is the reason the average data transfer rate in the A-E workload did not increase as significantly as the average request rate.

Performance Implications

Previous work on Web server workload characterization [2] analyzed access logs that predated the widespread use of browsers with persistent (i.e., disk) caches, proxy caches, and (transparent) network caches. Today, Web server workloads have changed due to the growth of a Web caching architecture. In particular, many Web server responses are Not Modified and contain no content data. Our more detailed workload characterization study [1] identified several other ways in which caching is altering Web server workloads. For example, by analyzing user sessions we determined that caching, in some cases, is reducing the number of requests which might utilize a persistent connection. Further research is required to determine if fewer requests per session is a growing trend, and if so, what are the implications on Web servers and on HTTP.

A second observation regarding the effects of caching on Web server workloads was made earlier. From the perspective of Web server performance, the main benefit of client, proxy, and network caching is the reduction in workload at the server, particularly during periods of extreme user interest. Our results above indicate that current consistency mechanisms prevent Web servers fully benefiting from caching. In other words, when portions of the Web site's content is extremely popular, the site's servers do not seeing a substantial reduction in workload. Instead of responding to a large number of GET requests, the servers must respond to a large number of cache consistency requests (i.e., GET If Modified Since requests). These validation requests are an improvement since they reduce network bandwidth utilization; however, they do not reduce the number of requests that a server must handle.

One of the improvements in HTTP/1.1 [7] is the addition of an expiration mechanism to reduce the number of requests that reach a Web server. Many of the requests for consistency information in the World Cup workload were caused by the caching of static image files. Much of this traffic could have been eliminated if the expiration functionality of HTTP/1.1 had been utilized by the site and supported by the caches. However, the consistency mechanisms in HTTP/1.1 are not adequate in all situations (e.g., the modification patterns of some files are unpredictable, such as those that report the score of a football match in progress). Furthermore, it remains to be seen whether this functionality of HTTP/1.1 will meet the needs of content providers or a new, more automated system will be required. Several research efforts have looked at alternative cache consistency mechanisms, including [10]. If a new system is indeed required, it should include a method of propagating only the changes to the cache storing the old version of a file, as suggested by Mogul *et al.* [11].

During the World Cup tournament an estimated 13 million cumulative users visited the France '98 Web site. During this same period an estimated cumulative audience of 40 billion watched the matches on television. While the gap between Internet users and television audiences is partially due to restricted Internet access in some countries, the main reason is clear: audiences preferred live (high-quality) video to still images and text descriptions. This suggests that the integration of video and the Web may be necessary to reach a much greater portion of the world's population. Adding high-quality video to a Web site would have significant performance implications.

Summary, Contributions and Future Work

This article presents a detailed workload characterization study of the 1998 World Cup Web site. The data set analyzed in this study contained 1.35 billion requests collected over a three-month period, making this the largest Web server characterization study to date. Throughout the article emphasis was placed on comparing the characteristics of the World Cup workload to those observed in other Web server workloads.

The results of our study revealed that caching at Web clients and proxies and within the network is changing the workloads seen by Web servers. Current cache consistency mechanisms are the main cause of these changes. While these mechanisms are reducing the total bytes transferred by Web servers, they do not appear to significantly reduce the number of requests Web servers must handle during periods of extreme user interest in the content on those servers.

This article presents preliminary results on many different facets of the workload of the World Cup Web site. More in-depth analyses are needed on many of the topics discussed in this article. For example, existing Web server benchmarks may need to be adjusted to reflect current workloads. Such benchmarks are needed to more accurately estimate the performance of a particular server configuration. Additional workload characterization is needed, particularly of sites on an ongoing basis, to determine if the characteristics observed in this data set are present in others and to understand how these characteristics change over time. In order to perform more accurate analyses in the future, more precise measurements of (server) workloads are needed. This may involve changing the data collected in access logs (e.g., store finer-grained timestamps) or utilizing alternative methods of data collection (e.g., system instrumentation). Finally, as we alluded to earlier in this article, a more efficient cache consistency mechanism, preferably one that requires little human intervention, is needed to further the scalability of the Web.

Acknowledgments

The authors would like to thank all the people who made this work possible. The authors are particularly grateful to the people at EDS who provided the World Cup access logs; Christian Hostelet, HP Technical Director of the World Cup; Joel Dubedat and Jean Le Saint for providing information on the World Cup Web site architecture; Katey Kennedy and Robert Slinn for providing statistics on the television audiences for the World Cup; Mark Crovella of Boston University, for his assistance with the statistical analyses; and Paul Barford of Boston University, Jim Pitkow of Xerox PARC, Sharad Singal and Gary Herman of HP Labs, and the anonymous reviewers for their constructive comments on the article.

References

[1] M. Arlitt and T. Jin, "Workload Characterization of the 1998 World Cup Web Site," Tech. rep. HPL-99-35R1, Hewlett-Packard Labs, Sept. 1999.

[2] M. Arlitt and C. Williamson, "Internet Web Servers: Workload Characterization and Performance Implications," *IEEE/ACM Trans. Net.*, vol. 5, no. 5, Oct. 1997, pp. 631–45.

[3] P. Barford *et al.*, "Changes in Web Client Access Patterns," *World Wide Web J.*, Special Issue on Characterization and Performance Evaluation, 1999.

[4] M. Arlitt, R. Friedrich, and T. Jin, "Workload Characterization of a Web Proxy in a Cable Modem Environment," *ACM SIGMETRICS Perf. Eval. Rev.*, vol. 27, no. 2, Aug. 1999, pp. 25-36.

[5] K. Thompson, G. Miller, and R. Wilder, "Wide-Area Internet Traffic Patterns and Characteristics," *IEEE Network*, vol. 11, Nov./Dec. 1997, pp. 10–23.

[6] P. Barford and M. Crovella, "A Performance Evaluation of HyperText Transfer Protocols," *Proc. ACM SIGMETRICS '99*, Atlanta, GA, May 1999, pp. 188–97.

[7] R. Fielding *et al.*, "RFC 2616 – Hypertext Transfer Protocol – HTTP/1.1," June 1999.

[8] V. Paxson, "Empirically-Derived Analytic Models of Wide-Area TCP Connections," *IEEE/ACM Trans. Net.*, vol. 2, no. 4, Aug. 1994, pp. 316–36.

[9] M. Crovella and M. Taqqu, "Estimating the Heavy Tail index from Scaling Properties," *Methodology and Computing in Applied Probability*, vol. 1, Nov. 1, 1999.

[10] H. Yu, L. Breslau, and S. Shenker, "A Scalable Web Cache Consistency Architecture," *Proc. ACM SIGCOMM '99*, Cambridge, MA, Sept. 1999.

[11] J. Mogul *et al.*, "Potential Benefits of Delta Encoding and Data Compression for HTTP," *Proc. ACM SIGCOMM '97*, Cannes, France, Sept. 1997, pp. 181–94.

Biographies

MARTIN ARLITT [M] (arlitt@hpl.hp.com) is a research engineer for Hewlett-Packard Laboratories in Palo Alto, California. His research interests include network traffic measurement, computer systems performance analysis, and workload characterization. He graduated from the University of Saskatchewan in 1996. He is a member of ACM and USENIX.

TAI JIN (tai@hpl.hp.com) is a research engineer at Hewlett-Packard Laboratories in Palo Alto, California. He was a key contributor to the HP-UX networking projects in the late 1980s and was involved in the creation of the HP intranet. During that time he developed a tool which revolutionized network software distribution within the company. He also created several useful services accessible through the World Wide Web. His interests include networked systems, exploiting the Web, performance tuning, creating useful tools, and the stock market. He graduated from Columbia University in 1984.