Journal of Universal Computer Science, vol. 4, no. 9 (1998), 763-778
submitted: 1/9/98, accepted: 16/9/98, appeared: 28/9/98 © *Springer Pub. Co.*

Web Traffic Latency: Characteristics and Implications [1]

Binzhang Liu and Edward A. Fox
Virginia Polytechnic Institute and State University, Virginia, USA
{bliu,fox}@vt.edu

Abstract: It is critical to understand WWW latency in order to design better HTTP protocols. In this study we characterize Web response time and examine the effects of proxy caching, network bandwidth, traffic load, persistent connections for a page, and periodicity. Based on studies with four workloads, we show that at least a quarter of the total elapsed time is spent on establishing TCP connections with HTTP/1.0. The distributions of connection time and elapsed time can be modeled using Pearson, Weibul, or Log-logistic distributions. Response times display strong daily and weekly patterns. We also characterize the effect of a user's network bandwidth on response time. Average connection time from a client via a 33.6 K modem is two times longer than that from a client via switched Ethernet. We estimate the elapsed time savings from using persistent connections for a page to vary from about a quarter to a half. This study finds that a proxy caching server is sensitive to traffic loads. Contrary to the typical thought about Web proxy caching, this study also finds that a single stand-alone squid proxy cache does not always reduce response time for our workloads. Implications of these results to future versions of the HTTP protocol and to Web application design are discussed.
Keywords: World Wide Web, Latency, Characteristics, Implications

1 Introduction

In the past several years the World Wide Web has experienced tremendous growth, and has become the dominant source of Internet traffic. Now millions are using the Internet for WWW traffic. The Web has become an important tool to access information. The Graphics, Visualization, & Usability Center's (GVU) *8th WWW User Survey* reported that "84% of the users report that they considered access to the Web indispensable, nearly the same percentage as those who feel email is indispensable, and 85% of the users use it daily." [GVU 97]

HTTP/1.0 protocol is a simple request/response protocol, not designed for heavy use. In order to accommodate continually increasing WWW uses, HTTP needs to be effective and efficient. Based on studies of persistent connections and pipelining, HTTP/1.1 was designed to improve the performance. However, HTTP/1.1

[1] This is an extended version of a paper presented at the WebNet '98 conference in Orlando, Florida. The paper has received a "TOP Full Paper Award". Binzhang Liu was a graduate student in the Computer Science Department at Virginia Polytechnic Institute and State University. Now he works for Northern Telecom at Research Trangle Park, North Carolina, USA.

has deficiencies of complexity, poor extensibility, lack of generality and poor scalability [Nielson 97]. To solve these deficiencies, in July 1997, the World Wide Web Consortium (W3C) started the HTTP-NG project to design the next generation of the HTTP protocol that fulfills these requirements and resolves deficiencies of HTTP/1.1. The W3C group initiated a wide range of Web characterization studies. The HTTP-NG activity statement indicates "It is important to understand the actual system and how it is being used before attempting to optimize it." [W3C 97] In the past, though many studies have been characterizing Web traffic, little is known about the characteristics of Web latency. Web proxy caching is widely used in the Web system, but little is known about the effectiveness of proxy caching in improving Web latency. Research on these two issues should enhance the understanding of the overall Web system. Hence, the main objective of our study is to characterize Web response time. Specifically this paper adopt an experimental approach to answer the following questions:

- What kind of distribution does response time follow?
- Does proxy caching improve response time?
- What is the effect of network bandwidth on response time?
- How does response time change with different levels of traffic?
- What kind of distribution is followed by the count of embedded images in a page?
- How much elapsed time can be saved by persistent connections?
- Are there any periodic patterns for response time?

2 Related Work

Improving Web latency is a major research area. Many studies report that speed continues to be the number one concern of Web users. Some people even have called WWW the "World-Wide Wait". [GVU 95] Web users do not like to wait for a Web page that takes a long time to retrieve. Web latency comes from many sources, involving the HTTP protocol, Web server implementations, network bandwidth, characteristics of Web documents, characteristics of Web clients and network topology. The W3C Web Characterization Group emphasizes characterizing "the kinds of tasks actually performed using HTTP, and the kinds of documents that are retrieved". [W3C 97] It also is important to understand the nature of WWW latency in order to properly design, implement, and improve the WWW system. Long latency is observed from some Web sites. But the causes of long latency are not known yet. A study by Manley and Seltzer concluded that the latencies can not be explained by server over-loading and suggests that the bottleneck lies in the network [Manley and Seltzer 97]. Another study found that bandwidth-related delay may not account for much of the perceived latency [Panmanabhan and Mogul 94]. One study by Touch, Heidemann and Obraczka reported that most users see end-to end latencies of about 250 ms and concluded that the persistent connections do not substantially affect Web latency for the vast majority of users [Touch et al. 96]. Viles and French studied the availability and latency of the Web and suggested shorter client-side time-out intervals than those used for TCP connection establishment [Viles and French 95]. The studies by the NRG group at Virginia Tech have shown that 30% to 50% hit rates can be achieved by proxy caching [Williams et al. 1997]. Prefetching attempts to predict future accesses, and pre-loads documents into the cache; this may significantly improve network performance [Crovella and Barford 1997]. Another study found that Web resources change frequently and so suggested limits on the utility of simple caches [Douglis et al. 98]. To our knowledge there was no study to characterize Web response time.

3 Workloads Used in the Study

Four proxy log files are used in the study. The America Online workload (AOL) is about 40 minutes worth of proxy log file from America Online's proxy server. The Boston University workload (Boston) is a proxy log file in the Computer Science Department at Boston University. The VT Campus workload (VT) is a proxy log file from the Virginia Tech campus-wide proxy server. VT Library (VTLIB) is a proxy log file from the Newman Library proxy server at Virginia Tech. Table 1 describes these workloads. For additional details, see [Abdulla 98].

Table 1: Workloads used in the study

Workloads	Periods	Total Accesses
`America Online`	Dec. 1, 1997	825,602
`Boston University`	Jan. 27 to Feb. 8, 1995	522,928
`VT Campus`	Sep. 28 to Oct. 5, 1997	696,975
`VT Library`	Sep. 28 to Oct. 5, 1997	1,014,875

4 Experiments

In our experiments, we use Webjamma [Johnson 97] to re-play log files. Webjamma replays a workload by reading a log file of URLs, sending HTTP queries in those logs, and timing the transfers. Since Webjamma just discards the transfered data, the only delay is from the transfer. Webjamma submits a configurable number of HTTP requests in parallel. In the proxy caching experiments, we used a modified version of squid 1.1.6 [Johnson 97]. *Connection time* is defined as the time between when a browser tries to set up a TCP connection to a Web server or proxy server and the first byte is received by the browser. *Transfer time* is the time between when a browser receives the first byte from a Web server or proxy server and the browser receives the last byte. *Elapsed time* is equal to connection time plus transfer time.

By varying five variables - proxy option, connection type, network bandwidth, number of Webjamma processes, and time ten experiments were designed to allow explantation of the problems listed before. The first factor, proxy option, is either none, where the HTTP queries are sent directly to the original server, or one, where the HTTP queries are sent to a proxy cache, which then sends them directly to the server. The second factor, connection type, represents the type of HTTP connections (persistent or non-persistent) used. The third factor, bandwidth, reflects the type of network connection between the the browser and Internet. The fourth factor, number of Webjamma processes, is used to simulate different load levels on the proxy server. The last factor is time. Re-playing log files at different hours of a day and different days of a week allows us to examine the periodicity of Web response time.

Table 2 lists detailed information for each experiment including workload used, sample size, network connection, number of Webjamma processes and proxy option.

Table 2: Detailed description of the ten experiments

Experiment	Workload	Sample Size	Network Connection	No. of Webjamma Processes	Proxy Option
1	AOL, Boston, VT, VTLIB	Full log	Ethernet	20	No
2	VT	10K	Modem	1	No
3	VT	Full log	Ethernet	1, 20	Yes
4	VT	10K	Modem	1	Yes
5	VT	10K	Ethernet	1, 10, 20 , 30 50, 70, 90	Yes
6	AOL, VT	1K	Ethernet	1	No
7	AOL, VT	1K	Modem	1	No
8	AOL, VT	1K	Ethernet	1	Yes
9	AOL, VT	1K	Ethernet	1	Yes
10	VT	10K	Ethernet	20	No

5 WWW Response Time Without a Proxy Cache

5.1 Response Time Via Switched Ethernet Without a Proxy Cache

From the data resulting from experiment one, average connection and elapsed time were calculated using the PERL scripts developed during this study. Table 3 gives average connection time, average elapsed time and ratio of connection time to elapsed time for the four workloads without a proxy. Table 3 shows that average connection time ranged from a low of 0.27 to a high of 0.54 seconds, and average elapsed time ranged from a low of 0.57 to a high of 2.0 seconds. The ratio of average connection time to average elapsed time ranged from a low of 0.27 to a high of 0.69. In all workloads, this ratio is higher than 0.25. It indicates that at least a quarter of the total elapsed time was spent in setting up the connection.

To compare the response time of local and remote accesses, the VT Campus and VT Library workloads were split into local accesses (.vt.edu domain) and remote accesses (domain other than .vt.edu). Table 4 lists the average connection time

Table 3: Average connection time, elapsed time and ratio of connection time to elapsed time

Workloads	Connection Time	Elapsed Time	Ratio
America Online	0.54	1.98	0.27
Boston University	0.39	0.57	0.69
VT Campus	0.27	0.73	0.36
VT Library	0.32	0.88	0.36

Note: All time is in seconds.

and average elapsed time for local and remote accesses. Both average connection time and average elapsed time for local accesses are shorter than for remote accesses.

Table 4: Response time for local and remote accesses

Workload	Average Connection Time		Average Elapsed Time	
	Local	Remote	Local	Remote
VT Campus	0.20	0.31	0.47	0.76
VT Library	0.22	0.34	0.34	0.98

Note: All time is in seconds.

Figure 1 shows that over 90% of the time the connection time is less than one second. Figure 2 shows that about 80% of the time the elapsed time is less than one second. The cumulative distributions of connection time and elapsed time follow Pearson distributions except the cumulative distribution of connection time of the VT campus workload follows a Weibul distribution. Tables 5 and 6 list the parameters of distributions for connection time and elapsed time. See [Law and Vincent 94] for a detailed description of various distribution functions.

Table 5: Parameter estimates of distribution of connection time

Workloads	Distribution	α_1	α_2	β
America Online	Pearson	0.88	2.90	1.00
Boston University	Pearson	1.03	3.53	1.00
VT Campus	Weibul	0.81	N/A	0.33
VT Library	Pearson	2.97	0.81	1.00

Table 7 lists the connection time under various cumulative frequencies. For all workloads, 99% of the time, the connection time is less than 10 seconds. This

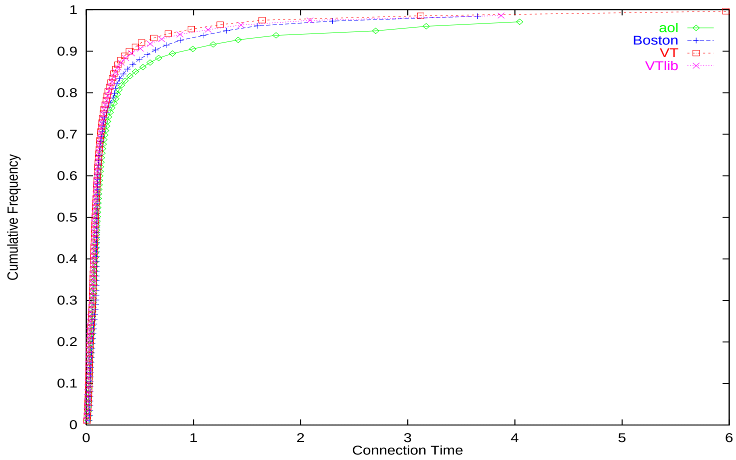

Figure 1: Cumulative distribution of connection time

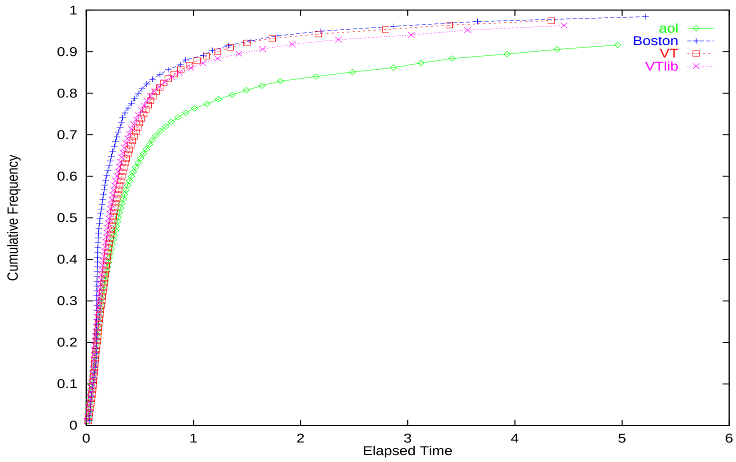

Figure 2: Cumulative distribution of elapsed time

result suggests that a Web client default timeout value should not be longer than 10 seconds.

Table 6: Parameter estimates of distribution of elapsed time

Workloads	Distribution	α_1	α_2	β
America Online	Pearson	2.35	1.15	1.00
Boston University	Pearson	3.36	1.16	1.00
VT Campus	Pearson	3.81	1.54	1.00
VT Library	Pearson	3.18	1.28	1.00

Table 7: Connection time for various cumulative frequencies

Workloads	90%	99%	99.9%
America Online	0.90	9.69	22.56
Boston University	0.62	5.14	14.12
VT Campus	0.40	3.74	13.98
VT Library	0.46	4.53	22.58

Note: All time is in seconds.

5.2 Response Time Via a 33.6 K Modem Without a Proxy Cache

To examine the effects of network connection on response time, in experiment two we chose a subset of the VT Campus proxy workload and re-played it using Webjamma via a 33.6 K modem connection. The average connection time from a client via 33.6 K modem network connection is 0.59 seconds (2.2 times longer than that from a client using switched Ethernet). Average elapsed time from a client via a 33.6 K modem connection is 2.33 seconds (3.19 times longer than that via switched Ethernet). Table 8 lists the connection and elapsed time under

Table 8: Connection time of modem users under various cumulative frequencies

Workloads	90%	99%	99.9%
VT Campus	0.79	4.46	13.12

Note: All time is in seconds

various cumulative frequencies. Table 8 shows that 99% of the time, connection time via a 33.6 K modem is less than 4.5 seconds.

6 Web Response Time with a Proxy Cache

In experiments three and four, the VT campus workload was re-played to a proxy server in the Computer Science Department. The results are summarized

Table 9: Response time of proxy caching using VT campus workload

Number of Processes	Network Connection	Average Connection Time	Average Elapsed Time	Ratio1	Ratio2
1	Ethernet	0.47	0.86	1.76	1.18
20	Ethernet	0.81	1.43	3.03	1.96
1	Modem	0.87	2.50	1.48	1.07

Note: All time is in seconds.

in Table 9, where ratio1 is the ratio of the connection time with a proxy to the connection time without a proxy. Ratio2 is the ratio of the elapsed time with a proxy to the elapsed time without a proxy.

Table 9 shows that for Ethernet users, if proxy traffic load is low (e.g., in the one process case) average connection time is about 1.76 times longer and the elapsed time is about 1.18 times longer than that without a proxy. If proxy traffic load is heavy (e.g., in the 20 process case), average connection time is about three times longer and the average elapsed time is about two times longer than that without a proxy. For a 33.6 K modem user, with a proxy the average connection time is 1.48 times longer, and average elapsed time is almost the same as the average elapsed time without a proxy. These results indicate that proxy caching increases both average connection time and elapsed time. It also indicates that increased traffic loads degrade the performance of a proxy caching server.

Table 10 lists the connection time with a proxy under various cumulative frequencies.

Table 10: Connection time of VT workload with a proxy under various cumulative frequencies

Option	90%	99%	99.9%
`proxy`	1.4	12.2	25.2

Note: All time is in seconds.

Figure 3 shows that over 80% of the time the connection time is less than one second and about 80% of the time the elapsed time is less than 1.3 seconds. Table 11 lists the connection time of VT campus proxy traffic for local and remote accesses under various cumulative frequencies.

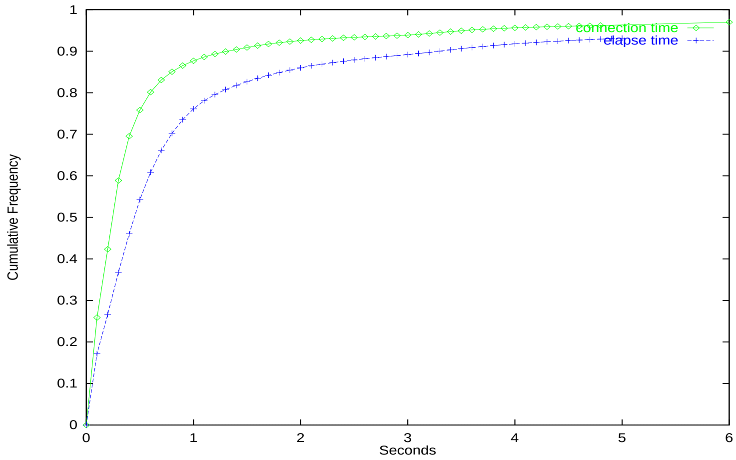

Figure 3: Cumulative distribution of connection time and elapsed time

Table 11: Connection time of VT campus proxy workload for local and remote accesses under various cumulative frequencies

Option	90%	99%	99.9%
`local`	1.2	2.2	9.9
`remote`	1.4	12.3	25.3

Note: All time is in seconds.

7 Response Time and Proxy Traffic Loads

In experiment five, the number of parallel Webjamma processes ranges from 1 to 90, whence the corresponding completed requests per second range from a low of 0.65 to a high of 20.83. Earlier, Slothouber [Slothouber 96] studied Web server performance using a queuing model and found that before a server reached its theoretical upper bound load, the response time was almost constant (a mere fraction of a second). When the server approached full utilization, response time grew asymptotically toward infinity. Contrary to his findings, our results show that proxy server performance is generally sensitive to traffic load. Even at very low request arrival rates, when the number of processes increases from 1 to 10 (equivalent to a range of 56,471 accesses/day to 436,363 accesses/day), the average connection time increases by 37%, and the average elapsed time increases by 38%. In Figure 4, Ratio represents the ratio of the average connection time to the average elapsed time. Figure 4 shows that response time increases with an increase in the request arrival rate, but when that rate goes beyond 16 per second (1.38 million per day), the response curve becomes steep and the ratio of average connection time to average elapsed time is increasing.

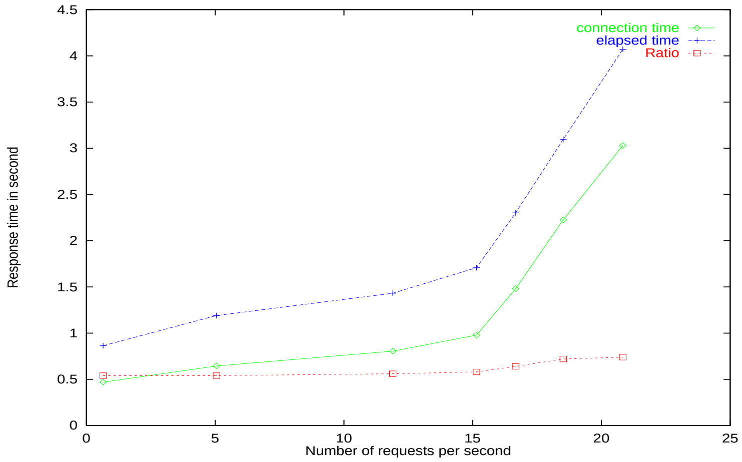

Figure 4: Response curve of response time to proxy traffic load

8 WWW Response Time With Persistent Connections

To examine how persistent connections affect response time, experiments six, seven, eight and nine were conducted. In this study, we simulate persistent connections from clients to Web servers for a page, which we define as having multiple objects including an HTML file and in-line image files. Connection time of a page for non-persistent and persistent connections is the connection time of the first object in a page. Elapsed time for a non-persistent connection is calculated by summing the elapsed times of all objects in a page. Elapsed time of a page for a persistent connection is the connection time of the HTML file plus the sum of the transfer times of each object in a page.

8.1 Distribution of Number of Embedded Images

The average number of unique in-line image files in a page for the AOL workload is 3.33; for the VT workload it is 2.56. About one third of our HTML pages do not have an embedded image. 65% of the pages contain at least one unique embedded image in the AOL workload. In the VT workload, about 46% of the pages do not have an embedded image and 54% of the pages contain at least one unique embedded image. The maximum number of unique embedded images in the AOL workload is 39; for the VT workload it is 36. Distributions of the number of unique embedded images in a page follow a Random Walk, which is a type of heavy tailed distribution. Table 12 lists the parameter estimates of the distributions. Figure 5 shows that for 80% of the pages the number of unique embedded images is less than 5, and for over 10% of the pages the number of unique embedded images is over 10.

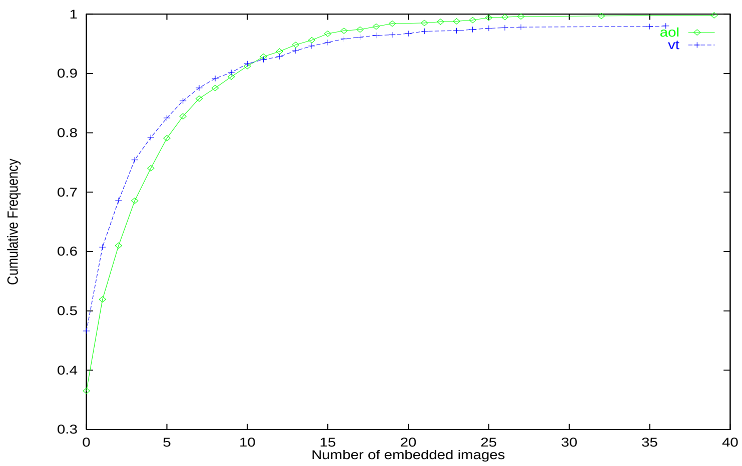

Figure 5: Distribution of number of unique embedded images in a page

Table 12: Parameter estimates of distribution of number of embedded images in a page

Workloads	Distribution	α	β
America Online	Random Walk	0.13	1724.8
VT Campus	Random Walk	2.13	173.06

8.2 Percentage of Elapsed Time Saving From Persistent Connection

Table 13: Percentage of elapsed time saving from persistent connection

Workloads	10BaseT Switched Ethernet		33.6 K modem	
	non-proxy	proxy	no-proxy	proxy
AOL	22.78	35.52	23.87	29.66
VT	26.70	49.97	30.51	36.38

According to Table 13, the percentage of elapsed time saving from persistent connections ranges from a high of 49.97% (10baseT switched Ethernet with a proxy for AOL workload) to a low of 22.78% (10baseT switched Ethernet without a proxy for AOL workload). These results show that elapsed time savings from using persistent connections are significant.

9 Periodicity of WWW Response Time

A study by Abdulla et al. found that Web traffic displayed strong periodicity [Abdula et al. 97]. In order to examine whether periodicity exists for Web response time, experiment ten was conducted. Two time series - average connection time and average elapsed time were calculated from the results of experiment ten. These two time series are used in the analysis to identify periodicity of Web response time.

9.1 Plot of the Response Time Series

To visualize the periodicity of Web response time, a plot of average connection time (ct) and average elapsed time (et) is drawn. Figure 6 shows the time plot.

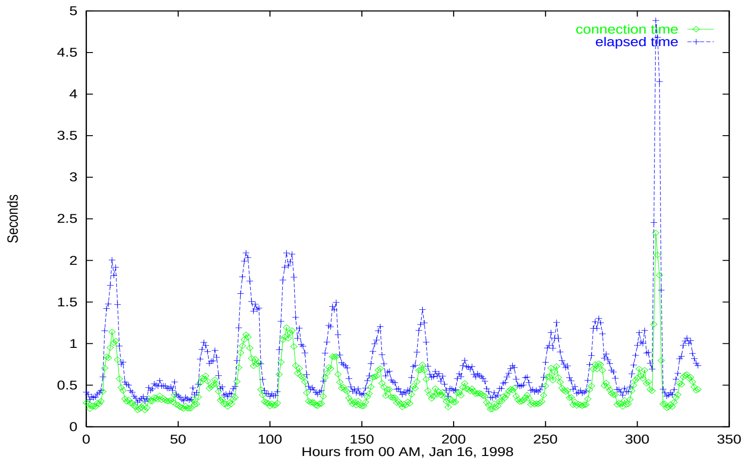

Figure 6: Time plot of response time series

In Figure 6, ct and et designate average connection time and average elapsed time. Figure 6 is the time plot of connection time and elapsed time over a 333 hour period. By counting the main peaks, it is noticed that there are about 14 peaks in the plot. This suggests that a period is approximately 24 hours (one day).

9.2 Spectral Analysis

Spectral analysis is often used in looking for periodicity in data. In this study, the SAS spectra procedure is used to produce estimates of the spectral densities of time series data, and then these estimates are used to obtain periodograms. The Fisher's Kappa test is also specified to test white noise. The Fisher's Kappa statistics for both series are larger than the 5% critical value 7.2, so the null hypothesis that the time series data are white noise is rejected. Figure 7 shows the plot of periodograms of both time series (connection time and elapsed time).

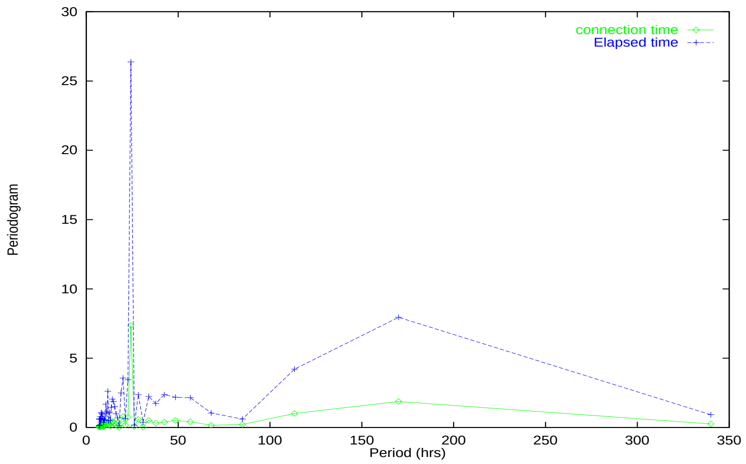

Figure 7: Plot of periodogram by period

From Figure 7 the periodicity in the time series is very clear. The peaks in the figure correspond to the major periods in the data. One main peak at around 24 hours corresponds to the daily cycle. Another small peak at around 168 hours corresponds to the weekly cycle.

9.3 Correlation of Response Time and Web Traffic

To find out whether this periodic pattern is related to the Web traffic, we extracted hourly accesses to the VT campus proxy from its log files for the period of Jan. 15, 1998 to Jan. 29, 1998. The regression analysis for average connection time and average hourly accesses to the VT campus proxy and the Computer Science courseware Web server ei.cs.vt.edu are shown in Tables 14 and 15.

Coefficients of the access count variable in Tables 14 and 15 are significant. The Web response time is highly correlated with the VT campus proxy traffic and the Web traffic to the ei.cs.vt.edu server during the same period (Jan. 15 to Jan. 29, 1998). It is interesting that the Web response time is more highly correlated with the Web traffic to the ei.cs.vt.edu than the VT campus proxy traffic. Since the coefficient of the access count variable is positive, an increase in the access count will lead to longer response time. This result indicates that Web traffic is at least partially responsible for the long Web latency.

Table 14: Regression analysis of the connection time and hourly accesses to the VT campus proxy

Variables	Coefficient	T ratio
Constant	0.31	6.64
Accesses	8.6E-05	3.89

Note: R^2=0.41, F=15.11

Table 15: Regression analysis of the connection time and hourly accesses to the ei.cs.vt.edu Web server

Variables	Coefficient	T ratio
Constant	0.16	4.06
Accesses	0.0003	8.24

Note: R^2=0.76, F=67.84

10 Conclusion

In this study using four workloads we conducted ten experiments. We characterized Web response time and examined effects on response time of proxy caching, network bandwidth, traffic loads, persistent connections for a page, and periodicity. The following conclusions are drawn from the results above.

1. Connection time is a major component of total elapsed time for HTTP/1.0. In all four workloads, for a non-persistent connection, the ratio of connection time to elapsed time is higher than 0.25. At least a quarter of the total elapsed time is spent in establishing a network connection.
2. Distributions of connection time and elapsed time follow Pearson, Weibul or Log-logistic distributions. For both low speed modem users, and users with switched Ethernet connection, 99% of the time, connection time will be less than 10 seconds. This result suggests that Web client timeout values should not be higher than 10 seconds.
3. Contrary to popular thoughts, results from several experiments indicate that at least in our cases proxy caching does not necessarily result in shorter Web response time. In our cases, it almost always increases response time. In order to achieve better performance in terms of response time, proxy systems should be better designed, and relevant HTTP protocol changes on proxy caching should be made. We found in the proxy log file over 10% accesses are "not modified" (304 status code). Proxies can be designed to allow for a distribution model to validate cache contents and hence "conditional Get" will not be necessary.
4. The response curve of a proxy server (squid 1.1.6) to traffic load is not flat even at a very low request rate. Response time increases with an increase in request arrival rate. This indicates that a proxy caching server is sensitive to traffic load. Contrary to our empirical results, using queuing theory Slothouber found that as the load on the Web server increases the time required to serve a file increases very gradually (almost imperceptibly) [Slothouber 96].

This may indicate that Web server performance can't be correctly analyzed using a simple queuing model.

5. The results show that using persistent connections can achieve a significant performance improvement in terms of elapsed time. A 23% to 50% elapsed time saving can be achieved by using persistent connections. The amount of elapsed time saving from persistent connections is a function of the number of embedded images in a page. For the VT workload, under persistent connections proxy caching can lower elapsed time.
6. Speed of network connection has a significant effect on the response time. Connection time from a client with 33.6 K modem is 2.2 times longer than that from a client with a switched Ethernet connection. For modem users, connection time still constitutes a large part of total elapsed time. In our case, connection time is still 25% of total elapsed time for modem users without a proxy. It is not necessarily the case that the percentage of time saved from using persistent connections for a page via a modem is less than that via high speed Ethernet. Contrary to Touche's result [Touch et al. 96], this result suggests that even for modem users, migrating client browsers to HTTP/1.1 compatible versions can achieve a significant response time improvement.
7. The distribution of the number of unique embedded images in a page is not normal; rather it follows a Random Walk distribution. The average number of unique in-line image files in a page for the AOL workload is 3.33. The average number of unique in-line image files in a page for the VT workload is 2.56. Over one third of pages do not contain an embedded image. In the HTTP/1.1 performance study by a W3C group [W3C 1997], they used a page with 42 in-line GIF images. This test page is not representative, so their HTTP/1.1 performance results may be biased.
8. There exist daily and weekly periodic patterns for response time. The periodicity of Web response time indicates that the present Web system is overloaded. This result also provides the rationale for timing of pre-fetching, Web crawler and Web indexing activities. These activities should be scheduled to run at low points in the cycle.
9. Web response time has a very high variability. Average connection time at the peak point is about 3.28 times longer than at the bottom point in the daily cycle. We found that Web response time is highly correlated with Web traffic. This suggests that Web traffic is at least partially responsible for high latency.

Acknowledgments

Members of the VT Network Research Group (NRG) provided helpful comments on the manuscript. NSF grants CDA-931261 and NCR-9627922 partially supported this work. IBM donated equipments used to collect and process the traffic log files.

References

[GVU 97] Graphics, Visualization, & Usability Center.: *GVU's 8th WWW User Survey*, College of Computing, Georgia Institute of Technology, Atlanta, GA October, 1997. URL: http://www.cc.gatech.edu /gvu/user_surveys/survey-1997-10/.

[Nielson 97] Nielsen, H. F.: *W3C Summary of Briefing Package for HTTP-NG Project*, August, 1997. URL: http://www.w3.org/Protocols/HTTP-NG/BriefSummary.html.

[W3C 97] W3C: *"HTTP-NG - The Next Generation"*, August, 1997. URL: http://www.w3.org /Protocols/HTTP-NG/Activity.html.

[GVU 95] GVU: "Graphics, Visualization, & Usability Center. *GVU's 5th WWW User Survey"*, College of Computing, Georgia Institute of Technology, Atlanta, GA, October, 1995. URL: http://www.cc.gatech.edu /gvu/user_surveys/survey-10-1995/.

[Manley and Seltzer 97] Manley, S. and Seltzer, M.: *"Web Facts and Fantasy, W3C HTTP-NG Reading List"*. August, 1997. URL: http://www.w3.org/Protocols/HTTP-NG/ReadingList.html.

[Panmanabhan and Mogul 94] Panmanabhan, V. N. and Mogul, J.: *"Improving HTTP Latency"*, 1994, URL: http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/DDay/mogul/HTTPLatency.html

[Touch et al. 96] Touch, J., Heidemann, J. and Obraczka. K.: *"Analysis of HTTP Performance"*, Information Sciences Institute, University of Southern California, August, 1996. URL: http://www.isi.edu/lsam/publications/http-perf/index.html.

[Viles and French 95] Viles, C. L. and French, J. C.: *"Availability and Latency of World Wide Web Information Servers", Computing System, 8, 1(1995):61-91.*

[Williams et al. 1997] Williams, S., Abrams, M., Standridge, C. R., Abdulla, G., & Fox, E. A.: *"Removal Policies in Network Caches for World-Wide Web Documents", Proceedings, ACM SIGCOMM,* Cannes, French Riviera, France(1997), 293-305.

[Crovella and Barford 1997] Crovella, M. and Barford, P.: *"The Network Effects of Prefetching", Technical Report TR-92-002*, Computer Science Department, Boston University, February 1997. URL: http://www.cs.bu.edu/faculty/crovella/papers.html.

[Douglis et al. 98] Douglis, F., Feldmann, A. and Mogul, J.: *"Rate of Change and other Metrics: A Live Study of the World Wide Web",* URL: http://www.w3.org/Protocols/HTTP-NG/ReadingList.html.

[Abdulla 98] Abdulla, G.: *Analysis and Modeling of World Wide Web Traffic,* May, 1998, Ph.D Dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.

[Johnson 97] Johnson, T.: *"Webjamma",* July, 1997, Blacksburg, VA. URL: http://www.cs.vt.edu /~chitra/webjamma.html.

[Johnson 97] Johnson, T.: *"Squid and Harvest Modifications",* July, 1997 Network Research Group, Virginia Tech. URL: http://www.cs.vt.edu/~chitra/squid_harvest.html.

[Law and Vincent 94] Law, A. M. and Vincent, S.: *"UniFit I, II, User's Guide".* 1994 Averill M. Law & Associates, Tucson, AZ 85717, 1994.

[Slothouber 96] Slothouber, L. P.: *"A Model of Web Server Performance",* 1996, StarNine Technologies Inc. URL: http://www.starnine.com/webstar/overview.html.

[Abdula et al. 97] Abdulla, G., Nayfeh, A. and Fox, E. A.: *"Modeling Correlated Proxy Web Traffic Using Fourier Analysis", Technical Report TR-97-19*, Blacksburg, VA, November, 1997.

[W3C 1997] W3C: *"Network Performance Effects of HTTP/1.1, CSS1, and PNG"*, June 1997, URL: http://www.w3.org/Protocols/HTTP-NG/ReadingList.html.