Web Response Time and Proxy Caching

Binzhang Liu, Ghaleb Abdulla, Tommy Johnson, Edward A. Fox
Computer Science Department
Virginia Polytechnic Institute and State University
Virginia, 24061-0106
Phone: (540)231-3615, Fax : (540)231-6075
email: {bliu,abdulla,tjohnson,fox}@vt.edu

Abstract: It is critical to understand WWW latency in order to design better HTTP protocols. In this paper we characterize Web response time and examine effects of proxy caching on response time. We show that at least a quarter of the total elapsed time is spent in setting up TCP connections. 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. Contrary to the typical thought about Web proxy caching, this study finds that a single stand alone proxy cache does not always reduce response time. Implications of these results to the HTTP-NG protocol and Web application design also are discussed in the paper.

1 Introduction

Recently, in an effort to design the next generation HTTP protocol (HTTP-NG), the W3C group initiated a wide range of Web characterization studies. In the HTTP-NG activity statement they stated "It is important to understand the actual system and how it is being used before attempting to optimize it." [1] Though many studies have been characterizing Web traffic, little is known about the characteristics of Web latency. While Web proxy caching is widely used in the Web system, little is known about the effectiveness of proxy caching. Characterization of Web response time will help understand the nature of WWW latency and give guidance to designers of Web system applications and HTTP-NG.

Research on the effectiveness of proxy caching is very active. A study at Virginia Tech has shown that hit rates of 30% to 50% can be achieved by a caching proxy [2]. However, one study found that Web resources change frequently and suggests that a simple cache may be of only limited utility [3]. In this study, we explore the effect of speed of network connection and using proxy caching on latency. Specifically this study will answer the following questions:

- What kind of distribution does response time follow?
- Does proxy caching improve response time?
- How does response time change with different levels of traffic?
- What is the effect of network bandwidth on response time?

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. The 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.

2 Experiments

Five experiments were conducted in the study. The first four experiments were run using two variables, each at two levels. See Table 1. The first factor, bandwidth, reflects the type of network connection between the the browser and Internet. The second factor, Proxy Cache, 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.

Each of the first four experiments consists of replaying four HTTP log files using Webjamma [4]. See Table 2 for a list of the log files for 4 workloads considered. In the proxy caching experiments, we used a modified version of squid 1.1.6.

Webjamma replays a workload by reading a log file of URLs, sending HTTP queries, and timing the transfer. Since Webjamma just discards the transfered data, the only delay is from the transfer. Webjamma maintains a configurable number of HTTP requests in parallel.

In the fifth experiment we vary the number of parallel Webjamma processes accessing the proxy to simulate different load levels on the proxy server. In this experiment, we use a subset of the VT Campus workload.

Table 1: Variables and levels used in the study

Variable	Low level	High level
Bandwidth	33.6K modem	Switched 10baseT Ethernet
Proxy Cache	None	One

Table 2: Workloads used in this study.

Workloads	Periods	Total Accesses
`America Online`	Dec 1 (several hours), 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

3 Response time without a proxy

The four workloads were replayed, recording the connection time and elapsed time. Average connection time ranged from a low of 0.2660 to a high of 0.7313 seconds. Average elapsed time ranged from a low of 0.5687 to a high of 1.9824 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, indicating that at least a quarter of the total elapsed time was spent setting up the connection. Although HTTP/1.1 supports persistent connections we found that most of the HTTP transactions in our VT Campus network were HTTP/1.0 transactions (about 89% still use HTTP/1.0). This may be one of the reasons why this ratio is high.

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

Workloads	Connection Time	Elapsed Time	Ratio
American Online	0.5385	1.9824	0.27
Boston University	0.3931	0.5687	0.69
VT Campus	0.2660	0.7313	0.36
VT Library	0.3181	0.8828	0.36

Note: All times are in seconds

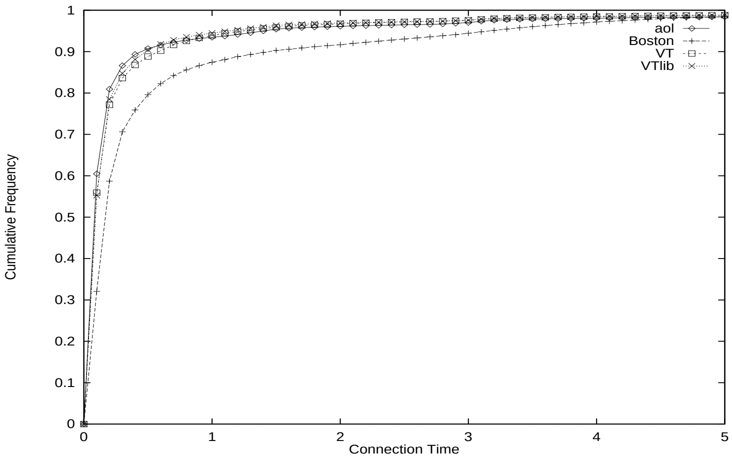

Figure 1: Cumulative Distribution of Connection Time

Figure 1 and 2 show that over 80% of the time the connection time and the elapsed time are less than one second. Table 4 lists the connection time given various cumulative frequency values. The results show that 99% of time, connection time is less than 10 seconds. This result suggests that a Web client's default timeout value should not be higher than 10 seconds. It is found that all

Table 4: Connection time for various cumulative frequency

Workloads	%90	%99	%99.9
American 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 times are in seconds.

of the above response times follow an Pearson distributions except the cumulative distribution of connection time of VT Campus workload follows Weibul distribution.

4 Response time with a proxy

The second two experiments were then performed, replaying the workloads through a proxy server this time to quantify the performance changes due to proxy caching.

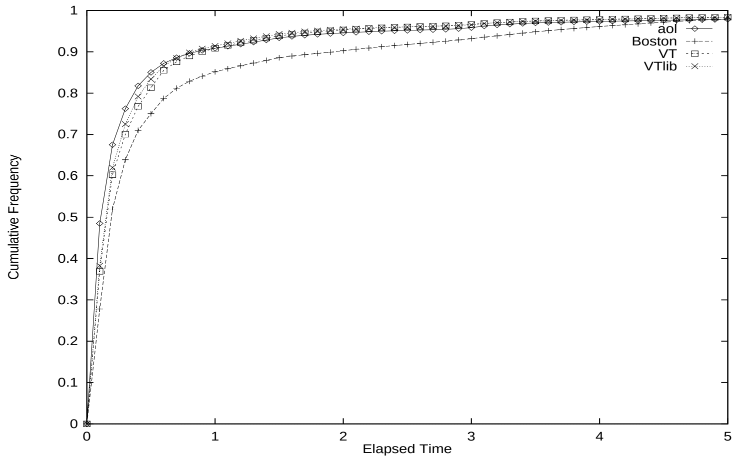

Figure 2: Cumulative Distribution of Elapsed Time

Table 5: Response time of proxy caching using VT Campus workload

Network Connection	Average Connection Time	Average Elapsed Time	Ratio of Connection Time to Elapsed Time
`switched Ethernet`	0.4689	0.8630	0.5423
`modem`	0.8668	2.4970	0.3471
`ratio`	1.8486	2.8934	N/A

Notes: All time is in second

Contrary to our expectations, Table 5 shows that with the proxy the connection and elapsed times are longer than those with no proxy for both switched Ethernet and modem connections.

5 Response time and proxy traffic loads

In this experiment, the number of parallel Webjamma processes ranges from 1 to 90, and the corresponding completed requests per second range from a low of 0.65 to a high of 20.83. The results are presented here.

Figure 3 shows that response time increases with an increase in the number of parallel Webjamma processes (equivalent to an increase in the request arrival rate). When the number of parallel Webjamma processes exceeds 50 (equivalent to 16 requests per second or 1.38 million requests per day), the response curve becomes steeper. This result shows that proxy server performance is very sensitive to traffic load and as it becomes overloaded the performance will degrade quickly.

6 Effect of network connection on response time

To examine the effect of network connection on response time, a subset of the VT Campus workload was re-played using two different network connections: PPP via a 33.6 K modem and switched 10baseT Ethernet. Table 6 lists the average response time and elapsed time from the experiment. In the Table 6, all time are in seconds, and the ratio is defined as the response time via a 33.6 K modem divided by the response time via switched Ethernet.

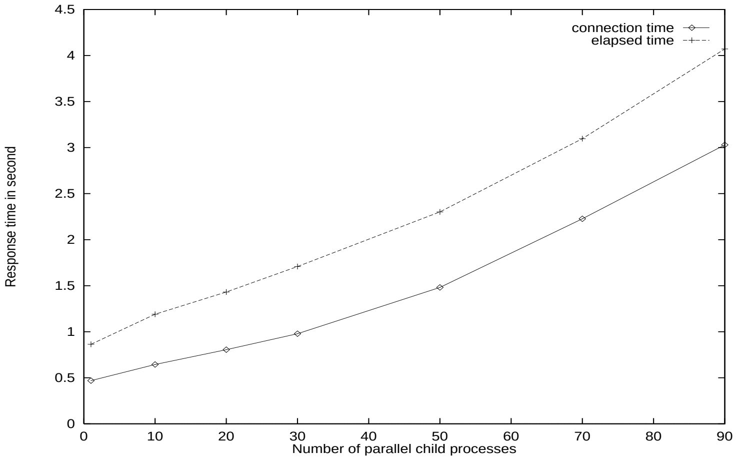

Figure 3: Response curve of response time to proxy traffic load

Table 6: Response time with different network connection

Network Connection	Average Connection Time	Average Elapsed Time	Ratio of Connection Time to Elapsed Time
Switched Ethernet	0.2660	0.7313	
33.6k modem	0.5870	2.3342	0.2515
Ratio	2.2068	3.1920	N/A

Average connection time from a client via a 33.6 K modem is 2.2 times longer than that from a client via switched Ethernet. Average elapsed time from a client via a 33.6 K modem connection is 3.2 times longer than that via switched Ethernet connection. Although Table 6 shows that end user network connection speed has a significant effect on response time, we expected the difference in performance to be higher. This shows that there is a fixed delay introduced in both cases from servers and network load. It is found that connection time via modem follows a Log-logistics distribution and elapsed time via modem follow Pearson distribution. Table 7 lists the connection and elapsed time given various cumulative frequency values.

Table 7: Response time of modem users under various cumulative frequency

Response Time	%90	%99	%99.9
Connection Time	0.79	4.46	13.12
Elapsed Time	4.67	20.09	59.85

Notes: All time is in second.

7 Recommendations

Based on the above results, we give the following recommendations:

- For both low speed modem users and users via switched Ethernet, 99% of the time, connection time will be less 10 than seconds. This result suggests that the Web client timeout value should not be higher than 10 seconds.

- Speed of network connection has a significant effect on the connection time. Average connection time from a client via 33.6 K modem is two times longer than that from a client via switched Ethernet. In both switched Ethernet and modem cases, connection time is at least a quarter of the total elapsed time. Contrary to Touch's result [5], our study suggests that even for modem users, migrating client to HTTP/1.1 browsers can achieve a significant response time improvement.

- Simple proxy caching does not always reduce response time. For switched Ethernet users, connection time with the proxy is 1.8 times longer than with no-proxy. For modem users, connection time with proxy is 1.5 times longer than that with no-proxy. Average elapsed time with proxy is a little bit longer than that with no-proxy for both switched Ethernet and modem users. The proxy caching should be better designed, and relevant HTTP protocol changes regarding proxy caching should be made. We found in the proxy log file over 10% accesses are "not modified" (304 status code). Proxy can be designed to allow for a distributed model to validate cache contents and hence "conditional Get" will not be necessary.

- In our experiment, when traffic load is above 16.7 requests per second, performance of a proxy degrades quickly. It shows that the performance of a proxy caching server is very sensitive to proxy loads. For a proxy server, arrival rate should stay below a suitable threshold in order to achieve acceptable performance.

8 Acknowledgments

Members of the VT NRG provided helpful comments on the manuscript. NSF grants CDA-931261 and NCR-9627922 partially supported this work. IBM donated the equipment used to collect the traffic log files.

References

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

[2] Stephen Williams, Marc Abrams, Charles R. Standridge, Ghaleb Abdulla, Edward A. Fox. *Removal Policies in Network Caches for World-Wide Web Documents. In Proceedings, ACM SIGCOMM*, August, 1997.

[3] Fred Douglis, Anja Feldmann and Jeffrey Mogul. *Rate of Change and Other Metrics: A Live Study of the World Wide Web.* URL:http://www.w3.org/Protocols/HTTP-NG/ReadingList.html.

[4] Tommy Johnson. *Webjamma.* URL:http://www.cs.vt.edu/c̃hitra/webjamma.html.

[5] Joe Touch, John Heidemann, and Katia Obraczka. *Analysis of HTTP Performance.* August, 1996 URL:http://www.isis.edu/lsam/publications/http_perf/